Constructing concepts and word meanings: the role of context and memory traces

Catalina Urquiza Arribas

Thesis submitted in partial fulfilment of the requirements for
the degree of Doctor of Philosophy

University College London
September 2014

I, Catalina Urquiza Arribas, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

_____

for Francisco

Abstract

The main aim of this thesis is to develop a new account of concepts and word meaning which provides a fully adequate basis for inferential accounts of linguistic communication, while both respecting philosophical insights into the nature of concepts and cohering with empirical findings in psychology on memory processes.

In accord with the 'action' tradition in linguistic theorising, I maintain that utterance/speaker meaning is more basic than sentence meaning and that the approach to word meaning that naturally follows from this is 'contextualism'. Contextualism challenges two assumptions of the traditional 'minimalist' approach to semantics: (i) that semantics (rather than pragmatics) is the appropriate locus of propositional content (hence truth-conditions); and, (ii) that words contribute stable, context-independent meanings to the sentences in which they appear.

I set out two stages in the development of an adequate contextualist account of utterance content. The first provides an essential reformulation of the early insights of Paul Grice by demonstrating the unavoidability of pragmatic contributions to truth-conditional content. The second argues that the ubiquity of context-dependence justifies a radically different view of word meaning from that employed in all current pragmatic theorising, including relevance theory: rather than words expressing concepts or encoding stable meanings of any sort, both concepts and word meanings are constructed ad hoc in the process of on-line communication/interpretation, that is, in their situations of use. Finally, I show how my account of word meaning is supported by recent research in psychology: context-dependence is also rampant in category and concept formation, and multiple-trace memory models show how information distributed in memory across a multitude of previous occasions of language use can come together to build an *occasion-specific* word meaning, thereby bypassing the need for fixed word meanings.

Acknowledgements

I would first like to sincerely thank Robyn Carston for the time and energy she has put into supervising my thesis. Her continuing patience and thoughtful comments on my drafts were the most helpful anyone could want. I also benefitted immensely from our long exchanges and will never forget them!

I am also grateful to many former professors and advisors who have inspired and encouraged me over the years. I am in particular grateful to Anne-Marie Houdebine, Marie Leroy, and Vincent Nyckees for fuelling my interest in theories of meaning.

I am indebted to countless fellow students and researchers with whom I have discussed ideas and shared thoughts. Independently of whether these ideas made it into the thesis, being given the opportunity to share my thoughts, and listen to others has been one of the most rewarding aspects of my project.

I also must thank my family and friends for their understanding and their efforts in adapting to my life as a doctoral candidate. A special thanks to Rafael and Maxi who often listened to my ideas with genuine interest, and to my worries with compassion. A very special thanks to Claudia and Neto for their faith in me.

My greatest debt of gratitude is to Francisco Lagos Dondé. I cannot imagine how I would have managed without his unwavering support, wisdom and patience.

Table of Contents

Chapter 1: Introduction

The main aim of this thesis is to propose a new perspective on word meaning and concepts that both respects philosophical insights into the nature of concepts and integrates contributions from theoretical and empirical studies in psychology and linguistics. My focus is what words mean *in their contexts of use*, so accounts of utterance comprehension processes recently developed within contemporary pragmatic theories of communication, such as relevance theory, are my point of departure. I take the 'action' tradition as my global framework within linguistics: its main tenet is that a more appropriate approach to the phenomena of language is what we *do* with language rather than the tacit knowledge people might have of the grammar of their language, as in the 'product' tradition.[1] An emphasis on 'speaker' and 'utterance meaning' instead of 'sentence meaning' naturally follows from this stance. My aim is to develop a corresponding account of concepts and word meanings as the constituents of these speaker and utterance meanings.

Among the claims of the account I seek to develop is that, *contrary* to the assumptions of traditional semantics, words do not contribute fixed, context-independent meanings to the utterances in which they appear; and, pragmatic contributions, which have long been assigned subordinate roles in fixing an utterance's content, are actually obligatory in arriving at a level of communicated content. I propose that these claims, now fairly consensual within contemporary pragmatic circles, represent an important enough shift

---

[1] A very similar division into two traditions can be found in the writings of most authors in this field. King and Stanley (2005), for instance, label them the 'expression-centered conception' and the 'speech-act-centered conception' of semantics. I adopt the terminology of Herbert H. Clark (1996) for its initial clarity. As the thesis progresses, it will become evident that many language theorists share Clark's thoughts and position. Their views will be presented alongside their own choice of terminology. An example in anticipation: Peter Bosch (2009) calls the 'product tradition' the 'linguistic knowledge paradigm', he explains:

> The central concept on which both syntax and semantics are built in the linguistic knowledge paradigm is the sentence, both as a basic notion of grammar and – in its guise as proposition or sentence meaning – as the basic notion of formal semantics (Bosch, 2009: 1).

I come back to a full depiction of the 'action' versus the 'product' tradition in chapter 3 (section 3.2).

to warrant positing a new framework for theorising on word meaning in context. This important shift in perspective can already be seen in the contributions of some pragmatists, but also of a few linguists and philosophers. I propose to bring their perspectives together under the label of 'contextualism'. I then argue that the logic of contextualism, which is based, among other things, on robust findings of the ubiquity of context-dependence in communication, should be pushed to its logical limit: a *radically* contextualist, 'eliminativist' view of word meaning. This account is radical insofar as it denies that words encode anything like a stable linguistic meaning and postulates that both the concepts that words express and the meanings we assign to words are constructed *ad hoc* through processes of general reasoning.[2]

To support my claims, I look to two well-established traditions in psychological perspectives on concepts that have come to similar conclusions as radical contextualism: context-dependence is rampant in interpretation; it can be found not only in utterance comprehension processes but more generally in *figuring out a scene*, as is evidenced in particularities of our categorising behaviour and in concept construction. Contrary to the traditional view according to which memory retrieves concepts *ready-made*, new 'exemplar models' posit that memory consists of largely 'undifferentiated' information and 'memory traces'; and, 'norm theory' posits that this information is scanned and summarised with respect to a particular context or task at hand so that, instead of outputting a fixed, context-independent concept, memory *constructs* a uniquely relevant ad hoc category or concept. Finally, this tradition assumes, like a certain number of linguists and philosophers, that the cognitive mechanisms responsible for

[2] On the issue of the role of general reasoning, I depart from the current relevance-theoretic position (see Wilson, 2005) insofar as I see language *in use* (i.e., conversational exchanges) as a *special domain* within general reasoning but I do not claim that pragmatic processes are carried out by a modular, special-purpose inferential mechanism, a submodule of the theory of mind module. For Wilson, pragmatic processes cannot be seen as 'a special case of a more general mechanism operating in broader domains' because, among other things, they are 'special-purpose mechanisms attuned to regularities existing only in the domain of intentional behaviour' (2005: 1132). My reasons for opting for a 'general reasoning' account of occasion-specific word meaning construction will become clear towards the end of chapter 3 and in chapter 4.

these constructions are part of *general* reasoning processes and common sense instead of specialised modules.

My aim is to bring these insights together on the topic of occasion-specific word meaning construction in order to make the case that standing, linguistically-specified, context-independent word and expression meanings are not needed as a point of departure in the processes of utterance comprehension generally offered by cognitive pragmatics.

The discussions in this thesis centre around four main topics (i) concepts (including contrasts between traditional and contemporary accounts); (ii) word meaning; (iii) psychological perspectives on concepts; and, finally, (iv) memory for language. Beginning after this introduction, chapter 2 first introduces the complex topic of concepts. In a thesis on word meaning, it might be surprising to find such detailed attention to theorising on concepts and topics related to concepts, like categorisation. The reason is simple: concepts are a central concern to any study of the mind but an even more pressing concern for the study of linguistic meaning since concepts are the constituents of the thoughts we aim to communicate when we speak.

The presentation of the topic of concepts starts in chapter 2 with two preliminary sections: a presentation of the traditional, classical theory of concepts (section 2.2), followed by a contemporary picture of the debate (section 2.3). Despite the fact that few would defend it today, the classical theory went undisputed for so long, and had such a profound influence, that it is an essential part of any discussion on concepts. Furthermore, since the very diverse theories presented in subsequent chapters are more or less direct *reactions* to the tenets of the traditional classical theory, an opening presentation is necessary as a point of departure for the discussion. This is followed by a discussion of present-day theorising focused on the fact that theories of concepts are currently developed not only by philosophers but also by psychologists. Since this introduces an important difference in perspective, I outline the points of contention that represent the main challenges of bringing philosophical and psychological perspectives together on the topic of word meaning.

The remainder of chapter 2 focuses on philosophical and relevance-theoretic perspectives on concepts. I give as thorough a presentation as possible of Jerry Fodor's detailed and influential theory of concepts. Fodor's early insights were largely adopted in configuring the framework of a then *new* science of the mind. He greatly influenced theorists in various fields, among them, relevance theorists Dan Sperber and Deirdre Wilson. Sperber and Wilson, however, as they brought their own expertise to bear on specific aspects of the theory, particularly the relationships between concepts and word meanings, felt the need to propose certain very important reformulations that are also the topic of my presentation in chapter 2.

Chapter 3 focuses on the complex set of contributions to the topic of word meaning from theorists working in cognitive pragmatics and related fields. These contributions take their starting points in diverse schools of thought and disciplines and as a result are not devoid of inconsistencies. Rather than a fundamental flaw, this simply follows from the novelty and intricacy of the issues at hand combined with the wide scope of influences taken to reflect on them. One of my objectives is to bring the diverse strands that result from this kaleidoscope of influences together on a particular topic: the viability of an account of word meaning *in context* that postulates a *new* framework.

The widely recognised complexity of word meaning makes this task challenging but hopefully not impossible. To help organise the presentation, I divide the chapter into sections and stages. A first section, entitled 'Two contrasting traditions' (3.2), sets the scene for the most critical disagreements between the established framework for theorising on word meaning (i.e. formal semantics) and the rival views emerging from contemporary cognitive pragmatics. The rest of the chapter is organised into two 'stages' with an intervening section entitled 'Philosophical foundations for radical contextualism' (3.4). I see this chapter as presenting different proposals which might at first seem fractured but which, I believe, all point in the same direction: towards an acknowledgement of the ubiquity of context-dependence in communication, an emphasis on *inference* based,

rather than code-based, models of communication and the role of general cognition and common sense in all sorts of interpretive tasks.

Chapter 4 is subdivided into 4 main sections. The discussion of the psychological perspective to concepts begins with a short discussion of some preliminary notions important to the debate: mental representation, abstraction and similarity. Then, a section entitled 'Concepts and categorisation' summarises the complex series of discoveries made by certain psychologists studying categorisation and categorising behaviour. Section 4.4 sets out the contributions of one particular psychologist, Lawrence Barsalou, responsible for reintroducing sensitivity to context into the debate with, among other things, the notions of ad hoc categories and ad hoc concepts. Finally, in section 4.5, I focus on the consequences of this new view of categorisation and concepts on theories of word meaning and, particularly, on how it supports meaning eliminativism.

The aim of chapter 5 is to complete this picture with a discussion of memory. Throughout the thesis I argue that instead of fixed, pre-existing forms as input to our everyday processes of interpretation, including utterance comprehension, we construct ad hoc concepts and occasion-specific word meanings by selectively reactivating memory traces in general (instead of *semantic*) memory. For this process, it suffices to scan and summarise previous episodes or occasions of use captured in 'memory traces' on our *episodic* memory. This directly challenges views of memory that assume its role is to retrieve pre-existing forms and replaces it with a dynamic view in which retrieval is a construction process that selects from past experiences with a particular purpose or context at hand. An important section focuses on Douglas Hintzman's multiple-trace memory model (5.4), and a final section on its implications for a positive account of meaning eliminativism.

In this last paragraph of the introduction, I briefly summarise the main claims of this thesis and how they relate to the chapters I have described above. My main objective is to put 'meaning eliminativism' forth as a viable option to current accounts of word meaning in context. In the meaning

eliminativist position I defend, words do not need to have fixed, context-independent meanings to serve as points of departure in utterance comprehension procedures. To argue for this point, I first present accounts of concepts from a philosophical perspective in chapter 2. My interest in concepts is justified not only because concepts are the constituents of the thoughts that we aim to communicate when we use words but also because, just as the received accounts presented here, I too hold that *word meanings are concepts*.[3] However, in my own conception, our understanding of both concepts and word meanings is transformed in light of the ubiquity of context-dependence, among other things. In chapter 3, after reviewing existing accounts of word meaning and identifying the logic behind the role context is increasingly taken to play, I claim that despite being on the right track, contemporary cognitive pragmatic accounts of word meaning in context are not radical enough. I argue that the evidence amassed by contextualists warrants the positing of a *new* framework which would leave the traditional semantic framework, also known as 'the modular view', behind. In chapter 4, I extend the analysis of context-dependence to concepts. Contrary to most thinking on concepts, and to key aspects of the views presented in chapter 2, I argue that concepts are not fixed, pre-existing mental entities simply reactivated in new settings. Tasks such as categorisation, decision-making and utterance comprehension can be explained without necessarily stipulating any kind of rigid, innate, atomic or anatomic concepts. I claim that memory does not store (and does not need to store) its database in the form of fixed concepts. Rather, as described in chapter 5, when a task is at hand, a uniquely context-specific category/concept emerges from memory, the product of a very powerful,

[3] As with the overwhelming majority of studies into the meaning of words, I am particularly interested in words that express natural kind and artifact kind concepts. I therefore mostly leave to the side discourse connectives, pronouns and other grammatical markers. Relevance theorists have developed the notion of 'procedural meaning' to account for these other types of word meanings (see Blakemore, 1987; Wilson, 2011), I come back to procedural meaning briefly in chapter 2 (§ 2.7.1).
Also, note that not everyone in contemporary cognitive pragmatics takes the position that 'open class' words, as they are sometimes called, to differentiate them from grammatical words such as 'the' and 'with', express *full-fledged* concepts instead of something more schematic, this is a topic I come to in chapter 3 (§ 3.3.3).

dynamic process of ad hoc construction. Thus, meaning eliminativism is vindicated by psychological models that offer alternatives to fixed, pre-existing forms and given substance by the descriptions of the mechanisms that construct occasion-specific forms.

Chapter 2: Philosophical and Relevance-Theoretic Perspectives on Concepts

2.1 Introduction

The main objective of this chapter is to set the scene for certain correlations between perspectives on concepts and theories of word meaning. The view that concepts are, or should be, *the* central concern of theorising on language and thought is largely compatible with the approach that is to be the central focus of this chapter: Jerry Fodor's philosophical perspective on concepts. I am interested in Fodor, not only because he is one of the philosophers who early on led the challenge against the assumptions of the classical theory of concepts and continues to be influential today, but because he exerted and continues to exert a decisive influence on relevance theorists like Deirdre Wilson, Dan Sperber and Robyn Carston, who have in turn greatly influenced me. Relevance theory has, from its very beginning, looked to Fodor's construal of language and thought and adopted his framework as a base for theorising on communication and pragmatics, but, at the same time, because pragmatics is a specialised discipline, it has contributed significant insights to some of the key issues in his construal, particularly with relation to word meaning *in context*. So, the long discussion of Fodor's contributions in this chapter, followed by the relevance-theoretic partial reformulation, is justified because, on the one hand, the basic outline of relevance-theoretic cognitive pragmatics was developed with Fodor's theory of concepts as background; and, on the other hand, because the arrival of more recent developments, particularly in lexical pragmatics, raises some thought-provoking questions as to the compatibility of the relevance-theoretic project with Fodor's framework. Fodor's views on concepts are also particularly relevant as part of my presentation because I arrive at a construal of concepts that is only very partially compatible with Fodor's, and ultimately at a construal of word meaning that departs significantly from that of relevance theory. As stated earlier, one of my claims is that *word meanings are concepts*; importantly, however, this is only partially compatible with similar proposals by Fodor and relevance-

theorists. In this chapter, I therefore set out Fodor's theory of concepts and relevance-theoretic pragmatics to reflect how they are a point of departure in my own theorising. Below, however, I first take up a common point of departure for all three perspectives: the traditional, or 'standard classical' theory of concepts.

2.2 The Traditional View of Concepts

To a first approximation, the main tenet of the standard classical theory of concepts is that concepts are structured. Firstly, what it means for a concept to be structured is for it to be decompositional. A concept's content is made up of 'primitive' features so that the meaning of the whole is a function of the meaning of the parts. For instance, the concept CHAIR would be said to 'compose' out of features such as *seat-for-one*, *with back-rest*, *four legs*, … .[4] Secondly, a concept's structure is said to be 'definitional' when it provides the necessary and sufficient conditions for the concept's application. A working classical account of the concept CHAIR would provide a definitive list of features for CHAIR that was both a true description of all existing (and possible) chairs and a reliable way of identifying any object as belonging or not to the class of chairs.

The advantages of such an account would be considerable. Its simple formulation and straightforward applications would greatly facilitate the study of our mental lives. For those of our mental capacities that involve the deployment of concepts, such as categorisation, language production and

[4] With regards to the conventions for representing concepts, words, and features in my text, and the use of single quotes, I have followed the usage in the literature as much as possible. I take this opportunity to spell out what the different formats mean in my text. I use simple quotes for technical terms when I mention them in passing or when I first introduce them. For instance, in the passage above I put 'primitive' in single quote marks to mean that, in this context, it is a technical term that I am assuming the reader is familiar with. I also use simple quotes for reporting a subjects' word(s), expression or utterance: for instance 'It's raining' expresses the thought that it is raining. I use small caps for quoting concepts: for instance, CHAIR in the passage above. Or, in another example further down the line: different shades and intensities of blue fall under the concept BLUE. For categories, when the distinction needs to be made between a category and a concept, I use small caps in italics, for instance, subjects judged peas better exemplars of the category *VEGETABLE* than brussels sprouts. Finally, I use italics for features. For instance, according to the definitional approach, features for the concept CHAIR were *seat for one*, *with backrest*, etc. I will repeat particular conventions as the need arises.

comprehension, decision-making, to name but a few, their study and description would be by the same measure facilitated. Enthusiasm for this possibility is probably what fuelled interest in the standard classical theory for so long. Inevitably, however, its shortcomings had to be faced. Perhaps surprisingly at first, and then steadily more and more resolutely, concepts resisted definitions and proposed definitions succumbed to counter-examples. Let us, for instance, return to the concept CHAIR, for which I proposed some possible features: *seat-for-one*, *with back-rest*, *four legs* and slyly added '…' suggesting that other features, or *different* features, would be needed to complete the analysis. In fact, some chairs are big enough, or some people thin enough, that two can sit in one chair, and the object in question is no less a CHAIR for that. The number of legs is not *strictly* necessarily four. Yet making the definition less precise by subtracting how many it sits or the number of legs is not a promising solution. The set of features has to be precise enough to describe chairs and only chairs, otherwise, the definition loses its explanatory power. Furthermore, even supposing that the problem of how many it sits and of the number of legs could be solved, the definition would still be unsatisfactory since *seat-for-x*, with *back-rest* and *with x-legs* also describes some BAR STOOLS. [5] This issue, which arose repeatedly, prompted philosophers, followed closely by psychologists, to take a closer look at the theory. The assumptions behind the definitional approach were then analysed and depending on each researcher's perspective, different possible revisions were suggested or different assumptions were outright rejected.

One of the main assumptions of the classical theory, which can be traced back all the way to Plato and Aristotle, is that each member of a category has some critical feature, or set of features, that somehow marks it as a member of its category.[6] A related assumption is that a concept just is

---

[5] This is but one example of many in the literature, I return to similar arguments against the definitional account and illustrate them with other examples later in this chapter when I present Jerry Fodor's criticisms of the definitional account and his proposed alternatives.

[6] Platonic idealism is the earliest version of the view that there are specific *defining* 'essences', 'forms' or 'ideas' behind the objects that surround us. I come back to this idea and related claims in chapter 4, § 4.3.4.

the mental representation of these features formulated as necessary and sufficient conditions. The standard classical theory of concepts holds that the rules for defining a concept obey two principles: 'necessity' and 'sufficiency'. Necessity refers to the fact that the conditions listed for category membership *must* be true of the entity in order for it to qualify as a member of the class. Sufficiency means that if something fulfils *all* the conditions listed for category membership, then it must be a member of that category. The conditions are said to be 'individually necessary' and 'jointly sufficient' to completely and unmistakably define a category. Following these rules would ensure effective definitions, immune to the criticisms expressed above concerning CHAIR.

There are two major problems with this. First of all, as already mentioned, the fact is that actually pinning down any definitions has proved elusive. Two possible responses to this problem initially emerged. The first observes that despite the fact that definitions prove elusive, they could be posited as *present* in the mind, but, like many other types of knowledge, perhaps residing just under the level of consciousness so that using them is unproblematic while making them consciously explicit remains a challenge. A second related point is that perhaps concepts do have definitions but they are simply not in natural language *form*, so again, despite the fact that they exist, putting them into words remains challenging. However, even taking these points into consideration, there was still the problem of the psychological reality of these definitions. What role were they playing in, for instance, language comprehension? Walter Kintsch (1974), for instance, made the following predictions based on the standard classical theory: a word such as 'convince' is more complex than a word such as 'believe' since under the definitional approach, it would make sense to define the one in terms of the other. 'Convince' could arguably be analysed as 'cause to believe'. Furthermore, since processing more complex concepts should require more cognitive effort, Kintsch hypothesised that if subjects were slower in a phoneme-tracking task when the word preceding the target phoneme was 'convince' rather than 'believe', then this could be interpreted as supporting the standard classical theory. The results, however, clearly

pointed to the opposite conclusion: there was no sign of (allegedly) more complex words having an effect on the speed of phoneme-tracking.[7]

Evidence against definitions accumulated quickly. Experimentalists rightly agreed that it followed from the assumptions of the classical theory that subjects somehow represented *as part of their conceptual knowledge* the necessary and sufficient conditions for something to qualify as a chair; but, *no evidence* could be found of this. Rather, it seemed that subjects possessed and used concepts independently of whether they *or anyone else* could provide necessary and sufficient conditions for those concepts. Concept possession could no longer be theorised as involving the mental representation of defining features for classes. In fact, the decades spent looking for definitions had produced more evidence against definitional structure than in favour of it. This launched the philosophical and psychological theories of concepts that I present in this and the following chapters. Views on the merits of the classical theory shifted drastically and it was at last unavoidable to abandon the main tenets of the theory. For some, such as Jerry Fodor, the evidence supported abandoning *any* decompositional account of concepts. For others, moving forward from the failure of the classical theory involved, among other things, rethinking decompositionality by questioning the need for strict necessary and sufficient conditions while holding on to the idea that concepts were composed of features.

2.3 Incompatible Perspectives on Concepts?

Differences between the psychological and the philosophical perspectives on concepts run deep. As illustrated above with definitional decompositionality, the general consensus that a new theory of concepts needed to be formulated was immediately followed by disagreements on which aspects of the theory constituted the mistakes to be avoided.

---

[7] Kintsch, 1974; discussed in Laurence and Margolis, 1999: 17-18; see also Fodor, Garrett, Walker and Parkes, 1980, discussed in this chapter, section 2.5.

But perhaps even more fundamentally, psychologists and philosophers, *in correspondence with their different methodologies*, envisioned their course of action and their aims differently. So, despite the fact that the questions they asked could be *broadly* the same, the way forward was unlikely to be shared. Roughly, philosophers examine existing arguments (be they philosophical or psychological) and employ logical reasoning to address questions such as *what is a concept?*; psychologists, on the other hand, rely on empirical methods and, in particular, the cognitive psychologists who addressed the issue of concepts privileged the experimental approach. Of course, ideally, psychological and philosophical approaches complement each other and it is uncontroversial to hold that a complete account of concepts takes both into consideration. Unfortunately, however, some collaboration is required for a unified account and, more often than not, what one side favours as a valid contribution the other judges detrimental; what one side deems of critical importance, the other largely ignores.

On this note, an important and recurring criticism by certain philosophers of the work of psychologists on the topic of concepts is their alleged disregard for the limits of psychological explanation.[8] They hold that an inescapable starting point for any discussion on concepts is the distinction between how the world *is* and how *we think or infer* that it is. For Jerry Fodor, for instance, for a cognitive theory of concepts, this means that a distinction must be made between the metaphysical issues of conceptual *identity* and the epistemological issues of conceptual *access*. Fodor (1998) makes this a distinction between questions concerning what a concept *is* and what it means to *have* a concept. Arguably, philosophers have the upper hand when it comes to theorising on concepts because psychological accounts only concern themselves or *should only concern themselves* with

[8] In this subsection, I focus on the criticisms of Jerry Fodor and Georges Rey, who are representatives of this philosophical perspective. Of course, not all philosophers would agree with them; in fact, as we'll see below, Fodor often deplores the ever-increasing popularity of the view he argues against.

the second question. Furthermore, the metaphysical question logically precedes the epistemological one:

> First you say what it is for something *to be* the concept *X* – you give the concept's 'identity conditions' – and then *having* the concept *X* is just *having whatever the concept X turns out to be* (Fodor, 1998: 2).

Until the recent interest among psychologists in concepts, these questions were exclusively addressed by philosophers and Fodor's view that an explanation of *access to* or *possession of* concepts would necessarily *follow* an explanation of *defining or identity conditions* for concepts was the norm. Fodor's concern is that recent psychological accounts appear to either inadvertently overlook the distinction and the principles it traditionally imposes, or, worse, make unfounded claims to the effect that this distinction has been fundamentally altered by certain experimental results.[9] In either case, philosophers who make this point (i.e., Jerry Fodor and Georges Rey) fear that the resulting psychological framework for the study of concepts would prematurely dismiss metaphysical issues in favour of a purely *epistemological* enquiry mistakenly believing that it could stand *alone* as a theory of concepts.

The situation is further complicated by the fact that, while it is this general disregard for the limits of epistemological enquiry and the resulting potential for confusion that attracts the criticism of these philosophers, they are at the same time happy to accept psychological research as an important contribution to theories of concepts. Far from denying the value of discoveries such as prototypes, their disagreement is with particular interpretations that, in their view, misrepresent or exaggerate the consequences of psychological findings (e.g., by such claims as that concepts are constituted by prototypes). Redressing the situation would therefore depend not on denying that psychology has revealed new data on how

---

[9] Fodor mostly aims his criticism of psychological accounts at 'prototype' theorists. As I'll show in chapter 3, however, prototypes are only one of the types of categorising *effect* that a more complete understanding of categorisation behaviour reveals. I address Fodor's criticisms in the section on categorisation in chapter 3.

people *access* their beliefs, for instance, but on maintaining a strict distinction between such findings and core metaphysical issues. This distinction is supposed to afford psychology an adequate framework for a more measured and controlled development of its contributions. But there is an immediate possible objection: that psychology is inherently limited does not necessarily mean that the philosophy of Fodor and Rey is apt and ready to complement it. In other words, it might be true that experimental findings cannot stand alone as theories of concepts, but it does not follow that an adherence to *existing* philosophical principles is the answer. For instance, while the point they make about the metaphysical/epistemological distinction is undeniable, this leaves ample room for disagreements on how psychological approaches should deal with it. In the remainder of this section, I look at two slightly contrasting philosophical accounts of this issue: the question is whether it is possible to bring psychological and philosophical perspectives together on the topic of concepts. Whether these suggestions are likely to be adopted by psychologists is an issue I return to in chapter 4.

In a review of Edward Smith and Douglas Medin's (1981) influential book on concepts and categories, Georges Rey (1983) gives a particularly clear account of how philosophers such as himself view the recent contributions of psychological enquiry to the study of concepts. He starts by pointing out that until the arrival of Eleanor Rosch and those who followed her, issues of conceptual identity and possession were the sole domain of philosophers. The contrast between the 'classical view' elaborated by this philosophical tradition and the evidence from psychology is quite direct: according to the traditional account, concepts can be *defined* by necessary and sufficient conditions which a user must grasp in order to be competent; Rosch and colleagues' proposal is that concepts can be associated with *typical* features and individual exemplars (Rey, 1983: 237-238). Importantly, however, for Rey, it is one thing to say that psychological enquiry *has* forced philosophers to re-evaluate their *epistemological* assumptions, it is quite another to claim that these findings warrant challenging or *abandoning* the core principles of

the 'classical' account. In arguing for this claim, Rey first reviews the key claims of Smith and Medin (1981) with the objective of ascertaining whether

> …people's responses to categorisation queries bear upon the question of the identity of concepts, or even on the conditions under which they are competent to use one (Rey, 1983: 240-241).[10]

Then, in his challenge to the authors' claims, he calls upon the functions concepts have traditionally been called upon to perform in order to conclude that:

> …as a theory of *concepts*, [Smith and Medin's] proposal hopelessly confuses metaphysical issues of conceptual *identity* with (roughly speaking) epistemological issues of conceptual *access* (Rey, 1983: 238).

The reason for this very negative evaluation, to the best of my understanding, is that according to Rey, Smith and Medin's 'conception of *concept*' does not serve any of the functions concepts have *traditionally* been called upon to serve. To illustrate these 'functions', Rey lists four non-exclusive, non-exhaustive functions of concepts: *stability*, *linguistic*, *metaphysical* and *epistemological* functions. The purpose of the *stability* function is to guarantee commonalities between contents so that two subjects or a single subject at two different times can be said to be in the *same* cognitive state. The fear is that without stability, we would not be able to say that two thoughts are about the same thing, with supposed catastrophic consequences for our ambitions to explain human mental life.[11] The *linguistic* function refers to the link between the words of our natural languages and the concepts they represent. Clarifying these two functions is not Rey's objective in this article; furthermore since much of the work in this

---

[10] Since I give ample coverage to claims such as those in Smith and Medin (1981) in chapter 4, I do not go into any detail on Rey's review here. Suffice it to say that categorisation theorists believe it is basic that subjects do not *know* defining conditions for the concepts they use but that, according to Rey, this has little bearing on whether there *are* defining conditions.

[11] This function is particularly important to Fodor who notes 'concepts are public' as one of his five non-negotiable conditions for a theory of concepts (discussed later in this chapter, section 2.8).

thesis is a development of the idea that words express concepts, and that stability is largely unproblematic, I leave these points to the side for now. I am particularly interested in Rey's account of the *metaphysical* function of concepts: he holds that concepts have often been asked to provide the basis for metaphysical claims (Rey, 1983: 243). Despite the fact that the distinction between metaphysical and epistemological is, as he puts it, 'not everywhere perfectly sharp' he argues that there *is* a sense in which whether something actually is out there in the world is different from whether anyone *knows* whether there is something out there. If a concept can be regarded as providing principles of classification, then it functions as support for a metaphysical claim such as *what is out there in the world is an X*. In case what is out there is a natural kind, such as *cow*, Rey adds, not just any fact about any cow will do, what is needed is a characterisation of the 'universal' COW, or, of the essence of COW (Rey, 1983: 243). The contrast with the *epistemological* function is given as follows: there are reliable indicators by virtue of which we *tell* if something is *X* which should not be confused with the *metaphysical* principles that make it that something is an *X*. Rey illustrates this with GENDER: there are well defined conditions by virtue of which something *is* FEMALE, but rather than '(impolitely) ascertaining these conditions in public', we go by superficial but reliable features. From this perspective the confusion of psychologists could be explained as one question being replaced by another: the question of 'How do you [strictly speaking] *know* something is an *X*?' by 'How can you *tell* something is an X? (Rey, 1983: 244).

Despite the fact that Rey's view of Smith and Medin's work is overall negative, precisely because he takes them to 'hopelessly confuse' the metaphysical with the epistemological, I would argue that the overall article, and some of Rey's following work (1985, 2010), actually points to a way of respecting the metaphysical/epistemological distinction that is *open to psychologists*: Rey's advocacy of Hilary Putnam's 'division of linguistic labour hypothesis' and *externalist* semantics. Following Putnam allows a clearer distinction than alternative accounts: not only is there a distinction between what a subject *knows* and how the world *is*, there is a further distinction

between knowing the defining conditions of a term or concept (which *may not be the case of anyone at all*) and being a competent user of a term or concept. In other words, you can claim that there are properties which something must, *as a metaphysical necessity,* have to be *X*, and, at the same time, that these necessities need not play any *epistemological* role. Furthermore, we are free to believe that science and experts provide 'optimal accounts' and either, if we are 'realists', we can believe that those accounts might be *strictly true*, or, if we are not, then simply that they are the accounts human beings will eventually agree upon (Rey, 1983: 255). Finally, Rey mentions the possibility that the defining conditions are replaced by empty slots in our representations. People are very willing to accept that they do not know the defining conditions for their terms, or that the experts could have got it wrong, but rather than giving up on defining terms, they adopt a flexible strategy that makes their knowledge *revisable*.

An approach largely developed after Smith and Medin (1981) but squarely within today's psychological approaches to concepts describes subjects' dispositions to maintain a slot, even in the face of changing contents for this slot, as *psychological essentialism*. Briefly, the essentialism expressed by Rey above when he states that 'not just any fact will do' as a defining condition, that what we need in support of our metaphysical claims is the characterisation of the 'universal' cow, is, I will argue, a deeply engrained, human tendency to *believe* and *act as if* natural kinds have essences. This is a human *psychological* propensity and *therefore* independent of whether kinds do in fact have essences. Therefore, independently both of whether natural kinds have essences, and of whether we can know them, we collectively *assume* that those essences exist; perhaps we are wired to *think* and *act* as if the world is made up of discrete kinds. I come back to this in chapter 3, where I give more ample coverage of Putnam's theory of meaning, and in chapter 4 where I discuss psychological essentialism. I now move on to Fodor's take on the possibility of interdisciplinary work on concepts.

Fodor (1998) is distinctly less optimistic than Rey; for him, not only have psychologists in general misapprehended or overlooked the metaphysical/epistemological distinction, but the reversal of the traditional order has come to be widely adopted, not only in the field of psychology and linguistics but also in philosophy of mind and cognitive science in general. According to Fodor, *the* current trend that is leading cognitive science astray is asking the epistemological question 'what is concept possession?' *before,* or worse *instead of*, asking 'what is a concept?'. In arguing for his own approach, Fodor (1998, 2003, 2004) presents what in this context he calls 'concept Cartesianism' (i.e., prioritising the question of what a concept *is*) as a viable alternative to the 'alarming' trend he labels 'concept pragmatism' (i.e., roughly, any account failing to prioritise the question of concept *identity*).[12] I come back to Fodor's proposal for a theory of concepts later in this chapter (section 2.6). Here I am interested in the role he sees psychology playing in an overall account of the mind. As for Rey above, for Fodor, opposing philosophical and psychological approaches does not necessarily imply a rejection of psychological enquiry as a whole, but it does call for strictly prioritising certain questions over others. He holds that the validity of the psychological question critically depends on rejecting the attempts that have become so common in cognitive science and philosophy of mind to directly answer questions about concept possession with the mistaken belief that 'having a concept is a matter of what you are able to *do*, it's some kind of epistemic 'know how'' (Fodor, 1998: 3). Importantly, what is rejected is not the possibility of explaining how we are able to reliably recognise something as *X* or how we are able to draw sound inferences about *Xness*, rather, it is the specific claim that having these capacities *is* having the concept. For Fodor, this approach *cannot* answer the question of what a concept *is* because it frames concepts as *capacities* and, traditionally,

[12] Elsewhere, Fodor characterises concept pragmatism as defining concept possession in terms of abilities to *do* certain things, like the ability to sort things, or the ability to draw inferences about things. Prinz and Clark (2004), however, reject this characterisation as too limited: they claim that for pragmatists, concepts are 'collections of action-oriented abilities' of the kind that allow us not only to sort things and draw inferences but, more globally, to 'coordinate our behavior with the objects of the world' (p. 60).

concepts are the kinds of things that underpin capacities; they serve to represent the things our thoughts are about. [13] This approach inadvertently replaces the leading question by the subordinated one; nevertheless, recognising this opens the possibility for an interdisciplinary account of concepts by stipulating that philosophers and psychologists ask different but complementary questions, this on the condition that the question of concept identity always precedes the question of concept possession.

> Why shouldn't someone who thinks, qua Cartesian, that having a concept is having something in one's head that serves to represent the objects of one's thoughts, also be interested, qua psychologist, in what we do, or can do, or should do with the concepts we have? Cartesians don't deny that it's the uses we put our concepts to that makes them worth the bother of having or of studying.
> What Cartesians deny is just that our putting our concepts to the uses that we do is constitutive of the concepts or of our having them (Fodor, 2003: 21).

Once more, however, while psychologists might agree that their perspective is limited and even that philosophical issues have a certain priority, the question remains open as to whether or not they should accept the specific frameworks set out for them by any given philosopher. They might agree with Fodor 'that it's the uses we put our concepts to that makes them worth the bother of having or studying' without thereby accepting Fodor's account of what concepts *are*. The question of what concepts are is, after all, still wide open, even according to Fodor (1998), and so psychologists would be justified in preferring to follow their own methodology, including a certain compartmentalisation of their issues, for the time being. This, at least, is the position I adopt as I shift my attention to the psychological perspective in coming chapters.

---

[13] Fodor offers various arguments in favour of rejecting the view that talk of 'epistemic' capacities can answer questions of concept identity: the latter logically precedes since, for instance, tracking something requires representing the trackee, or, in other words, epistemic capacities *presuppose* concepts, therefore, they cannot *constitute* them (Fodor, 2003: 20).

Perhaps the problem is that very few theorists are well enough versed in both philosophical and psychological approaches to launch effective interdisciplinary dialogues. Or perhaps, on a more positive note, interdisciplinary dialogue *has already begun*, but it is only in its early stages and therefore not yet widespread although ready to blossom. There are, after all, some bright lights on the map of interdisciplinary approaches to concepts which can be cited as examples: the philosophers Eric Margolis and Stephen Laurence have carefully read the experimental literature on concepts and convincingly argue that bridges between philosophical perspectives and empirical observations are possible (Laurence and Margolis, 1999).[14]Also, as seen above with the discussion of Rey and Fodor, philosophers might be critical, but they are no longer simply dismissive of psychological contributions. Finally, it is also important to recognise that psychological research into concepts is not always purely experimental and therefore devoid of philosophical considerations. Throughout this thesis, I present the work of influential psychologists like Lawrence Barsalou, Edward E. Smith, Douglas Medin, and Douglas Hintzman, among others, who do not lose sight of the theoretical issues while still developing mostly empirical accounts. They study concepts and conceptual processes by contrasting the available evidence with whatever the current standard account is in order to point out deficiencies and suggest improvements.

Yet, this optimism might be misguided; despite certain exceptions and a general pluridisciplinary *ambition*, it still seems to be the case that the differences in method, terminology and theoretical influences are so profound that interdisciplinary dialogue cannot be sustained and the two accounts will continue to follow separate trajectories. We must acknowledge that after the common rejection of the classical theory of concepts, philosophers and psychologists each followed a different logic in further

[14] Other philosophers, however, like Jesse Prinz and Andy Clark, take a less conciliatory approach. In their 2004 joint paper, they make a solid case *against* the Cartesian/pragmatist dichotomy as described by Fodor. They define concept pragmatism in positive terms by linking concept possession to action while rejecting the validity of Fodor's Cartesianism.

researching concepts. For philosophers, the constraints were to do with how human beings represent the world to themselves, how they 'lock onto' properties and compose thoughts involving them. For a philosopher like Jerry Fodor, the further main task at hand was to account for the productivity and systematicity of thought. For him, one of the basic observations was that thought is *systematic*. That is, anyone capable of having one thought, for instance that *Jim punched James*, is also capable of having another thought, namely that *James punched Jim*. A second basic observation is that the capacity for having different thoughts seems unbounded despite the fact that as a resource the mind must be finite. Thought is *productive* because it takes building blocks (that is, concepts) and recombines them in novel ways. To account for these observations, Fodor adopts the constraint that whatever concepts are, they must be *compositional*, as otherwise systematicity and productivity cannot be explained.[15]

Psychologists, on the other hand, see concepts as knowledge structures that allow us to understand and interact with the world around us.[16] Their main concern is to explain capacities such as identifying an object as a member of a class or generalising from a particular object or experience to a class. Critically, concepts are mental structures that are *inclusive* of information, that is, psychologists are happy to include all sorts of 'contingent' information in their knowledge structures and say of these knowledge structures that they are concepts. This contrasts with most philosophers for whom including 'contingent' information in concepts blurs the line between the concept itself and the encyclopaedic information that

---

[15] Fodor, 2001; Fodor and Lepore, 2002; Fodor, 2004; Fodor, 2008. The comments on Fodor's theory of concepts here are not meant to be a full portrait of his theory, as most of chapter 2 is devoted to that. I come back to a fuller discussion of compositionality, for instance, in section 2.6. Here, I just seek to highlight the differences between a typical philosophical approach and a typical psychological one in order to draw the contemporary picture of the debate.

[16] I am thinking of psychologists like Michael Posner, Eleanor Rosch, Douglas Medin, Edward E. Smith, Gregory Murphy and Douglas Hintzman, among others. As with Jerry Fodor, a large section of this thesis is devoted to presenting their contributions. Here I just wish to highlight the most notable differences in general perspective. The contributions of individual psychologists will be presented in chapters 4 and 5.

can be associated with it. Despite rejecting the definitional approach, psychologists hold on to *rich* conceptual representations. In prototype theory, concepts are knowledge structures that include the typical properties of the things that they apply to; in theory-theory, they contain something like a theory of what the concept applies to; finally, in exemplar theory, the representation of a concept/category contains individual exemplars having been judged to belong to the category. Chapters 4 and 5 are dedicated to a careful presentation of these and other related possibilities. Once both the philosophical and the psychological perspectives on concepts have been presented, two opposing perspectives on mental representation will have emerged. In Fodor's view, accounting for the productivity and systematicity of thought, among other things, entails that concepts are bare, *atomic* mental representations. For psychologists, the use we put our concepts to suggests mental representations that are *rich* (although *not definitional*) knowledge structures. I will often come back to this simple opposition as it reappears in various discussions in future sections and chapters.

2.4 The Fodorian Framework

This section first sets out a general framework for theorising on human cognition. It contains Fodor's very influential constraints on *how* best to construe, in a post-behaviourist era, general issues such as the basis for an internal system of representations and a construal of cognitive processes as computations, plus his contribution on how to include beliefs and other mental entities into our theories of mind and how to frame this science of the mind within a more general 'physicalist' framework. The following three sections focus on Fodor's most important contributions to the topic of concepts. In the final section before the closing remarks, I adopt the relevance-theoretic perspective on Fodor's theory. As announced above, I am most interested in certain reformulations that the pragmatic perspective affords, particularly with regard to the relation between concepts and words.

2.4.1 Representational Theory of Mind

In contemporary philosophy, there are basically two answers to the broad question of the nature of concepts. Concepts are either abstract objects (as in the tradition following Frege) or they are mental representations. The representational theory of mind, held by Fodor and the majority of the field, adopts the latter position and explains human cognitive processes in general on the basis of an internal system of representations and computations. Fodor's representational theory of mind holds that mental representations in this internal system of representations have semantic properties and functional roles. Consider the mental representation *Fang is ferocious*. If someone holds the belief that *Fang is ferocious*, this is a tokening of a mental representation. It is a propositional attitude (of belief that *Fang is ferocious*). It has a functional role (of *belief*, rather than, say, desire). Functional roles are distinctive; in other words, mental processes for belief are different from those for desire.

Fodor has been a key figure in developing representational theory of mind. In his *Language of thought* (1975), he persuasively argues that the internal system of representations is very much like a language, a language *of thought*. He develops in depth the idea that it has a language-like syntax and a compositional semantics. Following Margolis and Laurence (2007), Fodor's analogy with natural language can be understood as pointing out that the distinctions available in the internal system parallel those of natural languages: there is a distinction between predicates and subjects and logical devices are present. As for attributing a compositional semantics to the language of thought, this means that 'the content of a complex representation is a function of its syntax and the contents of the representations from which it is composed' (p. 562). The same arguments that support productivity and systematicity in natural language would also apply to thought. And explaining the productivity and systematicity of thought is, as stated earlier in this chapter, one of the main objectives that Fodor believes a theory of concepts should accomplish.

Finally, Fodor's overall project can be described as 'vindicating folk psychology within a physicalist framework' (Cain, 2002: 1). In order to understand Fodor's position on how (human) *behaviour* is to be explained, it is helpful to have in mind two basic commitments that underlie the diverse topics running through his extensive work: a commitment to folk psychology (also known as commonsense belief/desire psychology) and a commitment to physicalism (also known as materialist psychology).

2.4.2 Commonsense Belief/Desire Psychology

The basic assumption of commonsense belief/desire psychology is that sentient beings 'have, and act out of, beliefs and desires'. A further assumption is that humans have a 'theory of mind', in other words, the capacity to meta-represent our own and other peoples' beliefs and desires.[17] This higher-order capacity can be described as the compass by which human beings navigate in the social world. Furthermore, the range of mental states and events that fall under the description of commonsense belief/desire psychology ('folk psychology' for short) is so great that we can assume it is involved each time we interact with another social being. Out of all of these mental states and events, Fodor, and psychologists in general, are particularly interested in *intentional* states. Intentional (mental) states are of particular interest because they have 'semantic properties'. In other words, they can be differentiated by their meaning or content and, are critically *about* something so that they have truth, or satisfaction, conditions. This can be illustrated with an example: suppose I have the belief that Fang is ferocious. The identity of this state, its meaning or content is *Fang is ferocious*. It is an *intentional* state because it is *about* Fang; furthermore, it represents him as *ferocious*, therefore, it is true if and only if Fang really is ferocious. Yet the objects or phenomena that intentional states are about do not have to exist. It is thus possible to have beliefs about the tooth fairy.

[17] I would add that the basic and higher-order capacities proposed here also mean that for human beings, *having a theory of mind* means that besides inhabiting the physical world, we also have *mental* lives made up of *mental states*.

Intentional states must also be differentiated by the relations to the content that they involve. Just as I can be in a belief relation to the content *Fang is ferocious*, I can be in a desire relation to the same content. A mental state's *aboutness* differentiates it from other kinds of mental states such as emotional or affective states which, in contrast with intentional mental states, do not represent anything from the outside world.[18]

Finally, Fodor's insight is that the importance of 'commonsense belief/desire psychology' goes beyond offering the best explanations for our everyday interactions because folk psychology, as a theory, is akin to *scientific* theory. When a folk psychological generalisation is made explicit, it parallels scientific generalisations in that it takes on the deductive structure characteristic of scientific explanation; it uses technical terms (in this case, for instance, *belief*) to describe unobservable phenomena; and it makes (law-like) generalisations. Fodor's example is 'If *x* wants that *P*, and *x* believes that *not-P* unless *Q*, then ceteris paribus *x* tries to bring it about that *Q*' (1987: 13). That the generalisations of folk psychology are hedged by *ceteris paribus* clauses, as in the example above, does not detract from their validity; as Fodor points out, all special science generalisations are so hedged, and do not thereby cease to be informative.

2.4.3 Physicalism

Fodor's physicalism is based on the widely accepted assumption that since reality is ultimately physical in nature, a complete physics would provide the most basic explanation possible in science. This explanation would use the vocabulary of physics, into which all other special science vocabularies can, in theory, be translated. Physicalism is mostly unproblematic when it comes to the phenomena of '*non*intentional' sciences such as biology since the terms of this science (e.g., *living organism*, or *gene*) are physically instantiated. It is arguably straightforward to see how the objects, states, events and processes of most special sciences imply causal processes and

[18] For a detailed characterization of intentional mental states see Fodor (1987) and Cain, (2002, chapter 1).

laws that are underpinned by *physical* causal processes and the laws of *physics*.

This is much more complex, however, when it comes to 'intentional' sciences, which involve mental phenomena. Physicalism would imply that even *mental* states are somehow to be described in terms of physical systems. In other words, at the most basic level of explanation, a person's mental states would be described as the products of her physical nature and its interactions with her physical surroundings (Cain, 2002: 16). Can physicalism be true of intentional states? Traditionally, this question has been answered in the negative. Roughly, prior to the arrival of cognitive science, the accepted position on this issue was 'dualism': the mind is a *nonphysical* substance; it can have 'no position in physical space' (Fodor, 1980: 114). This implies the denial of mental states since, if a physical system cannot have mental states, then as far as scientific explanation goes, they must be nonexistent. Fodor explains the challenge as follows: if psychologists are justified in using the experimental methods of the physical sciences in the study of the mind, then it must be that mental processes are no different from physical processes; justifying the experimental methods of psychology and cognitive science in general therefore depends on finding an alternative to dualism (Ibid).

Fodor proposes physicalism and embarks on a career-long defence of the possibility of offering a physicalist framework for intentional mental states. Much of Fodor's insistence on informational semantics and conceptual atomism can be best understood in the light of his commitment to physicalism. He chooses informational semantics as the framework for his theory of content because it is a *naturalised* semantics: it accounts for mental content in terms other than those of the *intentional* sciences (i.e., in the terms of the *physical* sciences).[19]

[19] Fodor's commitment to a 'naturalised causal' theory of meaning is reiterated in his recent *LOT 2: the language of thought revisited* (2008: chapter 7).

I now turn my attention more specifically to Fodor's theory of concepts. Limited time and space do not allow a complete overview of Jerry Fodor's entire project. For my purposes, the most important topics are his rejection of the traditional theory of concepts, which I take up below, section 2.5; his arguments against empiricism and in favour of atomism, section 2.6; his proposal of an information-based semantics, section 2.7; and, finally, in section 2.8, Fodor's five criteria for a theory of concepts presented in the late 1990s and some of the work that followed.

2.5 Arguments Against the Definitional Account

In Fodor, Garrett, Walker and Parkes (1980), Fodor and his colleagues argue for an extensive revision of what they call 'the standard picture'. The standard picture has important overlaps with what I have been calling the classical theory of concepts. Both theories not only subscribe to the existence of definitions but also make them play a central role in their explanations. The only difference would be that Fodor et al. are particularly interested in language and consequently argue against definitions by challenging diverse positions held in classical theories of word meaning. At the heart of the problem, according to Fodor et al., is that adherence to the definitional account seems to depend not on direct evidence *supporting* definitions but rather on as-yet-unchallenged assumptions concerning definitions. Notice that Fodor, as so many other researchers interested in theories of concepts, is particularly interested in lexical concepts. The article in question is exclusively focused on (monomorphemic) words and their definitions.

In this section, I first present three of the 'theoretical positions' on word meaning that Fodor et al. draw our attention to. In order to revise 'the standard picture', the authors propose to take a closer look at the exact role that definitions are supposed to be playing within these theoretical positions. Their first objective is to show that they share a common weakness: their reliance on definitions. A second objective is to present a meaning postulate approach as a viable alternative to the definitional

account. Despite the fact that Fodor later rejects meaning postulates, at this stage of his thinking they are quite important.

The first theoretical position Fodor, Garrett, Walker and Parkes (1980) look at claims that the relationship between language and the world is illustrated by the way definitions fix extensions. For instance, suppose we accept that the definition for the lexical term 'bachelor' is 'unmarried man'. The extension of 'bachelor' would be given by the intersection of the extensions of 'unmarried' and 'man'. This, however, is not necessarily a huge step in fixing the extension of the lexical term 'bachelor' since we now have to repeat the process with 'unmarried' and 'man'. Suppose that 'man' could further be analysed into 'human' and 'male' and that 'male', for instance, could no longer be broken down with further definitions. What would fix the extension of 'male'? The definitional account in itself is not equipped to answer this question, which suggests to Fodor and colleagues that the first weakness of this account is that the answers offered are incomplete. As it is habitually presented, the theory would have to be combined with the empiricist claim that all of the lexical terms in a natural language (except those that are already 'primitive') break down into primitives that express sensory-motor properties.[20] However, applying this to existing definitions is far from straightforward. If KNOWLEDGE is JUSTIFIED TRUE BELIEF, how is BELIEF any more sensory-motor than KNOWLEDGE? And could BELIEF be broken down into exclusively sensory-motor features? In fact, both the analysis into primitives and their organisation have proven elusive, so that Fodor and colleagues conclude that definitions actually fail to provide analyses of the vocabulary of a language like English. Perhaps, Fodor et al. suggest, contrary to what the standard picture maintains, there is no difference between words considered definable, such as 'bachelor' and 'knowledge', and the words used to define them, such as 'man' and 'belief'. Perhaps all of these words are equally 'primitive'. This idea is at the heart of Fodor's conceptual atomism, to which I return below (section 2.6).

[20] For empiricists, only sensory concepts are 'primitives', Fodor's proposal distinguishes itself primarily because he proposes that lexical terms *in general* are primitive despite the fact that they might very well be non-sensory. See Fodor 1981(b).

Another theoretical position closely related to the idea that the definition of a word determines its extension holds that to know the meaning of a word is to know its definition; and, similarly, that to understand a word is to have this definition available. This, in turn, presupposes that understanding a sentence involves creating a mental representation of its contents in which expressions are replaced by their definitions. In other words, it holds that definitional analysis is necessary for the process of decoding to take place; for instance, failing to derive UNMARRIED MAN from the word 'bachelor' would result in not understanding a sentence containing this expression. Fodor, Garrett, Walker and Parkes (1980), however, note that there is not much evidence to back up this claim. An experiment by Walter Kintsch, mentioned earlier, compared alleged mental expressions according to whether or not they contained a definition. The idea is that an utterance such as 'John is unmarried' would be processed faster than 'John is a bachelor' since UNMARRIED is part of the definition of BACHELOR. But the results showed that there is no difference between processing one or the other, and, even if a difference were to be found, since the experiment does not control for heuristic shortcuts it would be inconclusive at best.

Finally, Fodor and colleagues consider a third and final 'theoretical position' which holds that definitions underwrite the validity of informally valid arguments. Standard logic is commonly assumed to provide a rational reconstruction of validity intuitions for arguments such as (i) '*John left and Mary wept,'* (ii) '*Mary wept*'. The validity of this argument depends on the definition of 'and', which, as part of the logical vocabulary, is well-defined. The standard picture's proposal is that this treatment could be extended from arguments whose validity turns on the meanings of their logical vocabulary to '*in*formally valid arguments' whose validity turns on the meanings of the non-logical vocabulary. To illustrate this, consider once more 'John is a bachelor'. As stated above, definitional theory holds that understanding a sentence involves replacing expressions with their definitions at a designated level of representation, the 'semantic level'. The resulting expression could be something like '*John is a man and John is*

unmarried'. The same principles of standard logic which licence '…[*therefore*] *Mary wept*', i.e. the conjunction-elimination rule, would apply to these representations, thereby extending standard logic's treatment of the validity of arguments from arguments which turn on logical form to, in theory, *any* complex expression provided it was defined. There would then be really little difference between the validity of an argument such as that from '*John left and Mary wept,'* to '*Mary wept*' and that from '*John is a bachelor',* to *'John is unmarried*'. Both are examples of the schema F & G ⟶ F. Provided that many expressions could be defined, this would offer a systematic approach to validity intuitions in general, or, in other words, a window through language processing to the workings of our inferential apparatus.

This possibility was explored by, among others, Fodor himself in Katz and Fodor (1963). There is, however, a critical flaw in this alleged mechanism: unlike 'bachelor', that can arguably be defined/represented as '*unmarried* ***&*** *male*', certain other relatively well-defined terms, such as '*kill*' ⟶ '*cause to die*' do not contain a term from the logical vocabulary. While above the conjunction-elimination rule licences '…[*therefore*] *Mary wept*', it is not clear *what* would account for the entailment relation between '*John killed Mary*' and '*Mary died*'. Despite the fact that the argument is intuitively valid, the standard picture can offer no explanation. It therefore seems that beyond its unfortunate dependence on definitions there is a further problem: there are *some in*formally valid arguments for which the standard picture account is insufficient. According to Fodor et al. (1980), a further mechanism, not contemplated by the definitional account, is necessary to account for those inferences that, like '…[*therefore*] *Mary died'*, turn on the meaning of a content word (i.e., '*kill*'). Fodor and colleagues consider providing this mechanism as a crucial part of a viable alternative to the definitional account (I come back to the formulation of this alternative later in this section).

To summarise where we have gotten to so far, Fodor, Garrett, Walker and Parkes (1980) present the definitional approach under the most

unfavourable light. Their arguments go beyond the usual acknowledgment that there must be something wrong with the core assumptions of this approach if definitions are so radically elusive. They argue that even if definitions could be found, they would fail to do the work that the definitional account set out for them. It is also argued that it is not the case that definitions straightforwardly fix extensions, that understanding an expression involves the availability of a definition and, finally, that definitions serve to underwrite the validity of informally valid arguments. For all of these reasons, according to the authors, moving past what they call 'the standard picture' critically depends on offering an alternative to definitions. Elsewhere, Fodor also elaborates on the reasons that generally justify abandoning definitional approaches and seeking alternatives. These range from the shortcomings of empiricism, which Fodor takes up at length in Fodor (1981b), presented below in section 2.4, to psychological *im*plausibility arguments that challenge the assumptions concerning the 'semantic level of representation' assumed in the two last theories above (see Fodor, Fodor and Garrett, 1975).

I now move on to Fodor's arguments in favour of meaning postulates as an alternative to the definitional account. Definitions are supposed to licence symmetrical inferences.[21] Given the definition of 'bachelor,' UNMARRIED can be derived from 'bachelor' with a conjunction elimination rule such as (i) BACHELOR ⟶ UNMARRIED & MAN; (ii) BACHELOR ⟶ UNMARRIED. The constraint on this logic is that what remains when a feature is extracted should still be a property. So when UNMARRIED is removed from the definition of BACHELOR, the property feature MALE remains, and when MALE is removed, UNMARRIED remains. The problem with this logic is that there are innumerable cases in which we do not have the *kind* of definition required for this to work. Consider 'red': for 'red' to entail 'colour' in the definitional framework there would have to be something left when you take COLOUR out. But there does

[21] Another way of formulating the same constraint is that implication in a definition must be bidirectional. UNMARRIED & MAN ⟷ BACHELOR.

not seem to be. There is no predicate *P* such that (i) RED ⟶ COLOUR & *P*; (ii) RED ⟶ COLOUR. This is known as the 'residuum problem'. Fodor et al. (1980) propose to solve this problem with the introduction of meaning postulates. Traditionally, an entailment such as RED ⟶ COLOURED or DOG ⟶ ANIMAL, a 'one-way' inference between two lexical concepts, is called a 'meaning postulate'.[22] Meaning postulates do not claim to offer definitions (i.e. necessary and *sufficient* conditions), which makes them immune to much of the criticism aimed at the classical approach. According to Fodor (1975), Carnap (1956) introduced them as a way to simply capture that *X* entails *Y*, independently of a definition, by adding an inference rule. Notice that meaning postulates and logical entailments share the form *X entails Y*; the difference between them is that logical entailments presuppose definitions while meaning postulates do not. Importantly, their dissociation from strict definitions should, at least initially, be seen as an advantage since that means that they can allow for lexically governed inferences, which are ubiquitous in natural language, without the constraint of having to provide the corresponding definitions. Furthermore, notice that 'red' is far from being an isolated example. In fact, there seems to be a correlation between the scarcity of definitions and the ease with which intuitions of validity, such as that 'red' implies COLOURED, and that 'dog' implies ANIMAL, can be captured using meaning postulates. Instead of trying to *define* CHAIR, as I did in chapter 1, I can simply postulate that CHAIR entails FURNITURE, and CHAIR entails SEAT. Finally, given an account integrating meaning postulates, a further argument against the definitional account appears: meaning postulates are more versatile than logical entailments based on definitions because any valid argument that can be captured by a definition can also be captured by a meaning postulate but some valid inferences, such as RED ⟶ COLOUR are problematic for the definitional framework. Definitions therefore appear not to be necessary in capturing formally valid arguments after all, and fail to capture informally valid arguments *while meaning postulates*

[22] Notice that meaning postulates hold between concepts *not words*. It is the concept RED that has a meaning postulate, or an inference rule, attached to it: RED ⟶ COLOUR. The underlying assumption is that words encode concepts

capture both. For Fodor and colleagues, this means that the meaning postulates approach could even be considered preferable to the definitional approach; they argue that,

> so far as questions of validity are concerned, definitions just *are* a special case of meaning postulates. Roughly, they're the symmetrical ones (Fodor et al., 1980: 274).

Occam's razor would recommend keeping only one of the two, so definitions, in the few cases where they are possible (e.g. 'bachelor'), would once again lose out.

Despite all of these advantages, Fodor later decided to abandon meaning postulates. The meaning postulates account is still of interest to my purposes in this chapter because, although it was later rejected by Fodor, relevance theory adopted it early on and continues to defend it today (Sperber and Wilson, 1986/95; Horsey, 2006). I come back to the disagreement between Fodor's current position and classic relevance theory later in this chapter (§ 2.9.3). In the above paragraphs, my objective has been to briefly present Fodor's early arguments in favour of meaning postulates as a background both for relevance theory's current position and Fodor's later change of heart. It is worth noticing that Fodor's reasons for abandoning the meaning postulate approach are, in a sense, the same reasons he had for advocating the approach earlier. In (1998), Fodor argues that an approach where you can simply postulate that *X* entails *Y* cannot be superior to an approach where inferences are *governed* by definitions. The meaning postulate approach extends accounts of validity judgments beyond their original scope, but at a price: the suspension of the constraints that actually *guarantee* the validity of the judgements. For the later Fodor, this means that the alternative ceases to be attractive.

2.6 Arguments Against the Empiricist Account

As mentioned above, one of the weaknesses of the definitional approach is to suppose that complex terms straightforwardly break down into primitives that express sensory-motor properties. This issue is not

addressed by the reformulation of 'the standard picture' above; rather, Fodor treats it as part of the 'innateness controversy' to which he dedicates an article in 1981. In this article, Fodor systematically confronts his own approach, which he now calls 'nativism', with the empiricist account that is supposed to complement the definitional account presented above.[23] He points out that there are several important points of agreement between the empiricist and the nativist accounts: both hold that concepts are either 'primitive' (that is *un*definable) or complex. Both hold that the primitive basis is *innate*, or, in other words, that it is not learned through rational processes. Finally, both hold that possible concepts are constrained by the basic components that can go into concept construction and a limited number of combinatorial devices. But agreement stops here. According to Fodor (1981b), the basic components are lexical concepts: something expressible in English by a 'monomorphemic' predicate term such as 'heavy' or 'chair' or 'green'; and the other type of concept is a 'phrasal' concept such as MEN WITH SHAVED HEADS. According to the definitional account, lexical concepts can be either primitive or complex. The role that definitions play is to express the relations between complex lexical concepts and the primitive basis of which they are constituted. Consequently, the crucial disagreement centres round concepts such as BACHELOR. For the definitional account, BACHELOR is definable and so falls on the 'complex' side of the divide while for Fodor, it falls on the 'simple' side of the divide since it is a lexical concept.[24]

From the perspective of theories of concept learning, this makes a difference. Fodor considers two possibilities: either the approach to concept learning is 'empiricist' or it is 'nativist'. According to the empiricist approach, the fact that BACHELOR is complex means that it can be learned by

---

[23] Fodor refers to this empiricist account as 'Empiricism' with a capital *E* and sets out its tenets with very few citations from its individual advocates (he does, of course, mention, Locke and Hume, for instance). For more contemporary versions of empiricism, he mentions particular examples, like George Miller's account of certain verbs used as nouns, see below.

[24] For Fodor, almost all lexical concepts fall on the 'simple' side of the divide - that BACHELOR is defin*able* is unimportant since this is not how it is represented in our cognitive systems. I take up further arguments in favour of Fodor's conceptual atomism in a later subsection.

somehow putting together other (simpler) concepts that already exist in the subject's conceptual system. This is a rational process because it involves formulating hypotheses and potentially confirming them.[25] Fodor illustrates how this is supposed to work with an in-laboratory discrimination-learning task. Suppose you are the subject of the experiment: you are presented with cards on which there is a geometrical form of a certain colour and told that the target concept you are to learn is 'FLURG'. A trial consists of the experimenter presenting a card, your judgement as to whether or not you think it is FLURG and feedback. Suppose that for the first trial you are shown a green triangle and you correctly guess FLURG. This gives you evidence for a certain range of hypotheses: for instance, 'all cards containing a green shape are FLURG' and/or 'all cards containing a triangular form are FLURG'. You are soon given the possibility to test your hypothesis on a new trial. The experimenter now presents a red square and since you see no green shape nor triangular form, you answer 'non-flurg' and are given 'wrong' as feedback. The idea is that through a series of trials, hypotheses are formulated and rejected or confirmed in response to experience until you reach 'criterion'. The problem that Fodor points to is that there is a certain circularity: if in order to confirm the hypothesis that the FLURG cards are GREEN OR SQUARE you need to *formulate* a hypothesis *containing* the concept GREEN OR SQUARE, then the concept is already available to you, *that is, you already have it*, before confirming the hypothesis. This means that these kinds of tasks, presented as paradigm cases of concept learning, are not really cases of concept learning. Fodor (1981b) describes them as cases of 'belief-fixation'. You surely already possessed the concept GREEN OR SQUARE before the experiment started, all you have learned is that the experimenter chose to call green or square things "flurg" *for the purposes* of the experiment.

[25] Fodor is here construing learning as a rational process of hypothesis formation. Other theorists in later sections will have very different conceptions of 'learning'.

What does empiricism say about primitive concepts? As mentioned above, empiricism holds that primitive concepts are *un*definable and that the primitive basis is innate. Since primitive concepts are by definition unstructured, they cannot be learned by the rational processes described above; rather, they are learned by a different kind of causal process that Fodor (1981b) describes as 'brute-causal'. The idea is that 'the structure of the sensorium is such that certain inputs trigger the availability of certain concepts. Punkt' (Ibid, p. 273). This position, Fodor notes, is actually quite close to nativism, which also holds that there is a mechanism behind concept acquisition that can be described as brute-causal. The difference, of course, is that for empiricism this mechanism is only at work for sensory concepts whereas for Fodor's nativism:

> the triggering of the sensorium is, normally, causally necessary and sufficient for the availability of *all concepts* except those that are patently phrasal (Fodor, 1981b: 273, my emphasis).

To illustrate, imagine the attainment of the concept TRIANGLE. The empiricist story would have the concepts LINE and ANGLE become available directly through sensory stimulation. TRIANGLE would then be built out of these primitive concepts. The nativist story is that there is no reason to believe that experience can occasion concepts like LINE and ANGLE and not occasion TRIANGLE. It holds that the brute-causal processes of the classic empiricist account can be extended to explain non-sensory vocabulary. Of course, this implies a parallel extension of the innate specifications that relate lexical concepts to their occasioning experiences. Trouble in determining what these might be is probably what later led Fodor to attempt a retreat from this position, labelled 'radical nativism', in his (1998) reformulation of innateness. However, despite these more recent reformulations, it is the earlier version that underlies Fodor's best-known account of concepts and so the one I choose to discuss.

Fodor (1981b) attracts attention to the advantages of his nativist story. First of all, it offers an approach to concept attainment that no longer depends on concepts having definitional structure. Thus, arguably, it

integrates some of the lessons learned from the failure of the standard classical theory of concepts presented in the previous chapter. Basically, the consensus regarding research into definitions was that subjects do not automatically have the necessary and sufficient conditions for concepts that were believed to constitute their definitions. Therefore, they could not possibly use them in the hypothesis formulation and confirmation paradigm above to learn new concepts. For instance, going back to the TRIANGLE example, the subject would have to formulate a hypothesis in which the definition of TRIANGLE appeared in order to learn the concept. This might not seem problematic for TRIANGLE since there is intuitively a way to build it out of primitive concepts such as LINE and ANGLE, but it would be quite challenging for tens of thousands of other words for which definitions would have to be postulated (e.g. SHIP, TRUMPET, BURGER, ELECTRON). Another advantage is that, from this perspective, it would naturally follow that the reason the conceptual repertoire is constant despite the large variation between individuals in a given species is that *the make-up of the sensorium* is common to the species. As Fodor puts it:

> the concept isn't coming from the environment, it's coming from the organism. All the environment does is provide triggers that release the information (Fodor, 1981b: 280).

Finally, Fodor (1981b) also argues against the contemporary version of the empiricist-plus-definitional account presented above. He acknowledges that there is a weaker version of the story in which the primitive basis is not strictly made up of sensory concepts. According to this perspective, the question of which concepts are primitive becomes an empirical matter and the theory simply states that it probably includes a logical syntax and a to-be-determined framework of basic concepts. Fodor is convinced that the evidence is overwhelming even against this weaker version and offers a worked out example in order to convince the reader. The example he takes up was originally proposed by George Miller in a paper published in 1978. Miller's claim is that when nouns such as 'butter', 'dye', 'grease' (named type *M* nouns) are used as verbs, the meaning of the

verb can be rendered as '*x* covers the surface of *y* with *M*'. A paradigm example of type *M* nouns is 'paint'. According to Fodor, this is a proposal for defining the verb 'paint$_{TR}$' in terms of the noun 'paint$_{N}$' together with part of the presumed framework: the concepts COVER, SURFACE and WITH. Using the noun as part of the definition is supposed to considerably simplify the task, yet Fodor holds that despite this, the definition does not work: '*x* paints *y*' is not coextensive with '*x* covers *y* with paint'. First of all because the candidate definition is true of an event in which a paint factory explodes and covers the spectators and surroundings with paint. But, according to Fodor, this is not a case of the factory, or the explosion, *painting* the surroundings. Perhaps this flaw could be corrected by adding the condition that '*x*' denote an agent or, in other words, that the covering aspect of the definition be *intentional*. But according to Fodor, this move would still be insufficient because there would still be cases where *an agent intentionally covers the surface of y with paint* without it being the case that *the agent painted y*. He gives the example of Michelangelo and the Sistine Chapel. Michelangelo covered the surface of the ceiling with paint in the process of painting a picture on the ceiling. Fodor finds it inappropriate to describe this as *Michelangelo painted the ceiling* suggesting that this simpler description is rather only suitable for a house painter. He suggests that to further patch the definition up, it would now be necessary to add the concept PRIMARY INTENTION OF AN ACT in order to distinguish between the great master and the common painter. But this last addition would surely prove too costly. Recall that the interest of definitions is, in part, to help concept learners (i.e., children) build their concepts up out of innate, or in general, simpler, elements. If the definition has to include PRIMARY INTENTION OF AN ACT, this advantage is seriously jeopardised. Furthermore, according to Fodor, even this complex definition would not necessarily be sufficient to correctly pick out the cases of 'x paints y' since Michelangelo can also be said to have dipped his paintbrush in cerulean bleu and therefore to have *intentionally covered its surface with paint* without having *painted his paintbrush*. Again, the conclusion is that we can give necessary *but not sufficient* conditions on

the meaning of the word and so the approach in terms of meaning postulates is preferable.

To sum up this section, Fodor has argued for a theory of concepts in which a great many concepts are unstructured. His suggestion for a new theory of concepts is that the divide between primitive and complex lexical concepts be abandoned following the observation that those lexical concepts ordinarily considered to form part of the group of complex concepts, are actually more compatible with inclusion in the primitive basis. This claim is central to Fodor's conceptual atomism, one of his most important contributions to theories of concepts. In upcoming sections, discussions of other aspects of Fodor's theory of concepts and meaning will make the claims above clearer. The section immediately following takes up Fodor's semantics.

2.7 Fodor's Information-Based Semantics

Fodor's information-based semantics builds on earlier views of semantics which hold that the referential power of concepts is due to the information they carry. In Fodor's own terms,

> "Carrying information" is a relation that is best introduced by examples, so here are some popular ones. In typical cases: smoke carries information that there is fire; a tree's rings carry information about its age; a falling thermometer carries information that it is getting cold; utterances of the form of words "that is a platypus" carry the information that that is a platypus; and so forth (Fodor 1990a, reprinted in Margolis & Laurence 1999: 514).

Moreover, the relations between symbols and things symbolised are relations of causal covariance. 'Dog' tokens are caused by dogs and we can therefore say that a symbol, or a token of a symbol, carries information about its reliable cause. The correlations are supposed to be lawful so that *all* 'dog' tokens mean dog.

Fodor cites the work of B. F. Skinner as a source for information-based semantics. As stated in the introduction to this chapter, Fodor is

strongly committed to physicalism and, consequently, he agrees with Skinner's account *in so far as it offers a naturalistic theory of meaning*, that is, a naturalistic way to break the intentional circle. For this purpose, Fodor offers a 'slightly cleaned up' version of Skinner's account of meaning. To the question of why the word 'dog', for instance, expresses the property of *being a dog*, this account answers that the 'behavioural disposition' to the verbal response 'dog' is 'under the control of' a certain type of discriminative stimulus (dogs, of course). Also, as would be expected, the probability (or frequency) of the response increases with the presence of this stimulus, which is in accordance with classic operant conditioning.

Fodor does not doubt that the behaviourist program, of which Skinner was a prime advocate, has been judiciously shelved following Chomsky's (1959) review of *Verbal Behaviour*. He simply points out that, while it was thoroughly effective in arguing against Skinner's learning theory, Chomsky's review leaves Skinner's semantics 'untouched'. Typically, Chomsky noted, utterances are not responses but actions and consequently uttering 'dog' depends more on the contingencies of a conversational context than on the presence of a dog. But this observation, according to Fodor, does no damage to the semantics underlying Skinner's view of language. These two aspects of Skinner's theory are logically independent: it is possible to give up his ideas on social reinforcement while holding on to the idea that tokenings of 'dog' express the property *dog* because they are under the control of instantiations of dog. A revised version of Skinner's theory of meaning would then hold that there are mental states with intentional objects. In the case discussed here, the mental state consists of entertaining the concept DOG and the intentional object is the property *dog*. The mental state *expresses the property dog*.

This 'updated Skinnerian semantics' is far from being in line with the behaviourist program, most notably because it postulates intentional mental states. As mentioned earlier in this chapter (§ 2.2.2), Fodor is strongly committed to commonsense belief/desire psychology, which gives a central role to intentional states, because he believes that there is no better explanation for our behaviour. Fodor's 'updated Skinnerian semantics' also

contradicts the basic behaviourist program by denying the importance of learning theory. For Fodor, but not for Skinner, it does not matter *how* the relation between the symbol and the property comes about:

> This account [updated Skinnerian semantics] isn't behaviouristic since it's unabashed about the postulation of intentional mental states. And it isn't learning-theoretic since it doesn't care about the ontogeny of the covariance in terms of which the semantic relation between dog-thoughts and dogs is explicated (Fodor, 1990b: 56).

In learning-theoretic accounts, on the contrary, the ontogeny of the relation is critical because *that* is what Skinner's theory of operant conditioning is explaining. But Fodor follows Chomsky in denying that operant conditioning is playing the role Skinner thought it was.[26]

Fodor insists that what does matter is that the relation between the symbol and the property, or, in Skinner's terms, the organism's 'response disposition', does not depend on the organism having any other response disposition. In other words, the organism that is disposed to utter 'dog' in the presence of – and only in the presence of – dogs need not have a disposition to utter 'cat' – even in the presence of a cat – or any other disposition. This point is crucial to Fodor's conceptual atomism. In the previous section, we saw the first tenet of conceptual atomism: most lexical concepts are unstructured and therefore *un*learned (which implies that they are, in some sense or other, innate). Now, we come to the second tenet: all that matters for meaning is 'functional' relations, that is, the relations of covariance between symbols and their denotations. The relations are deemed 'functional' in that they are limited to 'relations of nomic covariance' instead of *mediated* – as held by the semantics of, for instance, Quine (Fodor, 1990b: 56). Quine, along with many if not most philosophers of language, would argue that theoretical inferences mediate the application of our

[26] Notice the stark contrast between Skinner's account of learning in terms of operant conditioning and Fodor's account in terms of rational processes of hypothesis formation. A key aspect of Skinner's account was how he thought the relation between the symbol and the property came about, that is, how it was developed. However, for Fodor, Skinner was wrong about learning and, therefore, he rejects that part of Skinner's theory.

concepts. To illustrate, according to Quine, applying PROTON to protons would be 'theory mediated' since what one means by PROTON is at least partly determined by what one believes of protons. In the inferential-role semantics tradition that evolved from Quine's observations, this can be true even for patently simple concepts such as DOG. In the structuralist version, DOG is a symbol in a symbol system and only the system as a whole has meaning. Individual symbols get their meaning derivatively through differentiating relations with other symbols within the system. To illustrate, consider that DOG is partly individuated by *not* being CAT, PARROT, or BASEBALL. For inferential-role semantics, emphasis is placed on the inferences that a person must be disposed to draw if they possess the concept. For instance, from *x is a dog* to *x is an animal*. The empirical support offered in favour of this position is that we do not consider someone who does not know the difference between a dog and a parrot to possess the concepts DOG and PARROT just as we would not consider someone who does not know that a dog is an animal to be in full possession of the concept DOG.

For Fodor, however, it is of the utmost importance to distinguish between these 'truisms' and deeper issues of concept possession conditions. He holds that inferential-role semantics and any 'anatomistic' theory of concepts leads via a slippery slope towards 'ruinous' holism (Fodor and Lepore, 1992; Fodor, 2003). Any move towards accepting inferences as content-constitutive risks ending in having to accept endless possession conditions for even the simplest of concepts. So, for instance, Fodor's atomism holds that having the concept DOG is *independent* of having any other concept, even the concepts CAT or ANIMAL. Accepting that some inferences belong to the possession conditions for DOG, on the other hand, would see ANIMAL as *part of* the concept DOG; in other words, the inference DOG ⟶ ANIMAL would be considered (potentially) necessary to the possession of DOG. The worry is that if ANIMAL is necessary, FURRY, BARKS or BITES POSTMAN, and so on quasi-indefinitely, could be equally necessary. According to Fodor, the only way to block such a deluge would be by overcoming Quine's epistemological objections to the analytic/synthetic distinction. Or, in other words, by finding a principled distinction between

those inferences that are 'constitutive' of conceptual content and those that are merely 'collateral'; for instance, between 'no unmarried man is married', true in virtue of the meaning of 'unmarried' and, other inferences that do not so clearly express a truth, such as 'unmarried men are eligible for marriage', or 'some unmarried men are lonely' which are true or false depending on facts about the actual world. This is not a hopeful answer because not only does no one have such a solution, but, according to Fodor, 'no one has a clue how to put one together' (2003: 151, see also Fodor, 2004).[27] I come back to this issue when I present Richard Horsey's proposal for dealing with Quine's objections and re-establishing a modified (and psychologised) analytic/synthetic distinction (§ 2.7.3). For now, it is simply important to note that at least part of Fodor's argument in favour of conceptual atomism is his conviction that accepting any content-constitutive inferences leads inevitably to semantic holism which holds that *all* of the inferential relations a concept can enter into are constitutive of its content. To avoid this, Fodor stresses the fact it must be possible to have dog-thoughts which covary with dog instances *independently* of any other thoughts. The conditions for meaning must be able to be satisfied by symbols that do not belong to symbol *systems* (Fodor, 1990b: 56).

To summarise the preceding sections on Fodor's view of concepts, in the first part of the period covered so far, Fodor mostly contrasted his proposal with the view of conceptual content based on the alliance of the definitional view and classical empiricism. To recap, the key issue for Fodor is that, contrary to the received view, lexical concepts are not decompositional, and,

[27] As hinted here by Fodor, the issue of whether or not there is an analytic/synthetic distinction is far from a settled matter in philosophy. This is unsurprising given that establishing how knowledge is possible is at the heart of philosophical investigation. Traditionally, two camps have opposed each other: the rationalists who hold that knowledge is achievable through reason alone and the empiricists who respond sceptically to the rationalists and argue that only through experience can knowledge be attained. In the 18th century, Kant made a lasting contribution to the issue with, among other things, his notion of analyticity: not all knowledge derives from experience because there are self-evident truths that we can know *independently of experience*, in other words, a priori. For traditional accounts of concepts, this notion of a priori truths translated as content-constitutive inferences for concepts such as 'bachelors are unmarried'. In this view, failing to infer that *a bachelor is unmarried* is simply failing to grasp the concept BACHELOR.

therefore, proposing that lexical concepts are structured like phrasal concepts cannot be the answer to where conceptual content comes from. The arguments presented thus far come together as follows: the assumption that concepts are structured critically depends on concepts having definitions but the evidence against definitions is overwhelming. Perhaps more importantly, even if definitions were somehow to be found for our lexical concepts, this would still fall short of answering the question since the origin of the primitive concepts out of which complex concepts would be the constructs would remain intact. Fodor's proposal is to dismiss the assumption of decompositional lexical concepts and adopt atomism plus nativism. At this stage, this solution is supposed to work as follows: concepts are either primitive or complex. Complex concepts are 'logical constructs out of primitive concepts'; they are structured more or less as conceived by the received view but, contrary to this view, do not include monomorphemic lexical concepts such as BACHELOR which are unstructured and thus fall on the *simple* side of the primitive/complex divide. Primitive concepts are unstructured atoms so that, again contrary to the received view, they cannot be learned through hypothesis formation. On this basis, Fodor offers his first formulation of nativism: contrary to traditional thinking on concept acquisition, monomorphemic lexical concepts are *attained*, rather than *learned*; they are triggered by the environment as part of a process best described as 'brute-causal'.

In a second phase, roughly covering the decade between Fodor's criticisms of empiricism and his writings on conceptual content around 1990, Fodor's criticisms shift from the deficiencies of the definitional view to the dangers of inferential-role semantics. His focus becomes the development of his own information-based semantics. To this end, he proposes the replacement of content-constitutive inferences with an account of conceptual content the main characteristic of which is that the possession of any specific concept does not depend on possessing any other.

In the next section, the last one on Fodor's contributions, I focus on a third period in Fodor's development of his theory of concepts. This period can be

said to start with the publication in 1998 of *Concepts: where cognitive science went wrong*. In continuation with the second phase, Fodor argues against inferential-role semantics, but something critical has meanwhile changed: the view he rejects has been widely adopted by cognitive scientists and philosophers of mind. In response, Fodor reassesses his project and reformulates certain aspects of it. As before, limits of time and space do not allow a full review of the contents of this work and those that further develop its topics in the following years. I focus only on a couple of key points central to Fodor's project and important to my own.

2.8 Fodor's 5 Criteria on a Theory of Concepts

The publication of *Concepts: Where cognitive science went wrong* in 1998 is a key moment in Fodor's career-long contribution to thinking on concepts.[28] It's foremost objective is to develop the argument against inferential-role semantics; according to Fodor, the very wide adoption of this account of meaning in the cognitive sciences is, in fact, as worded in the title of the book, 'where cognitive science went wrong'. In pursuit of this goal, Fodor both reiterates his main arguments in favour of information-role semantics and conceptual atomism – which he now brings together under the label of *informational atomism* – and, importantly, reassesses what and who he is arguing against. As he has argued before against behaviourism, empiricism, and inferential-role semantics, he now argues against the version of inferential-role semantics having been widely adopted in cognitive science, which he labels *concept pragmatism*. The main aim of the book is therefore to mobilise much the same arguments as before while highlighting *how* they work to discredit a conglomerate of theories emerging as dominant, and thereby put the cognitive sciences back on the right track. Part of this disagreement was touched on earlier in this chapter (section 2.3 on

[28] Fodor's (1998) *Concepts: where cognitive science went wrong* was followed by a couple of other publications that also focus on contrasting conceptual Cartesianism and conceptual pragmatism, for instance, (2003) *Hume variations* and (2004) 'Having concepts: a brief refutation of the twentieth century'. Time and space constraints do not allow a full presentation of these publications, but they are mentioned when particularly relevant.

'Incompatible perspectives on concepts?'). There, I said that Fodor's dissatisfaction with the theories he labels concept pragmatism, the reason they cannot answer the leading questions regarding concepts, is chiefly to do with the way concepts are framed in these approaches: putting epistemological questions before *or worse* instead of metaphysical ones leads to conceiving of concepts as *capacities*, which they are not. The book argues for a different conception of concepts with the help of five 'non-negotiable' criteria *any* theory of concepts should respect but that are, in fact, only collectively satisfied by Fodor's own account. The accounts of prominent philosophers, linguists, and psychologists are presented with a focus on how they fail to meet these criteria; and this is followed, at the end of the book, with some revisions Fodor offers of his own previous position, namely an attempted retreat from radical nativism.

In a nutshell, Fodor's position is that, despite all the research on concepts carried out over decades, and the enthusiasm expressed by many in the cognitive sciences, little progress has been made: *we still need a theory of concepts* and 'none of the theories of concepts that are currently taken at all seriously either in cognitive science or in philosophy can conceivably fill the bill' (Fodor, 1998: 23).

Fodor's five criteria on an adequate theory of concepts are listed below, I have reordered them with respect to his presentation so as to first mention those points that are less controversial and with which most accounts, including my own, have little or no trouble complying, although compatibility is always only partial. Then I list those where more careful considerations would be necessary. (i) *Concepts function as mental causes and effects*; (ii) *Concepts are categories and are routinely used as such*; and, (iii) *Quite a lot of concepts turn out to be learned*. These first three criteria are quite frequent assumptions of contemporary theories of concepts and are, in my view, generally unproblematic. Future sections will reveal how my own approach deals with them where there is any substantial difference with Fodor's account. For instance, my own account of 'concepts are categories and are routinely used as such' is different from Fodor's but there is an underlying fundamental agreement; I devote a whole chapter (chapter

4) to categories and categorisation and return there to the discussion of the relation between concepts and categories. My overall view of concepts also has important consequences for issues of learning; I turn in particular to the issue of the acquisition of word meaning in chapter 3 (§ 3.5.4). Fodor's remaining two conditions are more frequently the nexus of disagreement: (iv) the 'publicity constraint': *concepts are* public*; they're the sort of things that lots of people can and do* share; and, (v) the 'compositionality constraint': *concepts are the constituents of thoughts and in indefinitely many cases, of one another* (Fodor, 1998). According to Fodor, alternatives to his account regularly fail to meet the compositionality constraint; I would argue, however, that this is so only because of Fodor's very rigorous conditions on concept individuation. It is Fodor's particular construal of (iv) which sets his general approach apart and makes compatibility between his approach and its alternatives only partial.

In Fodor's approach, following classic representational theory of mind means embracing the view that concepts are symbols and that they are therefore 'presumed to satisfy' type-token relations; it follows that saying of two people that they *share* the concept RAIN, for instance, means that they have *tokens* of literally the same concept *type* (1998: 28-29). This position explicitly denies any possible variation between two *shared* concepts: so for instance, my concepts RAIN, WATER, DOG, or TRIANGLE, and Fodor's have to be identical; anything less than this, is branded a sort of *relativism*. Any theoretical or experimental procedure that distinguishes between two subjects' corresponding concepts is therefore simply judged as having it wrong on concept individuation. Furthermore, Fodor has the validity of intentional explanation, *anywhere* it is found, depend on his rigorous construal of content identity: the generalisations of representational theory of mind (on the model of 'Thirsty people drink water') are only valid insofar as the mental contents of these generalisations (among them WATER, for instance) really are shared by those the explanation is designed to cover.

According to Fodor, if concept identity were in any way relaxed, then everyone else's concept WATER could be different from my own; furthermore, my own *now* could be different from my own *at another*

moment in time; it would therefore only ever be me *now* who, for instance, wanted a drink of water (1998: 29). Finally, Fodor holds that there is no way out of this; trying to replace content identity with content *similarity* only postpones the problem since any construal of *similarity* depends on an underlying notion of *identity* (Fodor, 1998: 30-34).

This very rigorous conception of concept identity has direct consequences for Fodor's view of publicity and compositionality: roughly, we all share the same concept RAIN and this concept is a constituent of the *belief* that it will rain, the *hope* that it will rain, and *any other* RAIN related thoughts. If in these thoughts, RAIN contributes the same contents, then accounting for the *productivity* and *systematicity* of thought is relatively straightforward. In a previous subsection of this chapter, I mentioned Fodor's particular interest in accounting for two observations: that thoughts (beliefs, hopes, etc.) are *productive* and *systematic*: that is, that 'there are an indefinitely many distinct [thoughts] that a person can entertain (given an abstraction from 'performance limitations')' and that 'the ability to entertain any one of them implies the ability to entertain many others that are related to it in content' (Fodor, 1998: 26, see also Fodor 2001). The reason having the thought that *Jim punched James* opens up the possibility of having the thought *James punched Jim* is, in this perspective, partly because *PUNCHED* can be counted on to contribute strictly *the same* content to both these thoughts. The unbounded capacity for different thoughts would follow from the fact that concepts like PUNCH are building blocks that can be recombined in ever novel ways. Likewise, intentional explanations on the model of 'If X is thirsty X seeks water', and 'X is thirsty, Y is some water, X will seek Y', are also given a solid foundation by the fact that the building blocks, such as THIRSTY, contribute the exact same content in all the thoughts they appear in. Notice the weight placed on conceptual content: thoughts can be productively and systematically recombined because, following Fodor's strict construal of concept individuation, they always contribute *the same* content to the thoughts in which they appear; and intentional explanation is justified because it can be coherently described within the representational theory of mind framework.

Fodor's account of the compositionality of thought certainly has the advantage of simplicity and elegance, but I am particularly interested in issues to do with word meaning and communication; and, naturally, I wonder whether Fodor's compositionality of thought translates into a compositionality of language. More specifically, how are the constituents of an utterance supposed to correspond to the constituents of the thought the utterance is intended to communicate? It seems that the very *nature* of language and communication imposes some constraints, so, for instance, communicative efficiency is such that a speaker can be inexplicit about part of her message and trust her hearer to fill in some blanks. For Fodor (2001), however, natural languages are not compositional in a more fundamental sense:

> My point is that a perfectly unelliptical, unmetaphorical, undeictic sentence that is being used to express exactly the thought that it is conventionally used to express, often doesn't express the thought that it would if the sentence were compositional. Either it vastly underdetermines the right thought; or the thought it determines when compositionally construed isn't in fact, the one that it conventionally expresses (Fodor, 2001: 12).

What's important here is the partial compatibility between this position (language is not compositional in the way thought is) and the approaches to utterance (and word) meaning of contemporary cognitive pragmatics that interest me in this thesis. Fodor's (2001) position can be said to parallel one of relevance theory's main tenets: linguistic underdeterminacy, or, in other words, the fact that there is an inevitable gap between what one means and what one says. Fodor, however, is less than consistent with his construal of possible differences in compositionality of thought and language: in a 2004 article he seems to revert to the claim that the principles of compositionality (productivity and systematicity) governing thought also apply in public languages:

> Language (/thought) is productive and systematic because it is compositional… if being able to say (/think) that John loves Mary implies being able to say (/think) Mary loves John, that's because both

> sentences (/thoughts) are made out of the same set of primitive elements by the application of the same constructive rules (Fodor, 2004: 37).

My own position is that building an account of how compositionality works in public languages requires adopting a truly *inferential* model of communication and taking a very different perspective on concepts. My starting point is relevance theory and so I turn my attention, in the remainder of this chapter, to their view on concepts and word meaning. Critically, relevance theory claims to work within the general framework set out for cognitive science by Fodor (i.e., representational theory of mind, modularity)[29] and explicitly adopts the atomic view of lexical concepts and broadly Fodorian semantics; yet, at the same time, as a specialist discipline, relevance theory brings its own insights into word meaning and concepts, and so parts company with Fodor on some key issues.

2.9 Concepts and the Inferential Model of Communication

2.9.1 The Correspondence Between Concepts and Words

The first question I address in this section is the correspondence between the concepts in our thoughts and the words we use to express them in our natural languages. And, in particular, the question of the nature of the word/concept mapping [one-to-many, one-to-one, or many-to-one].[30] These questions are complicated by the fact that words on the one side and concepts on the other are far from homogenous. On the side of words, it is obvious that a distinction must be made between content words such as 'chair' and 'apple'; pronouns such as 'you' and 'he'; and grammatical words

---

[29] Much could be said about relevance theory's stance on modularity, but since this is not the main topic of this section, I only briefly mention here the shift from a previous position that inferential comprehension involves no specialised mechanisms (Sperber and Wilson, 1986-95: 65), to their current position that there is a specialised pragmatics module (See Sperber 2001; Sperber and Wilson 2002; Sperber 2005; and, Wilson 2005).

[30] Ultimately, the question will be *whether* words <u>*encode*</u> concepts (see chapter 3, section 3.5 'The second stage: abandoning the modular view'). But in order to respect a certain logic of presentation, for the time being, the question is how words encode concepts, or, perhaps less tendentiously, the *correspondence* of word forms and concepts.

such as 'the' and 'a'. Of these, only content words seem to be candidates for encoding concepts. Diane Blakemore (1987) proposed a conceptual-procedural distinction to reflect the fact that some words encode conceptual content and others encode procedures. In other words, while content words encode concepts, discourse connectives such as 'also', 'so' and 'after all' seem rather to offer guiding support to the hearer for the inferential aspect of interpretation. This view has been substantially developed since Blakemore's early work and the case for a *procedural semantic* analysis of a range of linguistic phenomena including pronouns, mood markers, particles, demonstratives and interjections has been made (see Wilson & Sperber, 1993; Wharton, 2003; Scott, 2013; for a survey and update, see Wilson, 2011). On the side of concepts, a classic distinction involves differentiating between *lexical* and *phrasal* concepts. The first can be expressed in a natural language with a single word and the second are expressed by a phrase. This is usually illustrated with the example of SIBLING which in English would be a lexical concept and in French a phrasal concept FRÈRE OU SŒUR.

Another point of interest in an inferential model of communication is that while some words clearly do not encode a concept, such as the third person placeholder 'it' (much less the use of 'it' in structures like 'it seems that'), other words, which do encode some conceptual component seem nonetheless not to encode a 'full-fledged' concept.[31] Sperber and Wilson give 'my', 'have', 'near', and 'long' as examples of words that have a certain conceptual content but can be compared to pronouns in that they depend on the context in order for their contribution to be fully specified (1998, reprinted in Wilson & Sperber, 2012: 32).

Let's now set aside words that are claimed to encode procedural meaning or pro-concepts, and then, confining our attention to the remaining words,

[31] Sperber and Wilson (1998, reprinted in Wilson and Sperber, 2012) suggest these might be considered pro-concepts; Carston (2002) suggests that this incompleteness of atomic concepts could possibly extend to the majority of concepts. I come back to this point at length in chapter 3 (§ 3.3.2.4 and § 3.3.3).

consider the issue of word-concept mappings. What possibilities are there for the word/concept mapping? The first possibility to consider is based on the well-known, though often rejected, thesis of conceptual structure. It holds that most lexical concepts are structured, even *definitional*, so that the concept-word mapping is not one-to-one (e.g., 'bachelor' encodes UNMARRIED ADULT MALE, and 'spinster' encodes UNMARRIED ADULT FEMALE).

Fodor has long forcibly argued against this particular view of concepts, preferring to it the conceptual atomism I presented earlier in this chapter. According to Sperber and Wilson, in Fodor's view, mismatches between the concepts in our minds and the words we use to express them are so rare, that adopting a second possibility is warranted: roughly, that, despite the difficulties mentioned above, the mapping is one-to-one between concepts and words.

Sperber and Wilson further hold, however, that before accepting this proposal, important insights from pragmatics in general, and relevance-theoretic pragmatics in particular, should be weighed in. They claim that there is an unacknowledged middle step between Fodor's rejection of the first possibility, and his support for the second: he seems to ascribe a role to the encoded contents of the words uttered that reveals a remnant of the code model of communication. I would add that if this is so, Fodor can at least be taken to adopt an *updated* version of the code model of communication since he makes some room for inferences in completing what a sentence communicates. He explicitly cites H. Paul Grice's insights in outlining a theory of communication although he suggests that his account is 'Gricean in spirit though certainly not in detail' (Fodor, 1975, footnote 3, pages 103-104). Sperber and Wilson's point, however, is that he falls short of accepting the thesis relevance theory defends, roughly, that what is communicated simply *cannot* be fully encoded. They cite a passage from Fodor's (1975) *Language of thought* capturing his view of language:

> A speaker is, above all, someone with something he intends to communicate. For want of a better term, I shall call what he has in mind a message. If he is to communicate by using a language, his problem is to construct a wave form which is a token of the (or a) type

> standardly used for expressing that message in that language (Fodor, 1975: 106. Cited by Sperber and Wilson, 1998; reprinted in Wilson and Sperber, 2012:34).

In this quote, according to Sperber and Wilson (1998), there is something reminiscent of the old code model of verbal communication in which linguistic sentences correspond to thoughts in a straightforward manner.[32] The assumption is that a thought is transmitted in the message because it has been *entirely* captured by the signs of a particular code. In this perspective, communication is successful simply when both interlocutors share a single code. Updated versions of the code model allow for some inferential processes, but these are not seen as essential to the comprehension process, as they are in relevance theory, but rather only as convenient short cuts.

Basically, in the code model and in its updated version, all of the communicable thoughts must be *encodable* because it is primordially through decoding that a message may be reconstructed. Fodor further assumes that most single lexical items straightforwardly express atomic concepts (e.g., 'chair' expresses the atomic concept CHAIR, 'walk' expresses the atomic concept WALK) and, therefore, that most lexical concepts are 'simple'. For the traditional linguistic view, however, there are also lexical items which encode complex concepts (e.g., 'grandmother' is taken to encode 'FEMALE PARENT OF PARENT'). For Fodor, these latter types of examples can be merged with the 'simple' lexical items (so 'grandmother' encodes the atomic concept GRANDMOTHER). Finally, phrasal concepts, such as GRANDMOTHERS MOST OF WHOSE CHILDREN ARE DENTISTS, must be expressed by a phrase in a natural language, such as 'grandmothers most of whose children are dentists'. Based on this, the conclusion would be that most (non-phrasal) concepts are in a one-to-one relationship with words.

---

[32] It is worth noting that Fodor (2001) seems to move away from his (1975) view since he acknowledges underdeterminacy:

> … language is strikingly elliptical and inexplicit about the thoughts it expresses'; and he adds, 'though, to be sure, it manages to express them all the same (Fodor, 2001: 11).

To these examples, Sperber and Wilson add cases where a word communicates something different, perhaps something more specific, than it encodes. Imagine the following context: Peter and Mary are going to the theatre; it is now about time to leave:

(1) Peter: Are you ready to go?
Mary: I need a minute.

The first thing to notice is that Mary's answer does not directly answer Peter's question. Decoding the linguistic evidence provided by Mary will not get Peter very far. After all, a minute refers unambiguously to 60 consecutive seconds, but this is unlikely to be what Peter takes Mary to mean. It might occur to him that Mary needs more time *before she is ready to leave*. In the terms of relevance theory, Peter's expectations of relevance, awakened by Mary's ostensive communicative behaviour and constrained by the general presumption of optimal relevance carried by all utterances, yield assumptions that Peter can use to interpret Mary's utterance as an answer to his question.[33] It follows from the presumption of optimal relevance that he has the right to suppose that the extra cognitive effort being asked of him by Mary's indirectness will pay off in the form of increased cognitive effects.

Let's consider Peter's interpretation of 'minute'. Peter adjusts 'minute' in order to warrant the inferences he has drawn (e.g., *Mary needs more time; with more time she will be ready to leave*). He assumes that the amount of time she needs is fairly small and broadens the notion expressed by 'minute' to perhaps something like ENOUGH TIME TO FINISH WHAT SOMEONE/MARY IS DOING.[34] The role of the word 'minute' in Mary's utterance is no more than *a piece of evidence* that points in the direction of her intended meaning. This is the heart of relevance theory's inferential model of

---

[33] The 'presumption of optimal relevance' as defined by relevance theory:
(i) the utterance is relevant enough to be worth processing, and (ii) it is the most relevant one compatible with the communicator's abilities and preferences (Sperber and Wilson, 1995: 266-78).

[34] For a discussion on whether concepts such as I am proposing here (i.e., ENOUGH TIME TO FINISH WHAT SOMEONE/MARY IS DOING) are atomic or decompositional, see Hall (2011).

utterance comprehension. The consequences for theories of word-concept relations are important. It seems that the word 'minute', which can be said to encode (or logically entail) the concept 60 CONSECUTIVE SECONDS, can be used to communicate a very different conceptual content on a particular occasion. This has direct implications for the possible mappings between words and concepts discussed above.

According to the first option, word meanings are complex arrangements of semantic/conceptual features. These features are what organise the vocabularies of natural languages. For instance, dogs, cats and horses all have the features ANIMAL, FOUR-LEGGED, MAMMAL, and so on. The theory behind this assumed that if all the words could be analysed, it would be found that the set of features necessary to account for the whole vocabulary of a natural language would be smaller than the number of words in that language. For our purposes here, the important point is that a certain reading of this position blocks the possibility that the mapping between words and concepts is one-to-one because features/concepts are fewer than words. As stated above, this position has been scrutinized and criticised by Fodor and relevance theory follows him in rejecting it. The possibility that Fodor argues for, however, that excepting rare cases of mismatches, there is roughly a one-to-one mapping between concepts and words cannot be accepted either for, as shown above, it does not take pragmatic inference processes duly into consideration. Words are not limited to communicating the concepts they encode, as inference 'open(s) up new paths, to otherwise inaccessible end-points' (Sperber and Wilson, 1998, reprinted in 2012: 38) and makes it possible to communicate meanings in a way unimagined by theories under the influence of the code model.

Thus, Sperber and Wilson (1998) suggest a third possibility to characterize the mapping between concepts and words: it is concepts that greatly outnumber words since only a fraction of the concepts available to our minds is lexicalized (Sperber & Wilson 1998, reprinted in Wilson & Sperber, 2012: 35). For instance, the ad hoc atomic concept MINUTE*, paraphrasable as 'enough time to finish what someone is doing' is not

lexicalised. Thus the concept-word mapping is many to one, according to this account.

2.9.2 Relevance-Theoretic (and Fodorian) Semantics

According to standard relevance theory, despite this marked difference from Fodor's conception of the mappings between words and concepts, their account is still compatible with concept atomism, and is in line with his view that there are two kinds of semantics as endorsed in Fodor (1975).[35] Relevance theory follows Fodor (1975) in claiming that there are two types of semantics: translational linguistic semantics and 'real' semantics (Carston, 2002a: 56-61).[36] Translational linguistic semantics, as its name suggests, offers translations between natural language expressions and forms in the language of thought, roughly between the words of our languages and the concepts of our internal representational systems. A particular individual's translational semantics would then be the set of statements in the form: the public language expression '*abc*' *means (or encodes) the Mentalese form* '*ijk*'. Translational approaches to semantics have long been criticised as incomplete since instead of providing truth-conditions, meaning is given in terms of structured representations. The 'translation' it offers is from one kind of representation to another but not between a representation and what it represents. Relevance theory answers this challenge by complementing its translational linguistic semantics by a 'real' semantics. 'Real' semantics deals with the relations between the mental representations in our heads and that which they represent out there in the world. The contents of such a semantics are truth-conditional; a particular individual's 'real' semantics would be made up of T-sentences in the form '*hijk' means (is true iff) such-and-such.* A particularity of relevance

---

[35] Much could be said here regarding whether relevance theory really is compatible with the Fodorian view. My objective in this thesis is not to decide on these matters but rather to try to propose my own account. However, I will come back and add to this topic in the following chapter. For now, I can also refer to Assimakopoulos (2012) who takes the very interesting position that relevance theory and Fodor are rather *incompatible*.

[36] For a more detailed discussion of the 'two types of semantics' thesis, see Carston, 1988, 2002a: 56-61.

theory's own account of translation plus 'real' semantics is that for relevance theorists, translational semantics provides a semantics-pragmatics interface.

Furthermore, despite the fact that, when used in communication, single lexical items do not systematically map to *the same* atomic concepts, as Fodor's example of 'cat' to CAT would suggest but an example such as 'minute' to MINUTE or MINUTE* disputes, it can still be the case that the constituents of our mental representations are atomic. That is, if CAT and MINUTE are atomic concepts, the same may also be true of MINUTE*. The difference is that, in the relevance-theoretic account, arriving at MINUTE* is a fully pragmatic inferential process. This might be incompatible with mainstream semantics in that it allows pragmatic processes to contribute at a level traditionally assumed to rely exclusively on code, but this does not necessarily introduce an insurmountable antagonism between the relevance-theoretic and the mainstream semantic perspectives on the task of providing a semantics for natural language expressions. Yet, as I will discuss in chapter 3, there are very important consequences of dropping the assumption that *pragmatic processes make no contribution to explicit content* and that *any aspect of utterance interpretation in which pragmatic processes play a role is automatically an implicature* (Sperber & Wilson, 1995: 256), consequences which do impact on how we think about encoded lexical meaning.

2.9.3 Concepts in an Inferential Model

I now turn to another aspect of the relevance-theoretic account of concepts. According to Sperber and Wilson (1986/95), concepts can be usefully pictured as labels or addresses in memory.[37] As such, they serve two functions: they are headings or nodes at which information, particularly information pertaining to the concept's denotation (i.e., the things in the world that the concept applies to), can be stored, and they serve as

[37] A key aspect of my own account is a very different understanding of the workings of memory. I return to this topic briefly below and at length in chapter 5.

constituents in logical forms.[38] The information stored at a conceptual address need not be all automatically accessed so that it should not be confused with the concept's content. Sperber and Wilson first described the entries as following:

- *The logical entry consists of a set of deductive rules which apply to logical forms of which that concept is a constituent.*
- *The encyclopaedic entry contains information about its extension and/or denotation: the objects, events and/or properties which instantiate it.*
- *The lexical entry contains information about the natural-language counterpart of the concept: the word or phrase of natural language which expresses it* (Sperber and Wilson 1986/95: 86).

These entries give us an idea of the psychological objects concepts are taken to be. The logical entry contains meaning postulates in the form of deductive rules. These rules describe output assumptions on the basis of input assumptions. As mentioned earlier (section 2.5), another point of disagreement between Fodor and Sperber and Wilson is their current position on meaning postulates. I return to this issue below.

The encyclopaedic entry of a concept would contain assumptions (i.e., world knowledge) about the denotation of the concept (e.g. for the concept CAT, it would include general facts about cats such as what they look and sound like, their movements and place in human societies; it might also include random facts about cats such as the name of a famous cat or the fact that they were considered sacred in Ancient Egypt).

The lexical entry contains surface form information that it receives from the language module, typically restricted to schematic phonological and syntactic information and perhaps including orthographical information. Newer notions brought to the fore by the usage-based tradition, such as frequency ratings and co-occurrences have not been

[38] I return to the notion of concepts as constituents of logical form in chapter 3; here, I focus on the idea of concepts as 'headings or nodes in memory' which connect with *information.*

traditionally included as part of the lexical entry. This is probably because, until quite recently, it was assumed that memory stored as little information about surface linguistic form as possible, in order not to become overcharged. Recent research, however, has found evidence that even imperceptible variations in pronunciation are apt to be stored with surface linguistic forms (Bybee, 2000; Pierrehumbert, 2001).

This section on Fodor's influence on relevance theory and of Sperber and Wilson's adoption of some of his main tenets would not be complete without at least a brief look at meaning postulates. I therefore open a parenthesis here to present the arguments in favour of Sperber and Wilson (1986/95) and Horsey's (2006) position on this particular topic. Sperber and Wilson (1986/95) explicitly adopt meaning postulates to capture the logical content of the vocabulary of any particular natural language. Thus, they are in agreement with the earlier, but not the later, Fodor. Recently, Richard Horsey took up a systematic defence of this position as part of a PhD project. In his thesis, he considers whether Fodor was justified in rejecting meaning postulates.

According to Horsey, Fodor's latest position relies on the assumption that *only* analyticity can make an inference content-constitutive; in other words, the assumption that for an inference to be content-constitutive it must be beyond 'merely' necessary, it must be *necessarily true*. As hinted above, relevance theory cannot follow Fodor on this as it would block most of what has been proposed as logical entries on account of meaning postulates not being analytically true. As a way out of this, Horsey proposes to offer a notion qualitatively different from analyticity that is nonetheless potentially successful in determining which inferences can be considered content-constitutive. His solution is both simple and innovative: the minimum condition for an inference to be content-constitutive is not that it be *necessarily* valid but rather simply that the subject *considers* it valid.

Very briefly, Fodor's rejection of any account of concept content that involves content-constitutive inferences is based on Quine's rejection of the analytic/synthetic distinction and the (unjustified, according to Horsey)

assumption that content-constitutive inferences and analytic inferences are 'one and the same'. Horsey counters that an analytic/synthetic distinction is not *strictly* necessary because an underlying psychological mechanism, that of *validity for the user*, can do the work the analytic/synthetic distinction was called upon to do, at least with regard to human cognition. Quinean arguments against meaning postulates can thus be 'sidestepped' on the grounds that

> …it is perfectly possible for an inference to be content constitutive for a subject if the subject regards the inference as valid. But this in no way requires that the inference actually is valid, and therefore does not require that the inference is analytic (Horsey, 2006: 25).

To recapitulate, Horsey suggests that it is possible to provide a psychological basis for distinguishing between content-constitutive inferences and non-content-constitutive inferences; with this distinction in hand, he rejects Fodor's arguments for abandoning meaning postulates and argues that this position constitutes a 'moderate' atomism which 'allows an atomist to maintain the notion that content can be constituted in part by inferential relations' (Horsey, 2006: 25). In Horsey's view, meaning postulates *must be* included in order to deal with the logical content of the vocabulary at large because any other account falls short of capturing the canonical inferences some concepts enter into. He rejects Fodor's (2004) position that having the concept AND, for instance, is simply being disposed to think 'conjunctive thoughts' (Horsey, 2006: 60, see also Prinz and Clark, 2004).[39]

He likewise rejects that his 'moderate' atomism could be tantamount to accepting the inferential role semantics position because while the

---

[39] Fodor (2004) illustrates this idea as follows:

> Consider the concept TREE. It's presumably characteristic of concepts as such that they (can) occur as the constituents of thoughts. Clearly, the concept TREE often does. So one might think*: this tree is taller than that tree* or *some trees are deciduous*, or *there was an old woman who swallowed a tree*, and so forth indefinitely. I assume that all these thoughts share the very same concept TREE. Likewise, mutatis mutandis for logical constants (connectives, quantifiers and the like); there are indefinitely many conjunctive thoughts, and I assume that they all share the very same concept AND (Fodor, 2004: 33).

standard three rules are sufficient to *define* AND,[40] it does not necessarily follow that *grasping* all three rules is necessary since 'there might be other general considerations' that accomplish the same task (Horsey, 2006: 98). So, in accordance with Fodor, having the concept AND does not depend on *just* being disposed to accept its canonical rules; but, contra Fodor, this does not mean that explaining conjunction can stop at saying that it is having 'conjuctive thoughts' [or thoughts of conjunction]. According to Horsey, AND has truth-conditional content and Fodor's position does not address the critical question of its origin; it does not explain *how* AND expresses conjunction. Only reintroducing meaning postulates into Fodor's account can adequately fill this gap.

The interest of Horsey's proposal is not that it solves the problem of how to draw the analytic/synthetic distinction (which in fact it avoids), or that it gives a simple answer to the question of how to decide whether an inference is content-constitutive or not. Rather, its interest is that it makes a case for Quine's criticism of the analytic/synthetic distinction not being a barrier to integrating meaning postulates into a moderately atomistic theory of meaning. Horsey's thought provoking criticisms of Fodor's pure atomism and his very in-tune-with-the-times psychological basis for content-constitutive inferences are in his favour, but the success of his proposal also faces, to my view, two major challenges. The first is internal: supposing that Quinean criticisms are successfully sidestepped and meaning postulates are content-constitutive, how would these concepts still qualify as atomic? Horsey reservedly qualifies his atomism as 'moderate' but by Fodor's standards (as set out in Fodor 2003, for instance), it qualifies rather as 'inferential anatomism' and shares the lot with other anti-atomistic positions (p. 150). The second challenge possibly facing Horsey is more

[40] Consider the truth-functional connective 'AND'. This connective is governed by standard introduction and elimination rules set out by Richard Horsey, for instance, as follows:

a. p,q / pCq
b. pCq /p
c. pCq /q; (Horsey, 2006: 96-97)

general. Following relevance theory, he adopts the assumption that the information stored under a concept is organised in different entries (logical, encyclopaedic and lexical) corresponding to different functions (broadly, providing deductive rules, providing background information and interfacing with the parser). The differences are critical in selecting what is in the logical entry as constitutive of the content of the concept and excluding what resides in the other two. Some philosophers fear that anything less than a clear distinction between what *constitutes* conceptual content and what is only *contingent* information associated to a concept 'blurs the lines' between the well demarcated and stable mental entities concepts are taken to be and the information, often labelled 'encyclopaedic information' that they are 'merely' associated with. Horsey seems to seek a reformulation of this distinction more compatible with his subjective characterisation of validity while still answering to the fears of these philosophers; however, as I'll endeavour to show in the chapter on psychological perspectives on concepts and the chapter on memory, there is another alternative that, while remaining at least as psychologically plausible, *sidesteps* the distinction problem. The emerging understanding of memory's role in interpretation will probably result in *fundamental* changes to how we model 'information storage' in memory. This will, in turn, bring an understanding of *what* information is associated with a concept that is more interested in the *context* the concept appears in and in the information that is relevant *at a certain time*; this understanding, I will argue, has the potential to replace a certain aspect of the logical/encyclopaedic and synthetic/analytic distinction: the assumption that there is a permanent, context-independent division between what is *constitutive* of a concept and what is 'merely' contingent.

2.10 Closing Remarks

The topic of the kind of changes that the newly emerging conception of memory and of memory's role in interpretations bring to the topic of concepts and word meanings is a complex one and I will need to return to it several times in the coming chapters. In these closing remarks, however, I

can already briefly address certain aspects of how new research into memory possibly challenges some of the ideas presented in this chapter.

In chapter 1, I mentioned two traditions in psychology and suggested that they would play an important role in my own account of concepts and word meanings: 'norm theory' and 'exemplar theory'. I reserve a fuller presentation of their main tenets for later, but here, as a preview, consider that under the sway of the 'filing cabinet' metaphor, language theorists (and not only they) have assumed that memory stores the bulk of its information in the format suggested by the representational theory of mind discussed in this chapter (§ 2.4.1), that is, in the form of concepts, such as FEROCIOUS and *propositional* mental representations such as 'FANG IS FEROCIOUS'. These representations were assumed to exist in well-demarcated and stable forms. A further assumption was that they were retrieved from memory in these same forms, unchanged by any mechanism pertaining to retrieval itself. Information in memory was pictured to exist like written documents in files with memory as the filing cabinet that kept each piece of information in its place. Research into memory starting in the 1970s, however, has rather uncovered that the contents of memory are more correctly pictured as *imprints* left behind *by* our experiences than *conceptual* information that is organised, put in files and stored in cabinets. Information is in a much more 'undifferentiated' form than previously imagined; it can be correctly pictured as 'traces of episodes' rather than filed documents. The assumption that memory is a collection of mainly static recorded facts has, therefore, given way to the emerging picture of a *dynamic* memory made up of innumerable individual memories of events which processes of retrieval shape into the organised structures we are familiar with. If this is true, then the feeling of pulling information from memory is not the process of finding ready organised information in files but rather the process of *creating* a particular representation (called a memory echo) that results both from the particular cue used to probe memory (the cue given to the search engine, if you will) and the particular contents of memory activated by the cue. The selection of memories activated by the cue is best explained by an analogy to an echo: memory traces represent all the available information in memory

but only those that bear some similarity to the cue 'bounce back' as a result of cueing. The representation created carries both what you wanted to know and what you found. The great advantage of this is that it makes the results *distinctively context-sensitive*. The cue used to probe memory carries elements of the context in which it was created and naturally activates memory traces *according to their relevance to the particular task at hand*.

To illustrate how this new understanding of memory may challenge the idea that a concept's information is stored under different entries, consider, for instance, the case of frequency and co-occurrence 'information' related to words mentioned briefly above. How would this information be integrated into the account involving the three distinct entries (logical, encyclopaedic, lexical)? It seems that it should figure in the lexical entry together with information about the surface form. In the emerging picture, however, this is not information *about* words, rather it is information captured in the traces of the instances of use *themselves*. More frequent words and concepts are represented in memory by a larger set of traces and co-occurrences are captured as episodes or instances *of co-occurrence* so there is no need to postulate an entry of any kind for the storage of this information. I expand on this in chapters 4 and 5.

The first aim of this chapter was to give an introductory presentation of two perspectives on concepts, broadly speaking, philosophical and psychological, that the rest of the thesis takes as known. There are, then, three important points to carry over to the chapters that follow. First, relevance theory holds that words can communicate concepts *other than those they encode*. Relevance theory also proposes that words sometimes express 'ad hoc concepts' which are arrived at through a fully inferential process. It is my view that this introduces an irreversible incompatibility with traditional semantics in that once *pragmatic* processes are taken to contribute at a level traditionally assumed to rely exclusively on code, a paradigm shift ensues. The only justification other than tradition to keep pragmatics subordinated to semantics was that it could not contribute truth-conditional content; *but if it can*, then we can expect further

consequences in case the role played by pragmatics should further increase, for instance, if it were the case that pragmatic processes were intimately involved in any or all processes of interpretation.

My position will be that the way forward from this is not to try to reconcile with traditional semantics but rather to propose a new framework; critically, this new framework involves rejecting any aspect of traditional semantics that necessarily relegates pragmatic contributions to a subordinate role. In coming chapters, I develop the confrontations ensuing as a result of different possible positions on this issue: upholding traditional semantics, trying to reconcile or find a middle ground with traditional semantics, and finally, breaking with the traditional semantic framework.

The second point to take away from the discussions in this chapter is that relevance theory makes some very specific assumptions about how information related to concepts is stored in memory: very briefly, it holds that concepts can be pictured as nodes or headings at which information pertaining to the concept's denotation and linguistic status can be *stored* according to its form in different kinds of entries. This proposal inherently adopts a certain view of memory that, as I have suggested, new discoveries overturn. I would like to add that an account of memory as consisting of largely undifferentiated information and 'traces of episodes' which can be selectively activated is compatible with relevance theory in spirit if not in the detail. Relevance theory's proposal can be reformulated as follows: information is stored not at a particular node or 'conceptual address' but in *general* memory; however, it is true that this information is not all automatically activated each time the concept is deployed; rather, activation is *selective*, and selection *obeys the general principle of relevance*. Finally, this information is general, 'encyclopaedic' information that a subject can associate with any aspect of her experience and is therefore not to be confused with what philosophers take to be conceptual content.

The third and final point is to do with Fodor's 'criteria for a theory of concepts': In section 2.6, I said that Fodor advocates an approach to concepts according to which, in order to say that two people, or two time slices of the same person, *share* a concept, they must have tokens of *literally*

the same concept type. It is this position that then determines his 'publicity' and 'compositionality' constraints: briefly, publicity depends on no one else's concept WATER, being *any* different from my own; and, for compositionality to work as Fodor construes it, the concept RAIN, for instance, has to contribute *the same* content to any thought of which it is part. As announced, my own construal of concepts will not comply with these constraints as construed by Fodor. In chapter 4, I will develop an alternative to the idea of sharing concepts that will involve subjects using what they share, namely contexts, in order to *converge* on conceptualisations rather than share concept-types. For Fodor, this is exactly the kind of approach that must be avoided because it poses a danger to the possibility of accounting for the productivity and systematicity of thought. The question, however, is whether Fodor's worry is sufficiently, *independently* warranted, or whether it is worth exploring other approaches to concepts and trust that, in time, they will lead to adequate, alternative explanations for the productivity and systematicity of thought.

Chapter 3: Word Meanings and Pragmatics

3.1 Introduction

My first objective in this chapter is to bring together a varied set of contributions to the complex and rapidly evolving topic of word meaning in context. The disagreements between theorists I call on are large and small, but the issues raised represent, in my view, the biggest challenge facing cognitive pragmatic approaches to language processing today. This chapter directly follows from the previous chapter's discussion of philosophical perspectives on word meaning and concepts, and takes up where the relevance-theoretic reformulation of Fodor's views left off. Here I introduce some of the theoretical arguments in favour of the particular approach I advocate: 'meaning eliminativism'. This chapter is also a preparation for the upcoming chapter on psychological perspectives on the topic of concepts, and word meaning. The logic behind the particular approach to word meaning I suggest in this chapter is largely supported by research from the fields of categorisation and memory research that I will only very briefly allude to in this chapter but to which the two subsequent chapters are devoted.

My discussion of word meaning in context in this chapter begins with a brief outline of two contrasting general views of language that are to frame the discussion, not only in this chapter, but throughout the thesis. Within this section, I adopt Recanati's 'contextualism' and use it as an umbrella term to group together the views of philosophers, linguists, pragmatists, and psychologists truly open to taking the consequences of rampant context-sensitivity seriously. Then, in an effort to organise the diverse contributions and show how they can be taken to collectively, if not uniformly, move forward in a particular direction, I organise the chapter into three groupings of contributions, which include two 'stages'. This should allow a clearer picture of the points of contention within a certain picture of evolution.

In the first stage, the focus is on reformulating the insights of one of the most influential founders of contemporary pragmatics, the philosopher

H. Paul Grice. Contemporary 'cognitive' pragmatics emerges as a challenge to traditional semantics as formulated even by Grice. In this first stage, I review the contributions by relevance theorists like Dan Sperber and Deirdre Wilson, the contributions of the French philosopher François Recanati, and those of pragmaticist Robyn Carston; my aim is to show that even in this 'first stage' of the evolution of theorising on word meaning in context, these theorists are on the right track towards a *new* framework for word meaning. In this section, I also give an account of the relevance-theoretic ad hoc concept construction and utterance comprehension process that I take as my point of departure. My contention however, is that these proposals are not radical enough, and that the evidence supports going much further. For instance, the relevance-theoretic notion of 'code' and 'encoded concepts' seems to block a truly 'contextualist' approach to just what is associated with word forms if it is not context-independent word meanings. I argue that abandoning the 'code' model of *language* (not just of communication) and resolutely stepping outside of the traditional framework are widely justified by the evidence. To support this radical suggestion, in a section between the first and second stages, I go back to the initial inspiration for contextualism as contained in the contributions of philosophers like Friedrich Waismann, John Searle and Hilary Putnam. The second and final stage is an exploration of radical contextualism as expressed in the contributions of contemporary thinkers like Peter Bosch, Robyn Carston and François Recanati. The aim of this section is to begin to consider the possibility of an outright refusal to subscribe to the traditional division of labour between semantics and pragmatics.

3.2 Two Contrasting Traditions

The overarching topic of this chapter is word and phrase meaning *in context*. To introduce this topic, I propose to first briefly take a look at two contrasting traditions in language theorising whose disagreements punctuate the discussion in the rest of the chapter. According to Herbert H. Clark, an early and influential proponent of one of these traditions, it was the introduction in the 1960's of Chomsky's 'generative' linguistics that first

forged a division between two distinct but connected fields within language studies. For Chomsky, it was essential to explain linguistic *competence*, that is, the knowledge structures that underlie people's abilities to produce and recognise the sentences of their language. In this tradition, the linguists' primary concern was people's tacit knowledge of the grammar of their language (i.e. the rules of phonology, morphology, syntax, and semantics); their linguistic *performance*, or how they actually produced and understood sentences, was a derived and secondary interest. Clark (1992, 1996) calls this tradition the 'product tradition' and contrasts it with another approach that emerged simultaneously though independently of the first: the 'action tradition'. [41] Inspiration for the action tradition is often traced back to the writings of the later Wittgenstein and he is frequently credited with a decisive, although hard to define, influence on ordinary language philosophers like John L. Austin, John Searle, Gilbert Ryle, Peter Strawson and Charles Travis. Although, as the iconic pioneer of much of the action tradition's content, Wittgenstein would surely merit treatment on his own, I opt, like Recanati, rather to group him *together with* the representatives of ordinary language philosophy and set out their contributions at the level of a group. François Recanati cites as one of their main tenets 'in vacuo, words do not refer and sentences do not have truth conditions' (2004: 2). More on what this amounts to below. According to Clark, the action tradition initially developed closely around the work of Oxford philosophers John L. Austin, John Searle, and Paul Grice. [42] They focused on the actions that are

---

[41] A very similar division into two traditions can be found in the writings of most authors in this field. King and Stanley (2005), for instance, label them the 'expression-centered conception' and the 'speech-act-centered conception' of semantics. I adopt the terminology of Clark for its initial clarity. As the chapter progresses, it will become evident that many language theorists share Clark's thoughts and position. Their views will be presented alongside their own choice of terminology. An example in anticipation: Bosch calls the 'product tradition' the 'linguistic knowledge paradigm', he explains:

> The central concept on which both syntax and semantics are built in the linguistic knowledge paradigm is the sentence, both as a basic notion of grammar and – in its guise as proposition or sentence meaning – as the basic notion of formal semantics (Bosch, 2009: 1).

[42] It is important to note that Grice kept a certain distance from mainstream 'ordinary' language philosophy. Rather than adopt it as an *alternative* to 'ideal' language philosophy, he sought to reconcile the two, namely by showing how pragmatics could account for the differences in meaning in use from the logical semantics of logical words like 'and', 'or', 'but'

accomplished as communication unfolds. Austin famously developed the thesis that as people use language they are 'doing things with words'; not only *uttering* words and phrases but *making* assertions, demands, influencing opinions, and so on. Grice brought to the fore the fact that people depend on the recognition of their intentions in order to get their messages across. This tradition soon influenced a broad range of theorists and gave rise both to the usage-based tradition in linguistics and, particularly through a development of the ideas of Paul Grice, to contemporary (Anglo-Saxon) pragmatics.[43]

Herbert H. Clark was among those particularly influenced by the action tradition. For him, the differences between the two approaches are profound: in his 1992 book, he writes that the two traditions are built on 'very different foundations'. The product tradition's foundation is language *structure*, while speaking and listening are of interest only insofar as they are manifestations of this structure. The action tradition seems to approach language phenomena from the opposite direction: from what people *do with language* to how language works. His observations on these differences can be summarised in the three tenets of the action tradition: (i) in language use, utterances are more basic than sentences; (ii) speaker's meaning is primary, word and sentence meaning are derivative; and, (iii) speaking and listening are not autonomous activities but parts of collective (or 'joint') activities (Clark, 1992, see also Clark, 1996).

These points will appear in diverse forms throughout the discussion in this chapter. According to the first tenet, the disagreement between the product tradition and the action tradition concerns whether it is *sentences* or rather *utterances* that are more basic. The product tradition holds that sentences are fundamental because they are the basic units generated by a language's grammar. Its focus on grammatical structures identifies

---

(See Grice, 1975; Travis 1985; Recanati, 1994; Carston, 2002a). More on Grice's singular position below (§ 3.3).

[43] Continental 'enunciation' theories, also called 'French discourse analysis', following Emile Benveniste among others, also focused on language understanding and would surely be of great interest but, unfortunately, I cannot give them the time and space they deserve here.

sentences as the fundamental category of language. The action tradition argues, on the contrary, that it is utterances that are more basic. The approach focuses on language *use* and argues that since it is only ever utterances (not sentences) that we actually hear or produce, it is utterances that are the fundamental category of language. The action tradition considers it a category mistake to speak of *sentence* production and comprehension; again, because it is only *utterances* we produce and understand. The difference is important because sentences can be devoid of critical information necessary to identify referents: speaker(s) and hearer(s), time, place and other circumstances. But also because very frequently utterances are not tokens of sentences; rather, they are short phrases like 'Got a light?' or 'Never mind', individual words like 'Taxi' or 'Please' and other odd bits of language, gestures and actions, like waving and saying 'Good bye' or raising your eyebrows and saying 'Pardon?'. There is no justification for excluding these very familiar communicative acts from a theory of language.

According to the second tenet, the disagreement centres on whether it is word and sentence meaning or speaker meaning which is primary. Theories of language structure usually opt for making word and sentence meaning come first. They reason that just as the syntactic units of a sentence come together to form a more complex syntactic structure, the meanings of words also combine to form larger semantic representations. The first two tenets can be illustrated with the following examples:

(2) The game warden watched the poacher with binoculars.
(3) I shot an elephant in my pajamas.

The product tradition starts its analyses at the level of the structure of the sentences: it assumes that the meaning of the sentences (2) and (3) is the syntactical combination of the meanings of the words in each. In (2), for instance, it would be the conventional meanings of the words *game warden*, *watched*, *poacher* and *binoculars* and the way they are put together. (2) would be considered a structurally ambiguous sentence or a surface manifestation of two distinct sentences and the task of the linguist would be

to describe how speakers process (and perhaps disambiguate) such structures. In the action tradition, which puts utterances and speaker meaning first, (2) would not necessarily be considered ambiguous. In the context of an utterance, speaker and hearer share a common background so that the speaker, *who has a particular poacher in mind*, can reasonably believe that his hearer can identify this poacher (Clark, 1992: *introduction xiv*). If the speaker does not believe that his hearer can identify the particular poacher, then using 'the poacher with binoculars' would be anomalous.

In (3) a parallel aspect of how *potential* ambiguity is solved in natural language use is particularly clear. In Groucho Marx's 'One morning I shot an elephant in my pajamas', notice that what makes the phrase funny (and quotable) is not so much the fact that it is ambiguous but that, *despite a certain amount of structural ambiguity*, it is reliably processed in the same way. True ambiguity would predict that at least a significant number of tokens would be interpreted as the elephant being in pajamas; however, Groucho Marx's completion 'How he got into my pajamas I'll never know' suggests that it is only through a considerable deviation from normal understanding that we can imagine the elephant in the pajamas. It is having to re-process an utterance by reviewing tacit background assumptions (e.g. elephants do not wear pajamas) that characterises jokes. It's once we've found the 'right' way, the way the joker *intended* his utterance to be interpreted, that we exclaim 'Oh! I get it!'.

Finally, the third tenet – *speaking and listening are not autonomous activities but parts of collective activities* – represents Clark's particular contribution to theorising about language. Once we admit that utterances and speaker meaning are more basic than sentences and word/sentence meaning, it becomes evident that language is fundamentally a form of joint action. This idea is already present in Grice (1967 reprinted in Grice, 1989) who developed the conversational maxims after realising the importance of

cooperative interaction in language.[44] Clark's third tenet can be interpreted as a global guideline setting out how theorising about language should proceed. Other theorists have proposed their own guidelines and much of the rest of this chapter is a presentation of their views. As is often the case in the human sciences, there is a considerable issue with terminology: terms do not usually quite overlap from theorist to theorist and, while sometimes the differences are minor, many times they carry important theoretical implications that should not be overlooked or downplayed. For instance, in these paragraphs, I've used Clark's terminology to describe two traditions: the product tradition and the action tradition; but, despite Clark's impressive influence on the general field of linguistics (he has an h-index of 50 and his 1996 book *Using language* has been cited in Google Scholar over 5,000 times), his terminology has not been adopted.[45] Also, in an ever changing landscape of approaches and proposals, a single set of terms is unlikely to suffice. Therefore, in what follows, further terminology will be introduced as needed.

3.2.1 Challenges to Traditional Theories of Word Meaning

As mentioned above, the product tradition's focus is on *sentence* meaning and the action tradition's focus is on *utterance* meaning; a seemingly small difference resulting in substantial dissimilarities between their respective views of how to study language. In this subsection, I turn my attention to the product tradition's ensuing position on word meaning in order to set the stage for the current stark disagreements between theorists on the issue of contextual contributions to the meanings of words. Following mainstream philosophy of language, semanticists in the product tradition assume, first of all, that words have stable, context-independent meanings. This, in turn,

---

[44] The theorists influenced by these ideas are so numerous that they would be hard to list. One that deserves special mention is Michael Tomasello who adopts the *language in use* view of language as joint action and very thoroughly addresses the consequences of this for theories of language evolution and language acquisition (see Tomasello 2003b, 2009).

[45] It is very rare for a linguist to have a high h-index. Compare Clark's h-index of 50 with Noam Chomsky's now at over 100 and consider that the second is cited more often for his political than his linguistic writings.

leads them to suppose that it is generally unproblematic to take sentence meaning to be the syntactical combination of the (conventional) meanings of the words in the sentence.[46] Secondly, they assume that sentence meaning thus derived is propositional, or, in other words, that it determines truth conditions.

The alternative view, defended today by philosophers, linguists and pragmaticists of the action tradition and glossed above as that held by the ordinary language philosophers is that, *in vacuo, words do not refer and sentences do not have truth conditions*. In linguistics and pragmatics, this is often expressed as the 'linguistic underdeterminacy thesis', or the observation that *sentence meaning underdetermines truth conditions*. In stark opposition to the traditional view of sentence meaning as propositional, this position holds that it is only *utterances* which can be propositional since, in order to complete the truth conditional content expressed, indefinite numbers of background assumptions and contextual factors must be taken into consideration. Certain developments of this view, which I'll call 'contextualism', have led theorists to question the traditional semanticist's fundamental assumption that what words contribute are stable, context-independent meanings and, importantly, to develop alternative views.

The alternative views of contextualists have met with forceful resistance from the advocates of the traditional view. Minimalism or anti-contextualism, as we might call this response, might even call attention to the fact that challenges to the traditional view and effective resistance to change are not new to this debate. The *original* ordinary language philosophers Ludwig Wittgenstein, Peter Strawson, John L. Austin, and John Searle, among others, *already* presented strong arguments for a radical change in the conception of meaning in the 1950s and 60s, but the traditional view successfully resisted these challenges, and held strong. I will argue, however, that renewed interest in matters of meaning have

[46] They are, of course, forced to admit *some* context-dependence but seek to limit it as much as possible by claiming, for instance, that context-dependence is limited to a small set of indexical elements. This is a key issue I return to below.

rekindled old arguments and relaunched the debate. New elements, particularly by people working in my field, that of cognitive pragmatics, weigh in and will surely succeed, this time around, in dislodging the traditional view.

There is heartening evidence for this potential change of fates. Most important is the genuine interest, *or at the very least concern*, with which some proponents of the traditional view address the evidence brought forth by contextualists. It is no longer generally ignored as inconsequential; rather, it is increasingly agreed that it *must be* accounted for. A related encouraging fact for contextualists is that addressing the issue of context-sensitivity has already triggered possibly far-reaching changes. As different underlying assumptions are critically addressed, a clearer picture of the complexity of the issues can emerge. At first, this picture reveals inconsistencies which might lead to disagreements amongst theorists, but as these disagreements themselves are addressed, a new consensus can be built. Finally, while more formally inclined theorists today mostly endeavour to explain context-sensitivity away, perhaps this is only their first reaction and as their familiarity with the issue increases, they, *or the generation immediately following*, will come to see things differently; this recent episode of the debate is, after all, very new.

Emma Borg (2004, 2012) is an excellent example of a formal semanticist with a keen interest in both defending the traditional view and addressing the issues brought to the fore by contextualists. She recognises that the radical context-sensitivity of natural language provides a new type of challenge to standard formal theories like the one she proposes. Her approach is not to downplay the importance of speaker intentions in accounting for what is 'intuitively' said and communicated, rather, it is a redefinition of semantics, a thinning down of semantics into *minimal* semantics that integrates some of the lessons of contextualism. She has recently set out four claims to broadly define minimalism: (i) semantic content for well-formed declarative sentences is truth-evaluable content; (ii) semantic content for a sentence is fully determined by its syntactic structure and lexical content: the meaning of a sentence is exhausted by the

meaning of its parts and their mode of composition; (iii) there are only a limited number of context-sensitive expressions in natural language; and, (iv) recovery of semantic content is possible without access to current speaker intentions (crudely, grasp of semantic content involves 'word reading' not 'mindreading') (Borg, 2012: 4-5).

These claims clearly belong to the framework described above as the product tradition: sentence meaning, rather than utterance meaning, is considered as basic and from this fundamental stance, positions on semantic content and truth follow. Very generally, theorists identifying themselves as minimalists hold as in (i) above, that sentences express *complete* propositions, that is, that they have truth-evaluable content. Moreover, as in (ii), they also maintain that context-independent aspects of meaning such as 'syntactic structure' and 'lexical content' *suffice* to build up sentence meaning. There is an issue of this sentence meaning being propositional *relative to a context*. In most minimalist approaches, the belief is that indexicals and demonstratives set up 'slots' that are filled by contextual processes of 'saturation', that is, by bottom-up contextual processes that do not call on speaker intentions. Mandatory and optional contextual processes are then distinguished: the former are supposed to contribute to propositional content and, therefore, fall in the domain of semantics, while the latter are contextual processes which are *not* triggered by an element in the sentence and, therefore, are not supposed to contribute to propositional content; rather, they are implicatures, the domain of pragmatics. More on what this amounts to below. For now, it is important to note that this implies that the kind of context-dependence insisted on by the action tradition is secondary to a purely semantic, formally-derived content. Borg takes these first two claims as fundamental to formal traditions and rightly adopted by most theorists in her field. They basically echo the positions of the pre-contextualist era of the early Wittgenstein, Gottlob Frege, and Bertrand Russell, among others.

Claims (iii) and (iv), on the other hand, are best understood as responses to contextualist challenges. Against contextualists, who claim that context-dependence is rampant in natural language, most minimalists claim

that the number of context-sensitive expressions in a natural language is actually quite limited and proceed to offer an account of how even these expressions can receive formal treatments; I briefly take this up in the following subsection on indexicals and demonstratives. Here, I rather make the point that, on certain issues, theorists *within* formalism disagree: with regard to (iii), for instance, Borg tells us that taking it as 'the most salient feature of semantic minimalism' as Cappelen and Lepore (2005) do, instead of as secondary, possibly weakens the formalist position.[47] An even more crucial disagreement emerges if we interpret 'context-dependent expressions' in (iii) as referring not only to indexicals and demonstratives but to *all* potentially context-sensitive expressions in a natural language (from more commonly accepted context-dependent expressions such as 'tall' and 'ready' to 'red', and, *following the much feared slippery slope logic*, in fact, to any and all natural language expressions). Minimalist theorists can be differentiated by how they address this last issue: some, like Stanley (2000), opt for enriching the underlying syntactic form (the logical form) in order to claim that *apparent* contextual-dependence is really linguistically mandated, while others, like Borg (2004, 2012), opt rather for minimising what we take semantic content to contribute, slicing the contribution of what is linguistically mandated thin enough as to avoid stepping outside of formalist approaches. Finally, a last important disagreement centres around the notion of speaker intentions in (iv): King and Stanley (2005), for instance, seem to agree with minimalism on the importance of upholding a strict division between semantics and pragmatics, but disagree on whether speaker intentions may be allowed on the semantic side of the divide. They hold, against most minimalists, that speaker intentions are constrained by the standing meanings of expressions and therefore represent

---

[47] Cappelen and Lepore (2005) cite the following as the first of three important features of minimalism:

> The most salient feature of semantic minimalism is that it recognizes few context sensitive expressions, and, hence, acknowledges a very limited effect of the context of utterance on the semantic content of an utterance. The only context sensitive expressions are the very obvious ones [listed by Kaplan (1989)] plus or minus a bit (p. 2, see also Borg, 2012: 10).

unproblematic contributions to semantic content. Unfortunately, limits of time and space do not allow for a detailed presentation of each of the approaches mentioned above. In these paragraphs, I have sought only to flag key issues that I come back to throughout the chapter and to attract attention to two related facts: that these approaches seek to address the issues brought forth by contextualism and that there is no widespread agreement amongst them as to which strategies will work against contextualism and which minimalist claims are non-negotiable.

Now, hopefully safe from underestimating the divergences amongst the representatives of the traditional view, I close this section with a brief summary, in the form of a characterisation of the *standard* minimalist view. One of the central contentious claims is that, in general, sentences express complete propositions. Context-dependence or sensitivity is downplayed and an effort is made to offer formal solutions to issues of identifying referents and resolving ambiguity. The belief is that a formal account can be given of the kind of contextual parameters required so that the mental process of arriving at truth-conditional content 'runs exhaustively along syntactic tracks' (Borg, 2004: 84).

The traditional view proposes a certain division of labour between semantics and pragmatics. In language interpretation, semantics is responsible for sentence meaning, its output is a proposition (i.e. a truth-conditional content) that the speaker can be taken to have *said* (if not *meant*). Pragmatics takes this truth-conditional semantic output and, roughly, expands on it as dictated by the demands of a particular context of utterance to arrive at a full understanding of what the speaker *meant*. The traditional approach offers many arguments for keeping these two levels of meaning separate (see Borg, 2004, 2012; Cappelen and Lepore, 2005; King and Stanley, 2005. 2006; among others), but, arguably, there is no a priori deciding argument and no reason not to let the evidence weigh in.

I now turn to a brief discussion of indexicals and demonstratives. There is wide agreement that indexicals and demonstratives necessitate a certain

kind of 'contextual intrusion' into the specifications of truth-evaluable meaning and so they have often served as test cases for theories addressing context-dependence (Borg, 2004; King & Stanley, 2005; Bosch, 2007, 2009; Recanati 1994, 2001a, 2004, 2010a). I take them simply to illustrate how the traditional approach presented above *would* work if the aspects of the context that contribute to even the most minimal content *could* be treated formally; that is, in total abstraction from what the speaker intends to convey.

3.2.2 The Example of Indexicals and Demonstratives

Following David Kaplan (1989 (1977)), the reference or content of context-sensitive expressions, such as indexicals and demonstratives, is mediated by their linguistic 'character'. This affords a distinction between two aspects of their meanings: the first is context-independent; it attaches to expressions at the level of the expression type. Thus, the type meaning (character) of the indexical 'I', for instance, is always *the speaker of the utterance*. The second aspect of meaning (content) varies across utterances, attaching itself at the level of the expression token. This first distinction is largely uncontroversial; I am more interested in the further distinction involving whether or not the expressions require a 'demonstration' in order to fix their referent.[48] Kaplan explains that a demonstration, such as a pointing gesture, accompanies the 'true' demonstratives. This kind of indexical would then contrast with the 'pure indexicals' which do not require any demonstration. In this framework, the personal pronoun 'I' is a pure indexical because generally no pointing or demonstration need accompany an occurrence for 'I' to refer to the speaker of the utterance. From the contextualist perspective, however, the case of demonstratives is fundamentally *different*. While *theoretically* formal semantics could model certain features of a context in order to fix which aspects of a context a character or linguistic rule needs to pick out in order to provide a *pure indexical* with a referent, such rules are unlikely to

[48] For an in depth discussion of Kaplan's general proposal, see Borg, 2004; for an opposing view, see Bosch, 2007, 2009.

be sufficient to determine the referent of a *demonstrative*. According to contextualism, what Kaplan's distinction reveals is that part of the truth-conditional content of the utterance/sentence is unavailable to a purely formal analysis because the notion of demonstration opens the doors to a non-formal aspect of meaning: the speaker's intentions.

Maintaining its claims puts formal semantics under pressure to offer an account of the truth-conditional content of indexicals and demonstratives appearing in utterances. I propose to look at indexicals first. With most minimalists, a formal approach to indexicals would consist of modelling certain features of a context in order to fix those parameters a linguistic rule needs to pick out in order to provide a pure indexical with a referent. For 'I' this would be the speaker of the utterance every time, for 'here' the place in which the utterance takes place every time, and so on for 'now', 'today', 'tomorrow', 'yesterday', etc. It is commonly assumed that such descriptions represent the straightforward way in which 'character' or linguistic rules can capture the formal aspect of these referring terms. However, does the linguistic rule really suffice to fix the referent for an occurrence of 'here' in an utterance? Consider 'here' as the response to the question 'Where should we have dinner?' Replacing 'here' with 'the place in which the utterance takes place' really only narrows down the reference to a place and says that that place should bear a relation to the place where 'here' is uttered. But such a description *by itself* is no guarantee of successfully fixing a referent because *the relation* will depend upon the context of utterance in such an unpredictable way that any predetermined choice of contextual features is always at risk of falling short; the number of factors to consider is simply too open-ended. This is why Carston (2002a) describes indexicals as 'overt indicators that a pragmatic process of contextual specification is obligatory' (2002a: 328). Imagine the exchange has taken place outside a restaurant: 'here' would then mean 'in *this* restaurant, the one we are standing outside of' where 'this' is a demonstrative. Is this result due to the particularity of this occurrence? Is it an abnormal use of 'here' that strangely refers not to the place where the occurrence has taken place but rather to a place in the proximity of where the utterance has taken place? Let's consider another

example. If I ask my overseas friends if they are coming 'here' for their Easter break, 'here' would naturally be taken to mean London, or perhaps the UK but it could conceivably mean Europe. The same goes for 'now' which depending on context can mean this second, today or any other amount of time including the present. The action tradition suggests a solution: take the linguistic rules as no more than guidelines and make *speaker meaning* primary; this authorises pragmatic inferential processes to pin down what persons, places, times, etc. interlocutors are jointly attending to and thereby assign referents. This solution, however, is rejected by proponents of the traditional, 'product' tradition in favour of approaches that fix referents for indexicals like 'here' and 'now' independently of speaker's intentions.

Finally, even if it were the case that a formal treatment worked for pure indexicals, there would be a considerable added difficulty in offering any such satisfactory account of a demonstrative such as *that*. As an illustration, consider that while visiting London with my overseas friends I offer the following advice: 'I wouldn't do that', 'I wouldn't eat that' or any such phrase. In order to fix what 'that' refers to in these phrases it would be necessary to imagine them as utterances in particular contexts. For instance, one of my friends, unaccustomed to authentic Indian restaurants, has just naïvely put the spiciest sauce on a piece of naan and is taking it to his mouth. How does he know what 'that' refers to in my utterance? Untroubled by the dictates of traditional linguistic theory, he simply assumes that it means the piece of bread in his hand. We are jointly attending to it as I utter the phrase and so I do not even need to point to it, although I could. Demonstratives, because they involve such 'pointing' and *convergence* between speakers are even less amenable to formal treatment than pure indexicals, although this, of course, has not stopped some formalists from trying.[49] But this seems ill judged since (i) the elements of a context that a formal semantics would have to model would only with great difficulty include a sufficiently wide scope to pick out a correct referent for *that*, and (ii) the solution proposed

[49] Borg (2004), for instance, maintains that her approach can deliver a truth-conditional semantics for demonstratives without the involvement of speaker intentions.

by the action tradition above is equally adequate for pure indexicals *and* demonstratives (and other troublesome cases of reference assignment) so that not adopting it would have to be duly justified.

To sum up, it seems that there are two alternatives to account for indexicals and demonstratives. One involves ad hoc rules and rule-following proportionate in complexity to the difficulty and awkwardness of accounting for convergence in reference without speaker intentions. The other takes very natural (and independently motivated) mind-reading abilities and only schematic or flexible language conventions and arrives at the same result. Furthermore, as I'll argue, the case of indexicals and demonstratives is only the tip of the iceberg when it comes to making a case for contextualism. In the remainder of this chapter, I present many more contributions of contextualist approaches. They are organised into two stages, each progressively calling for more radical changes to the traditional view of sentence and word meaning.

3.3 The First Stage: From Philosophical Pragmatics to Cognitive Pragmatics

The focus of this 'first stage' is the ideas of Paul Grice and the reformulations of his key insights jointly adopted by Recanati and the proponents of relevance theory.[50] The section includes presentations of selected contributions together with a discussion of the differences between Grice's and contemporary cognitive pragmatics' views on levels of meaning. The objective is to view arguments in favour of abandoning Grice's notion of 'what is said' for relevance theory's notion of 'explicature'. Among the contributions of cognitive pragmatics, I am particularly interested in the relevance-theoretic utterance comprehension procedure, especially as it relates to ad hoc concept construction since it is, for my own account, a

[50] Space and time do not allow a look at all the different schools of pragmatic theory that follow Grice, I therefore only briefly present the two that most concern us here: relevance theory and the approach of François Recanati.

point of departure. In a second half of this chapter (sections 3.4 and 3.5), I first look back to some early prescient work in philosophical semantic/pragmatic theorising and argue that the positions in the 'first stage' were not radical enough, particularly with respect to word meaning. Then, I move on to a 'second stage', which, in my view, represents the frontiers of pragmatics today.

3.3.1 Philosophical Sources

Contemporary pragmatics generally focuses on the *cognitive* mechanisms underlying language comprehension and interpretation. It is widely acknowledged that, as a field, pragmatics has progressively moved away from philosophical *formulations* and towards increasingly more cognitive frameworks. But it is also important to recognise that theorising in general, whether it comes from psychologists, philosophers or linguists has moved towards more cognitive formulations as part of the cognitive revolution which has affected not only pragmatics, linguistics and psychology but also philosophy. Notably, moving away from philosophical *formulations* does not mean turning one's back to philosophical *influences*. The constraints on formulating solutions to the issues which arise in pragmatics are simply increasingly those self-same constraints of the general framework of cognitive science. Philosophical insights were key to pragmatics' beginnings and are still critical today, as evidenced by the number of philosophers cited in the previous chapter and to be cited in this chapter.

A good example of a proposal completely reformulated under the influence of the cognitive revolution is that of Herbert Paul Grice. For contemporary pragmatists, Grice was a *philosophical* source of insight into *psychological* mechanisms of language comprehension. Grice's key contribution was the recognition of the central role played by speaker intentions. In his 1957 article 'Meaning', Grice first formulated an account of what it was for an individual to mean something by an utterance in terms of the intentions that were thus expressed and recognised. Grice's genius was to show that what a speaker can do beyond 'saying' is to mean something by 'implicating' it. This saying-implicating distinction has proven to be

indispensable in pragmatics irrespective of differences between interpretations and how often theorists have found it necessary to reformulate it.

Grice described the level of 'what is said' as speaker *m-intended*, that is, as part of the content the speaker can be taken to have overtly intended the hearer to understand, and, at the same time, as 'closely related to the conventional meaning of the words (the sentence) he has uttered' (Grice, 1989 (1957): 25). But this characterisation is problematic: as Robyn Carston remarks, Grice's two requirements sometimes pull in opposing directions (2010b: 220, see also Carston 2009a). For them to be met, the interpretation would have to stay close to the literal meaning of the constituents of the utterance and the way they are put together while, *at the same time*, be speaker intended. In most cases, however, speaker intentions are incompatible with literal interpretations and the interpretation cannot 'stay close to the literal meaning'.[51] Suppose that my overseas friends have now spent some days in London and have noticed the price of food in restaurants and shops. One of them says:

(4) Aqui los pobres no comen.

This utterance in Spanish translates word-for-word as 'Here, the poor do not eat'. If, in my efforts to interpret this utterance, I stay close to the literal meaning of its constituents and the way they are put together, I arrive at a certain construal of what is said, something along the lines of '[In – the touristy neighbourhoods of – London] the poor do not eat [anything]'; but this, perhaps because of the scope of the negation, is unlikely to be what the speaker m-intended. The problem is that with most examples, as with this one, it is hard to identify just what a speaker *could* m-intend *without* departing from the literal meaning of his words and how they are put together. In the above example, for instance, I would probably let elements of the context guide me in figuring out the content of the speaker's

[51] Carston captures this as 'no single level of meaning can do double duty as both sentence semantics and speaker-meant primary meaning' (2013: 176).

communicative intention (I might consider, for instance, the frequency with which people overstate their opinions or make hasty generalisations in their unguarded everyday speech); this influences the literal aspects of the utterance, which are put in parenthesis, so to speak: the negation and/or the meaning of the constituents would likely undergo some modulation in order to construe what is said. The explicature of this utterance would be something like 'In – the touristy neighbourhoods of – London the not so well-off do not eat good/nice food'. This, in any case, is the solution Carston and others have opted for and the one I present in what follows.

From the start, relevance theory chose a distinctively *cognitive* reformulation of Grice's key insight and of his framework in general. His co-operative principle and conversational maxims, for instance, were judged too normative to reflect natural psychological constraints. Relevance theory reformulated them as principles arising from features of the human mind (*its predisposition to maximise relevance*) and consequent constraints on the processing of information, and particularly *language*, (*the interaction between ostensive stimuli and optimal relevance*). The cognitive perspective critically reveals the central role played by relevance, defined as a potential property of both any external stimuli and any internal representation that provides input to cognitive processes (Wilson and Sperber 2004: 608). These *cognitive* and *communicative* principles have received thorough attention in the literature and I take them as known in what follows (for a complete presentation of these principles see Sperber & Wilson, 1986/95 and Wilson & Sperber 2004; for a comprehensive look at relevance theory, see Clark, 2013).

Another attentive reader and influential commentator of Grice is French philosopher François Recanati. His contributions are also key in the current reformulation of the saying/implicating distinction. There are many parallels between proponents of relevance theory and Recanati and some

important differences. These will emerge as the different contributions are presented. [52]

3.3.2 The Contributions of Cognitive Pragmatics

For relevance theorists and Recanati, very much like for Clark, the heart of the issue is the gap between sentence and utterance meaning, also often referred to as the 'linguistic underdeterminacy hypothesis'. Among contemporary pragmatists and some linguists the hypothesis is well-accepted but this is not the case amongst mainstream semanticists and philosophers of language. Thinking back to Grice's construal of 'what is said', it is important to keep in mind that he would *not* have endorsed linguistic underdeterminacy since in seeking to keep 'what is said' as close as possible to the conventional, encoded meaning of the sentence, he would have rather opted for a different explanation for any gap between the conventional context-invariant meaning of an expression and the occasion meaning. [53] On a par with the minimalists of the previous section, he only acknowledged appealing to context for cases of ambiguity and indexicals. But if this stance is untenable, as the contemporary contextualist view of pragmatics holds, then reformulating Grice's original saying/implicating distinction involves rejecting the minimalist view and building a new contextualist approach.

The first three subsections below set out the different contributions of cognitive pragmatics concerning Grice's 'what is said'. I first present Recanati's suggestion that a more careful look at Grice's saying/implicating distinction reveals that other distinctions need to be made and that this is an argument in favour of a contextualist perspective and against minimalism. In the section on relevance theory's 'enriched explicit content', I first complete the sketch begun in the previous chapter of Dan Sperber and

---

[52] To give an example, relevance theorists and Recanati have serious disagreements with respect to how 'what is said' is arrived at. For Recanati, purely associative processes are sufficient and the interpretative work is done by the dynamics of concept activation. While for relevance theorists, it is important to recognise the inferential driving force behind even these basic tasks (for a discussion, see Carston, 2007).

[53] In fact, when Grice did consider the kind of evidence that points to underdeterminacy in the contextualist framework, he sought to explain it as cases of implicature (Carston, 2013).

Deirdre Wilson's theory of concepts and ensuing positions on semantics. Then, I move on to contributions made towards defining a speaker-meant notion of explicit utterance content. The consequences, not only for theoretical pragmatics, but also for concrete theories of communication are explored: I underline the fact that the relevance-theoretic account of how we arrive at 'what is said' not only rejects minimalism on *theoretical* grounds but also provides a detailed account of how *pragmatic* processes are involved. Finally, some problems do arise with this account, particularly with respect to the code model *of language* implicit in the relevance-theoretic proposal; I close the section with a look at these and proposed solutions.[54]

3.3.2.1 Recanati's Triad

According to Recanati (2001b), bringing clarity to the original two-level distinction *and thereby to the issue of levels of meaning* calls for two further distinctions – beyond those given by Grice – to be made:

> Anyone who has reflected on the sentence meaning/utterance meaning distinction knows that a simple distinction is in fact insufficient (Recanati, 2001b: 75).

The first distinction he calls for is between the linguistic meaning of the sentence-type and the proposition expressed by an utterance of the sentence. Recanati illustrates this with the contrast between the sentence-type 'I am French' and an utterance of 'I am French'. The sentence-type has a certain linguistic meaning *as a sentence-type* which *because it is a sentence and not an utterance* is context-invariant. On the other hand, the proposition expressed by an utterance of 'I am French', or, in other words, 'what is said' has a context-dependent meaning.

The second distinction is between what is 'actually' said by an utterance and what is 'merely' suggested. What is actually said by an

---

[54] Notice that in chapter 2 (§ 2.9.1) the issue was relevance theory's rejection of the code model of *communication*; the rejection of the code model of *language* is a distinct issue.

utterance of 'I am French' depends on who says it. If Recanati says it, it means RECANATI IS FRENCH; but this utterance can also convey much more. Imagine you ask the philosopher whether he can cook, in this case, an utterance of 'I am French' would clearly count as an *implied* affirmative answer.

Since 'what is said' appears twice in these distinctions, the result is a triad:

> sentence meaning
> *vs*.
> what is said
> *vs*.
> what is implicated

The importance of this triad is what it reveals about sentence meaning and propositionality. As expected by the product tradition, meaning at the first level is composed of the conventional, linguistic meaning of the words and represents context-*in*dependent meaning. Two of the features that we saw were important for formal semantics are present here: literal conventional meanings and context-independence. A third feature, however, also of great importance to traditional semantics, *being propositional*, is not yet present at this level. In Recanati's triad, in fact, propositionality starts only at the second level, with context-dependence.

Consider once again the sentence-type 'I am French'. In the view that emerges from revisiting Grice's distinction, it is 'skeletal' because, as it stands, it *constrains* how the context should intervene to make an utterance of it truth-evaluable (generally, the speaker of the utterance must be French in order for the utterance to be true) but, short of a full situation of utterance (someone actually producing the utterance), it is not yet properly 'enriched' (Recanati, 2001b, 2003). The conventional meanings of the constituents of the phrase do have the *potential* to say something true or false, but to do so, there must be an act of *utterance* (or at least of *thought*).

Finally, consider the processes at work at each level of meaning. What is said is a 'fleshing out' of sentence meaning. The propositions arrived at through this process are virtually indefinite but they are strongly

constrained by the sentence meaning. For instance, 'I am French' can express that Recanati, or *anyone else who utters the sentence*, is French; but, it would not likewise be able to express the proposition that kangaroos have tails. This changes with the mechanism of implicature introduced in the third level. At this point, an utterance can pretty much communicate anything, simply because inference chains can join up quite distant meanings so that 'I am French' uttered by Recanati can mean that Recanati can cook or, provided the right context, just about anything including that kangaroos have tails (Recanati 2001b: 76).

3.3.2.2 Minimalism vs. Maximalism

Recanati's two distinctions can be mapped back onto a single distinction between sentence meaning and utterance meaning. Deciding which levels go together results in two very different interpretations of the triad. If, for instance, the notion of literal meaning is privileged over the notion of speaker meaning (as in formal semantics) then the two levels would look something like this:

Literal meaning {sentence meaning / what is said}
vs.
Speaker's meaning

If pragmatic processes are privileged, then the division would look like this:

Sentence meaning
vs.
Speaker's meaning {what is said / what is implicated}

Recanati (2001b) gives each of these interpretations a name: 'pragmatic minimalism' and 'maximalism' respectively. The first is the widely held view that 'what is said' results from taking a sentence-type meaning and filling the slots with certain contextual elements, for instance, choosing an actual referent for the personal pronoun *I*. Pragmatic minimalism corresponds to the view presented above as formal semantics in which the role played by context is minimal. Critically, in pragmatic minimalism, it is necessarily something *in the sentence* that triggers the process whereby the sentence is

made propositional. Recanati calls this type of process 'saturation' and contrasts it with other contextual processes that can intervene without the need of such a trigger. These latter processes are, according to pragmatic minimalism, external to 'what is said'.[55] Bach (1994), who also holds that sentence-type meaning needs to be completed before becoming propositional, has provided some (now well-known) examples to show how minimalism breaks down:

(5) I've had breakfast[56]
(6) You're not going to die.

Arguably, (5) expresses the proposition that the speaker has had breakfast before the time of the utterance, *at any time in her life, at least once*. Formal semantics would allow that an act of utterance is necessary in order for the slot provided by 'I' to refer to someone in the world. Suppose the phrase has been uttered by my neighbour in response to my offer of fresh-baked blueberry muffins. Despite this completion, *the only one formal semantics calls for*, the truth-conditional proposition expressed could not be what we would take the speaker to have communicated since it is vacuously true: my neighbour has had breakfast at least once in her life before her utterance; a vacuously true statement (hence one that is neither informative nor relevant) is unlikely to be taken as the 'intuitive' content of what she has said.

In (6), imagine a mother is comforting a child who has just scraped his knee. Under the pragmatic minimalism construal, what she has said is blatantly false, as the child is mortal and will someday die. But this is not

---

[55] As is often necessary, the picture I am giving here of minimalism and maximalism is somewhat simplified since theorists working on either side are bound to disagree on the details. One particular approach deserves to be mentioned: Jason Stanley (2000) holds that there is something like a hidden indexical in the logical form of phrases such as 'I've had breakfast' that triggers the saturation 'this morning' thereby eliminating the need for other (freer) pragmatic processes such as suggested by Recanati. The problem with Stanley's solution is that even if there were hidden indexical-like constituents in the logical form, which is debatable, there would still be the issue of whether (and how) indexicals are in fact interpreted without taking speaker's meaning into consideration, that is, in the absence of full-fledged pragmatic processes.

[56] I return to this example, which originally appeared in Sperber and Wilson (1986) and discuss it from a particularly relevance-theoretic perspective below, § 3.3.2.3.

what anyone would take her to have meant. Intuitively, what the mother means is that he is not going to die *from that cut*, just as my neighbour means that she's had breakfast *that morning*. The problem is that in neither of these cases is there something in the sentence itself which calls for these elements, so under minimalism, the two aspects which actually make these utterances meaningful, that is 'this morning' for (5) and 'from that cut' for (6), cannot be considered as part of 'what is said'. This is problematic since, arguably, one of the motivations behind Grice's distinction is that what is said should serve as the input to what is implicated. Notice that what my neighbour implies with her utterance of 'I've had breakfast' is that she is not hungry and I can only infer this if I take her to have said that she had breakfast *this morning*; just as the mother implies that the child should not make such a fuss because he is not going to die *from that cut*. Minimalism construes these two completions as merely implied thus likening them to the case of implying that someone is a good cook by uttering that that person is French. On these grounds, Recanati rejects minimalism and argues for maximalism.

Relevance theory and Recanati's respective reformulations of Grice coincide on most important points: both favour a fully pragmatic (speaker-meant) notion of explicit utterance content so their notions of 'what is said' (or 'explicature') are similarly opposed to a purely semantic construal of 'what is said'.[57] They therefore coincide on what the linguistic underdeterminacy hypothesis amounts to: in Recanati's terms, it is the gap between the truth-conditional proposition expressed by an utterance of 'I've had breakfast', for instance, and the context-independent meaning of the sentence-type. In the

[57] The difference between the two approaches is that while, for Recanati, the two different levels of sentence meaning and speaker's meaning are connected to different types of processing: associative and inferential, for relevance theorists, there is no purely associative level of meaning. This is the single most important difference between relevance theory's and Recanati's views on 'what is said'. For the former, deriving what is said is a full-blown pragmatic process throughout, while for the latter, the associative level is simply a matter of undirected concept associations and activations with contextual best fit determining the one that is accepted. Since the parallels between the two perspectives are far greater than the differences, I do not focus on this issue here (but see Recanati 2001b, 2004; and Carston 2007, 2013).

following two subsections, I adopt relevance theory's perspective on these issues and give a succinct account of some of relevance theory's most important contributions to cognitive pragmatics. I focus on those that have led to particular developments in my own field: lexical pragmatics. First, I give an account of the division of labour between semantics and pragmatics from a relevance-theoretic perspective. Most importantly, this involves putting emphasis on pragmatic processes of enrichment that affect not only the implied content of an utterance but also the explicit content.

3.3.2.3 Relevance Theory's Enriched Explicit Content

The starting point for Sperber and Wilson (1986/95) is their goal of explaining how speakers communicate their thoughts. They take it as given that the overall process is *inferential*, and that it depends on the recognition of the speaker's communicative intention (a complex higher order intention), which is achieved by following their relevance-based principles. The interpretation process they describe includes, like all the others we have seen so far, reference assignment and disambiguation; but, critically, it also includes otherwise completing or enriching the semantic representations recovered from linguistic decoding.

Let's begin by looking back at example (5)

(5) I've had breakfast.

I said above that the completions authorised by a strict division of labour between semantics and pragmatics resulted in supposing that my neighbour's utterance semantically expresses that *she has had breakfast at some time in her life, at least once*. The further content, namely that she has had breakfast *on that very day*, would, in the Gricean, and in many contemporary minimalist semantics frameworks, be considered to be pragmatically implied. The reason for this is that there is, arguably, no linguistic element in the sentence-type that calls for such a completion. This is supposed to contrast with indexicals that can be thought of as slots calling for saturation. In the minimalist conception, if there is no slot, the suggested

completion would be an *unjustified* pragmatic *intrusion*. Rather than allow such intrusions, minimalism opts for a strict observance of lexical decoding and syntactic structure to arrive at a purely semantic construal of 'what is said'. A purely semantic construal, however, can completely miss what the speaker explicitly *meant* by his utterance.

To address this, relevance theory embraces the notions of 'free' pragmatic enrichment (originally introduced by Recanati in 1993) and 'unarticulated' constituents (originally introduced by John Perry in 1986). Adoption of these notions constitutes a rejection of the semantic doctrine according to which anything not articulated in the linguistic form falls outside of what can be taken as *explicitly* expressed. A counterintuitive consequence of the minimalist position is that speakers of everyday utterances like (5) above would turn out not to explicitly say much at all. According to Carston (2009a), if one adopts the minimalist perspective, the speaker in the situation described could not be taken to have explicitly communicated (i.e. 'meant' or endorsed) any thought at all by uttering 'I've had breakfast'; she would not have made an assertion, but merely implicated something. Furthermore, cases like this one abound: many utterances involving quantification, definite descriptions or vagueness would need to be similarly analysed. Consider the following utterances:

(7) Everyone cried when Brazil lost.
(8) The bakery is closed.

And, the well-known example,

(9) There's milk in the fridge.

Briefly imagine a suitable context for each: suppose that the first speaker meant that 'Everyone *who supported Brazil* cried when Brazil lost *the 2014 World Cup*', the second that 'The bakery *in our town* is closed', and the third that 'There is *sufficient* milk *and of suitable quality for coffee* in the fridge'. These additions would all be considered merely implicated so that, although the speakers did not mean that *everyone in the universe cried*, or that *there is one and only one bakery in existence and it is closed*, this is what they would

be taken to have *said* on those minimalist accounts that equate 'what is said' with the proposition semantically expressed (Carston, 2009a; Carston and Hall, 2012; and references therein). Furthermore, because minimalism holds that this minimal level is propositional and truth-evaluable, the proposition semantically expressed by an utterance of 'Everyone cried when Brazil lost' would almost certainly be false since it is unlikely that *everyone in the universe cried when Brazil lost*. Intuitively, however, it is not that *everyone in the universe cried*, but rather, through contextual sensitivity, that *everyone in a certain group cried*. Crucially, the speaker commits herself to the truth of the *enriched* proposition. In the 'There's milk in the fridge' example, if all that remains is a stale drip of milk, the hearer is justified in feeling misled. Furthermore, back to example (5), it would be the enriched content 'I've had breakfast *this morning*' that provides the crucial premise for further inferences such as *my neighbour is not hungry* and, therefore, it is deriving *this* content that should be represented as a key subtask of any psychologically plausible comprehension process (Carston 2009a, section 3).

The above observations led relevance theory to abandon Grice's distinction between what is said and what is implicated and propose instead a distinction between what is *explicitly* and *implicitly* communicated: in (5), between, for instance, 'I've had breakfast *this morning*' and 'I am not hungry'. The introduction of this amended distinction is important because it openly challenges the Gricean and minimalist assumption that pragmatic principles cannot make any contribution to what is explicitly communicated. Furthermore, it neatly captures the fact that, from a psychological point of view, the only relevant levels of propositional meaning are what the speaker can be said to have explicitly communicated, labelled the 'explicature', and what she implicitly communicated (or implicated), labelled the 'implicature(s)'.

Ideas introduced in the above section are more fully explained below in discussions of relevance theory's utterance comprehension process (including processes of ad hoc concept construction). First, however, I

would like to touch on some theoretical and terminological issues related to the way the notion of 'explicature' has been presented in the literature: namely, its connection with notions such as decoding and logical form. In an early presentation of explicature, Sperber and Wilson (1986/95) write:

> an assumption [or proposition] communicated by an utterance *U* is *explicit* if and only if it is a development of a logical form encoded by *U* (p. 182).

To best capture what this means, emphasis must be placed on the fact that explicature is a *development* of logical form. Logical form is traditionally defined as a configuration of lexical and structural information associated by automatic parsing (the language module) to an utterance. Jason Stanley, for whom this configuration is quite rich, equates postulating logical form with adopting the assumption that

> syntax associates with each occurrence of a natural language expression a lexically and perhaps also structurally disambiguated structure which differs from its apparent structure, and is the primary object of semantic interpretation. […] In accord with standard usage in syntax, I call such structures logical forms (2000: 393).

In standard syntax and the traditional view of the division between semantics and pragmatics, the explicit content of the utterance is equated with a *strict* decoding into logical form and *semantic* (not pragmatic) interpretation. But adopting this view would lead back to the minimalist conception of a purely semantic construal of 'what is said'. To avoid this, relevance theory amends the notion by stipulating that *the logical form of an utterance underdetermines the propositional form expressed* (Sperber & Wilson, 1986/95) and that it, therefore, must be *developed*. The development, of course, includes fully pragmatic processes. Furthermore, in the last decade or so, relevance theory has expanded this idea of 'development' from the level of the 'phrase' to that of individual words which are now said to be 'pragmatically adjusted and fine-tuned in context, so that their contribution to the proposition expressed is different from their lexically encoded sense' (Wilson and Carston, 2007: 230).

3.3.2.4 Relevance Theory's Ad Hoc Concept Construction and Comprehension Procedure

In this subsection, I briefly present the relevance-theoretic pragmatic process of adjusting or *modulating* word meaning known as ad hoc concept construction; I focus on the role it plays in the global relevance-theoretic utterance comprehension procedure. But, because I later come to reject one particular aspect of the account, *namely* its adherence to the code model of *language*, I start this section, a continuation of the above paragraphs, with some words on relevance theory's use of the terms 'code' and 'encoded'. I suggest that a key consideration in understanding the use of these terms is that Sperber and Wilson very explicitly reject the traditional code model of *communication* which assumes that linguistic sentences correspond to thoughts through straightforward matching, that the content of a thought is transmitted in the message because it has been *entirely* captured by the signs of a particular code. Sperber and Wilson hold, on the contrary, that utterances are 'pieces of evidence about the speaker's meaning, and comprehension is achieved by inferring this meaning from the evidence provided' (Sperber and Wilson, 1998 reprinted in Wilson and Sperber 2012: 61). Yet, the relevance-theoretic comprehension process is still often framed as involving *decoding* which suggests that while rejecting the code model *of communication*, Sperber and Wilson endorse the code model *of language*. The main reason for this choice is perhaps Sperber and Wilson's adherence, following Fodor (1975), to the idea of two *kinds* of semantics. Wilson (2003) explains that the relevance-theoretic linguistic semantics model she supports

> …will adopt a simple model of linguistic semantics that treats words as encoding mentally-represented concepts, elements of a conceptual representation system or 'language of thought', which constitute their linguistic meanings and determine what might be called their linguistically-specified denotations (p. 273).

It seems that conceiving of language as a code goes together with the adoption of the division discussed in chapter 2 (§ 2.7.2) between linguistic

or 'translational' and 'real' semantics. In this perspective, 'code' is unproblematic because, despite suggesting unchanging, air-tight pairings, the pragmatic processes described by relevance theory account for the fact that the concept *communicated* is not necessarily the concept *encoded*. So, Sperber and Wilson's support for the notion of code is relativised by the fact that they disagree with Fodor on the prevalence of one-to-one mappings between words and concepts, and further by the general emphasis on inference-based comprehension processes that is characteristic of their approach.

To illustrate, consider the following: according to Wilson (2003: 273), the word 'drink' has a 'linguistically-specified', 'literal' meaning; in accordance with the idea of core meanings, suppose this means that 'drink' generally translates into the Mentalese DRINK; notice that so far the account is compatible with Fodor's 'disquotational lexicon'. But now imagine 'drink' in an utterance: at a party, suppose I say of a friend 'I hope he doesn't drink'. Relevance theory suggests that the interpretation of the word 'drink' in such an utterance depends on pragmatic processes intervening partly *as the word 'drink' is being translated into the language of thought*. These pragmatic processes modulate DRINK to create an occasion specific concept DRINK* which in this particular context denotes drinking *alcoholic* drinks (and more specifically, *significant amounts of alcoholic* drinks) rather than liquids in general. So it is important to note that the code model of language is complemented with the processes of lexical pragmatics. According to Wilson:

> the goal of lexical semantics is to investigate the relations between words and the concepts they encode, and the goal of lexical pragmatics is to account for the fact that the concept communicated by use of a word often differs from the concept encoded (2003: 273-274).

This process, labelled 'ad hoc concept construction' is described in detail in Wilson and Carston (2007).[58] At the end of this chapter, however, I argue

[58] See also Carston, 2002c, 2010a, 2010c.

that, despite these adjustments, the notion of 'code' and related notions of 'encoded concepts' are still problematic.[59] I argue, furthermore, that given the compatibility of the alternatives to 'code' with the relevance-theoretic project, and given the most recent developments in lexical pragmatics, a positive move forward for relevance theory would involve rethinking its commitment to the notion that words *encode* concepts.[60] But before this, I briefly focus on one more very important contribution of relevance theory: as stated already in the introduction to this section on the contributions of cognitive pragmatics (§ 3.3.2), relevance theory is one of the very few approaches to not only *theoretically* reject minimalism but to actually propose a material account of how arriving at 'what is said' involves pragmatic processes.

The relevance-theoretic utterance comprehension process was first described fully in Sperber and Wilson's 1986 publication so it is important to note that the original picture predates reflections on ad hoc concept construction. However, ad hoc concepts are such an important part of the current picture, and so naturally follow from the original outlook on the comprehension process, that this point, once taken note of, can be put aside. What I offer here is a presentation of the *standard* relevance-theoretic picture that includes an account of ad hoc concept construction. There is also an issue of the *evolution* of ideas about ad hoc concept construction *within relevance theory*; these differences are important since serious doubts about certain aspects of the standard account, as represented by Wilson (2003) and Wilson and Carston (2007), for instance, have arisen within relevance theory. I come back to these objections at the end of this

---

[59] Assimakopoulos (2012) has also recently discussed whether relevance theorist's commitment to Fodorian semantics is compatible with new developments within the theory, particularly ad hoc concept construction and Robyn Carston's position on whether modulation is optional or mandatory.

[60] Spelling out an alternative to the 'code' model of language and to the idea that words encode concepts involves more arguments than I could mention here. I come back to possible alternatives to the 'code' model shortly with references to the notion of 'entrenchment', as developed by Beckner et al (2009) and Hans-Jörg Schmid (2008), at the end of this chapter (section 3.6 'Closing remarks') and in subsequent chapters after a discussion of psychological factors. For the issue of whether words encode concepts, a whole section (3.5 'The second stage: abandoning the modular view') is dedicated to presenting an alternative.

section; but first, it is necessary to present the *standard* version of the comprehension procedure which integrates a certain vision of ad hoc concept construction since it is this version that is the best known.

The process of utterance comprehension is generally described as a non-demonstrative inferential process. It is triggered by an utterance (or ostensive stimulus, more generally) and is guided by the presumption of optimal relevance conveyed by all ostensive stimuli.

Relevance-guided comprehension heuristic:

(a) *Follow a path of least effort in constructing an interpretation of the utterance (and in particular in resolving ambiguities and referential indeterminacies, in going beyond linguistic meaning, in supplying contextual assumptions, computing implicatures, etc.*

(b) *Stop when your expectations of relevance are satisfied* (Wilson and Sperber, 2012: 7)

The process is considered effective when it has constructed appropriate hypotheses concerning the explicit and implicit content. A step-by-step description of the whole process is as follows: the *language* module intervenes first, it recovers the 'linguistically encoded meaning' of the utterance and feeds it to the *pragmatic* module. The *pragmatic* work is construed as involving sub-tasks which are carried out in *parallel*, not sequentially. In the picture of the standard position I am drawing here, once the linguistically encoded meaning has been recovered, the pragmatic module follows a path of least effort, enriching it both at the explicit and implicit level until expectations of relevance are met, or abandoned.

Arriving at an overall interpretation which satisfies the hearer's expectations of relevance involves a process of parallel mutual adjustment of the explicit content with the contextually derived assumptions and the cognitive implications; this process describes *how* expectations of relevance constrain and guide interpretive outcomes, or 'cognitive effects' in relevance theoretic terminology, so it is the 'central feature of relevance-theoretic pragmatics' (Wilson, 2003: 283).

To illustrate this complex process, suppose that I receive the following message on my phone:

(7) Be an angel and pick up some bread on your way home.[61]

Central to the *standard* relevance-theoretic account of how such an utterance would be interpreted is the idea that the words 'angel', 'pick up', 'bread', etc. have 'linguistically-specified meanings', or, in other words, that they *encode* the concepts ANGEL, PICK UP and BREAD. Recovering these encoded 'meanings' or concepts is necessary in order to access the various entries (the logical, encyclopaedic and lexical entries discussed in chapter 2, § 2.7.3) 'filed' under, or associated with, these concepts and proceed to their *modulation*. According to Wilson and Carston (2007), 'angel', for instance, encodes the concept ANGEL which activates a certain 'range of logical properties', among them, possibly 'an angel is a SUPERNATURAL BEING OF A CERTAIN KIND'. In their (2007) example: 'Sally is an angel', this activation would enable certain deductive inferences to be drawn; for instance, 'from the proposition that Sally is an ANGEL, it is deducible that Sally is a SUPERNATURAL BEING OF A CERTAIN KIND' (Wilson and Carston, 2007: 247). The encoded concept also allows access to the encyclopaedic entry, that is, a particular subject's wide collection of information related to angels, everything from scientific and culture-specific beliefs to personal and idiosyncratic representations.[62] Accessing ANGEL would activate properties of different subsets of angels such as 'good angels', 'guardian angels', 'avenging angels', 'dark angels', and so forth, thereby enabling further possible conclusions. But because a context such as (2) would carry the assumption of 'good angels', Wilson and Carston list the following as the encyclopaedic properties which are plausibly the *most highly activated* in this context:

(i) EXCEPTIONALLY GOOD AND KIND.

---

[61] I have adapted this example from Wilson and Carston's (2007) 'Sally is an *angel*'.

[62] In fact, this collection of information includes anything stored in memory that could be relevant to this particular interpretation. Carston (2002a) describes the encyclopaedic entry as containing, among other things, scientific information, commonplace assumptions, culture-specific beliefs, and personal experiences; this information, she adds, is stored not only in propositional form but also in scripts and scenarios, and in 'analogue form', perhaps as mental images of some sort (p. 321). The importance of this very wide construal of the encyclopaedic entry will emerge later in this thesis.

(ii) WATCHES OVER HUMANS AND HELPS THEM WHEN NEEDED.
(iii) VIRTUOUS IN THOUGHT AND DEED.
(iv) MESSENGER OF GOD, etc.

The context would further selectively activate certain properties more than others. I am being asked to *do* something, so any properties related to *helpfulness*, as in (ii), or *goodwill* as in (iii), would receive additional activation. Here, however, we come to a crux in the account: using (i) – (iv) above and following the comprehension process, I, as the interpreter of this message, am supposed to tentatively consider some contextual implications such as 'I AM BEING ASKED TO BE EXCEPTIONALLY GOOD AND KIND', 'TO WATCH OVER HUMANS AND HELP THEM WHEN NEEDED', 'TO BE VIRTUOUS IN THOUGHT AND DEED', etc. But these contextual implications, Wilson and Carston tell us, are 'not yet properly warranted'; the problem seems to be that *actually*, I am *not* a supernatural being and therefore *not* an ANGEL. However, by narrowing and broadening ANGEL so that it is limited to good angels and extended to *people* who share properties with good angels (as in (i) – (iv) above), I can take the message as asking me to be an ANGEL*, a modulated version of the concept ANGEL that could be paraphrased as SOMEONE PARTICULARLY HELPFUL (AND SOMEHOW ANGEL-LIKE).[63]

At this stage, through a process of mutual parallel adjustment, an ad hoc concept has been created: parts of this process can be construed as *forward* inferencing, for instance, the inference that I am being asked to be a GOOD ANGEL, and some are *backwards* inferences, I go *back* to the proposition derived from the 'linguistically decoded meaning' and replace ANGEL with ANGEL* (Wilson and Carston 2007: 248). I proceed in a similar way with the other constituents of the utterance. BREAD, for instance, would arguably be narrowed to a specific sort of bread.[64] At this point, the modulated ANGEL*,

[63] It should be noted that, according to Alison Hall (2011), ANGEL* is an atomic concept. The lengthy description is intended only to reflect the fact that this atomic concept is not lexicalised in English.
[64] I come back to BREAD just below, but for now, I would just like to point out that showing up with just *any* kind of bread would not count as complying with the demand expressed, whether you consider that it was *explicitly* expressed or not.

BREAD*, etc. can be plugged back into the comprehension process: the ad hoc concept construction process is a process embedded within a process.

Another relevance-theoretic contribution now plays its role: the ad hoc concepts ANGEL*, BREAD*, etc., other concepts in the utterance, contextual information and pragmatic expectations interact to produce an enriched conceptual representation which is the *explicature* of the utterance: what one interlocutor can take the other to have explicitly communicated; for (7), this can roughly be paraphrased as '*Be very kind/helpful and buy from a shop the bread we usually get and bring it home'*. It is this enriched representation that serves as a premise in the deduction of implicatures, but, just as with ad hoc concept construction, this does not mean *sequential* processing; explicatures need not *precede* implicatures in the comprehension process. Rather, the process depends on explicatures, implicatures and contextual assumptions being mutually adjusted in parallel 'until they form an inferentially sound relation, with premises (explicature, contextual assumptions) warranting conclusions (implicatures)' (Carston & Hall, 2012: 69).

3.3.3 Problems with the Code Model of Language and Possible Alternatives

As announced earlier, I am particularly interested in this account's reliance on the *code* model of language, and concerns this might raise. I therefore now turn my attention to some very interesting remarks in Carston's (2002a, chapter 5) section on 'word meaning and concepts' where she presents and discusses some of the very issues that my account raises. Her starting point is the possibility that the assumption that *words encode concepts* could be wrong. More precisely, that lexical items such as 'cat', 'open', 'raw' and 'happy', to which we could add 'angel' and 'bread', *despite explicit claims to the contrary*, do not encode concepts (whether atomic or not). As stated above, relevance theory's espousal of the idea that words encode full-fledged atomic concepts is in accord with the Fodorian picture of both concepts and word meaning. The danger I see in this is that ad hoc concept construction could then be interpreted as a *complement* designed to handle *those* cases in which lexical items communicate a concept different

from the one they encode. For instance, in the example above, there is a clear distinction between a 'literal' and a 'figurative' (or *loose)* construal of 'angel' suggesting that it is *these kinds of cases* of 'loose use' that call for modulation. According to Robyn Carston,[65] however, this is not what relevance theorists intended when formulating their account of ad hoc concept construction and the topic of ordinary modulations has received some attention in relevance theory, notably by Carston (2002a). Yet, I would argue that a more thoroughly contextualist view would need to insist on the point that even non-figurative, straightforward uses of words such as 'bread' in the example are adjusted in their contexts.

Carston's (2002a) section on 'word meaning and concepts' presents a related but different challenge: she tentatively proposes, as an alternative to the idea that *it is full-fledged concepts that words encode*, that what they encode are *concept schemas*. These are described as 'pointers to a conceptual space, on the basis of which, on *every* occasion of use, an actual concept (an ingredient of a thought) is pragmatically inferred' (Carston 2002a: 360). Yet, she also finds some aspects of this proposal problematic and so the issue of a definite alternative to what I called the *standard* account above is, at the end of Carston's section, left largely open.[66] Nevertheless, the arguments presented *for* concept schemas *and against* them are well worth close inspection.

Carston reflects on the word 'happy' and the concept that it is supposed to encode. According to the *standard* account, 'happy' both encodes a general and abstract concept HAPPY and provides the basis for arriving at much more specific concepts that 'happy' can be used to communicate, for instance, HAPPY* 'a moment of intense joy', or HAPPY** 'a steady low-key contentment'. Yet, if all the thoughts we have, and therefore all the thoughts *we communicate* regarding HAPPY are *specific sorts* of HAPPY, rather than general and abstract ones, what does it mean to have the general and abstract 'Fodorian' concept HAPPY? Does HAPPY, rather than HAPPY* ever

[65] Robyn Carston, personal communication.
[66] Very recently, Carston has come back to these issues with force in two papers (Carston 2012, 2013). I discuss the contributions of these papers later in this chapter (§ 3.5.3).

actually figure as a constituent of thought?[67] The alternative is to say that "happy' does not encode a concept, but rather points to a conceptual region' to *something* in memory (Carston, 2002a: 360).

Much the same thing can be said about the verb 'open': trying to have a thought with the very general concept OPEN, instead of with one of the specific understandings, such as 'open one's mouth', 'open a discussion', or 'open a can', is an odd experience 'as we seem to have no thought at all' (Carston 2002a: 361). Furthermore, if it were true that 'open' encodes a concept, would we not be able to construct *some* proposition, bizarre but *evaluable* for Searle's well-known examples: 'Bob opened the grass' and 'Chris opened the fork' ?

A good alternative seems to be to stop looking to a general and abstract concept OPEN as *the* element that provides access to the information needed for interpretation, and rather look to *the word* 'open' as a gateway to our vast stores of information related to *opening* in memory. The claim would be that the lexical form maps to a 'conceptual address' in memory to which are attached packages of information. Carston credits this view, which will receive much more detailed attention throughout this thesis, to the psychologist Lawrence Barsalou. Roughly, a selective process would pick out of the packages only those bits of information relevant to the current context and the result would be a concept.[68] For lexical pragmatics, this would mean that the constituent of thought that the speaker is communicating would be built not from any decoded meaning of the word but *entirely ad hoc* from very general information in memory.

Arguably, this proposal goes *beyond* limiting what a word encodes to a concept schema instead of a full-fledged concept. What is actually being

---

[67] The assumption of the classic account is that there *is* a core meaning of HAPPY that somehow exists in our minds. Whether this core is somehow *innate* or rather *learned*, for instance by extraction from contextual meanings, is the topic of future chapters. Here the issue is not whether this core/abstract meaning exists but *assuming it does*, would it play the role of a constituent of our thoughts?

[68] Notice that *this* is what Barsalou (1987) calls ad hoc concept construction. In his original proposal, there is no assumption that words encode concepts, so the two conceptions of ad hoc concepts are in fact very different (see chapter 4, section 4.4 for more on Barsalou's ad hoc concepts).

considered is a much more radical rearrangement of the process of occasion-specific word meaning construction itself. This rearrangement involves at the very least an important shortcut: it would seem that instead of going from 'happy' to HAPPY to HAPPY* *via the activation of encyclopaedic information filed under HAPPY*, 'certain bundles of information' are 'pointed at' by *the word* 'happy' and processed in the relevance-constrained way so aptly described by relevance theory to arrive at the conceptual unit communicated by the speaker (i.e., HAPPY*) (Carston 2002a: 360-361).

Notice that, clearly, in this version, the Fodorian HAPPY is left without a role to play in the construction process. It is less clear, however, whether Carston's proposed *concept schema* plays a role. If a concept schema is still considered some sort of encoded meaning, that is, something *encoded* by the word, then it seems to play no role in Barsalou's ad hoc concept construction since a conceptual content is the *end result* of the process rather than what the word 'points to'. Alternatively, if we say that the 'space' the word points to is a space *in memory*, then there is no longer anything necessarily *conceptual* or *schematic* about Carston's alternative.

There are other related objections to consider. Carston recognises that the acquisition story for concept schemas is unclear. In the Fodorian framework, learning the word 'open' would be a matter of learning to associate this word with the property (or properties) it refers to in the world. There would be two acquisition processes: one for the concept OPEN, that expresses the property of, say*, openness;* and one for the form-concept link, the English 'open' or the French 'ouvrir' for OPEN. Accounting for this is already complex, now imagine that the child must acquire an abstract entity (a *schema*) that is *other than* that concept (i.e., OPEN*) which actually plays a role in her thoughts. This would be the case if, as suggested above, rather than having a general concept OPEN, the child had stored memories of a variety of specific cases of actions, such as 'opening her mouth', 'opening the door', and 'opening a carton of milk' from which she somehow had to *extract* an adult *schematic* meaning for OPEN. Importantly, this does not answer the question of how this general schema is supposed to arise and *become* the

lexical expression *type*. In her closing to this section, the problem is rephrased as follows:

> There must be some process of abstraction, or extraction, from the particular concepts associated with the phonological form /open/ to the more general 'meaning', which then functions as a gateway both to the existing concepts of opening and to the materials needed to make new OPEN* concepts which may arise in the understanding of subsequent utterances (Carston, 2002a: 364).

But *must* there really be a general, 'all-purpose' meaning acting as gateway to occasion-specific meanings? As Carston mentions in a footnote, Recanati (1998), following the work of Douglas Hintzman (1986), has a radical alternative solution, ominously labelled the 'eliminativist' approach, according to which the idea that words encode anything stable must simply be given up:

> Hintzman's model does not appeal to the notion of the literal meaning of the word-type. Words, as expression-types, do not have 'meanings' over and above the collection of token-experiences with which they are associated. The only meaning that words have is that which emerges in context (Recanati, 1998, section 16 'Cognitive science and contextualism', cited by Carston 2002a: *n*. 16 p. 375).[69]

On this view, the child learns *what she is exposed to*, that is, particular pairings, and figures the rest out *on the fly* without ever having to construct a *stable encoded* meaning. This is the alternative I take very seriously in this thesis.

To summarise where we have gotten to so far, the accounts presented in this section on 'The contributions of cognitive pragmatics' are a considerable improvement with respect to accounts prior to the advent of contemporary

[69] Recanati (1998) presents eliminativism as an approach emanating from the cognitive sciences and constituting a return to the radical positions of early ordinary language philosophers. In his (2004) book *Literal meaning*, Recanati describes eliminativism as the most extreme position on a gradient. I return to the other possible positions on the gradient at length towards the end of this chapter (§ 3.5.1).

cognitive pragmatics. They offer an undoubtedly clearer picture of utterance understanding and move decidedly forward on accounts of the comprehension of individual words and expressions within utterances (lexical pragmatics).

However, I believe that the most important contribution is the recognition of the pervasiveness of context-dependence in this picture. The traditional view of context-dependence was mostly dismissive: a certain view of the semantic-pragmatic divide mistakenly paired context-dependence with *non*-truth conditional content and therefore labelled context-dependence of only *secondary* interest. As a result, it was generally downplayed and relegated to a 'lesser', subordinate discipline (i.e., pragmatics). The only imaginable advantage of this solution is that no *fundamental* changes to the discipline of semantics would be required. Nevertheless, the picture drawn by pragmatists of the importance of context-dependence and the sheer volume of evidence brought forward to support a more in-depth analysis makes this position untenable. Even where resistance to any fundamental changes is dogmatic, context-dependence has stirred undeniable interest. Of course, this does not mean convergent interpretations of the evidence, but, at the very least, an acknowledgement of a need to address the issue. Cappelen and Lepore, 2005, 2007; King & Stanley, 2005; and Borg, 2004, for instance, might only address context-dependence in an attempt to reject contextualism, but in doing so they let it take centre stage.

At the beginning of this chapter, I announced that after presenting the more established results of contemporary cognitive pragmatics, which I have done in section (3.3) entitled 'The first stage: from philosophical perspectives to cognitive pragmatics', I would move on to arguments calling for more extreme positions. This begins with the contributions I present as part of the section entitled 'Philosophical foundations for radical contextualism'. The main idea is that, free from the constraints previously imposed by semantic theorising, context-dependence can be seen as quite

the opposite of a *marginal* phenomenon and much more like a *defining* property of meaning in context.

3.4 Philosophical Foundations for Radical Contextualism

The traditional and most widely held position on word meaning in context puts semantics as the main discipline for issues of content and truth-conditions and *subordinates* pragmatics to semantics. Furthermore, as is usually the case when arguing against an established tradition, the burden of evidence is on those who disagree; understandably, the burden of evidence is also in proportion to the consequences of adopting the proposed changes. In the case of semantics and pragmatics, the minimum change calls for reconsidering the extent to which pragmatics 'intrudes', or *should be allowed to 'intrude'*, into semantics. The term chosen by the semanticist is deliberately laden with negative connotations: an intrusion is an *illegal* and *forced* entry. It suggests that 'allowing' pragmatic processes to make contributions to truth-conditional content would disrupt the order and must therefore be avoided. Calling these contributions 'intrusions' is clearly a cry of alarm on the part of semanticists. But what does this kind of talk reveal? If pragmatic contributions are necessary, then it is of little consequence to call them unwelcome; rather than a persuasive argument in favour of the established order, this seems to reveal a certain unease with a changing picture.

The amount and robustness of the evidence brought forth by contextualists even seems to point in the direction of a reversal of the hierarchy, if not a disappearance altogether of semantics as traditionally construed. The defining characteristic of the new picture is that context-dependence is no longer a problem that semantics solves by calling on pragmatics. In the new framework, context-dependence is not a problem at all, it is not even a 'marginal' phenomenon encountered in *some* language situations but rather is *itself the norm*. Acknowledging this would involve rejecting some very established positions in semantics. Very few theorists are ready to adi such a radical position. But the more evidence that is amassed by pragmatists, the more context-dependence seems to be

ubiquitous and deeply tied to how language works. This section takes up the challenge of presenting arguments in favour of such a move, which I propose to call '*radical* contextualism' (to differentiate it both from anti-contextualism and from more conservative versions of contextualism which maintain that words have fully semantic, context-independent, standing meanings). Radical contextualism experiences resistance from outside of pragmatics, particularly from semantics, and from within pragmatics, from theorists comfortable with one of the current positions on the semantics/pragmatics divide. There are strong arguments, however, as I hope to show in this section and the next, to support radical contextualism. They come from many sources: not only pragmatists, but linguists more generally, philosophers, psychologists, and artificial intelligence theorists. Understandably, semanticists are the most reluctant since accepting radical contextualism would involve acknowledging that semantics does not deliver truth-conditional content and thereby force them to completely rethink their discipline. The issue will not be settled soon, as, I believe, the phase for presenting evidence is far from over. But I also believe that simply framing the question differently can make a huge difference in winning support for radical contextualism. Because the discussion was previously always framed in a particular way, namely as pragmatics subordinated to semantics, certain insights have not received their full share of attention; if we accept the logic behind the kind of context-sensitivity illustrated in the previous section, labelled 'the first stage', now becoming part of the consensus, we should explore the logical limits of context-sensitivity to see what further consequences it could bring. I am particularly interested in the consequences for conceptions of word meaning. A fresh look at word meaning, I believe, involves breaking free from the framework of traditional formal semantics and pushing the limits of context-sensitivity beyond making a place for pragmatics *alongside* semantics.

With respect to the two contrasting traditions in language theorising described earlier in this chapter (section 3.2), radical contextualism is clearly in line with the action tradition; it inherits some of its inspiration from the later Wittgenstein (i.e., after his 'anthropological turn'), and from

the ordinary language philosophers of the 1950s and 1960s. A careful presentation of precise influences and connections amongst the Vienna circle, the Oxford natural language philosophers and present day theorists would undoubtedly enrich the discussion in this section but, unfortunately, is beyond the scope of this thesis.[70] I rather focus on a particular issue present across key contributions in the various traditions above: the unguarded *and systematic* underestimation of context-sensitivity in the traditional view of meaning. I begin with Friedrich Waismann's notion of 'open texture' and John Searle's 'background'. I move on to a brief discussion of Hilary Putnam's 'externalist semantics', which I justify as an attempt to address some issues left in suspense in chapter 2. My objective is to present, in as clear a form as possible, the arguments in favour of these notions and how they disprove the assumed irrelevance of context-sensitivity; I try to anticipate, as much as possible, the objections they might raise. A more detailed account of how I see word and utterance meaning being built in context must wait until after the 'second stage' in this chapter (section 3.5); and more detail will be added after the presentations in chapters 4 and 5.

3.4.1 Waismann's 'Open Texture'

A good place to start a chronological review of the ideas of radical contextualism is Friedrich Waismann's (1951) article introducing the notion of 'open texture'.[71] While commenting on a now widely discredited view of

---

[70] For a more in depth discussion of Friedrich Waismann, John Searle, Hilary Putnam and others having made significant contributions to contextualism, see Recanati's (2004) seminal book *Literal meaning*.

[71] In his 1951 article, Waismann, a prominent member of the Vienna circle, returns to one of the principles championed by the group initially under the sway of Wittgenstein during what Monk (1991) calls his 'verificationist phase' (circa 1929). Waismann is the only member to have engaged in long and frequent discussions with Wittgenstein on a range of topics including the 'principle of verification' which is still present in the theses Wittgenstein dictated to Waismann as part of their project for a book. But the book and theses are soon after abandoned as the transitional Wittgenstein morphs into the later Wittgenstein. Among possible speculations as to what brings about the change between the early and the later Wittgenstein, there are at least two important influences worth noting: the first was probably key in Wittgenstein's, at first hesitant, disowning of the *Tractatus*, and the second in his taking a clear anthropological turn. Frank Ramsey, a mathematician and philosopher at Cambridge, was among the first few who understood the *Tractatus*, and addressing one of his points of criticism was the objective of the only paper published by

meaning (i.e., verificationism), Waismann notes that in our everyday life, the way we understand sentences has little to do with purported methods of verification; rather, most understanding is straightforward. When it is not, the methods of verification called on in the literature fall short because we lack the set of predetermined rules that would make them efficient. Waismann holds that what has been overlooked is the 'open texture' of concepts:

> The fact that in many cases there is no such thing as a conclusive verification is connected with the fact that most of our empirical concepts are not delimited in all possible directions' (1951: 119-120).

To illustrate, he first asks us to imagine a dog owner encountered in a public park who declares 'My dog is intelligent'. If he had said 'My dog barks' or any other common, dog-related quality, we would not think twice about *what he meant* by his statement; but, according to Waismann, this question is immediately raised by the combination of DOG + INTELLIGENT *because of its novelty*.[72] In this case, Waismann tells us, we would surely ask the owner for an explanation and *with this explanation*, we would build a specific context for the understanding of this specific utterance. Critically, for Waismann, we cannot rely on definitions since a term is defined only when an exhaustive description can be given for how it may be used and such descriptions/definitions are beyond our reach because not only can we never be sure that we have taken every possible detail into consideration

---

Wittgenstein after the *Tractatus*, a defence no sooner begun than abandoned by Wittgenstein for whom the cracks in his earlier positions progressively become apparent. Wittgenstein also explicitly recognises the influence of an economist, Piero Sraffa, in the introduction to *Philosophical Investigations*; he mentions their discussions as the stimulus 'for the most consequential ideas in this book'. According to Monk (1991), it is from these discussions that Wittgenstein gained the 'anthropological' way of looking at philosophical problems which so characterises his later writings (p. 261).

[72] In today's context of greater understanding of the parallels between humans and animals, whether it be regarding their intelligence or emotions, Waismann's example might not seem novel in any interesting way. It must be understood in its 1950s context: it seems, although this is speculation on my part, that Weismann has never heard this adjective applied to anything other than humans and that the speaker *means* intelligent in the way Waismann has never conceived of it (i.e., as applying also to dogs), hence his need to *figure out* what the speaker meant. I come back to very similar examples in the section on 'norm theory' in chapter 4.

but it is quite impossible to predict future changes to actual conditions. Waismann's is one of the first accounts of linguistic underdeterminacy. His message will reappear in the writings of Searle and numerous contemporary pragmatists. The key point is that the meaning of everyday expressions like 'intelligent' do not have determinate conditions of application since these must be set *with respect to* a particular situation.

3.4.2 Searle's 'Background'

Searle's objective in his (1978) and (1980) articles is very clear:

> I want to challenge one aspect of [the] received opinion, the view that for every sentence the literal meaning of the sentence can be construed as the meaning it has independently of any context whatever (Searle, 1978: 207).

Searle clearly foreshadows the position taken today by theorists like Recanati who claim that it is not sentences but utterances that are propositional and therefore truth-conditional.[73] But Searle's contribution also has a direct impact on word meaning. As I will try to show, his message that contexts come with background assumptions carries important consequences for lexical pragmatics.

Searle's starting point is that, contrary to the received view, the difference between sentences and utterances is not the same as the difference between types and tokens. This confusion, like many others concerning meaning, stems from underestimating context-sensitivity. The received view holds that sentence-types have context-free meaning and that it is this context-free meaning that determines the applicability (i.e., the truth-conditions) of the sentence. Furthermore, it assumes that context-sensitivity is simply a question of *tokening* these sentences. The type remains context-free while the token takes on the contextual features of its

---

[73] This is also the position adopted by Moravcsik (1994) in his influential article 'Is Snow White?'. Moravcsik very clearly agrees with the action tradition's claim that only *utterances* have truth conditions.

tokening. For Searle, however, context-sensitivity is present at a much more basic level: it is not the result of superficial indexicality, of utterances of 'I am hungry' meaning different things depending on who utters them or the time of day at which they are uttered; rather, *context-dependency is a fundamental characteristic of language*, due to the fact that the conditions of application of words (and groups of words) are relative to a set of contextual, 'background' assumptions which vary from context of utterance to context of utterance.

To call attention to the importance of background assumptions, and demonstrate that they are wrongly taken for granted, Searle calls on the out of the ordinary, much like Waismann. In Searle (1978), he has his readers imagine the case of 'The cat is on the mat'. The illusion of a single, stable literal meaning for this sentence is produced by assuming that it describes the everyday situation of a household cat sitting or sleeping on a household mat and *that the addition of these background assumptions is only a function of interpreting the sentence 'literally'*. That this is not the case is readily made evident by the fact that the sentence can stay quite literal while the background assumptions change. Searle further has his readers imagine that the cat and the mat are floating freely in space, they are disposed as described by 'The cat is on the mat' but there is no force of gravity to this 'on'. Is the force of gravity something we should add to the semantic stipulations, that is, the conditions of application, of 'on'? This would hardly solve our problem since there are an indefinite number of such assumptions that we would have to deal with. The alternative proposed by Searle is to acknowledge the role of background assumptions and accept that 'the notion of the literal meaning of a sentence only has application relative to a set of background assumptions' (Searle, 1978: 214). It is only once the sentence is used, only once it is an actual utterance, that we can reflect on which background assumptions its truth-conditions depend on.

Finally, this context-dependency extends to words since words, like sentences, would also make different contributions to truth conditions depending on their use. Searle is aware that this challenges the accepted tradition since Frege: classic compositionality stipulates that each

constituent of a sentence has a definite semantic content in such a way that the literal meaning of a sentence is made up of the meaning of its constituents and the way they are put together.[74] Searle's point can be illustrated with some simple examples, adapted from Searle (1983): 'Open your eyes', 'Sam opened his book to page 37', and 'The surgeon opened the wound'. Searle argues that to understand the contribution made by 'open' in these sentences, one must both consider the semantic content that all of the occurrences share (i.e., the 'core' meaning of 'open' – which is conditioned to an actual application but not non-existent) and recognise that in each case 'open' contributes something different to the truth conditions. That there is a distinction to be made between the more literal and the derived or figurative meanings of 'open' is evident when we consider phrases like 'Fred opened a restaurant', and 'The artillery opened fire'. However, differences in truth conditions need not be a function of the difference between the literal and the non-literal; rather, truth conditions may be different among occurrences that are equally literal. Consider what it means to 'open a wound' and compare it with what it ordinarily means to 'open one's eyes'. One would involve cutting, and the other, *in the absence of an extraordinary context*, would not. These observations lead Searle to conclude that there is more to understanding than grasping the meaning of the words in a sentence. After all, he adds, grasping the meaning of 'open' and the meaning of 'mountain', for instance, is of no help in grasping the meaning of 'Open the mountain'.

As with Waismann, it is the familiarity of certain situations that hides what is involved in their interpretation. For occurrences of 'open' in 'open your eyes' and 'open the book', we rely on *what we know* about our internally generated eye movements and about our interactions with books. For 'open a restaurant', and 'open fire' we rely on what we know about common practices in relation to businesses, restaurants, war, and so on. Notice the striking compatibility between Searle's position on word

[74] Notice that Searle disagrees with the classic stance on the compositionality of *language* and that there is a potentially important difference between the compositionality of thought and the compositionality of language, as discussed in chapter 2 (section 2.6).

meaning and Carston's (2002a) concept schema proposal. Both claim that whatever the 'core' meaning of a word such as 'open' might be, it is insufficient, its *contribution* cannot be 'grasped' independently of context and thus *calls for* pragmatic processes.

Another important point to take away from Searle's notion of 'background' and Waismann's 'open texture' is the way they potentially turn the tables on traditional views of the division of labour between semantics and pragmatics. If conditions of application are not *fixed* prior to an interpretation and the set of background assumptions that a specific utterance's truth conditions depend on can only be determined once it has appeared in context – *after the fact*, then it seems that the classic picture of pragmatics subordinated to semantics must be rethought.

3.4.3 Putnam's 'Externalism'

In the seminal 'The meaning of "meaning"' (1975), Putnam explicitly declares his focus to be 'the concept of word-meaning' calling it 'more defective' than our concept of sentence meaning (p. 132). I could not hope to give anything like a full account of Putnam's contributions to the topic of *meaning*, and so my brief discussion in this subsection is limited to some key points and to how Putnam's views fit with the general picture I am drawing of radical contextualism. My aim in presenting Putnam's contributions is to further question the assumptions of the traditional stance on 'conditions of application' of a word and, by drawing a contrast, to continue to make the case for the ubiquity of context-dependence. Additionally, this subsection should help by complementing the discussion begun in chapter 1 (section 1.3) and thereby address some possible objections to the account I have put forth so far.

A classic distinction, between 'extension' (i.e., the reference of a word in a particular world) and 'intension' (i.e., that component of meaning that determines extensions), is commonly assumed to provide clarity to the issue of word meaning. Putnam (1975) however, begins his article with the claim that it actually fails, since only one of the two terms makes an aspect of meaning any clearer (i.e., 'extension'), while the other rather simply replaces

one vague term for another (i.e., 'meaning' for 'intension'). Worse, 'extension', which might seem straightforward since it is simply the set of things of which a term is true (e.g., 'rabbit' is true of all and only rabbits), actually depends on hidden *idealisations*. First, it is not the term *itself* that 'has an extension' but rather, 'strictly speaking' it is *an occasion of use* that has an extension. Furthermore, in mathematical terms, a 'set' is definite, something either belongs to it or does not, but natural language is teeming with 'borderline cases'. A final idealisation of 'extension' involves overlooking the consequential philosophical complexities it inherits from the notion of 'truth' (Putnam, 1975: 131-133).

These issues, however, appear minor when you consider that *even if* we grant that meaning is somewhat clarified by the notion of extension, there still has to be a second vital component to meaning if the account is to explain why two terms with *the same* extension can *differ in meaning*. Putnam's example is 'creature with a heart' and 'creature with a kidney', since every creature with a kidney has a heart, these expressions have identical extensions and yet *very different meanings*. The assumption has always been that an effective completion for the notion of 'extension' is the notion of 'intension' but, Putnam argues, if the account of intension is *vague*, because it simply calls on 'something like the *concept* or the *intension* of the term' when attempting to account for the meaning of a term *beyond its extension*, then the distinction has fallen short of *actually* clarifying meaning. Of course, a clear account of the notion of 'concept' would be an ideal solution *if any were available*, but this 1975 article appeared just as the traditional theory of concepts, with its claims of defining concepts through individually necessary and jointly sufficient conditions of application, was falling irremediably into disrepute. Furthermore, moving forward from this point simply cannot be put on hold until a working notion of 'concept' becomes available because of how intertwined 'meaning' and 'concept' are; attempting an initial disentanglement does not seem a promising strategy.

What Putnam proposes instead is a fresh angle on the complex issue in the form of the now well-known and influential 'Twin-Earth scenario'. A long-discussion of how Putnam makes his point would take us too far from

the precise contributions that most interest me here; so, I limit myself to a very brief presentation of the thought experiment. Putnam invites us to imagine a planet exactly like the Earth that has lakes and rivers filled with a liquid just like ours. The inhabitants are duplicates, 'Doppelgängers', of the Earth's inhabitants, they have the same thoughts regarding water as we have, they quench their thirst with this liquid and the English-speakers among them call it 'water'. Now imagine this scenario taking place circa 1750 when, through the advances of science the composition of water is finally known and, as it turns out, the liquid on Twin Earth is not H_2O but a *different* compound XYZ with superficial features *like those* of H_2O. The question is, upon learning that the liquid on Twin Earth is not H_2O, would we still call it 'water'? The general intuition is that people from Earth would say that Twin Earth 'water' is not *really* water and vice versa. Or both would say 'water' over there does not *mean* the same thing as here.[75] But because two twins, one on Earth and the other on Twin Earth share *all* their thoughts as they think about what 'water' means, there is nothing *in their thoughts* that is different and 'water' *means* something different here and there nonetheless. For Putnam, this reveals the major underlyng flaw in the traditional view of meaning: the idea that meanings are in the head, often phrased in the literature as '*meanings are mental entities*' or '*extensions are determined by intensions*'. All there is to meaning cannot be what subjects have in their heads since my Doppelganger and I have the same water-thoughts and 'water' does not mean the same thing here and there nonetheless.

Putnam's proposal for dealing with this meaning/reference problem, and some very similar ones, involves adopting the 'division of linguistic labour hypothesis'. He illustrates this with the example of GOLD, already mentioned in relation to Georges Rey's references to Putnam in chapter 1. According to Putnam, we dissociate acquiring the word 'gold' and the concept/intension *or meaning* from acquiring the *method of recognising*

[75] The two words 'water', corresponding to Earthian English and Twin-Earthian English, can be seen as homonyms, words that sound the same but refer to different things: in this case either to Earth H_2O or to Twin Earth XYZ.

whether something really is gold. So there are both superficial ways of recognising gold and necessary and sufficient conditions for something to really be gold (what Rey refers to as the 'epistemological' and the 'metaphysical' functions, respectively, of the concept GOLD). The 'division of linguistic labour hypothesis' means that meanings are only present in the linguistic community as a *collective* body. So, for instance, the criteria I gave for GOLD in chapter 1, (i.e., an atom of gold always has 79 protons), would typically be known by chemists but not necessarily by the general public. If I need to identify something as gold, I might be satisfied with my own imperfect method of recognition, (i.e., its very superficial traits), or I might ask a jeweller, but depending on what the identification puts at stake, I can choose to consult ever more authoritative experts: gold being one of the pure elements, there is a branch of science, namely chemistry, that would probably be the most authoritative. In every language community, there is a specific subset of speakers who know, perhaps not perfectly but *authoritatively*, what the associated 'conditions of application' or 'criteria' are for a certain term. Other speakers' competent use of this term depends on a structured co-operation between speakers closer and farther away from the authoritative information that we all suppose exists *somewhere* within the community. For Georges Rey, this idea is key to maintaining a clear distinction between the metaphysical and the epistemological functions of concepts since it makes a distinction between knowing the defining conditions or criteria for a term (being an expert) and being a competent user of the term.[76]

In a very recent publication, Putnam (2013) succinctly sets out how his externalism should be understood: first, there is the now very widely accepted point that '*nothing* that is in the head of the average speaker suffices to determine what the word *gold* refers to' but, importantly, this *does not* mean that what goes on in the brain does not fix the meanings for the words we encounter in our everyday conversations. Meanings are not *in*

[76] In the following chapter, I come back to this issue once again to give a specifically psychological take on this distinction.

the head but a brain does go through 'all sorts of 'maturation' and 'acculturation' in order for it to *know a natural language*' (Putnam, 2013: 197). A related point in Putnam's proposal is the particular role he ascribes to scientific theories in fixing the meaning of our terms. He disagrees with the logical positivists who assumed that scientific definitions *fix* the references of our terms. He calls attention to the fact that it is a *combination* of theories and *experiments* that tell us what our terms refer to, and stresses that experiments depend on the external environment. Putnam's point seems to be that while the contents of our brains, including our scientific theories, obviously do play a role in fixing the referents for our terms, the time has come to recognise that there are two other factors that traditional semantic accounts have too long ignored or downplayed: *other people and the world*.[77] This is very close to the contextualist claims I defend in this thesis; I portray meanings as *constructions* that occur *in context* between people who *naturally cooperate, converge, and negotiate when using language*. In Putnam's terms:

> Our verbalized thoughts have meaning only in conjunction with our transactions with objects in our environment and with other speakers (2013: 201).

He is complementing his (1975) construal of meaning *by what is not* (i.e., 'meanings ain't in the head') with a positive construal of what meaning *is*:

> It is in the context of a network of social and physical interactions, and only in such a context, that I can do such a thing as 'think that the price of gold has become very high in recent years' (Ibid).

---

[77] Traditional semantic theory has never denied the importance of the world; in fact the language-world relation is what traditional theories of reference are about. Putnam's point is rather that what previous views might have downplayed is the fact that the meaning of 'gold', 'water' or any other such term does *not* change with every scientific discovery made because of the role the *world* plays. For Putnam, the world is *the collection of paradigms*, it is our examples of gold and water, for instance, that fix the references of our terms. This is what Putnam means by 'extensions being determined indexically', discussed below. What he denies is that the meaning of a term like 'gold' or 'water' is determined by a (scientific) definition because this would imply that the meaning of 'gold' or 'water', or any other such term, would *change* with every scientific discovery made.

He goes on to add: 'The thought is no more simply in my 'head' than the meaning of the word 'gold' is' (Ibid).

The division of linguistic labour changes the picture of 'extension' and 'intension' drawn at the beginning of this subsection. Putnam's findings complement the original description of extension as 'the set of things a term is true of' with some very interesting insights. First, as discussed just above, extension is fixed *socially*. Social and physical interactions are just as important as what is in our minds, and what happens in our minds depends on a process of acculturation so interaction is found at every step. Secondly, extension can be determined *indexically*: going back to the Twin Earth scenario, before the discovery of the composition of water, or in *any expert-free scenario*, a valid meaning explanation for 'water' on Earth (or on Twin Earth) is to point to a glass of it and say 'This is water'. Putnam (1975) attracts our attention to the fact that for words, such as 'now', 'here' and 'this', no one has ever suggested that 'intension determines extension'. But if indexicals have long been recognised to vary in their extension across contexts why would terms like 'water' not do the same? The validity of the description given above, the fact that I can point to a glass of water and say 'This is water', means, according to Putnam that 'indexicality extends beyond the obviously indexical words and morphemes' because water is *that* stuff which bears a similarity relation to the stuff *around here*. This is why, as soon as Earthians learn that the stuff on Twin Earth is not H_2O, they no longer accept calling it 'water', and furthermore, consider that the word 'water', previously thought to mean the same thing on Earth as on Twin Earth, no longer does under these new circumstances.[78]

To summarise this section, Putnam claims that a term's extension is not fixed by a concept (or an *intension*) an individual speaker has in his head.

---

[78] In Putnam's words: 'Water at another time or in another place or even in another possible world has to bear the relation $same_L$ to *our* 'water' *in order to be water*. Thus the theory that (1) words have 'intensions' which are something like concepts associated with the words by speakers; and (2) intension determines extension – cannot be true of natural-kind words like 'water' for the same reason it cannot be true of obviously indexical words like 'I'' (Putnam, 2008: 312).

Rather, as revealed by the division of linguistic labour hypothesis, extensions are determined socially and indexically, with no need for the speaker to be in full possession of complete or exact conditions of application for the concepts he competently uses. With regard to the heightened importance this puts on cooperative activity, Putnam writes:

> We may summarise the discussion by pointing out that there are two sorts of tools in the world: there are tools like a hammer or a screwdriver which can be used by one person; and there are tools like a steamship which require the cooperative activity of a number of persons to use. Words have been thought of too much on the model of the first sort of tool (Putnam, 2008: 310).

Finally, anticipating a possible objection, of both Putnam's proposal and the contextualist proposal I seek to build on it, it is important to clearly highlight that Putnam does *not* hold that 'meanings do not exist'. On the contrary, he very explicitly claims that they might not exist *in the way we thought they did*, but that they are nonetheless very real (Putnam, 1975: 132).

3.5 The Second Stage: Abandoning the Modular View

The aim of this second stage is to explore the consequences of following the leads set out in the first stage and the philosophical arguments of section 3.4. To do this, I bring together the contributions from cognitive pragmatics discussed in the first stage with some more radical claims and propose an account of word meaning that no longer accepts the dictates of traditional semantics.

3.5.1 From Quasi-Contextualism to Radical Contextualism:

Recanati (2004) describes one quasi-contextualist position, 'strong optionality' and three contextualist positions ranging from moderate to radical. The first, the 'strong optionality view', holds that modulation is optional because there is potentially a context in which the sense expressed by a word is simply the sense that that word possesses by virtue of the rules of the language; in other words, modulation is optional in the most

straightforward sense: it either takes place *for contingent reasons* or does not take place at all (Recanati, 2004: 137). This position is much like that described earlier in this chapter (§ 3.3.2.4) with regards to relevance theory's standard position on ad hoc concept construction and the utterance comprehension procedure which holds that while modulation is very prevalent, it is *not* obligatory.

There is only a fine distinction between this view and the next view on the gradient: in the 'pragmatic composition view', there is still a sense that words could express the senses they possess by virtue of the rules of the language; but, it is also granted that literal senses undergo modulation simply as a result of *composition*. This would mean that the idea that a word such as 'drink' has a literal sense is only an illusion prompted by considering it in isolation. Recanati points out that when individual words become part of a whole, they must *cohere* and to cohere, they inevitably undergo a process of adjustment.

Radical contextualism begins with the third position on Recanati's gradient: the 'wrong format view'. He describes it as the view that

> words have meanings, but these meanings don't have the proper format for being recruited into the interpretations of utterances; they are not determinate senses but overly rich or overly abstract 'semantic potentials' out of which determinate senses can be constructed" (Recanati, 2004: 141).

Finally, the most radical position is 'meaning eliminativism', defined as going farther than the wrong format view 'in the same direction': it denies that *what* words have *as* linguistic types is anything like *meanings* in the traditional sense (Ibid). The most important difference between these two radically contextualist views and the pragmatic composition view is that while the latter still considers that the input to the pragmatic process that constructs an occasion-specific sense is itself a 'sense' that *could* figure in an interpretation without the need of modulation, the other two views deny that this kind of meaning exists. The wrong format view does not eliminate meanings altogether; rather, they are either too abstract or too rich to go directly into an interpretation. Meanings still exist in some sense since it is

assumed that the pragmatic process of occasion-specific meaning modulation is a process of *elaborating* or *delimiting* these meanings to suit a particular context. This contrasts with the more extreme view that 'eliminates' meaning. According to meaning eliminativism, there is simply nothing like a linguistic meaning associated with a type that could serve as input for a *modulation* process. Eliminativism pushes the idea that meanings only exist as the occasion-specific *senses* of particular tokens to the extreme. It postulates that, in order to arrive at an interpretation, we do not need anything like a linguistic meaning to kick-start the process.

> The senses that are the words' contributions to contents are constructed, but the construction can proceed without the help of conventional, context-independent word meanings (Recanati 2004: 147).

To illustrate how this particular construction process works, Recanati refers to an unstated but commonly assumed picture of where context-independent linguistic meanings are supposed to come from: they are allegedly the *products* of a certain induction process. Imagine a child learning a new word meaning: she is only ever exposed to the specific senses the word expresses (or is taken to express) on actual occasions of use. The child is supposed to extract the context-independent meaning of the word from this collection of specific senses. Once this is accomplished, the context-independent linguistic meaning can be the *input* to another process: that of meaning modulation. Imagine we line these processes up: we begin with contextualised senses, those to which the child is actually exposed; the first process is one of induction or abstraction that has linguistic meaning as its output. When somewhere down the line an utterance activates this linguistic meaning it goes through a process of modulation to arrive once more at a contextualised sense that can go into the interpretation the child builds of a new utterance.

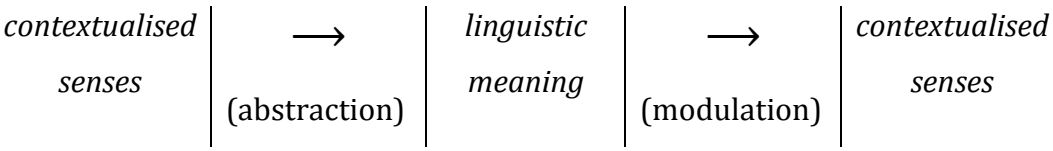

Figure 1. 'Abstraction and modulation' (Recanati, 2004: 147).

That this series of steps is necessary in first learning a word and then using it is a common, albeit perhaps tacit, assumption of most approaches shy of meaning eliminativism. Recanati's insight is to reveal how in these views 'both contextualised senses and context-independent linguistic meanings are input, and both are output in some construction process' (2004: 147). Notice that 'linguistic meaning' appears as a middle step between two instances of the only place where meaning is irrefutably found: *in context*. This suggests a possible simplification of this line-up: why not simply skip the intermediate step of creating a context-independent linguistic meaning and suppose that the computations involved in constructing the occasion-specific meaning expressed by a word (or expression) *in context* takes as input the contextualised senses that that word (or expression) actually had on previous occasions of use? This is the meaning eliminativism position: there is no need for an abstract context-independent linguistic meaning because the process of constructing occasion-specific senses for words and expressions can *merge* the two processes (i.e., abstraction/induction and modulation) into a single process that takes previous uses as input and yields as output occasion-specific senses perfectly adapted to the context at hand (Recanati, 2004: 147).

Fleshing out meaning eliminativism is among the principal aims of two chapters to come and sections in the remainder of this chapter. Part of my contribution in this thesis is to take this very radical proposal seriously; I propose to explore how far contextualism can be taken, in great part by following the logic of 'eliminating' linguistic meaning. I am open to the possibility that in the end meaning eliminativism will probably be deemed too radical; it is, after all, construed as the most extreme contextualist position *possible* by Recanati. His aim in describing the possible positions

along the gradient does not seem to be to unambiguously endorse one particular position, but rather, to identify the most important differences between possible positions and explore their advantages and disadvantages as a way of framing his own, and his readers', thinking on this topic. I, for instance, use the framework to argue in favour of moving as close as possible to the extreme radical contextualism described in meaning eliminativism and as far away from moderate, or 'modular', forms of contextualism. Such an extreme position faces many objections, some of which are addressed in the closing remarks of this chapter; I discuss detailed solutions in chapters to come. But first, I endeavour to illustrate radical contextualism with some concrete proposals. Any particular proposal in lexical pragmatics is unlikely to fit neatly with one specific position on Recanati's gradient; rather, different proposals probably naturally fall somewhere *along* the gradient, closer or farther away from the benchmarks described. In the remainder of this section, I try to spell out particular approaches that 'abandon the modular view' insofar as they are more or less in line with wrong format and/or meaning eliminativism. I also present a detailed account of 'semantic potential' and 'contextualised senses', in my view, the most important notions available in pursuing a truly radical contextualist approach to word meaning in context.

3.5.2 Bosch's 'Contextual Concepts'

I include Peter Bosch among the radical contextualists for two main reasons: first, he explicitly defends the position that we can simply dispense with semantic contents that are not context-relative (2007: 59); secondly, he looks to frame his approach within the action tradition. As mentioned earlier in this chapter (section 3.2), Bosch joins Clark in viewing the field of language studies as divided into two traditions: a 'product', or *sentence*-based tradition and an 'action' or *utterance*-based tradition. Bosch, a cognitive scientist and computational linguist, joins Clark in pointing out that the tradition focused on sentences works with the assumption that all the speaker needs is knowledge of the primitive expressions of his language (the lexical items) and the rules for their combination. Speakers can produce

and understand sentences generated by these rules because their meanings are built up compositionally from the meanings of their parts. Following this tradition, much contemporary syntactic theory has focused on determining just how syntactic rules determine the way constituents come together to form larger syntactic and semantic representations. But, adopting the action tradition's perspective, Bosch points out that there is an important hidden assumption in this program: it is assumed that lexical items *contribute the meanings* that are then combined into larger structures. Yet the accounts by lexicologists of just what these word meanings are, what should be included in the meaning of a word, and what the representations look or function like, have not been forthcoming. Bosch argues that the reason for this is a certain denial of the *kind* of solution required for the problem. He observes that contemporary linguistics is focused on *linguistic* knowledge and that the treatment required, at least for some cases of 'productive language use' is conceptual. As Waismann and Searle before him, he sees world knowledge and common sense as desirable elements in an account of language. Instead of limiting the information accessible to the language user to *linguistic knowledge*, and privileging *linguistic processes* of interpretation, as the product tradition would do, he embraces the more global approach exemplified by the supporters of the action tradition. He explicitly agrees with Searle that it is only the *statement* made by uttering a sentence on a particular occasion that can be said to be true or false; sentences do not have truth-conditions (Bosch, 2009).

Bosch adds more radical claims to this now relatively standard contextualist position by elaborating on some of the consequences this new perspective has for word meaning. His starting point is that the traditional picture of meaning variation (labelled 'polysemy' in most of the literature) is misleading. Even more modern accounts that endeavour to include as many aspects of context-dependence as they can (like Pustejovsky's 'generative

lexicon'),[79] fail to recognise the true role of context in polysemy.[80] Bosch's claim is that what is missing from these accounts is the realisation that language processing is not only the processing *of language*. In other words, it does not only involve the processing of information from linguistic sources. On this point Bosch's perspective follows that of most contextualists and seems nearest to what Recanati calls the 'wrong format' view since he holds on to the idea of something like a lexical entry, while considering that the contents of this lexical entry cannot themselves be used in the interpretation. Very much like in the relevance theoretic account of ad hoc concept construction presented in the previous section, and particularly in tune with Carston's (2002a) concept schema proposal, Bosch suggests that the lexical entry is rather only a pointer to a lexical concept that pragmatic processes calling on *all types of information including non-linguistic sources of information* must complete. Also, in Bosch's account it is *pragmatic* processes that fix the referents or denotations for at least some of the expressions in our everyday utterances. At the same time, Recanati's gradient seems to serve to place Bosch's perspective farther out than standard relevance theory's quasi-contextualism and somewhere on the radical contextualist side of the divide. His mention of a lexical entry might suggest compatibility with the description given of the wrong format view; but other considerations, such as his construal of 'polysemy', discussed below, might suggest a more radical position.

Bosch's original contribution is the deconstruction of the traditional notion of polysemy; in its place, he introduces the notion of 'context dependence of predicate expressions', a solution to meaning variation that insists on the ubiquity of context-dependence by making it *characteristic* of predication rather than marginal or secondary; in other words, it is the rule rather than the exception. According to Bosch (2007) previous accounts of

---

[79] I would add Nicholas Asher's (2011) book *Lexical meaning in context: a web of words* as an attempt to account for context-dependency that completely misses the point made by Bosch.

[80] It would undoubtedly be of interest to present these accounts and detail Bosch's criticisms but limits of time and space do not allow for this.

context-dependency were too limited. Kaplan proposed a short list of *explicit* indexicals, which depended on the context of the utterance for their truth-evaluable content, and this was expanded by Perry to some *implicit* constituents (also called 'unarticulated' constituents) that functioned *like* indexicals.[81] According to Perry (1998), for instance, an utterance of 'It's raining' contains an *implicit* place reference, since arguably, unless we know *where* it is supposed to be raining, we cannot say if the statement made by the utterance is true.

According to Bosch, however:

> the situation may actually be even more difficult than Perry's argument suggests. Not only are there *implicit* indexical constituents that make the semantic value of a sentence depend on the utterance situation, much in the way that Kaplan proposed for explicit indexicals, but also a large proportion of the *explicit* constituents that are not in Kaplan's class of indexicals depend on properties of the utterance context in the contribution they make to the truth-evaluable content of the sentence (Bosch 2007: 60).

To illustrate this, Bosch suggests considering the 'value' of the verb 'rain' in an utterance 'It's raining' as a response to the following:

(8) a. Why did you call a taxi?
b. Can we go for a walk now?
c. Are you saying it's still drizzling?
d. Is it pouring like this morning?

In each of these, 'It's raining' communicates an entirely different thought, or, in Bosch's terms, constitutes a completely different 'conceptual representation' and this is not because it is raining in a different place or *raining* any differently. The traditional account holds that 'It's raining', at a basic level of communicated content, expresses the speaker's belief that *it is*

[81] According to Perry (1998), 'since rain occurs at a time in a place, there is no truth-evaluable proposition unless a place is supplied' (p. 9; see also Perry and Blackburn, 1986). According to Recanati, however, 'It is a metaphysical fact that every action takes place somewhere' (Recanati, 2002: 306) so the relevant notion of unarticulatedness should be reserved for cases where the place is provided in virtue of features of the context only because comprehension requires it, or in other words, for purely pragmatic reasons, not because it is linguistically required.

raining and that any additional content is inferred, and therefore justifiably considered *secondary*. In (8a) and (8b), for instance, the hearer's interpretation involves ascribing the belief that *it is raining* to the speaker and *inferring* that this is given as a reason for calling a taxi, or declining an invitation for a walk. But this does not seem to be the thought expressed by *it is raining* as a response to (8c) and (8d). For (8c), does the hearer first ascribe the belief that *it is raining* to the speaker and then amend that to *[the speaker] would say it's raining rather than say it is drizzling* and for (8d), *the speaker does not believe that it is pouring, he believes that is [just] raining*? Instead, Bosch suggests, we could entertain the possibility that 'rain' does not have a 'lexically fixed constant semantic value', and, what meaning it does have cannot be defined independently of speaker's intentions (Bosch, 2007: 59-60). In this particular set of examples, we can even imagine that it is raining in exactly the same way in all four contexts and that there is nonetheless a difference in what the verb 'rain' contributes. From such examples, Bosch concludes that the observed variations in truth conditions (previously explained as cases of polysemy) are in fact due to the combination of underspecified lexical representations, on the one hand, and the effects of the context, on the other. The most direct consequence of this view, he argues, is that explaining cases of meaning variation *including* cases of productive language use (as in figurative language) requires a treatment at the conceptual level. Instead of trying to specify lexical semantics in terms of the truth conditions words contribute independently of the context in which they appear, Bosch suggests *attributing* the variations in truth conditions directly to 'the differences in conceptual representations that result from differences in the utterance context' (2007: 58). Critically, this implies dispensing with semantic contents that are not context-sensitive instead of taking 'core' meanings and modulating them. For another illustration of this, consider the following example:

(9) Charley isn't working.

If 'Charley' is a device, (9) means that it is malfunctioning; if 'Charley' is a human being, it means something quite different, that he has the day off, for

instance. We could say that there are either two entries representing two very different sets of information associated with two different concepts: WORK$_1$ and WORK$_2$ (briefly, WORK$_1$ to refer to the functioning of *machines*, and WORK$_2$ to the employment of *human beings*), or a single entry associated with a concept that covers both WORK$_1$ and WORK$_2$; but, *either way*, in order to choose between the two conceptual representations of 'working', access to the *intended reference situation* is required.[82] There is nothing in the entries that allows us to differentiate them if we do not know whether 'Charley' is a computer or an employee: *that* relevant information can only come from the utterance situation. Furthermore, the features of context that are relevant, like whether the subject is human or not in the example above, are not easily predefined. Consider an example where knowing that we are talking about our friend Fred would not suffice to differentiate between two semantic values for 'He's working':

(10) Where is Fred?
(11) How can Fred afford such expensive holidays?

An utterance of 'He's working' as an answer to (10) carries information about Fred's location; and, as an answer to (11), about his finances. Just as in the relevance-theoretic approach, fixing the denotation of 'work' is a fully pragmatic process, in Bosch's description, 'the denotation of work can only do this if it is enriched by contextual knowledge, and in different ways for (10) and (11)' (2007: 62). Further agreement is in the fact that for both approaches it is the enriched 'work' that is truth-conditionally relevant. Bosch finds this evidenced by how the two utterances of 'He's working' licence *different* inferences. What is less clear is Bosch's position on concepts themselves: it is different conceptual representations of 'working' that guide reference assignment in (9) and create context-sensitive semantic values for 'He's working' as a response to (10) and (11). But little is said about these

[82] Of course the lexical entries could be made much more complex than the simplified example given here, but the point would be the same. It does not matter because deciding between them requires knowledge of the intended reference situation and if the intended reference situation is already known, then the entry is superfluous.

conceptual representations. Against more traditional approaches, Bosch claims that the interesting part of working out a conceptual representation is a process of conceptual processing and *general reasoning* rather than a process of linguistic semantics. He means that the information necessary to know what a concept denotes does not come primordially from what is commonly considered *lexical* information, suggesting instead that *contextual* information concurrent with utterances is the key source. In (2007), he proposes to call the contextual referents or context-sensitive semantic values of expressions 'contextual concepts', but again, nothing more than a 'rough sketch' is given of these '*contextual* concepts' *as conceptual constructs*.

As suggested above, while it is easy to see signs of the wrong format view in Bosch's proposal, because he contemplates different possible contents for lexical entries in the WORK example and explains polysemy as the product of the combination of underspecified lexical representations and contextual effects, I would argue that his position is too radically contextualist to be best described as a wrong format view. It is important not to miss the fact that, as Bosch discusses the notions of lexical entries and lexical representations, he rejects them, at least as they are *traditionally construed*. He proposes an alternative that might use some of the *terminology* of more conservative accounts (he seems to borrow the notion of enrichment from relevance theory, for instance), but adds a measure of more radical contextualism in certain key points. He does not go so far as to advocate a meaning eliminativist position, but, I would argue, his proposal does fall somewhere farther on the gradient than typical cases of wrong format.

To see this, consider Bosch's double claim that the construction process of occasion-specific senses requires *general* reasoning and that contextual information concurrent with utterances is a key source *over* 'lexical' information. If the occasion-specific sense construction as *described by meaning eliminativism* were added to this, in other words, if the process of building contextual concepts were as described by Bosch with the added assumption that the computations involved take as input the contextualised

senses that the words or expressions actually had on previous occasions of use, then it would effectively be the case that, as Bosch defends, the variations in truth-conditional content would not be attributable to semantic values that are not context-dependent. This is not Bosch's position, he explicitly claims that his approach can dispense with semantic contents that are not context-dependent, but he does not explicitly say *what* takes their place. However, this addition to his proposal seems compatible with his approach, as he does say that his proposed construction process implies dispensing with a certain type of input and finding an alternative solution to variations in truth conditions that no longer implies taking 'core' meanings and modulating them. The claim that there is *nothing* in the lexical entry for 'work', for instance, that allows us to differentiate between possible conceptual representations of 'working' in the 'Charley isn't working' examples is eliminativist in at least this sense.

3.5.3 Carston's 'Non-Conceptual Word Type Meaning'

To complete this discussion of moderate and radical contextualist positions, I now briefly turn my attention to Robyn Carston's recent publications on the nature of word meaning (Carston, 2012, 2013). As these articles arguably revisit and extend on her (2002a) 'concept schema' proposal (§ 3.3.3), I begin this section with a short review of that proposal. Very briefly, the standard relevance-theoretic linguistic semantics model holds that literal, 'core' word meanings exist and that their role is to determine the *linguistically* specified denotations of words. This, together with the idea that words encode (atomic) concepts like those that serve as the constituents of our thoughts (in *Mentalese*) is challenged by the observation that the thoughts we communicate with use of words like 'happy' are not ever general and abstract thoughts but rather always 'specific sorts' of thoughts that have 'ad hoc' concepts (i.e., HAPPY*, HAPPY **) as their constituents. The standard account holds that a general and abstract concept HAPPY is the (partial) input to the pragmatic process that has the specific sorts of HAPPY as output, this however, assumes that there is a context-independent abstract concept HAPPY, and a *conceptual* standing

meaning for 'happy', *other* than the specific concepts HAPPY*, HAPPY **, and their matching pragmatically derived occasion-specific meanings. Yet, if the abstract concept HAPPY is not *ever* a constituent of our thoughts, why assume that words encode (atomic) abstract *concepts*, or, in other words, that words have conceptual standing meanings?[83] Among the alternatives, a promising lead points to an occasion-specific word meaning construction process that builds word meaning *entirely ad hoc* from very general information in memory. In light of this, and other considerations, Carston proposes that what words encode might be more akin to a concept *schema* that serves as a pointer to a 'conceptual region' in memory; but ultimately, the question is left open because the idea of *concept schemas* presents its own set of problems.

Carston's point of departure in her more recent articles is still the standard relevance-theoretic model, but it now takes a lexical pragmatics angle:

> The concept expressed by the use of a word in context often diverges from its lexically encoded context-independent meaning (2012: 607).

Her objective in these articles is to revisit the underlying assumption that this lexical meaning *modulated* by pragmatic processes is *conceptual*, or otherwise, 'directly expressible'. She proposes instead that context-independent word meaning is *non-conceptual*, in other words, she proposes that words encode *something* but that it is 'intrinsically underspecified with regard to content' and must therefore be pragmatically completed before it can be taken to be what the speaker expressed or communicated with the use of a word or expression (Carston, 2012: 622). This constitutes a relevance-theoretic version of what in Recanati's terminology is the *wrong format* view since its main claim is that words have meanings but that they must undergo some transformation in order to be in the form required by

[83] In my own account, there is no abstract, context-independent concept HAPPY because concepts themselves are context-dependent (as I discuss in chapter 4, sections 4.4 and 4.5), but I also explain why we are convinced *and act as if* 'core' and abstract concepts, or 'concept *essences*' exist in a subsection on 'psychological essentialism' (chapter 4, § 4.3.4).

recruitment into an interpretation. This proposal is contextualist, under the description I have developed here, insofar as it does not construe modulation as optional; and it is a *moderate* contextualism insofar as Carston (2012, 2013) still assumes that context-independent word meanings *exist*: she cites as one of her objectives the exploration of 'the nature of the context-free word meaning' that serves the pragmatic process of occasion-specific meaning construction as input. Naturally, much as in the (2002a) section on 'word meaning and concepts', in the course of the discussion, multiple possible answers to the question of the nature of word meaning are discussed, and Carston seems drawn to options that point to occasion-specific meaning construction processes that completely bypass the need of anything like a stable meaning associated with a word; these options, however, are in conflict with the standard relevance-theoretic position and ultimately, she does not advocate their adoption. Also as in (2002a), however, the insightful arguments presented in their favour are very much worth a closer look.

Carston (2012) organises the possibilities she examines along a gradient much in the same way as Recanati. She describes the first possibility as evidenced by Jerry Fodor (1998) and what I have called the *standard* relevance theoretic position (Sperber and Wilson, 1986/95, 1998): word meanings are concepts and words *can and do sometimes* simply express the concepts they encode. The second position introduces the idea that words encode *underspecified* word meanings: pro-concepts and schemas, among others, are the possible *forms* these meanings could take. The version of this position that Carston defends, assumes, like other wrong format views, that whatever these underspecified entities are they require some transformation to become complete. The distinctiveness of Carston's version is that these forms are 'something less than conceptual'. Finally, the last position described roughly matches Recanati's meaning eliminativism:

> words (lexical forms) do not encode concepts or abstract schemas or constraints, but are associated with something else altogether, something that does not qualify as a meaning of the expression type (Carston, 2012: 608).

After introducing the gradient of possibilities, Carston focuses on the last two, most contextualist, positions: a relevance-theoretic version of wrong format and meaning eliminativism. Since my own interests are for a position very close to that of meaning eliminativism, I am most interested in the second of the two. She gives two possibilities as to what the 'something else altogether' that does not qualify as word-type meaning could be.

The first follows Recanati's own suggestion that perhaps explaining the construction of occasion-specific senses involves adopting a specific model of memory. In meaning eliminativism, context-independent word-type (linguistic) meaning does not exist; only *contextualised* senses exist (see figure 1, § 3.5.1); the multiple-trace memory model Recanati refers to holds that we keep these contextualised senses in memory in the form of *traces of episodes* of distinct occasions of language use. As described in chapter 2 (§ 2.7.3), these 'memory traces' can be *selectively* activated and thus serve as input to an occasion-specific sense construction process. Spelling out *how* traces of previous episodes can be selectively activated to build an occasion-specific word meaning is one of the overarching objectives of this thesis; a fuller presentation of this, however, must wait until chapter 5 where the notions of memory traces and memory trace activation are fully discussed. I come back to more specifics on the notion of contextualised senses in the next subsection (§ 3.5.4).

The second possibility mentioned by Carston (2012) for what a meaning eliminativist approach could take as input for its construction process is 'bundles of contingent encyclopaedic information'. She illustrates this possibility with Agustin Rayo's recent 'plea for semantic localism' or 'grab-bag' model (Rayo, 2013). Adopting key contextualist ideas such as the ubiquity of context-dependence and explicitly opposing the modular view,[84]

[84] Rayo (2013) begins with

> The purpose of this paper is to defend a conception of language that does not rely on linguistic meanings (p. 647).

He contrasts two possible answers to the question of what a subject must associate with the words and expressions of his language to count as a competent speaker: either the subject's use of his basic lexicon is guided by *semantic rules* which determine a given word's range of application *independently* of the subject's general capacity for reasoning and common-

Rayo suggests that, instead of semantic rules or conditions of application, subjects associate 'grab-bags' with their words and expressions. He argues that instead of context-independent *meanings*, what speakers have at their disposal is their 'sensitivity to context and common-sense' which allows them to build reasonable interpretations for words and expressions *for the purposes at hand*. Rayo describes these grab-bags as containing 'mental items: memories, mental images, pieces of encyclopaedic information, pieces of anecdotal information, mental maps, and so forth' (2013: 648).

In view of the contents of grab-bags, I would argue that these two seemingly different possibilities for what constitutes the input to construction processes are actually *complementary aspects* of a single *essentially* meaning eliminativist approach. As the chapter on memory will show, multiple-trace memory models have the advantage of construing something like *contextualised senses* very broadly so that *any* 'mental item' stored in memory that could be considered relevant to the context at hand could be included; in the course of that discussion, it will emerge that the *mechanics* of how memory stores and retrieves episodes can offer much substance to the idea that contextual senses, and in general *any previous experience*, can be captured in the form of traces of episodes and retrieved to serve as input in an occasion-specific meaning construction process. Given such a framework, theorising on the *kind* of reasoning processes that such construction processes imply can be seen as an equally important part of defining the approach. To the various contextualist proposals already presented in this chapter, Rayo's emphasis on the 'reasoning and common-sense' abilities that assemble the 'bundles of contingent information' in the grab-bags would need to be added.

Carston (2012) considers that talk of 'common sense and sensitivity to context', as in Rayo's proposal, is precisely the 'cognitive interpretive process' (i.e. pragmatics) that she and relevance theorists in general have attempted to elucidate. With this, Carston seems to envisage the possible

---

sense, or, instead of any specialised, specifically linguistic knowledge, subjects associate 'grab-bags' with the expressions of their language.

compatibilities between approaches such as Rayo's, and others calling on *general reasoning* such as Bosch, and the relevance-theoretic approach to ad hoc concept construction. For instance, the grab-bag model calls on 'contingent' encyclopaedic information, which is much like the information in the relevance-theoretic encyclopaedic entry. The difference, of course, is that, in relevance theory, *decoding* provides access to this entry, while this is not the case in Rayo or Bosch's accounts. Carston also mentions a key advantage to Rayo's approach (that has its parallel in Bosch's account): 'an immediate and simple solution to the polysemy/metonymy problem'. This solution involves grab-bags that are put together *differently* on each occasion of use. Instead of the words or expressions pointing to anything like a context-independent meaning as the (even partial) input to the construction process, in the grab-bag model, the subject *freely* selects what comes to mind in accordance with 'common sense and sensitivity to context'. A look at how this approach handles examples such as 'stop' and 'novel' suggests a very direct occasion-specific construction process. For instance, interpreting the word 'stop' in an utterance would not necessarily involve any grammatical information associated with 'stop' (that is, *different* information depending on whether it appears as a noun, or as a verb); rather, representations that bring *interfering*, *preventing* and *obstructing* to mind would be put in a grab-bag and the rest would be left to reason and common-sense (Rayo, 2013: 673; Carston, 2012: 621).

Carston (2012) concludes that such an approach would mean that language theorists could finally give up grappling with what semantic values or conditions of application to attribute to context-independent meanings and abandon the 'futile' attempts to select a single sense among a set of possible senses to be the basic, core meaning of a polysemous word (Carston, 2012: 621). This would have direct and considerable consequences for any approach still postulating fixed, context-independent meanings, including relevance-theoretic lexical pragmatics.

Carston (2013) suggests that if words *behave as if they do not encode concepts*, we should question assumptions that they do. She proposes instead that they encode something more schematic that 'merely constrains

or guides' the hearer's pragmatic process of recovering what a speaker has expressed (p. 187)[85]. Unfortunately, this proposal has the same problems as the proposal Carston has already considered in her section on 'Word meaning and concepts' (2002a), discussed above (§ 3.3.3); among other things, the acquisition story for concept schemas is still unclear: how is a child supposed to learn an abstract schema (for HAPPY, for instance) that is other than the concepts (HAPPY*, HAPPY**, etc.) that actually figure in her thoughts? This would presuppose that the child extracts an adult schematic meaning from her concrete experiences and that this then serves as her lexical expression type in subsequent experience with language. In the next chapter, I discuss the processes of abstraction and extraction in depth. In anticipation, I can say here that the problem is not so much with whether the child can extract type meanings from concrete experiences with occasion-specific senses; the problem is that, even supposing she can, it still does not mean that these capacities are employed towards creating an abstract schematic word-type meaning that serves 'as a gateway' in understanding occasion-specific meanings. Rather, I will defend the view that, much as in Recanati's meaning eliminativist position, processes of abstraction and extraction are part of an occasion-specific word meaning construction process that takes 'contextualised senses' as input without the need for anything like a word-type context-independent (schematic) meaning.

Carston's cautious restraint toward such an approach is justified by the difficulty of the questions it raises and the scope of its consequences. She feels that more research is needed before deciding on the path ahead; the question, of course, is how drastic a change the influences of more radical perspectives would impose on relevance theory if adopted. On the one hand, Carston seems to suggest that the way forward is to continue to 'take seriously' the idea that words come with 'meaning-relevant components'

[85] Carston (2012, 2013) proposal can be summed up as advocating a 'wrong format', instead of a 'strong optionality', view for relevance-theoretic lexical pragmatics.

instead of mapping directly to concepts or encoding concepts. On the other hand, Carston's (2012) closing suggestion is that these possibilities be investigated 'within the explanatory pragmatic account provided by the relevance-theoretic framework' (p. 622). It seems that the way ahead must walk a fine line between opening up to evidence and insights from outside of relevance theory and remaining within its trusted framework. It is left up to the reader to decide whether giving context-independent word meaning up is fundamentally *in*compatible with relevance theory or if another solution yet to be envisaged must be found. Carston and other theorists' preference for caution is also supported by the fact that it is much easier to argue against the existence of context-independent word meaning than to formulate a detailed alternative. So it is one thing to reject the more traditional outlooks on word meaning in context *in view of the evidence*, but it is quite another to propose an alternative that can resist *counterarguments*. That, however, is what I propose to do. In chapter 4, I revisit the notion of 'concept', through the research done by psychologists on categorisation among other things, and I offer an account of ad hoc concepts in line with the most radical contextualists presented in this chapter. Discussions in chapter 4 will reveal the need for a chapter-long discussion of memory (chapter 5). What I propose in the last subsection of this chapter is a detailed presentation of François Recanati's notion of 'semantic potential'. My aim is to show that the eliminativist approach proposes an alternative to traditional approaches to word meaning beyond 'eliminating' the notion of a context-independent, standing, linguistically-specified word meaning.

3.5.4 Recanati's 'Semantic Potential'

François Recanati is an important theorist comfortable with radically contextualist ideas and an essential source for developing a meaning eliminativist position. As discussed at the beginning of this chapter (section 3.3 on 'the first stage' of contextualism), he has made important contributions to cognitive pragmatics and has proven key in the on-going challenges to the framework established by formal semantics. In this

subsection, I focus on one of his most important contributions for my purposes: the notion of 'semantic potential'. As I develop it here, it is largely compatible with contextualism in general and with the notions presented so far in this third stage. It is part of the action tradition/contextualist framework insofar as it stresses the importance of utterance and speaker meaning and rejects the idea that pragmatic and semantic processes are insulated from one another; it is a radical contextualism insofar as it espouses a more complete dissociation between lexical forms and fixed meanings.

Recanati (2001a) is a careful reader of Searle's (1978, 1980, 1983) notion of background and readily adopts its conclusion that words are not linked to fixed sets of conditions of application, but rather, that truth-conditions *are arrived at as part of*, or better, *as a result of* the interpretation. The enormous consequences of positing that truth-conditions – or *any* norms employed to evaluate a situation, for that matter -– are arrived at *as part of an interpretation or evaluation*, instead of being retrieved, have been explored most notably by Daniel Kahneman and Dale Miller in their work on 'norm theory' (Kahneman and Miller, 1986).[86] I introduce their theory here as a brief detour because, despite not being mentioned by Recanati, it is particularly useful in giving a clear account of his notion of semantic potential and the larger philosophical and psychological framework of which it is a part. Norm theory's principal claim is that 'norms' are computed *after the event* rather than in advance. It has generally been taken for granted that interpreting and evaluating 'events in the stream of experience' requires consulting pre-computed norms, schemas and frames of reference. To this, Kahneman and Miller oppose a view that has each stimulus selectively activate its own 'alternatives' and *generate its own norm*. For Kahneman and Miller, 'norms' serve two functions: to represent knowledge and to interpret experience (1986: 150). In other

[86] Norm theory is only a small part of Daniel Kahneman's extensive and very influential work on human behavior, particularly on judgments and decision-making; work for which he received a Nobel Prize in 2002, (for a general introduction, see Kahneman, 2011 *Thinking, fast and slow*).

words, a stimulus event, which can be a word, or 'category name', is encountered in the stream of experience and according to norm theory it activates a set of representations relevant to *this* word in *this* context; from the *selectively* activated representations, '*generic properties*' are summarised *online* creating a uniquely context-sensitive, ad hoc *means* of interpreting that particular word in that particular context. Instead of *consulting* a norm in order to interpret an object or event, words, objects, and events *generate* their own norms or *frameworks of evaluation* 'after the fact', only once we have encountered them in context.

To illustrate the generation of norms and how they could be involved in language processing, let's think back to the 'intelligent dog' example. Waismann chooses the combination of DOG + INTELLIGENT precisely because to him it is surprising, I can now add that surprise, as defined by norm theory, is *the failure to make sense of an experience*. The combination is so novel to Waismann, that he feels compelled to ask the speaker for clarification. With this clarification, and some quick thinking on his own, Waismann constructs a set of norms, or a framework, that make the utterance less abnormal and therefore more understandable to him in that context. Norm theory posits that this is achieved by 'aggregating' a set of representations selectively recruited *for the context at hand*. Notice the similarities between relevance theory's ad hoc concepts, Bosch's contextual concepts, Rayo's grab-bag and 'generating frameworks of evaluation' for the purposes at hand. In chapter 4, much more will be said about how aggregating selectively recruited representations can answer questions about categorisation and concepts. For now, the important point is that meaning eliminativism is potentially supported by a theory (i.e., 'norm theory') which holds that it is not so much that norms, *or meanings*, do not exist, as that they are created *after the fact*, more as a consequence of processing words in context than *as a means* to process words in context.

To summarise, the assumption that interpretations depend on the retrieval of pre-computed expectations is challenged and a clear alternative emerges: rather than retrieving ready-made context-independent meanings, cores, schemas or ready-made *anything* in fact, computations occur as a

scene is taken in and concepts, norms, and senses are constructed, using *general* reasoning processes and any information in memory that is relevant to the purpose at hand. So, for instance, interpreting 'My dog is intelligent' would involve creating an ad hoc concept for INTELLIGENT,[87] which in turn implies the creation of an ad hoc framework for evaluating DOG + INTELLIGENT in a particular context (i.e., the owner's claim that his dog is intelligent). That Waismann would not have considered INTELLIGENT applicable to DOG illustrates the need he *now* has to construct a framework of evaluation that somehow makes the utterance understandable. All of this results in an interpretation that arguably has an occasion-specific sense for 'intelligent' as a constituent.

The challenge is to bring all of these different contributions together in a new notion of just what *is* associated with a word or expression of a particular language. Following Recanati, I suggest that addressing this involves two notions he has introduced: contextualised senses and semantic potential. Contextualised senses, presented above (§ 3.5.1), are those occasion-specific meanings a word (or expression) assumes or expresses (or is taken to express) in a particular context. Critically, contextualists in general, not only those of a more radical stripe, accept the existence of such occasion-specific, or contextualised senses. In a wrong format view, *contextualised* senses are supposed to serve as input to those extraction processes implied by the idea that there is a context-independent (*conceptual*) *schematic* standing meaning for words. In Carston's (2002a) proposal, for instance, these contextualised senses are the input the child would have at her disposal for the task of extracting an adult *schematic* word-type meaning, something problematic by Carston's own admission. In meaning eliminativism this problem and related problems are simply avoided. It is not assumed that it is possible or necessary to extract a sense

[87] Importantly, this ad hoc concept construction process can follow Barsalou's original notion of ad hoc concept construction in which a conceptual content is the end result and the input is unconstrained by encoded concepts or anything that qualifies as linguistic, context-independent word meaning.

that could somehow serve as *the* basic meaning a pragmatic process of modulation could reliably take as its starting point. Rather than assuming that these contextualised senses have an essential hidden core that is the output of the extraction process and the input to the modulation process, words only have *semantic potential*. A first description of semantic potential equates it with the notion of contextualised senses: it is nothing like a definition or conditions of application and it integrates *anything* arrived at as part of an interpretation. Recanati suggests that it is the collection of situations in which a speaker has observed the particular word or expression used in his lifetime experience with the language. Instead of instructions on how to use the word, this amounts to saying that what we know when we know how to use a word is a collection of legitimate uses.

To explain his proposal, Recanati invites us to imagine what it would mean for someone to learn a new word or predicate according to this new framework. At the beginning, the language learner would hear the predicate *P* used in a particular situation *S* and would associate *P* to *S*. At this early stage, the semantic potential of *P* for the language learner is simply *S*, the situation in which he has heard this predicate used. Now suppose he encounters it in novel situations: S_1, S_2, S_3. These situations are added to the semantic potential for *P*. When the language user wants to use the predicate himself, he will try to judge whether the situation at hand is sufficiently similar to the situations in which he has heard *P* successfully used. This is a process of learning so, of course, the language learner can misjudge which similarities between situations are relevant for the application of *P*, but in these cases the language community will offer corrections that he can use to guide him in the future. Completing the learning phase consists of amassing a large enough number of situations in which the use of *P* is justified so that he no longer mistakes situations that are similar, but not similar *in the right respects,* to the situations calling for the predicate *P*. Here the two notions begin to disentwine: contextual senses are still the occasion-specific senses words express in their contexts of use, and semantic potential is what the language learner (or everyday speaker/reader) *learns* as his experience with language increases. Recanati calls the collection of legitimate, valid

uses of the predicate *P* which represent the semantic potential for a particular predicate the 'source-situations' and the situation the speaker may want to apply the predicate to the 'target-situation'. With these two further notions, a very interesting explanation of context dependence can be given: a *target-situation* must present certain features for a particular predicate to apply to it; that is, it must be sufficiently similar to the source-situations of that predicate to justify its use, but the features that justify the use are not fixed in advance so they can vary according to the context. In the terms of norm theory, different 'sets of representations' are activated in different contexts by the same words and expressions. Recanati's account adds some detail as to *which* representations can be considered relevant for a certain context at hand: it is those situations where the target-situation is similar to the source-situations *in the right respects*.

There are several technical issues, implied by the account given so far, that need to be discussed before a full defence of an eliminativist position can be offered. One particularly thorny issue, the notion of similarity, will be addressed in chapter 4, 'Psychological perspectives on concepts'. Anticipating that discussion, a target situation has to be judged sufficiently similar to a source-situation in order for the predicate to apply, but there is fierce disagreement amongst theorists on the topic of similarity. Some of them deny that it can be made to do any work as a technical term, pointing out that without saying in what respect something is like something else, similarity can be quite empty. In Chapter 4, however, I introduce Tversky's notion of similarity and argue that *adequately contextualised*, similarity is well constrained. In that chapter, I also define and discuss the notions of abstraction and extraction to which I only referred in this chapter. Together with other considerations, in that chapter, a broader contextualist framework hopefully arises in which word meaning is only part of what is context-dependent.

Another technical issue left to the side for now is the memory model assumed by Recanati when he proposes that subjects access contextualised senses when constructing occasion-specific senses for the words and expressions of the utterances they encounter. As mentioned already (§

3.3.3), Recanati (1998, 2004) refers to Douglas Hintzman (1986) to claim that a psychological model, Hintzman's (1986) *memory* model, can provide support for his views. Parts of chapter 4 and a long discussion in chapter 5 take Hintzman's contributions up and spell out the relationship between contextualised senses and memory.

What should already be clear is that context-dependence results from the fact that, on one occasion of use, a particular set of features might be selected to justify the similarity between source-situations and target-situation, and a different set on a different occasion resulting in slightly different senses. Recanati does not offer any examples to illustrate semantic potential in action, but below I offer some possible applications that might serve as examples. First, imagine a child learning the word 'bath'. He has heard this word in situations where he is interacting with an adult and playing with his toys while wet or during his daily routine just before these activities. Notice that the fact that 'bath' involves being washed might not be the most salient feature of these situations for him. The semantic potential of 'bath' for this child simply is the collection of source-situations in which he has encountered the word 'bath'. And, at an early stage, the features that justify applying 'bath' in his mind might not be those of the language community at large. But, as his experience with language grows, he will accumulate source-situations eventually resulting in correct use and this usually by the time production begins. Many researchers have now recognised the importance of modelling child language acquisition as *gradual*; grasping the meaning of a word is not an all-or-none accomplishment but rather accumulative over exposures. The semantic potential notion offers a way of conceptualising, not only why learning must be progressive, but also the kind of associations that the child accumulates when he learns a word (for accounts of accumulative learning of word meaning and shifts in children's word meanings, see Levy and Nelson, 1994; Bloom, 2000; Tomasello 2003b, and references therein).

Another potential application for the notion of semantic potential is in modelling figurative language use. A *truly* novel and figurative use is difficult to interpret, or purely 'evocative', insofar as there are few or no

source-situations to call on when interpreting it. The interpretation of established or 'conventionalised' metaphors, on the other hand, can rely on vast amounts of prior uses captured as the semantic potential of this metaphor. Importantly, this does not mean that conventionalised metaphors have fixed meanings, as their meaning must also be built in context; rather, it means that the interpretation is more constrained than in cases of novel metaphors. These examples are not meant as an exhaustive catalogue of possible applications of semantic potential, but are intended merely to clarify the notion by illustration.

3.6 Closing Remarks

The aim of this chapter has been to bring together the complex set of contributions to the topic of word meaning from theorists working in cognitive pragmatics, philosophy of language, and related fields. These contributions take their starting points in diverse schools of thought and disciplines and as a result are not devoid of inconsistencies. Rather than a fundamental flaw, this simply follows from the novelty and intricacy of the issues at hand combined with the wide scope of influences taken to reflect on them. To facilitate my presentation, I divided the chapter up into sections and adopted only a partially chronological order. I also adopted the umbrella term, 'contextualism', to refer to the group of approaches that I argue best fit the evidence. My aim was to highlight the radical changes that occur in theorising on word meaning when, first, the ubiquity of context-dependence is acknowledged, as in even very moderate contextualism, and, second, when this leads to framing the question of the meaning of words in a completely different way, as I suggest is justified in radical contextualism. I argued that just as the ubiquity of context-dependence is becoming the new consensus within the field, we must explore the logical limits of this phenomenon, and stay open to the radical consequences it might bring. I am particularly interested in whether the notion of context-independent meanings or schemas is still warranted when the question of word meaning is asked outside of the traditional semantic framework.

My own approach to language comprehension processes is largely indebted to the relevance-theoretic account, but I also seek to move away from the assumption that, while words *in use* have contextual meaning, when out of context, they have stable, fixed, static, literal, core, context-independent meanings, whether fully conceptual or schematic. Insofar as relevance theory describes meaning modulation as taking encoded linguistic meanings, that is *specifically linguistic* knowledge as the input to pragmatic processes, there is a division between what is touched by pragmatic processes and what is not: between the encoded concepts that remain fixed and the ad hoc concepts that are occasion-specific. Only a strict division between semantics and pragmatics, with pragmatics *subordinated to semantics*, can justify such a division. Perhaps the work on meaning modulation has revealed, not that linguistic meanings are static and fixed *before* they are used in context, but rather that context-independent meanings are a chimera, real only in the linguist's and language enthusiast's metalanguage. And, just as with definitions, that other handy construct theorists so begrudgingly gave up, perhaps the more we look for perfectly context-independent meanings, in any shape or form, the more elusive they will prove.

Anticipating a possible objection, I should add that challenging the existence of context-independent meanings in no way equates communicating with words to communicating with 'kicks under the table and taps on the shoulder', in the words of Stanley (2000: 396). My view is 'eliminativist' in that it challenges the assumption of strictly linguistic information/meanings perfectly separate from the use to which they are put. But it is multiplicativist, if I may, in that, following the action tradition, it approaches the phenomena from the opposite direction: it starts with language *in use* where meanings multiply as occasions of use multiply; the extent to which they can vary is limited only by speakers' tendency to converge and conform. If we see convention as no more than 'a regularity of behaviour' and entrenched form-meaning pairs as 'solutions to recurring coordination problems', as suggested by Beckner et al (2009), Hans-Jörg Schmid (2008), and the action tradition in general, we see that when we

communicate using words, we not only benefit from past regularities produced in the course of countless interactions, but we also greatly benefit from the possibility of flexibly creating new conventions to suit present needs. These solutions are alternatives to the 'code' model of language.

As already mentioned, relevance theory has also greatly supported my own thinking on word meaning because, beyond rejecting minimalism for theoretical reasons, relevance theory is among the few approaches to propose a detailed account of the pragmatic processes involved in building context-dependent, occasion specific word meanings (i.e., the relevance theoretic ad hoc concept construction and utterance comprehension procedure). While I reject the advocacy of a 'code' model of language that is part of this proposal, I have argued in this chapter, and will continue to argue in subsequent chapters, that there are viable alternatives to the code model. Furthermore, these alternatives do not disrupt the foundational principles of relevance theory, as psychological plausibility is still considered a major factor in justifying explanatory constructs, and it is assumed that cognition is driven by relevance, defined in terms of optimal cognitive gains in proportion to cognitive effort. The new account could be generally compatible with relevance theory, and, critically, the pragmatic processes so aptly described by relevance theorists, would still be part of the construction processes the new account would postulate.

There are still considerable challenges in developing this new account. Support seems dispersed in the writings of very different theorists: Recanati, Rayo, Barsalou, Bosch, Kahneman and Miller seem to arrive at proposals compatible with certain radically contextualist views *independently* of each other. Recanati's meaning eliminativism, the only one of these that explicitly adopts *eliminativism* vis-à-vis any kind of context-independent linguistically-specified word meanings and discusses the consequences for theorising on word meaning in context, has not often been discussed in the literature and, to my knowledge, no one within pragmatics, not even Recanati himself, is explicitly and unequivocally calling for its adoption. Among the reasons given, by Carston for instance, is that meaning eliminativism is *too* radical an approach and that it is hard to see how it

would actually work in practice. I do not anticipate that breaking from the traditional semantics/pragmatics divide and winning support for eliminativism will be easy, but there is one major objection that I believe I can address: namely, the worry that, without *some kind* of context-independent input into the construction process, whether it be context-independent word meanings or something less conceptual or more schematic, the construction of occasion-specific word meaning becomes too unconstrained. I hope that some of the arguments in the last section of this chapter have shown that this is not necessarily the case. I continue in the following chapter to present evidence in support of the idea that our cognition is well-equipped with construction processes, which we can describe as pragmatic *or as part of our general reasoning*, that produce not only the occasion-specific meanings for our language comprehension processes but all sorts of context-dependent, occasion-specific structures that support our interpretations and processing of all the objects and events in the stream of our experience. Among the structures created are categories, concepts, norms, and scripts. In this framework, occasion-specific word meaning construction becomes just one of the products of our relevance-driven cognitive processes.

Chapter 4: Psychological Perspectives on Concepts

4.1 Introduction

Chapter 2 focused on a detailed presentation of Jerry Fodor's particularly influential theory of concepts and its adoption (with some adjustments) by relevance theory. The discussion ended with an open question regarding the relation between concepts and word meaning to which I promised to return in this chapter. The main focus of chapter 3 was the contributions of contextualist approaches to word meaning in context. In view of the results, I made a case for a particularly radical form of contextualism: meaning eliminativism. The main aim of this chapter is to provide a detailed account of the psychological framework in which I see my own radically contextualist, meaning eliminativist, approach working. As mentioned before, with the exception of relevance-theoretic lexical pragmatics, attempts to develop the theoretical reflections on word meaning presented so far into full-fledged accounts are rare. This is particularly true of the more radical approaches closest to my own; for instance, as far as I know, no one has proposed a fully worked out account using Recanati's notions of 'contextualised senses' and 'semantic potential'; this is probably due to the considerable consequences implied in adopting such a radically different view on meaning. Understandably, the more an account seeks to challenge received views, and propose a new framework, the greater its dependence on supporting evidence. To tackle this challenge, I propose, in this chapter and the next, to carefully consider contributions from the study of concepts and categories from a psychological perspective, and, from a related field, memory models for selecting and retrieving relevant information for on-line processes of occasion-specific word meaning construction and utterance comprehension. My account postulates, much like Fodor's, that word meanings (in context) are concepts, that is, words are used to express

concepts.[88] But I propose that this solution to word meaning requires an approach to concepts *significantly different* from any other I am aware of. [89] Critically, following the ubiquity of context-dependence uncovered in the previous chapter, I suggest re-examining the context-dependence of concepts *themselves* as a necessary step in a full account of the context-dependence of word meaning. Finally, as suggested at the end of the previous chapter, I consider it important that this account integrate the critical role played by general reasoning processes and common sense.

A closer look at categorisation research is vital since it is there that I found the most complete account of the kind of construction process that applies general reasoning processes and common sense in a context-sensitive way to interpretation and outputs ad hoc concepts and occasion-specific word meanings. The foremost objective of this chapter is to defend the idea that concepts themselves can be constructed *for particular purposes* and to argue that this has significant consequences for how we think about word meaning. To make a solid case for this, however, the evidence collected by psychologists needs to be presented and some key re-interpretations of classic notions need to be covered.

I divide the chapter into this introduction followed by three major sections (4.2, 4.3, 4.4): the first covers preliminary notions that serve as a bridge between topics already discussed in previous chapters, and as an introduction to the contributions of psychologists in this chapter. Section 4.3 attempts a brief but thorough presentation of the most important contributions in the study of categorisation for my purposes. Lastly, I propose a comprehensive account of categories and concepts that joins key contributions from the philosophers and language theorists of previous

---

[88] I also take it as uncontroversial that concepts, *insofar as they apply to things*, are, or function as, categories (Fodor, 1998). When I apply the concept ELEPHANT to Dumbo, I include Dumbo in the extension of the category *ELEPHANT*. The discussion in this chapter is not of this point of contact between concepts and categories, but rather of an interpretation of categorisation *research* that suggests that categories are unstable and context-dependent instead of fixed.

[89] There is one major difference between my approach and Carston's (2002a, 2012, and 2013) position: she claims that words encode concept *schemas* or *non*-conceptual contents. I claim, that what words do cannot be equated in any way to encoding. Both accounts agree on a second related point: that fully conceptual contents are what words *express*.

chapters with the contributions of psychologists (section 4.4, 'Barsalou's comprehensive account of categorisation'). One of the main challenges of this chapter is that technical terms at the heart of the discussion are taken to mean very different things depending on the perspective adopted. In an effort to give as clear a presentation of my account as possible, the first section of this chapter therefore offers a brief discussion and *disambiguation* of central notions; namely, 'mental representations', 'abstraction', and 'similarity'.[90]

My aim, briefly anticipating that section, is to highlight substantial differences in what these notions are taken to mean and what role they are supposed to play in our general cognitive processes according to different perspectives. Consider the notion of 'abstraction', it has generally been assumed that memory stores what a subject knows about the world in the form of 'abstract' concepts, for instance, an abstract concept BLUE is taken to somehow represent *BLUENESS* or the property of *being blue* and *thus* serves to classify different shades or intensities of blue as falling under BLUE. This assumption was part of the definitional account presented in the first *introductory* chapter. For reasons discussed in that chapter, psychologists trying to spell out the cognitive processes involved did not find this to be a good point of departure. I can now also add that an underlying assumption of this view is that a speaker is competent insofar as she applies words to those things that fit the criteria; for instance, only applies 'chair' to *seats for one, that have a back rest and four legs*. The problem that brought about the rejection of the definitional account was that these criteria (or necessary and sufficient conditions) proved so elusive that most, if not all, philosophers, psychologists, and language theorists gave up trying to pin them down for concepts. I would like to stress, however, that the rejection could have been even more complete; arguably, it could have gone to the point that future approaches no longer held that what we know about the world is in *abstract* form; in other words, that our long-term memory

[90] As a reminder of my conventions, I put technical terms in single quote marks when I first introduce them and thereafter only when needed. Small capitals are used for concepts, small capitals in italics for categories and italics for criteria or features.

contains stable, invariant representations of categories and perfectly delimitated concepts instead of undifferentiated information. Despite the rejection of the traditional account, it was still accepted that abstract concepts *as construed by the traditional account* are somehow still the base for our knowledge of the world, *despite the fact that we cannot define them*; and, to a lesser degree, that speakers have abstract *knowledge* associated with their concepts that makes them competent users. A critical question that rethinking the traditional view of concepts and the definitional account should have brought to the forefront is *where* abstract ideas or concepts come from, or how *and when* abstract ideas and concepts are created. Answering these questions is particularly important to building an eliminativist approach to word meaning since, on the one hand, eliminativism denies that the construction process outputting occasion-specific senses of 'blue' needs an abstract concept BLUE as input and, on the other hand, claims that it can take particular contextualised senses stored in memory as input and create a *new* occasion-specific sense for 'blue' as output.

A very important part of this chapter is, therefore, to show that, instead of taking 'abstractions' as the input to our construction processes, as previously assumed, we can begin to see that our cognitive systems are set up in such a way as to *build* interpretations by employing construction processes capable of *producing* whatever structure we need to make sense of a scene. Abstractions are, as traditional theories have always held, a very important part of our mental lives, but if norm theory and contextualism are right, they are not arrived at in the way previously assumed, and, even more importantly, they do not undertake the roles previously assumed. This possibility was already evoked in the very last section of chapter 3 where I presented recent thinking on reasoning, decision-making and language interpretation suggesting that making decisions, evaluating a particular scene, and making sense of an utterance, for instance, do not depend on the retrieval of pre-computed 'norms', but that the process of interpretation generates its own structures (frameworks of evaluation, categorical knowledge, concepts, and, *I contend*, occasion-specific word meanings).

Much of this chapter is intended as further support for this idea, the main difference is that I now adopt the point of view of psychologists working on categorisation and concepts.

The second section of this chapter presents a roughly chronological guide through the research into categorisation that covers notions such as 'prototype', 'exemplar' and 'psychological essentialism'; all have played an important role in various theories of concepts and are therefore pertinent to accounts of word meaning. Also, covering the evolution of categorisation research is relevant, not only because it helps to dispel some misunderstandings regarding possible interpretations of this research, but because the field has evolved in such a way as to point to an account of concepts that is highly compatible with, and *thus offers support to,* the contextualist and radical contextualist accounts presented in the previous chapter. Moreover, arguably, a radical contextualist account of occasion-specific word meaning would not be complete without an exploration of at least some of the mutual consequences certain construals of word meaning in context and certain construals of concepts have on one another.

Finally, there are some foreseeable objections to the overall account I present and so I have made the discussion of its supporting evidence very detailed. First, I focus on prototypes: among the possible interpretations that would make prototypes singularly important to theories of word meaning (in context) is if concepts are represented by prototypes. As discussed in chapter 2, the definitional approach assumed that each member of a category had some critical feature or set of features that marked it as a member of its category. These criteria were formalised as a set of necessary and sufficient conditions for category membership. Thereafter, however, two difficulties emerged: first, even after careful thought and long deliberations, philosophers found it difficult to come up with strict criteria for more than a handful of terms; secondly, it seemed that in their categorising behaviour, competent users did not actually rely on such defining criteria. Defining criteria simply did not seem to be playing the role originally envisaged for them in the traditional account. With the arrival of prototypes, a possible solution to this failure was envisaged: since

prototypes are very much like definitions in that they name criteria for membership to a class, the logic of definitions could in part be *preserved* in prototypes. A category would still be represented by an abstract set of features, the difference would be that, rather than postulating individually necessary and jointly sufficient features, features would only be *statistically prevalent*. This interpretation was immediately attractive to a great many but, ultimately, as I argue in this chapter, it should be abandoned.

A second subsection on categorisation (§ 4.3.2) covers the research that helped some psychologists move past their initial interpretations of prototypes to consider other phenomena in categorisation such as exemplars and, *of particular importance to my account*, the effects of context. Other important evolutions in the field of categorisation research are also presented in section 4.3 of this chapter, with special emphasis on accounts that integrate evidence from different approaches to offer a more complete and unified view of categorisation behaviour. Finally, to close this long second section of this chapter, I focus on the notion of psychological essentialism. A particularly apt criticism of prototype and exemplar accounts was that they dealt only with superficial perceptual features and 'statistical prevalence' and so were fundamentally incomplete, much in the way suggested by Rey in the introductory chapter (§ 1.3). The discussion in that section will allow me to respond to Rey's objections among others.

A third major section of this chapter will then bring the diverse contributions together and offer a comprehensive account of concepts and categories. This section will take up where chapter 3 left off. There, I claimed that subsequent chapters would offer support for the kind of radical contextualist account of word meaning in context under consideration in the section entitled 'the third stage'. For instance, Rayo's 'grab-bag' account was mentioned as an example of an eliminativist construal of word meaning. With the help of psychological construals of concepts, particularly Lawrence Barsalou's, I hope to give a more detailed explanation of how information at our disposal gets organised into the bundles mentioned by Rayo and others in chapter 3. Beyond that, my aim is that the presentation of categorisation research in this chapter reveal that the mechanisms underlying

categorisation do much more than just help us decide whether something is an *X* or a *Y*. Much like the meaning of a particular word might need to be constructed ad hoc for a particular purpose, as in the case of a stranger in a park uttering 'My dog is intelligent', categories are also frequently *constructed ad hoc for particular purposes*. In the above example, an ad hoc category for *INTELLIGENT BEINGS* might need to be created that critically does not exclude dogs. If Waismann's only existing category for intelligent beings *excludes* dogs, he would simply fail to understand the utterance. What he has at his disposal, however, is the means of creating an ad hoc category *INTELLIGENT BEINGS* that does include them, for the purposes of the context at hand. Notice that the process assumed here is very much like the relevance-theoretic process of widening and narrowing of denotations, but the difference is that the input is not any *linguistically*-specified standing meaning of 'intelligent'. Although the output is ultimately an occasion-specific meaning of the word 'intelligent', as in relevance theory, it is important to note that the overall process of constructing an interpretation explicitly claims to involve not only the creation of this occasion-specific sense, but the parallel creation of an ad hoc category, an ad hoc concept and a 'framework of evaluation' for constraints on these creations.

4.2 Preliminary Notions

The main purpose of this section is to give a preliminary introduction of important notions in an effort to set the scene for the rest of the discussion in this and subsequent chapters.

4.2.1 Mental Representation

Jerry Fodor credits Zenon Pylyshyn with important insights into foundational questions regarding cognitive science. Two of these questions directly regard mental representations: 'What kinds of things are mental representations?' and 'How do mental representations have content?' (Fodor, 2009: ix). There are broadly two stages in Pylyshyn's research into mental representations, the first spanning from the 1970s to the 1980s and

the second culminating with the 2009 publication just cited. In this subsection, which focuses on the first stage, the key question asked is *what kind of things mental representations could be*. In the early post-behaviourist era, the idea that mental representations were mental *images* was very popular. Pylyshyn (1973), however, found that, despite its acceptance and the amount of research it inspired, closer inspection revealed that the evidence offered in its favour was misleading, or at least in need of reinterpretation. In the second stage, to which I briefly come back near the end of the section on 'abstraction' (§ 4.2.2), the question is to do with how mental representations have content.

In the 1973 article, Pylyshyn starts by pointing out the ways in which mental representations could not possibly be 'images' in any intelligible sense. He wonders, for instance, if we could conceive of two images of identical chessboards that differ only in that one image somehow contains the relation 'is attacked by' and the other does not. The point here is that, while an image could conceivably code certain geometrical distributions and sensory attributes, it is more difficult to see how it could represent the types of relations that we associate with the scenes, in other words, *what we know about them*, and that mental images as representations are supposed to capture. The only evidence backing up the mental imagery view is that subjects report experiencing images as they introspect on what they know. Yet, that this subjective experience can be identified with actual processing is no more than an assumption; particularly since mental processes are generally not available to conscious introspection.

Pylyshyn also reveals important flaws in the arguments of theorists favouring mental images as the form of mental representation. For instance, they might argue in favour of their own accounts by pitting them against equally erroneous accounts that propose 'mental words' instead of 'mental images'. According to Pylyshyn, at most, these studies suggest that cognition is *mediated*, but the form underlying this mediation is unlikely to be images or even words (1973: 4). Mental representations, he concludes, must be in some 'common format' that encompasses different types of representations. He calls on the fact that we can both use words to describe pictures (mental

and other) and effortlessly associate pictures with words to justify the need for postulating *abstract* representations inaccessible to conscious experience. To illustrate, he offers the following example: 10 milliseconds suffice for a subject to identify a letter flashed before him on a display, but it takes more than 300 milliseconds for the subject to name the letter. He interprets this as showing that while it takes only 10 milliseconds to extract enough information from a visual display to identify a letter, this information, arguably because of the format it is in, is not immediately available to speech processes.

This meshes quite well with the then received view of knowledge as qualitatively different from perception in that it is an ordered system of propositional representations and not a '*montage* of sticks, stones, colour patches and noises' (Norwood Russell Hanson, 1958: 26; cited by Pylyshyn, 1973: 6). This, of course, will have to be hedged in stage two of Pylyshyn's work which involves reconsidering some of the assumptions adopted in the 1973 article, particularly with respect to the necessary *otherness* or 'abstractedness' of mental representations.[91] But at this stage, circa 1973, Pylyshyn reasons that the best way to guarantee that mental representations accurately capture *what we know* about the world is to separate them from our experience by underlining their *conceptual* or *propositional* nature, or, in other words, whatever separates them from the raw data. In stark contrast to the 'picture' metaphor, he concludes that the representations that arise from experience with the world must not correspond to raw sensory patterns but rather be *abstracted* and *interpreted*. Furthermore, they must be *symbolic* structures, 'no different in principle from the kind of knowledge asserted by sentences, or potentially assertable by some sentence'. So, in response to the question of what kinds of things mental representations are, Pylyshyn answers that they are

[91] This development in his thinking is critical to answering the second question stated above: *how do mental representations get their content?* Pylyshyn will then adopt the view that knowledge is <u>*not only*</u> a montage of sticks, stones, colour patches and noises (see § 4.2.2).

symbolic structures with the *abstract* qualities of propositions instead of the particular qualities of pictorial images (Pylyshyn, 1973: 7).

Following mainstream thinking in cognitive science at the time, Pylyshyn holds that the interpretive process proceeds via type-token pairings: types corresponding to features, associations, relations, objects and events are stored in memory and the specific instances that make up our experience token those types.[92] Two related assumptions complete this picture: first that the 'common format' deemed necessary to cognition is achieved by reducing 'raw' data into types; and, second, the widespread belief that this reduction is necessary in order not to overload memory. As a result, the types in Pylyshyn's (1973) account represent the *abstractions* by which all sensory experiences are interpreted and the results of these interpretive processes are themselves *abstract* because the particular instances in our experience token *abstract* types. It is important to remember that this combined view of *mental* representations and type-token pairings leading to *abstract* representations became, for some time, the widely accepted position in cognitive science.

The particular construal of the notion of abstraction it assumes, however, has recently begun to be challenged; *it is these challenges*, and their implications for a new framework for word meaning in context, that particularly interest me. I propose a brief re-examination of the assumptions of the general notion of abstraction: the issues involved will not be resolved, but I try to justify the direction I take in the remainder of this chapter and subsequent chapters. As mentioned already, both the traditional and the eliminativist view of word meaning, plus any approach in between them on the gradient, crucially rely on some notion of abstraction. What's more, it seems that *wherever* we look in the study of how the mind works, this notion appears as central. It is therefore truly puzzling that it receives relatively

[92] Notice the continuity between Fodor's position that adopting representational theory of mind implies construing concepts as symbols that satisfy type-token relations (as discussed in chapter 2, § 2.6) and the position that the interpretive process consists of type-token pairings.

little *direct* attention and that so few theorists choose to give explicit explanations of how they construe it.

4.2.2 Abstraction

The first central issue seems to be whether abstractions are *pure forms* in the Platonic sense, or whether they are created from experience with the world. To illustrate the difference: imagine in your life you have only ever come across one apple; roughly, according to one possibility, there is a perfect, 'abstract' form APPLE that allows something like recognition to take place: the type APPLE is tokened by the active representation of the experience of encountering the object in front of you. In this possibility, whatever is *familiar* in the second encounter with an apple is *predetermined*, pure forms pre-exist their instantiations. According to the other possibility, nothing is available to get tokened on your first encounter with an apple, you must yourself create *whatever* it is that will foster the feeling of familiarity, whatever it is that allows a subject to generalise from experience with individual objects to thinking, of the second, third, or hundredth apple encountered, 'this is an APPLE'.

Many who today are wary of the Platonic ideal opt for construing the notion of abstractions as the *creations* of observers and reject the idea of *perfect*, *pre-existing* forms. But this direction is not free from problems. Abstractions continue to be construed as central to cognition: the consensus is that they make certain higher cognitive processes such as generalising a feature to a class, making inferences or focusing attention possible; yet, if it is granted that subjects can *create* abstractions from their experiences with the world, it seems surprising how little underlying assumptions about *how these abstractions are formed* are discussed in the literature. Lawrence Barsalou (2003) suggests some further terminology to help remedy this situation. The starting point is that most theorists could be taken to share the assumption that creating an abstraction implies what he calls 'categorical knowledge': experience with the world brings us into contact with category members and their settings; categorical knowledge is created when we abstract away from these particular experiences with objects and

events to create classes and extract properties from their contexts of instantiation. To illustrate, categorical knowledge is created when, for instance, instead of seeing two sortally distinct items before you, you see two items and see them as APPLES. Or you see one *apple* and one *orange* and can think '*FRUIT*'. Consider how much more complex this becomes as the concepts that you are supposed to create become more 'abstract', or *less concrete*. For instance, the perceptual features shared by APPLES are the *perceptual* features of the individual items labelled as APPLES, so more readily available to our senses. Compare this with FRUIT. Superficial perceptual features are not helpful in the same way for FRUIT as they were for APPLE; and this complication is only magnified in *immaterial* abstract concepts like PEACE.

The categorising behaviour illustrated above is thought by some to depend on what Barsalou calls 'summary representations'. A summary representation for APPLE and a summary representation for ORANGE, for instance, are assumed to share something that supports their coming together under FRUIT. 'Summary representations' in Barsalou's terminology roughly correspond to what Zenon Pylyshyn (1973) calls a 'propositional' or 'conceptual representation', and Edward Smith and Douglas Medin (1981) call an 'abstract summary'. Depending on the approach, summary representations can range from very strict declarative rules held in long-term memory to more flexible forms such as statistical prototypes and connectionists attractors. For Barsalou, those who claim that there are underlying 'summary representations' that we base our categorising behaviour on would need to provide an explanation of where they come from; assuming that they are pre-existing, pure forms would take us back to the Platonic ideal.[93] The alternative is to follow exemplar theorists who, as we'll see in this chapter, postulate models that accomplish categorisation

[93] I do not discuss Barsalou's proposal for an alternative to summary representations other than to say that, at this point in his career, he believes everything needed to accomplish these tasks is present in the context broadly construed. Later in his career, he adopts the 'grounded' or embodied cognition paradigm (Barsalou, 2008). But not following him on grounded cognition does not interfere with adopting his views on context-dependence.

tasks while largely bypassing the need for pre-existing, pre-computed 'summary representations'.

The next key issue is that, until quite recently, it was assumed that once a 'summary representation' had been achieved, the particular information that was not retained did not need to be stored and so was discarded. In other words, in what can be called the pre-exemplar era, it was simply assumed that once cognition had formulated a rule as to what makes all apples *APPLES*, or somehow captured this in a 'prototype' or 'attractor', the particulars of the experiences with apples would become irrelevant, burdensome to memory and therefore, for reasons of economy, better lost than stored. The issue of what motivated such assumptions is given thorough treatment in the following chapter on the role of memory. For now, I can say that work on memory has brought to light evidence of *instance memorisation* (Medin & Schaffer, 1978) which directly challenges the belief that once a summary representation is achieved, details of the individual instances are lost. Starting in the 1970s and continuing to this day,[94] evidence showing how specific training items influence subsequent tasks of categorisation is routinely discovered. Part of this evidence is presented below, in the section on concepts and categorisation (§ 4.3.2). In the interest of clarity, however, and to anticipate objections, I would stress that only a very extreme exemplar model would claim to function without *any* summary representations.

Finally, postulating that the details of individual instances are stored in memory instead of discarded opens up a possibility that traditional models had ignored: namely, that, whatever processing or interpretation consists of (whether it is abstracting away from context or extracting features from instances), creating an abstraction, or 'summary representation' is not necessarily only done *while* input is being processed. The more traditional, pre-exemplar era models of abstract representations assume that *all* the relevant information of a situation can be extracted *as the situation is being processed*. In this view, mental representations become abstract *as a*

[94] From Medin and Schaffer (1978) to Nosofsky et al. (2011)

consequence of being created.[95] This can only be successful, however, if we are somehow pre-equipped with infallible organisational principles that allow us to selectively identify *at the time of initial processing* what information will be relevant for future categorisation and generalisation tasks. This is problematic because the variability in our experiences *and the uncertainty of the future* is such that no fixed set of criteria can be sufficient to produce reliable results. Once again, the answer seems to be to postulate, as in norm theory, that deciding whether a particular odd looking apple is in fact an apple, for instance, typically involves coming up with ad hoc criteria for classifying it thus. Accordingly, what is needed is a *dynamic* process of creation that waits until the categorisation or transfer task *is at hand* before generating the necessary framework of evaluation. In the approach I will be arguing for, the creation of the 'summary representation', if it happens, is *after the fact*. The biggest up side to this contextualist stance is that postulating *a priori* relevant criteria is avoided by letting the context at hand play a bigger role. One of the main objectives of this chapter is drawing a full picture of how context can play the role I am suggesting here.

Before moving on to the subsection on similarity, I briefly return to Pylyshyn's second stage of research focused on the issue of how mental representations have content. In his early work on mental representations, he assumed that whatever mental symbols we had were connected to the world *evidentially,* or in other words, via semantic relations of satisfaction. But as he further worked in the area of visual perception, he realised that there must also be a connection between the mind and the world *prior* to

---

[95] This is the case in Pylyshyn's view insofar as he considers that interpreting or processing a scene comes down to tokening *pre-existing* types. His reasons for insisting on this point have to do with his opposition to the imagery metaphor of mental representation. He thought that the fundamental misleading implication carried by the notion of 'mental images' was that what we retrieve from memory when we activate 'mental images' is like what we receive as input from our senses, that is, completely *undifferentiated*, raw data (Pylyshyn, 1973: 8). He only considered two extremes: either memory stored only the results of his type-token pairing interpretive process or an image-based memory could call images out as if replaying a recording. The information was either completely processed according to pre-existing types stored in memory, or completely uninterpreted, like images on a video recording.

any descriptive representation. At the primitive stage of processing a visual scene,

> we can, under certain conditions, also refer to or represent some things without representing them in terms of concepts (Pylyshyn, 2009: 7, see also 2007).

A type-token pairing requires that the individual visual object be seen *as falling under a class*, but in the early stages of processing a scene, the properties that would warrant this classification might not yet have been themselves recognised. In other words, there must be a way to treat an individual visual object as a token before its type can be identified. It occurred to Pylyshyn that there must be a way of tracking an individual instance while further information about it is gathered. This is compatible with other observations in the field and the general approach of exemplar theorists today. It is generally accepted that mental representations build up over time, so an attachment has to be established in our minds between the individual thing in the world and the budding representation as changes in the scene occur so that changes can affect the thing itself that is tracked, transforming some or all of its properties, without interrupting the connection. Pylyshyn's proposal is that some aspect of our perceptual processes allows us to think about something and keep track of it without the need of a description or classification of it. Something in the world is selected by our perception 'because it drew attention to itself', or, as Pylyshyn prefers, because it 'grabbed' one of the indexes called FINSTs, for FINgers of INSTantiation (Pylyshyn, 2009: 5).

Work on this proposal has led Pylyshyn to reconsider the conceptual/symbolic mind-world relation prevalent in cognitive science and in his own earlier approach. Clearly, if something can be tracked and information about it can build up over time *before* that something is represented as falling under a description or class, then mental representations are not always necessarily *abstract* and *symbolic*; Pylyshyn is forced to accept the thesis of nonconceptual representation that he had

previously rejected. [96] He still rejects the position that mental representations are images in any intelligible sense, but he now admits that you can refer to something without referring to it as a thing that has particular properties and that, therefore, warrants a particular categorisation. This is an advantage if you want to offer an account of how mental representations get their content since, at least initially, it grounds the mind-world connection in a *causal* relation, instead of an *evidential/semantic* one. Once the connection is established, it is easier to see how the conceptual information associated with the object can be accumulated.

4.2.3 Similarity

The development of the theoretical notion of similarity owes much to Amos Tversky's (1977) influential article. This seminal piece on similarity statements and judgements gives similarity a central role in human cognition: 'It serves as an organizing principle by which individuals classify objects, form concepts, and make generalizations' (p. 327). According to Tversky, however, despite the importance of such a notion for psychological theory, the models in vogue at the time overlooked crucial characteristics of similarity statements and judgements and consequently were in need of revision.

The axioms of the 'geometric' approach, which describes similarity roughly in terms of distances in dimensional space, treat similarity as symmetrical. This, however, points out Tversky, can easily be proven false. When we compare the similarity of two items in a statement of the form '*a* is like *b*', the less salient of the two tends to be placed in the subject, or *a*, position, and the more salient item, or 'prototype' (as in model), in the *b* position, it takes the role of 'referent' (as in point of comparison). So, for instance, we say that *the daughter resembles the mother*, and *the portrait*

[96] These nonconceptual mental representations still differ in important ways from the kinds of representations proposed by defendants of mental imagery, so Pylyshyn maintains his position on mental imagery (see Pylyshyn 2003a, 2007).

resembles the person. Our similarity judgments are therefore directional and '*a* is like *b*' is not equivalent to '*b* is like *a*'. This is particularly evident in our figurative analogies: *Turks fight like tigers* and not *Tigers fight like Turks*. Another incorrect prediction of the account dominant at the time is that if *a* is judged similar to *b* and *b* to *c* than *a* cannot be judged too dissimilar to *c*. Yet, Tversky, citing an example from William James (that has lost some of its relevance since the fall of communism), reminds us that Jamaica is similar to Cuba, because of their geographical proximity and Cuba is similar to Russia, because of their political affinity, but Jamaica is not at all similar to Russia.

Tversky's alternative is a model of *contextualised* feature matching. The model assumes that objects are represented by sets of features. These features are not necessarily those associated with the necessary and sufficient conditions of the classical definitional account. They can be features that subjects would tend to associate with an item and not strictly necessary of the category. Tversky further assumes that we have a kind of data base where each particular object, person, place or event is represented by its appearance, function, relations to other elements of the data base and, in general, any information relevant to the item that can be deduced from our general world knowledge. A related idea is present in the work of Roger Schank and Robert Abelson (1977). They are famous for the introduction of the notion of a script, but they are also remarkable for having recognised early on that judgments of similarity are contingent on general reasoning or 'cognitive capabilities'. The subject must have some sense of the *significance* of particular features in order to ascertain its relationship to the rest of the system. To illustrate, imagine that we encounter a number of zebras and that, among their features, we might identify *4 legs*, *stripes*, *has a mane* and *eats grass* as features likely to serve in the identification of a future zebra. Now, imagine a *stripeless* zebra comes along: in the classic model, the options seem to be either to remove stripes from the definition or to reject that this animal is a zebra. In Schank and Abelson's account, on the contrary, there is another option: a pragmatic, as in efficient and practical, decision can be made with respect to the weighting of the *stripes* feature: The *stripeless* zebra can be considered an exception to

the rule, or the decision of whether or not it is a zebra can be postponed pending other *stripeless* zebras, or the incident can simply be ignored, judged as a fluke. This decision would be based on how *useful* the subject finds stripes as a feature in identifying zebras.

The validity of similarity as a theoretical construct has come under attack and been defended, by turns, over the decades.[97] Some critics call for abandoning the notion altogether. Nelson Goodman, for instance, in an often cited criticism, argues that similarity is a meaningless notion, 'invidious, insidious, a pretender, an imposter, a quack' (1972: 437, cited by Medin, Goldstone and Gentner, 1993: 254). But the reason given by Goodman to justify giving up on similarity is that it depends on a frame of reference; he claims that, since with each similarity judgment, it is necessary to specify *in what respect* a comparison is carried out, it is not similarity that is doing the actual work; rather, similarity is like a blank, or a slot filled out by a reference frame of specific respects. Another frequent way of putting this same criticism is to say that the explanatory work in tasks of similarity is done by the processes selecting attributes (Murphy and Medin, 1985). Consider for instance two similarity judgments: 'cats and dogs are very similar' and 'cats are nothing like dogs', the point is that the content of these judgments depends on which attribute is selected: 'mammals', 'pets', 'loyal', and so on. The perceived problem is that if this attribute is not *given* as part of the similarity judgments, then similarity *on its own* is left emptied of its force. These criticisms, however, are without effect on Tversky's notion of similarity since, in general accord with norm theory (discussed in chapter 3, § 3.5.4), although his article predates Kahneman and Miller's by 9 years, Tversky is interested in similarity *judgments* and so adopts the assumption that some kind of framework for each judgment is generated *ad hoc*, in context; similarity is only problematic if considered *out of context*.

---

[97] It is often the same authors who recognise a certain weakness in superficial treatments of similarity who try to offer more in depth treatments, see for instance: Murphy and Medin, 1985; Medin, Goldstone, and Gentner, 1993; and, more recently, in Edelman and Shahbazi, 2012.

4.3 Concepts and Categorisation

This section takes up where the section on philosophical versus psychological theories of concepts in chapter 2 left off. In that section, I gave some of the reasons the classical approach was found wanting and I stated that philosophers and psychologists who had initially jointly agreed to abandon it later disagreed *fundamentally* on the way forward in building an adequate alternative. In the remainder of that chapter, I then focused on one particular philosopher's theory of concepts and on a reformulation proposed within my field: lexical pragmatics. I now turn to psychological theories of concepts and categories.

As mentioned in chapter 2, psychologists' interest was initially caught by the fascinating function of 'carving nature at its joints', or in other words, by the everyday but nonetheless astounding human capacity to group together and label the varied objects and events of the flow of experience. According to the classical definitional account, this was a matter of the object or event satisfying particular sets of individually necessary and jointly sufficient conditions and thereby 'falling' under a concept or into a category. The underlying (metaphysical) assumption was that there is one, and *only one*, way in which the world can be correctly divided up into classes. This tradition can be traced all the way back to Platonic pure forms and the Aristotelian notion of *essences*. Psychologists, however, observed two things. Not only had the definitional account failed to provide working sets of criteria by which the *essence* could be *defined* but it seemed that there was an even deeper flaw in the thinking behind this classical account: it conceived of category membership as an all-or-none phenomenon, while subjects found it quite natural to rate some members of a category as 'better exemplars' than others. For instance, in a rating task (Rosch & Mervis, 1975: experiment 1), peas were judged better examples of the category *VEGETABLE* than brussels sprouts and beets better examples than potatoes despite the fact that all four belong to the category. Under the classical definitional account, all instances of the category *VEGETABLE* are supposed to share a common essence and thereby possess the same 'criterial attributes'; they

should therefore be equal members of the category. Explaining the very robust evidence of what appeared to be *graded membership* called for a radically new approach to categorisation. A new area of research was born and, in a very short space of time, an impressive amount of empirical evidence was amassed and competing interpretations began to appear.

Categorisation researchers were prolific: there are extensive lists of theoretical and experimental articles and countless reviews (besides the references given throughout this chapter, see Smith and Medin, 1981; Laurence and Margolis 1999; Murphy, 2004, and references therein). Limits of time and space allow only a mention of some the most important works pertaining to categorisation research and, unfortunately, not a comprehensive review. My objective in this section is double: to give a brief and clear account of categorisation research, *particularly* leading up to Barsalou's proposal of ad hoc categories and concepts, and to underline the ubiquity of context-dependence.

The story is not one of homogenous straightforward progression; rather, different researchers propose very different theories to account for similar phenomena and consensus is far from immediate. The disagreements are apparent in the frequent distinctions between labels: 'prototype', 'exemplar', and 'theory' theorists or theories. I argue, however, that from a certain distance, these different approaches can seem like the succeeding stages of on-going research. Rather than representing completed analyses that are irremediably opposed, they represent competing interpretations and approaches that progressively take more data into consideration and are fine-tuned as our understanding of concepts and categorisation grows over time.

To see this clearly, I follow a more or less chronological order. The first subsection, entitled 'prototype theory' (§ 4.3.1) covers some of the earliest and best-known research with some of the first experimental results and interpretations. In the second stage, I move on to the 'context theory of classification' better known today as 'exemplar theory' (§ 4.3.2), which gives an alternative, competing interpretation of comparable evidence and rejects certain foundations of prototype theory on the grounds that context is not

sufficiently taken into consideration. A third stage, the 'dual' model (§ 4.3.3) represents a first encompassing model that attempts to offer an integrated approach that is comfortable with the evidence that had fuelled otherwise opposed interpretations. The fourth and final subsection (§ 4.3.4) introduces another set of related data that proves vital to categorisation: similarity judgments do not, *and there is no reason why they should*, depend exclusively on surface, perceptual features. With this last stage, what a subject *believes* about the objects and events she encounters plays a role in how they are processed and categorised.

4.3.1 Prototype Theory

4.3.1.1 Eleanor Rosch

Eleanor Rosch is widely recognised as the founder of prototype theory. In articles published in 1973 and 1975, she first described the phenomena of saliency and centrality that were to become the notion of 'prototype'. The work she did on her own and with colleagues like Carolyn Mervis, resulted in an impressive bank of data that came to be known as 'prototype effects'. Rosch's interpretations of this data were also particularly compelling and found widespread acceptance. In fact, prototype theory became so dominant that it partly shaded alternative interpretations. Until recently, for instance, audiences outside of psychology were mostly only aware of prototype theory and rarely exposed to parallel developments in categorisation research, like, for instance exemplar models. I come back to this point later in this section. In what follows, I present the key points of Rosch's contribution.

The starting point for Rosch was the observation that it was quite natural for subjects to rate different members of a superordinate semantic category like *FRUIT* on what seemed like their degree of membership. This so completely contradicted the classical account of categorisation that it sent Rosch off on a mission to collect as much data as possible. Over hundreds of trials, she had undergraduates assign ratings to the different members of everyday categories like *CLOTHES*, *FURNITURE* and *VEGETABLES*. They were

instructed to judge on a grade from 1 to 7 'how good' an exemplar (or member) of the category a particular item was. Not only did the task seem quite natural to the subjects but she found overwhelming agreement on the ratings among subjects.[98] Rosch also checked that the ratings did not correlate with the frequency or familiarity of the items on the lists. Her objective was to establish that ratings are robust rather than accidental so as to warrant a psychological explanation of their causes and consequences. In other words, she set out to demonstrate that the phenomena are 'psychologically real'. She predicted, for instance, that, in a particular reaction time task, subjects would respond faster to items rated higher on the list. Subjects were required to answer 'true' or 'false' to statements in the form: 'A (member) is a (category)', for instance, 'An apple is a fruit'. Subjects *did* answer faster to members rated high on the lists, which supported the interpretation that not all instances of a category are equivalent and that some members 'represent the core meaning of the category' (Rosch, 1973: 135). Rosch concluded that this inequality amongst the members of a class revealed that categories had internal structure.

Rosch (1975) and Rosch and Mervis (1975) further investigated 'internal structure'. They had the undergraduates list features or 'attributes' for the members of the categories rated in previous experiments. Contrary to the predictions of the definitional approach, there do not seem to be specific sets of features that subjects systematically list for any particular category. Unfortunately, Rosch and Mervis did not publish the actual feature lists or give many examples of what their subjects actually listed as features. But, suppose we were to list attributes for vegetables. In the minute and a half accorded in the experiments, we might come up with features such as *orange*, *fresh, crunchy, eaten cooked or raw* for carrot; *green, savoury* and *(most) kids do not like them* for *brussels sprouts*. *Green, savoury* and *eaten*

[98] Barsalou (1987) will later argue that while between-subject agreement *is* significant, the very high numbers reported by Rosch (a correlation of 0.9) is an effect of inadequate statistical tools and that agreement is much lower (around 0.5). Even these lower correlation scores, however, still warrant explanation and so this flaw in Rosch's reporting is of limited importance. I come to what I find to be the real issues with Rosch's account later in this subsection and later in this chapter (section 4.4).

cooked for *peas*, etc. Notice that even if we continued this for a list of 20 vegetables, as on Rosch's trials, there would not be a single feature that, because it *characterises the essence of vegetable*, would necessarily appear for all the members of this category. Notice also that the features subjects would tend to list in these tasks are non-defining features. The features are shared by other edible things, and, in general, would not allow a clear distinction between vegetables and fruit, for instance. On the other hand, what these lists of features do reveal is that both lettuce and brussels sprouts are *green* and that lettuce and carrots *can be eaten raw*, and so on. Rosch and Mervis (1975) propose to take Wittgenstein's (1953) concept of 'family resemblance' to describe these relationships between features and members of a class. They formulate the 'family resemblance hypothesis': natural semantic categories are networks of overlapping features, and a particular member of a category will come to be viewed as prototypical of that category to the extent that its properties are common among the other members of that category. So, in our reduced example, peas would be particularly prototypical because they are *green*, *savoury, eaten cooked* and *(most) kids do not like them*. In a test such as '*X*s are vegetables' we could expect subjects to answer significantly faster for peas compared to mushrooms. This is the *effect* of the *prototypicality* of peas. In other words, prototypes are those items in a category that are most like all other category members and this makes them relatively salient or representative. The organisational principle at the heart of this 'internal structure' is that items in a category group around particularly central and representative members. This also means that categories are continuous, they grade off from better to poorer examples (Rosch & Mervis, 1975). Rosch, Simpson and Miller offer further evidence in their (1976) review of 'typicality effects'. There is, for instance not only a reliable correlation between how central or typical an item is with how soon it is learned; there is also evidence that adult speakers name central members of a class before less typical ones in a production task. Finally, there is some evidence that category names trigger mental images of typical rather than atypical members of that category (Rosch, Simpson & Miller, 1976). In this article, prototypes are also

described as 'maximally informative', or as the 'best exemplars' of their class.

The success of Rosch's account was immediate. It effectively, if not completely or permanently, replaced the 'criterial attributes' account of the classical definitional approach. It offered a much more psychologically plausible account of the relations between features and categories. It did away with the

> tenacious tradition of thought in philosophy and psychology which assumes that items can bear a categorical relationship to each other only by means of the possession of common criterial attributes. … formal criteria are neither a logical nor psychological necessity (Rosch & Mervis, 1975: 603, see also Laurence and Margolis, 1999: 28).

Of course, Rosch's account also met some opposition and criticism; Jerry Fodor's, for instance, would be particularly pertinent since his theory of concepts is taken as a point of departure in this thesis. Nevertheless, insofar as Fodor's criticisms (1996 *with* Ernest Lepore), and (1998) were aimed at a particular *interpretation* of Rosch's experimental results, I would argue that discussing them in detail would only be justified if I were defending that particular interpretation, which I am not. In fact, as already mentioned in the introduction to this section, my view is that categorisation research advanced in stages and that Rosch's account is only part of the first stage in a relatively long list of contributions by psychologists. I claim that, because of this, interpretations of the experimental results that focus on only prototypes are best considered *partial* until later stages when related phenomena, such as exemplars, the effects of context and psychological essentialism are recognised. The particular interpretation of prototypes that was heavily criticised was the one that took it that *since prototypes can represent conceptual information*, prototype theory should be taken as advocating the notion of *a prototype* as a replacement for that of *a definition* in the theory of concepts. This interpretation has taken many forms in the vast amount of controversy, criticisms and responses it has generated. Fodor and his colleagues argue against it by pointing out all the

shortcomings of equating prototypes with concepts; most importantly, prototypes cannot be concepts because prototypes do not compose and, as discussed in chapter 2, compositionality is a sine qua non condition of concepts in Fodor's view. Arguably, however, a more modest interpretation of Rosch's results is available: there is, as such, no autonomous and complete theory of concepts and categories under the term 'prototype theory', and a completed theory would *not* give prototypes the role Fodor objects to; once more, any early account of prototypes is only a small part of a greater picture and so full evaluations are best left off for the time being.

Taking a more modest interpretation as a starting point is the best way, I believe, to salvage the valuable contributions of even the earliest stages and avoid throwing the baby out with the bathwater. A Fodorian might still have criticisms of the account I arrive at, but these criticisms would conceivably be very different from the criticisms aimed at prototype accounts. The interpretation I suggest is that what Rosch uncovered were prototype *effects*; it is an uncontroversial position since the reality and ubiquity of prototype effects are accepted by all:

> The discovery of the massive presence of prototypicality effects in all sorts of mental processes is one of the success stories of cognitive science. I shall simply take it for granted in what follows [...] (Fodor, 1998: 93).

From a theoretical point of view, Rosch's contribution is to illustrate how the difficult case of criterial attributes in the classical account can simply be avoided by relaxing the constraints of category membership and adopting instead the organisational principle of family resemblance, but this is not only still far from a complete account of categorisation and concepts, it does not necessarily yet provide the right foundations. In my view, providing these foundations depends on reconsidering the role played by context-dependence, something that takes us far away from prototype theory. This, however, is not the position most widely adopted: linguists like Dirk Geeraerts, philosophers like Jesse Prinz, and psychologists like James Hampton, among many others, prefer to keep prototype theory as a base

and offer their insights on how to develop it without calling for its dismissal.[99]

Another line of criticism is that much detail is missing from Rosch's early observations, for instance, prototypes and category gradience are not necessarily linked: classes with *and without* degrees of membership exhibit prototype effects. For instance, although '7' and '15' are equally clearly members of the category of *ODD NUMBERS*, still '7' is judged to be a 'better' member than '15' (Armstrong et al. 1983). Also, Rosch's account is vague on the *nature* of a prototype: it is either the most central member, as when she uses the term 'best exemplar', or it is an abstraction from central members, that is, a conjunction of the prototypical *features* of a class. The list of such criticisms is long and has been given ample coverage in the literature on prototypes.[100] Instead of addressing these criticisms individually, I opt for directly moving forward to an alternative built to avoid what I consider the major flaw: in Rosch's account, context plays no role in selecting which exemplar of a class is most representative *for a particular purpose*, or which features best represent a class *given the task at hand*. Her account of prototypes construes them as *fixed* and *stable*. Given the kind of construction processes suggested earlier in this chapter and in the previous chapter, this is a major flaw. If, instead of maintaining that a certain exemplar of a class is *pre-determined* to serve as the 'best exemplar' or that the most representative features of a class have been aggregated *once and for all*, Rosch had instead considered, as norm theory does, that prototypes are like norms, in that they are *means* generated for particular purposes, she would not construe prototypes as fixed. In other words, if the contexts in which prototypes are created, and the purposes for which they are created were taken into consideration, significant variation would appear. In the following subsection, I present an account that stresses the fact that

---

[99] Discussions of the contributions of the many other prototype theorists are unfortunately beyond the scope of this thesis, but see Geeraerts (1997, 2010); and various papers in Geeraerts and Cuyckens (2007); Prinz (2002, 2012); and Hampton (2006, *forthcoming*).
[100] For reviews see Smith and Medin (1981); Laurence and Margolis (1999); and Murphy (2004).

prototypes are abstractions *created* by the subjects in the countless experiments of categorisation research; in later sections, I add context-dependency to the considerations of how these abstractions are constructed in dynamic processes of categorisation. There are significant parallels between my views on word meaning and on categorisation and mental representation. I hope that as this chapter advances it will become evident that the radically contextualist view of word meaning presented in the previous chapter is supported by a certain interpretation of the most up-to-date findings in categorisation.

4.3.1.2 Posner and Keele

In two articles, published in 1968 and 1970, Michael Posner and Steven Keele offer one of the first reviews of evidence leading to the claim that subjects commonly *create* an *abstract* prototype to help them correctly categorise and generalise from previous experiences. Interestingly, their earliest article already identifies what will become one of the main points of disagreement between prototype and exemplar theories, namely, the issue of whether *the individual instances* or rather *only the prototypes* are stored in memory when subjects learn to group a set of stimuli together in a class.

They report two previous results. First, that the speed with which subjects learn to categorise patterns is a function of the degree of distortion between, on the one hand, the patterns the subjects are exposed to and, on the other, the *original pattern* from which they were generated (also called the prototype).[101] Second, that subjects can learn to discriminate patterns even when they have not seen the *original* patterns/prototypes from which the test patterns were generated. This was interpreted by some as indicating that the subjects were *creating* schemas/prototypes from the series of patterns to serve as reference points. This would mean that once

[101] In Posner and Keele's work, there are two uses for the term 'prototype'. First, a prototype can be the 'original pattern' from which distortions are artificially created, a prototype in this sense is a pattern from which the experimenters create other patterns. A 'prototype' is also the supposed schema that the subjects create when they encounter the individual patterns during learning. I will signal this distinction in the presentation in the interest of clarity when necessary.

such an abstract prototype had been extracted from the evidence given, their subsequent behaviour during categorising tasks could be explained as simple similarity judgments comparing any one test pattern and the prototypes stored in memory, (which would serve as reference points or 'norms'). But for Posner and Keele, this is but one possible interpretation, which they mark as the strong interpretation, perhaps because it is the one that makes the strongest assumptions. It holds that the *commonalities among a set of patterns are extracted from the individual instances during learning* and that *they alone are stored*. A weaker interpretation does not require that only what a subject was able to extract during learning to create an abstract prototype be kept in memory; rather, *the individual instances* of the patterns can also be recorded in memory. Notice that this last hypothesis would not require that the extraction process take place only during learning (Posner & Keele, 1968: 354).[102]

The authors designed materials and conducted a series of experiments to help distinguish between these two hypotheses. The materials are the now famous 'random-dot patterns' constructed by randomly positioning 9 dots in a 30 X 30 matrix and then applying a distortion metric. Posner and Keele's (1968, 1970) experiments mostly include a study or 'learning' phase and a test or 'transfer' phase.[103] The first experiment from the 1968 article, for instance, compared subjects' performance when the patterns they saw were low-level distortions with subjects that saw high-level distortions. The prediction was that if a clearly defined schema/prototype was necessary to accomplish the transfer task (that is, the correct classifications of stimuli into classes), then the group who saw the low-level distortions would be at an advantage. On the other hand, if being exposed to more variable stimuli in the learning phase is key, then the group with the high-level distortions would show better transfer

[102] This is important because only the second interpretation will prove compatible with norm theory, which is individually motivated.
[103] The 'transfer' task tests how well the original patterns were learned by measuring how well *new* patterns, that is new distortions of the same original patterns/prototypes, are grouped into classes.

skills. As expected, the results showed better transfer from experience with a 'broader' class of stimuli. However, it might be argued that subjects with the high-level distortions had longer study phases because their patterns were harder to learn and that this might explain, at least in part, their results. To test this, a second experiment replaced the study phase with a recognition task. The results of the first experiment were confirmed and based on these results, plus those of previous studies, the authors abandon the stronger hypothesis which states that only the schema/prototype is stored in favour of the weaker one according to which the original instances are also recorded.

A third experiment was then designed to determine what subjects are learning about the original patterns/prototypes that generated the patterns. They again had subjects study lists of patterns of distorted original patterns/prototypes and, once they could complete two errorless classifications of the lists, they had them look at a new list of patterns to test their transfer performance. The patterns on this list included the original 'prototypes', that is the original patterns from which the distortions were created; some 'old distortions', that is, patterns that the subjects had memorized during the study phase; some 'new distortions', never seen before but generated from the same prototypes at either the same level of distortion or a slightly lower level of distortion; and, finally, some new random, unrelated patterns. Transfer was tested twice, once immediately after the learning phase and then again twenty-four hours later.

When transfer was tested on the same day, there was little difference in error rates between the patterns just memorized and the original patterns/prototypes that had never been seen before. Both of these were classified faster than the new distortions despite the fact that some of these distortions were objectively closer to the original patterns/prototypes. They represented a lower-level of distortions of the same prototypes as those that generated the patterns subjects were familiar with. Among the new patterns, low-level distortions were more accurately categorized than higher-level ones. Also, old distortions are classified faster than original patterns/prototypes. Twenty-four hours later, error rates are slightly higher

across the board but the time difference between the classification of old distortions and original patterns/prototypes disappears.

Posner and Keele reflect on possible explanations for this. It seems clear that old distortions are accurately and quickly classified because they can be directly recognized. Original patterns/prototypes, on the other hand, cannot be recognized immediately, hence the slight delay. Nevertheless, they can still be classified accurately, perhaps based on a simple calculation of similarity to the stored patterns, that is, the individual exemplars from the study phase. Alternatively, the calculation could be based on information extracted from the study phase and the delay could be due to the fact that this information is not as clearly or completely defined as that of the studied patterns. In either case, it would still be true that if the abstract schema or 'prototype' is not created *during learning* it can still be created during transfer and, once encountered during testing, it can be stored as a particularly good example of its class. This would explain faster recognition twenty-four hours later.

Another two experiments in the 1970 article further test whether the prototype is created during learning or at time of recognition. Following Bartlett (1932), Posner and Keele reason that forgetting should affect peripheral information more than central information. The prediction is that, if the creation of the prototype takes place during the learning phase, then a longer delay should affect how well the individual patterns seen during learning are remembered since they have already served their function of input to the creation of the prototype and have consequently become peripheral. Alternatively, if the creation of the prototype took place during recognition, then losses in memory for the old distortions should correlate with decreased prototype recognition. The results of the two experiments show large losses in recognition of the old distortions but little change in the recognition of prototypes. The authors interpret this as supporting evidence for the hypothesis that prototypes are created during learning, a position that tends to imply that the individual instances are not kept in memory. Nevertheless, as I will discuss in the next section, future

findings would lead to a reinterpretation of Posner and Keele's (1970) results by other researchers, completely *reversing* this conclusion.

As with early interpretations of Rosch's data, these results are important as part of a bigger picture. I have included them in my overview, despite the fact that I consider their final results irrevocably overturned, because it is *in response* to evidence on differential forgetting that Douglas Hintzman, a psychologist specialising in memory, formulated his most influential arguments in favour of exemplar models. The following section presents one of the first proposals to challenge prototype theory and formulate the basis for a contrasting approach: exemplar theory. In this subsection, I return, among other things, to the idea of the activation of particular instances (or exemplars) stored in memory that I referred to earlier when I said that evidence for *instance memorisation* began to emerge in the 1970s and has been accumulating ever since.

4.3.2 Exemplar Theory

In this section, I present one of the first proposals to have given rise to 'exemplar theory': the 'context theory of classification'. A closer look at this proposal will clarify what differentiates prototype and exemplar theory and how the latter fits in the wider framework I advocate. In contrast to the theories presented above, the context theory of classification holds that classification judgments are based not on prototypes created *during* learning but rather on the stimuli themselves through a process of selective activation. The resemblance between this proposal and 'norm theory', presented in the previous chapter (§ 3.5.4), is not coincidental. Kahneman and Miller were careful readers of the theorists I present in this section and explicitly adopt tenets of their views. From a combination of very early small-scale demonstrations like those of Medin and Schaffer (1978), and Hintzman and Ludlam (1980), both presented below, and the thinking of people like Lawrence Barsalou, and norm theorists, a new approach to categorisation, the 'exemplar model' was born. The main claim is that classification judgments like those described in prototype theory can be

based on the selective activation of *instances, or exemplars,* stored in memory. As indicated in the section on 'abstraction' above, this critically depends on rejecting the assumption that the stimuli a subject uses to form 'categorical knowledge' are subsequently lost. Exemplar theorists claim, on the contrary, that our experiences are captured in representations of the individual instances or episodes of our experience and stored in memory in the form of 'memory traces'. Norm theory formalises this idea with the claim that because each object or event can *selectively* activate a set of representations, categorical knowledge need not be pre-computed; a category judgement 'can be derived on-line by selectively evoking stored representations of discrete episodes and exemplars' (Kahneman and Miller, 1986: 136). Exemplar theories have an important memory component, and so much depends on how memory is construed that I have dedicated a whole chapter to that discussion.

Here, I can say that in the new 'context theory/exemplar' framework, in stark contrast to previous assumptions, category knowledge is not retrieved from memory ready-made. Rather it is assumed that retrieval is a *dynamic* process that selectively activates certain representations ('memory traces' or 'exemplars') from memory to 'represent knowledge and interpret experience'.

In order to model this dynamic mechanism of retrieval, Douglas Medin and Marguerite Schaffer (1978) adopt Roger Ratcliff's (1976) suggestion of using a *resonance* metaphor; in other words, instead of imagining that items are called out of memory serially, we should imagine that they are *evoked* on the basis of their similarity to the probe (Medin & Schaffer, 1978: 210; see also Ratcliff, 1978). A probe can be anything in a subject's present experience, like an object or an event *and its context*; words, or 'category names' and their contexts (linguistic or non-linguistic) are included. The assumption is that this probe, or 'cue', carries information itself and interacts with the information in the items stored in memory. This new outlook on memory is vital not only to the context model presented below but to *all* modern dynamic theories of memory. It allows the integration of variables such as the strategies that subjects use, the effects of

attention, and the fact that memories are not necessarily veridical. Incidentally, it also integrates the great insight of Pylyshyn that mental representations are not like bits of film that once recorded can be replayed. For Pylyshyn, it was not possible to delay the interpretation of sensory input and he warned against the metaphor of recorded film kept at hand in case it was later needed. Replaying or re-viewing is still not possible in this approach. We have no conscious access to the information in our memory. We create mental representations that *use* memory traces but that are not those images themselves; rather, they are the constructions of an interpretive cognition. Finally, notice that here the effects of interpretation are considered, not only when sensory data is first being interpreted during input, but also when it is reinterpreted in light of new evidence in a new context.

4.3.2.1 Medin and Schaffer

Douglas Medin and Marguerite Schaffer (1978) illustrate how classification in the context model would work with a simplified example. Suppose the stimuli for classification are patterns that can be described by their binary values on a set of four features or dimensions: *colour*, *form*, *size*, and *number*. The notation coding these four features could be something like this: for *colour*, 1 represents RED and 0 represents BLUE; in second position, *form* is either 1 for TRIANGLE or 0 for CIRCLE; next, *size* is either 1 for LARGE or 0 for SMALL; finally, *number*, 1 for ONE and 0 for TWO. 0010 would read 'TWO LARGE BLUE CIRCLES'. Now suppose the following training set:

> Two patterns that the subjects learn to class in category A: A_1=1110 and A_2=1010. That is, TWO LARGE RED TRIANGLES and TWO LARGE RED CIRCLES; and two patterns that belong in category B: B_1= 0001 and B_2=1100. That is, ONE SMALL BLUE CIRCLE and TWO SMALL RED TRIANGLES.

Furthermore, suppose that the subject in this particular experiment has selectively attended to *colour* and *form* so that his mental representations of the stimuli are as follows:

111? A (A_1) 10?0 A (A_2)

00?1 B (B_1) 110? B (B_2)[104]

Now suppose a new pattern is presented: 1101 (ONE SMALL RED TRIANGLE). The context model holds that this pattern serves as a probe that selectively activates the subject's representations of previous patterns according to their similarity. The most likely match is with B_2; therefore, according to the context model, this particular exemplar is retrieved and since it is classed B, the new pattern would be associated with this class. 1101 and 111? are identical on the two values attended to by the subject (*colour* and *form*) and different by only one other value (*size*). 1101 is also highly similar to 111?, the model therefore predicts that this would be the next most likely exemplar to be retrieved.

Notice that the generalisation necessary here to extend category membership from a set of patterns to a new pattern does not involve the creation of a prototype. The authors then review the evidence that has been presented in favour of classification based on the creation of a prototype during learning and give alternative explanations. They also carry out four experiments with carefully designed stimuli in order to generate contrasting predictions for the two alternative views: prototype theories and the context model. According to the prototype models, a pattern that is closer to the central tendency for a class should be classified faster and with fewer mistakes than a pattern that is farther away from the central tendency. According to the context model, faster and more accurate classification is explained by the number of patterns displaying very high similarity with the new pattern. Stimuli were designed that created conflict between these two possibilities and the context model's predictions were more often verified.

The claims made by Medin and Schaffer mark an important moment in the development of this new field of study. Only fourteen years after the heyday of the definitional approach (Katz and Postal, 1964), not only are they denying that natural concepts can be defined by necessary and sufficient

[104] Since the subject did not necessarily attend to *size* and *number* the representations in the model replace this data with question marks.

conditions, they also challenge the novel and very well received notion of a category prototype. They call for a more parsimonious model where *the same mental representation created at the time of the event* (the memory 'trace' of the event), and not necessarily an abstract prototype, can be called upon to solve a new classification task. From their publication on, this approach to classification is called 'exemplar theory', or, less often, 'instance theory of classification'. The individual representations or memory traces of training stimuli are referred to as exemplars. I adopt this terminology from now on. Medin and Schaffer's (1978) article was widely read and has generated much debate. The next two sections present two of the many responses. First, there is some supporting evidence from Douglas Hintzman and Genevieve Ludlam (1980). As a top researcher in the field of memory, Douglas Hintzman is well placed to note both the consequences of the exemplar-context model on the field of memory and to assess what memory research can contribute to the debate on concepts and categorisation. In the second article, Donald Homa, Sharon Sterling and Lawrence Trepel (1981) challenge both Medin and Schaffer (1978) and Hintzman and Ludlam (1980).

4.3.2.2 Hintzman and Ludlam

Among the most persuasive evidence in favour of prototypes is differential forgetting. That is, the observation that performance in classifying the prototype of a category seems to suffer less from the effects of a retention interval than performance on the exemplars themselves. This was the main result of Posner and Keele (1970), which I announced would be reinterpreted by subsequent research. The clear objective of Douglas Hintzman and Genevieve Ludlam (1980) is, in fact, to argue that Posner and Keele's results do not necessarily lead to the postulation of prototypes. The alternative explanation they offer to differential forgetting rather supports exemplar models and further develops the consequences for theories of memory. The authors create a computer model that both records only exemplars, instead of exemplars plus prototypes and can simulate differential forgetting. The model clearly belongs to a particular class of

exemplar models, called 'multiple-trace memory models' which hold that each experience gives rise to its own memory trace. Experiences are configurations of primitive elements or properties (not only features but also relations). For example, if the stimuli are a yellow triangle and a blue square and the yellow triangle is bigger than the blue square, the features are *yellow, blue, triangle* and *square*. The relation in this case is *bigger than/smaller than*. Both new stimuli and memory traces are represented as property and relation strengths; in the examples handled by Hintzman and Ludlam, these have arbitrarily been set at 1 for properties and .5 for relations. The above stimulus would be coded as follows: number of objects = 2 (*e* and *f*); properties of object *e*: *colour* = yellow (1); *shape* = triangle (1); *relation* = bigger than (.5). Properties of object *f*: *colour* = blue (1); *shape* = square (1); *relation* = smaller than (.5). Furthermore, supposing that this stimulus is old and has already been categorised, the representation also includes category membership = A. Part of the elegant simplicity of this model is that the memory trace corresponding to a particular stimulus is simply the same as the description of the stimulus. One is simply a copy of the other. When a new stimulus is fed into such a system, it is assumed that it would simultaneously be matched for similarity with all of the other traces in memory. The model employs a modified version of Tversky's (1977) formula. The exact details of how similarity is calculated are too technical for this discussion, suffice it to say, the computation proceeds not only according to shared properties but also to their relative configurations. This set up causes retrieval to function in one of two ways: either retrieval pinpoints the single most relevant 'memory trace', or, alternatively, not necessarily a single trace but rather all those, *and only those*, stimuli that are relevant (i.e., that scored positive with Tversky's formula) to activate for the correct categorisation of a new stimulus. These two procedures, which represent the two alternative explanations, exemplar and prototype theory, are vital to Hintzman and Ludlam's demonstration. But first, the simulations must be set up as follows: start out with two prototypes (or *patterns)* that can be described as the configurations of 7 properties. From these, generate 7 exemplars by changing one of the properties or relations at a time. The

complete set of stimuli comprises 14 exemplars and two prototypes. Take out the prototypes and two of the exemplars and let the remaining set represent what the subject has in memory. The configurations are each marked with a category corresponding to the prototype that generated them, say A and B. This method is repeated to create 7 sets of stimuli. Forgetting is simulated by deleting one of the values at random. Testing begins before the first cycle of forgetting and is repeated after one and then after three cycles. The complete set of 16 stimuli are tested each time, the results of one such trial are pictured in figure 2. [105]

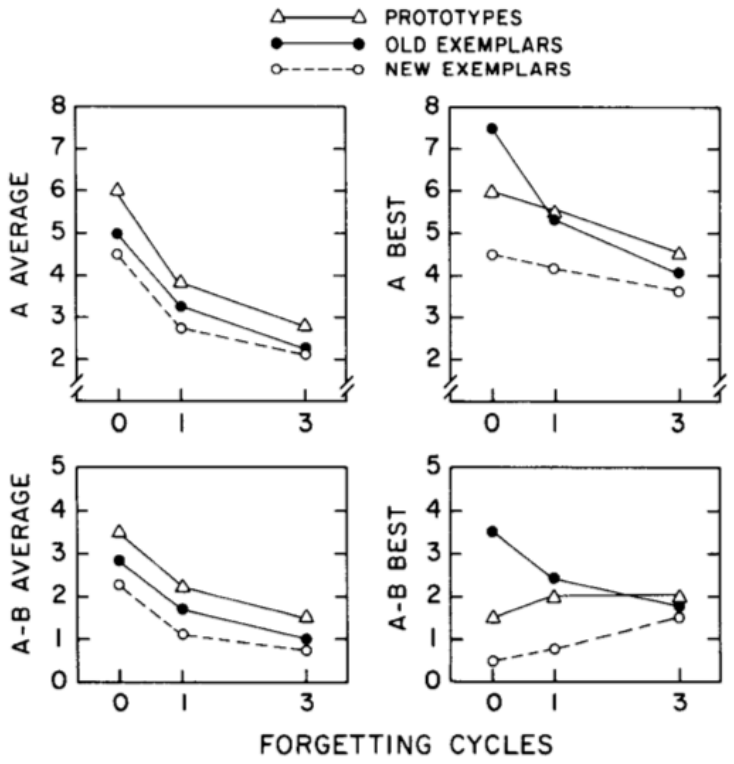

Figure 2: 'Differential forgetting' Hintzman and Ludlam (1980).

The four graphs represent the four possible measuring procedures. On the left, classification is based on average matches, that is, the average between

[105] Hintzman and Ludlam (1980), Differential forgetting of prototypes and old instances - simulation by an exemplar-based classification model, figure 1, page 380. Reproduced with permission from Douglas Hintzman.

all the traces with positive A values and on the bottom average A values minus average B values. These measures show consistently more accurate classification of prototypes than of exemplars despite cycles of forgetting. For Hintzman and Ludlam, the reason is obvious: prototypes here have a statistical advantage. They are similar to all the traces, so most similar to the average of these traces. Meanwhile, an old exemplar is perfectly, or, after forgetting, quite similar to itself but this contributes little to an average.

When retrieval pinpoints only the most relevant exemplars for identifying a new stimulus, two options are possible. Either only one trace that is the most similar to the new stimulus is activated (in the example from figure 2, that happens to be a trace marked A) or both the best A and the best B are activated and B is subtracted from A. In figure 2, both of these options clearly show that old exemplars have an initial advantage which they lose as forgetting cycles intervene. Performance on prototypes is comparatively stable and, at the end, can exceed performance on old exemplars. Again, this is due to the fact that the prototype has a statistical advantage. Being identical to one exemplar is only helpful when the information is intact; as soon as information begins to be lost, it is better to be similar to 6 exemplars than identical to only one (Hintzman and Ludlam, 1980: 381).

What we can conclude from this is two-fold. First of all, if an experiment is set up assuming that subjects average the information out, then the results will most likely support prototype perspectives (like the two graphs on the left in figure 2). But, if this assumption is dropped, and a 'best-match' technique of measuring accuracy is adopted, the results change drastically. If this is possible, reason Hintzman and Ludlam, then the evidence brought forward in favour of prototypes (e.g., Homa et al. 1973; Posner and Keele, 1970) cannot be taken as conclusive proof that subjects create the prototype they use for subsequent tasks during the learning phase. Second, the results using the best-match measurements offer added support to the context model since that model posits nothing other than exemplars and can simulate differential forgetting. Differential forgetting can thus no longer be considered unequivocal proof that classification

learning is based on prototypes. Nor do the results suffice as a demonstration that prototype theory is wrong. Rather, as the authors say, so far not enough evidence in favour of it has been found.

4.3.3 The 'Dual' Model (Prototypes and Exemplars)

The results of Hintzman and Ludlam did not stand unchallenged for long. In 1981, Donald Homa, Sharon Sterling and Lawrence Trepel published a long article carefully outlining and criticising the methodology and interpretations of Hintzman and Ludlam (1980). The first point of criticism for the exemplar theory (or, context theory) of classification is that the categories it uses in its experimental designs are not ill-defined categories. This is an important point, first of all, because research into ill-defined categories has included various manipulations of the stimuli that it would be hard to imitate with well-defined categories (Homa, Sterling and Trepel, 1981: 419). In ill-defined categories, the number of instances is potentially infinite; compare this to Hintzman and Ludlam's categories where the population of a category was limited to 6 members and the complete set of stimuli consisted of only 16 units. Furthermore, the dimensions underlying ill-defined categories are, by definition, unknown or at least imperfectly known. The categories in Hintzman and Ludlam were comprehensively describable by the values on 7 dual relational properties. Finally, ill-defined categories probably reflect real-life experience with learning better than artificially simple categories. These simplifications were, of course, a methodological necessity since the objective of the article was to contrast predictions by contrasting measurements using Tversky's (1977) formula which requires defined feature sets. However, according to Homa, Sterling and Trepel, they are far from trivial because they obscure learning variables that are important in the explanation of categorical knowledge, or, in their terms, of 'category-level summary representations'. In previous research, Homa and his colleagues have demonstrated that if the number of members in a category is increased, performance on immediate and delayed transfer is improved. The number of categories is also of importance since when more categories are learned, generalisation to new stimuli is improved.

Finally increased variance within a larger category can also lead to improved generalisations. These criticisms are well taken, but, to the extent that they only touch the methodological aspects of the research supporting exemplar models, they are also limited.

According to exemplar theorists, their contribution is to show that individual instances are not discarded as previously thought. They do not hold that higher-level generalisations are never created, rather simply that prototypes are not the whole story. To bring this point home with a demonstration of force, they have designed models that can do what prototype models do *without the prototypes*. These demonstrations are not intended to offer psychologically plausible processes of category learning and generalisation; rather, they seem to be especially designed to counteract perspectives relying too exclusively on prototypes.

In fact, Homa, Sterling and Trepel's own experiments use statistically distorted forms and manipulate learning variables yet their results *parallel* those of exemplar theorists in that they show that old information is available for subsequent tasks rather than lost. For example, in the experiments in Homa and Vosburgh (1976), performance accuracy on the old learning stimuli exceeded that of the new patterns at the same level of distortion by 10% to 20% after 10 weeks. This means that old patterns played a role when it came to transferring what had been learned about a category 10 weeks before to a current categorization task (cited in Homa, Sterling and Trepel, 1981: 419). However, their main objective is still to point out the limits of exemplar-based generalisations and the advantages of postulating the abstract category-level summary representations. They adopt the 'mixed' or 'dual' model that calls for three sources of information: a prototype (created during learning), specific exemplar information, and the breadth of the category (i.e., the number of instances in the category). They predict that the subject's representation of a category is only predominantly made up of particular exemplar information if that category was represented with only a few patterns during learning. But, as soon as experience accumulates, the representation would become dominated by a prototype together with information about the breadth of the category.

Additionally, time would also favour the dominance of a prototype in the representation of a category since differential forgetting is supposed to affect exemplar information to a greater degree than it does prototype information.

So, it seems that despite finding evidence in favour of the availability of particular exemplar information at the time of transfer, Homa, Sterling and Trepel still consider that the most reliable heuristic for categorization is the creation of a prototype during learning. Yet, insofar as they recognise the presence of exemplar information at the time of transfer and argue for a mixed model, their results support my view that ultimately a better understanding of categorisation involves bringing contributions from different and *even opposed* perspectives together. A complete theory of categorisation will offer explanations of prototype effects *and exemplar* effects, among other phenomena I have yet to present in this chapter. Before moving on, however, a brief summary and discussion is needed to set the scene for what follows.

From a certain point of view, prototype and exemplar models have much in common. Both developed as alternatives to a failing classical account. Both postulate that mental representations are created as a direct consequence of a subject's exposure to different stimuli in the flow of experience and both assume that responses to further stimuli are a function of their similarity to stored representations or exemplars. There are, nonetheless, some important differences between prototype and exemplar approaches. The most significant difference, in my view, is that the former can still be grouped with the classical account in assuming that the representations built up from the instances encountered are *fundamentally* different from the instances themselves (Posner & Keele, 1968: 353). As expounded by Pylyshyn (1973), the process of interpretation was believed to *abstract away from the particular details of specific stimuli by creating type-token pairings* that would be stored as sentence-like propositions. In prototype theory, the creation of an abstract prototype proceeds in much the same way. The idea that the prototype represented a category because it brought

together the features of the most central or 'typical' members was quickly adopted. As illustrated with the accounts presented above, it was assumed that subjects subsequently based their categorising behaviour on similarity to this abstract prototype. This, in conjunction with the belief that stable prototypes were theoretically preferable to variable ones, resulted in a certain repetition of the mistakes of the classical account: instead of acknowledging that perhaps there are *no stable conditions of application* and looking for an explanation of the relative regularity of our behaviour elsewhere, theorists held on to pre-existing conditions of application and conceded only that, instead of being *necessary,* they were merely *statistically prevalent.*[106]

The alternative, that is, acknowledging that, to a certain degree, conditions of application, or norms, vary with the contexts and purposes in which they arise, was banned as a sort of relativism. Notice that the assumption that Homa, Sterling and Trepel were reluctant to accept was that their subjects' ability to correctly classify a stimulus did not depend on underlying summary representations. Notice the similarity in the explanation of categorisation between, on the one hand, necessary and sufficient conditions as in the classical definitional account and, on the other hand, a statistical prototype. What sets exemplar models apart is that, instead of consulting anything like a pre-computed, pre-existing norm or prototype, the task at hand, whether it be constructing an occasion-specific word meaning, or any other, triggers a retrieval mechanism that scans and summarises online the relevant exemplars. If a framework of evaluation, or any ad hoc structure is necessary, it is through interaction between the probe, or 'cue', and the information in memory that it is generated.

Compared to Pylyshyn's early view of mental representations, where the reduction is maximal since all sensory events must be classified into a 'finite, and even relatively small, number of descriptive propositions'

---

[106] Regarding the stability of prototypes, in the early stages of research, it was firmly believed that prototypes were widely shared before further research uncovered cultural and contextual variability. I return in length to this topic in section 4.4 of this chapter on Lawrence Barsalou's contribution to concepts and categorisation.

(Pylyshyn, 1973: 7), representations in exemplar models are only minimally reduced. I follow exemplar theorists in underlining this point as critical for a system bound to face considerable variation. In fact, failure to provide clear-cut definitions for a multitude of everyday concepts had just raised awareness of the true diversity of our experience. Designing a system that could cope with the diversity and unequal distribution of properties plus truly novel experiences was among the top motivations for developing radically new views on mental representation and categorisation. And I believe that significant progress *has been made* on this issue. As I leave the discussion here it seems that opinions are still divided. Exemplar theorists find it a more compelling choice to 'scan and summarise' exemplars when the task is already at hand; they construe retrieval as a dynamic process that benefits from being driven by a particular task and the context at hand to selectively activate only those representations relevant to the task at hand, whatever it is. Given the evidence, this seems more psychologically plausible than to believe that an abstract prototype can be created as the input is first being processed, and that this prototype need only be 'retrieved' ready-made to accomplish the task at hand. The seeming lack of consensus however is in part due to the fact that I only presented a small fragment of the research on these issues. I concentrated on landmark articles that defined the debate *in its early moments*. During the period I covered, the field was just starting and it is still now very young. In chapter 5, I will draw evidence from memory research into the general debate on the issue of when abstract or summary representations are created. In the next section, I present another part of the story on categorization research born from the aforementioned need to take ever widening quantities of data and diverse perspectives into consideration. As announced in the introduction, a fault that prototype and exemplar theories share is an overreliance on superficial, perceptual features; in the subsection below, I present proposals from various psychologists, and one evolutionary biologist, to remedy this situation.

4.3.4 Psychological Essentialism

An alternative way of interpreting the failure of the classical definitional approach is by pointing out its dependence on essentialism. Roughly, *metaphysical* essentialism, is a philosophical doctrine claiming that individuals instantiate *essential* properties that are logically prior to them. These properties are *essential* in that they are independent of *and pre-exist* that which makes them known to us. In this view, correctly classifying the world into kinds depends on recognising which sets of properties are necessary of which kinds. In the introductory chapter (section 1.2), I briefly presented the standard 'classical theory of concepts' view that essential properties are individually necessary and jointly sufficient to identify members of a class. My critique of the classical account *then* was that these conditions were not really playing the roles postulated for them since categorising, and in general, competently using concepts, seems not to depend on having anything like necessary and sufficient conditions. In this section, I am more specifically interested in the issue of essences themselves, independently of whether definitions capture them.

The question is whether the objects and events we group together in classes have common essences. There are strong arguments in favour of the definite and complete abandonment of this idea but the debate is very old and the trenches deep. Few venture to launch an all-out attack, Richard Dawkins, however, one of the most prominent voices in the scientific community today, is an exception;[107] he has recently suggested in this year's *Edge* forum that *essentialism* is *the* scientific idea ready for retirement.[108] He calls it 'the tyranny of the discontinuous mind' and following Ernst Mayr, blames it, among many other things, for humanity's late discovery of

[107] He is not the only exception; Paul Bloom has also taken up the topic of our mistaken construal of essentialism in a recent book (2010) and made an important contribution to psychological essentialism of artifact concepts in an article (1996).

[108] Every year, the science and technology think tank *Edge* asks its members a different general question and contributions are posted on *Edge.org*. The forum's 2014 question is 'What scientific idea is ready for retirement?'; Richard Dawkins, who generally advocates a more insightful conception of kinds that shrugs off Platonic ideal forms, takes the opportunity to summarise his arguments for the retirement of essentialism.

evolution. In a nutshell, holding essentialism puts us at risk of dismissing the gradual and the variable.[109] As an alternative to essentialism, prototype theory proposed probabilistically verifiable feature correlations instead of essential properties and replaced necessary and sufficient conditions with an empirically observed organisational principle (family resemblance). But, according to some influential voices in the categorisation community (Murphy and Medin, 1985; Medin and Ortony, 1989; Barsalou, 1987; among others), this is not enough. The criticisms centre round the worry that probabilistic theories of loosely correlated features are too unconstrained to give a full picture of how properties come together in our mental representations; this objection deserves our full attention since failing to answer it can inadvertently lead to readopting the assumptions of metaphysical essentialism. In the remainder of this subsection, I argue that the best way to escape the mistakes of metaphysical essentialism, particularly in the field of language studies, is to replace it with *psychological* constraints, or '*psychological* essentialism'.

The contributions leading to the proposal of *psychological* essentialism, begin with a certain dissatisfaction among some psychologists with the accounts presented in previous subsections (from Rosch's 'prototype theory' to the 'dual' model that incorporates prototypes and exemplars) and their tendency to rely almost exclusively on surface features. Gregory Murphy and Douglas Medin's (1985) influential article identifies the main common flaw of accounts presented to that date as their overreliance on similarity among perceptual features. According to Murphy and Medin, similarity-based accounts of how features are correlated have

---

[109] Notice the parallel between ignoring the gradual and the variable in our general experience of the world and ignoring the gradual and the variable in word meaning in context. Another way of phrasing the belief that words have 'core' meanings is to say that a word's basic meaning captures the core *essence* of the object or event that the word is used to refer to. Essentialism is so pervasive, such an integral part of our thinking, that it appears even amongst contextualists. Searle (chapter 3, § 3.4.2), for instance, despite his insistence on context-dependence, considers that there must be something all the instances of 'open' share: a 'core' meaning which is conditioned to an actual application but nevertheless exists. This section addresses this point; but in anticipation, I can say that there does not have to be something all instances of 'open' share so long as users *believe* that there is, or so long as it makes sense to users that there is, or might be.

not really addressed the question of whether it actually is *similarity* which underlies the grouping of entities together in a category, or whether it is *belonging to a category* that makes entities seem similar. In other words, similarity could be a by-product of categorisation, rather than the force determining it. Importantly, however, they do not argue for abandoning similarity-based accounts. Rather, they explore the possibilities of complementing them with an account of what makes concepts and categories 'coherent'. By coherent categories, Murphy and Medin mean groupings of objects that 'make sense to the perceiver' and one of the principles they suggest can underlie *making sense of categories* is the notion of a 'theory'.

Basically, these authors hold that 'theories', or, rather, small sets of more or less loosely integrated *beliefs*, are the glue that holds some concepts together, giving them coherence, while other groupings seem improbable because they are not backed by any such theory. In their view, accounts to that date had focused on treating concepts as collections or correlations of features and, while they appreciate the improvement embodied in models that amend the traditional account of concepts and progressively take more into consideration (as shown in the progression presented in this chapter), they hold that these accounts generally exclude *theoretical* connections, despite being broad enough to include them.

> When we say that concepts are organized by theories, we use theory to mean any of a host of mental 'explanations' rather than complete, organized, scientific accounts (Murphy & Medin, 1985: 290).

As an illustration, consider dietary rules: what makes the class *FOOD* cohere? An ancient biblical tradition proposed a distinction between 'clean' and 'unclean' animals: the *Leviticus* lists criteria for deciding whether different creatures are to be regarded as 'clean' or 'unclean' and therefore not fit to be eaten. The criteria among livestock, for instance, are a parted hoof and 'that they chew the cud'; so, sheep and goats are clean but pigs are not. Most flying insects are regarded as unclean; but, surprisingly, there are some exceptions: locusts, katydids, crickets and grasshoppers are among the

'clean' animals disciples were allowed to eat. We can only wonder what the logic behind these distinctions could be, in Murphy and Medin's terms, what makes these classes 'cohere'? A first reaction might be to consider the distinctions in *Leviticus* simply arbitrary, although for some, of course, they might be dogma. Yet, if we give it more thought, an intuition appears that the distinctions and rules were probably guided by some unstated theory, perhaps health concerns. Notice that not knowing what differences in beliefs explain the differences in what is considered edible across cultures does not stop us from assuming that in the UK we are justified in accepting as edible certain creatures (e.g., pork) and equally justified in rejecting others (e.g., 'chapulines', a kind of cricket eaten in parts of Mexico).

According to the authors, any intra- or inter-conceptual relations can be included in the theoretical connections that afford coherence. The scripts we use in everyday interactions with the world, can also be included; they embody causal knowledge, implicit theories of causes and effects that underlie our understanding of the concepts involved. For instance, it might be more important to our concept of GOLD that *we consider it rare and (therefore) precious*, *that it is often used to craft jewellery, and (therefore) an appropriate material for an engagement ring* than the fact that *an atom of gold always has 79 protons*.

> The keystone of our explanation is that people's theories of the world embody conceptual knowledge and that their conceptual organization is partly represented in their theories (Murphy & Medin, 1985: 280-290).

As Murphy and Medin, concede, this proposal is only a first step in integrating world knowledge back into our psychological theories of concepts and categories. Another contribution, presented below, is pivotal in explaining how *beliefs about kinds* can replace *metaphysical* essentialism.

A very concrete contribution in the same direction as that of Murphy and Medin came in the form of Frank Keil's influential 1989 book *Concepts, Kinds and Cognitive Development*. Among other things, Keil focuses on studying the role played by factors *other than surface perceptual features* in

categorisation and reasoning about categories. He designed an ingenious experiment with children and adults: first, he came up with pairs of natural animal kinds for which a description could be given that would mention all the surface perceptual features and behavioural characteristics of one animal and say that it has 'the insides' of the other.[110] These descriptions were then read to the subjects. To illustrate, I have abridged the description given of the horse with cow insides:

> *These animals live on a farm, people put saddles on them, they eat oats and everybody calls them horses. But some scientists went to this farm to take a closer look, they did some blood tests, took some X-rays, and looked deep inside the animals with microscopes. They found that these animals had the inside parts of cows. They had cow blood and cow bones. Their parents were cows and their babies were cows (*Keil, 1989: 162).

As they heard these stories, pictures of a horse and a cow were on display to aide in explaining that the animal looked like the one but had the insides of the other. Finally, the children and adults were asked (one by one) what they thought these animals were; and a follow up question asked them to explain themselves. If the child said that the animal was a horse, the experimenter might go back to the description and repeat the part about the animal having cow parents and babies, for instance, to make sure that these details were being considered. The article offers some example transcripts of how these conversations went and an account of the precautions taken when designing the materials (these included interviews designed to identify the best possible wording for the descriptions). The results are quite straightforward: kindergarteners mostly assumed that the scientists' discoveries were irrelevant to what kind the animal in the picture belonged to. The majority of older children and the adults, however, did find the discoveries relevant to what the animal that looked like a horse *really* was (Keil, 1989: 168). This suggests that not only children of a certain age but adults in general categorise objects in terms of underlying *beliefs about*

[110] Keil (1989) also investigated plant kinds and artefact kinds; I focus here on animals kinds for brevity and because they produced the clearest results.

kinds. Notice that in this case, as in most cases, the beliefs underlying these judgements are no more than *folk* theories of biology.

This brings us back to Hilary Putnam's externalist semantics. There are several points of convergence between Keil's findings and Putnam's proposal: notice that although kindergartners are overly reliant on superficial features and do not yet participate in the division of linguistic labour, *as soon as they are more grown-up*, they join the adults in finding discoveries made by other people relevant to their own categorisation behaviour. Much could be said here about how this is relevant to children's language development, but the important point for my purposes is that, at some point, trusting others and delegating on issues of meaning become indispensable in achieving competence as a communicator (see Sperber, at al. 2010). Another point of convergence is that for both Keil and Putnam, meaning is much more complex than what is 'in the head' of the speaker. Speakers have folk theories, or ideals associated with types, but these folk theories or ideals are not *necessarily* what defines the types *scientifically*. In the example above, for instance, it seems that a certain *belief about natural kinds*, namely, to what type of animal the bones and blood inside the pictured animal would normally belong, can influence how that animal, inside and out, is classified. Similarly, consider that it was very important for the older children and adults of Keil's experiments that the animals in question had cow parents and cow offspring.

Perhaps we could say that older children and adults take these to *be essential* properties of being a certain type of animal? But rather than confirming *metaphysical* essentialism*,* this result only confirms that we humans have a very strong tendency to *act as if* natural kinds had essential properties, and this is an argument for *psychological* essentialism. Going back to Dawkins, the idea that cows have cow parents and cow babies, confronted with a specific branch of science, namely evolutionary biology, needs to be hedged. From one generation to another, this is absolutely true, only animals of the same species can mate and have offspring that can themselves mate. But change must be possible otherwise we, Homo sapiens would not have evolved from Homo erectus. Dawkins (2011a) illustrates the

gradualness of change with a thought experiment: imagine you have a picture of each and every one of your ancestors going back to your 185-million-great-grandfather (who, by the way, was a fish). You arrange these pictures on a 3 mile-long bookshelf. As you walk along the bookshelf back in time, any one member of your chain is of the same species as its neighbours, You would need to walk a hundred thousand years back, for instance, to start to see a slight difference in appearance between yourself and this ancestor; and you would probably still call this 4,000-greats-grandfather a 'man'. Pushing back to your 50,000-greats-grandfather, the differences are enough to count as a different species: Homo erectus; whether you still want to call him a 'man' or not, is, as they say, a question of semantics. Furthermore, consider that there would be as little difference between you and this 'man' as there would be between him and his 50,000-greats-grandfather. This can go on for quite some time, with very slight differences building up. At 250,000 generations, your ancestor might look a bit like a chimpanzee: actually, he would be the common ancestor we share with chimpanzees. Again, if we look on either side of this ancestor's photo, we see indistinguishably-like animals, for tens of thousands of years in each direction (Dawkins, 2011a, chapter 2).

What does this have to do with concepts and word meaning? As mentioned in the section on the traditional view of concepts (chapter 2, section 2.2), which set the foundations for the definitional account, held that a class, such as 'man', or 'Homo sapiens', possesses *defining essences* that make the members of that class *fundamentally different* from those of neighbouring classes. But what Dawkins's thought experiment shows is that a *flaw* in our thinking, he calls it 'the tyranny of the discontinuous mind' blinds us to the continuous and to intermediates; in fact, there is no single point on the imagined list of ancestors at which one of your ancestors suddenly acquires the essence of manhood, there is no point at which a Homo erectus gives birth to a Homo sapiens, and so; once more, the assumption that there are dividing lines, absolute definitions and *defining essences* is challenged. As discussed in chapter 2, among certain philosophers, one of the main functions concepts have been called to serve

is as the basis of metaphysical claims (Rey, 1983: 243, section 1.3). Expecting concepts to serve this function is justified by the belief that *although the metaphysical/epistemological distinction is 'not everywhere perfectly sharp', there* is *a sense in which whether something actually is out there in the world is different from whether anyone knows that it is*. It is not difficult to grant that there is a difference between something actually being out there in the world and anyone knowing whether it is; the issue, I would argue, is whether, in light of Dawkins's remarks regarding essentialism, a *concept* can support a metaphysical claim. According to Rey, if it's a natural kind term, such as COW, 'a characterisation of the 'universal' cow, or of the essence of COW' would serve to distinguish the conditions by virtue of which something *yes-or-no is* a cow from how we can tell that something is a cow; likewise for 'man'. According to Dawkins, however, this type of metaphysical essentialism is deeply flawed, reality is not black-and-white and we would be far better able to comprehend the world around us if we accepted 'life's grey areas' (2011b: 54).

If we are willing to rethink the 'metaphysical' function concepts are supposed to serve, we could explore basing this function on something else, not on concepts *tout court* but perhaps on scientific theories, although we would be forced to accept that they too evolve. Alternatively, we could adopt an entirely different explanation for what is happening when we use our concepts (or the words that express them) to make claims about what is out there in the world. Psychological essentialism suggests that we decide or learn to use the word 'man' or 'cow' and *assume* that we have good reason for doing so. If asked to justify our categorisation, we cite our folk-theories of biology, or evolution, or whatever. *Metaphysical* essentialism is probably not true, but what Dawkins deplores as a tyranny could perhaps be what explains our strong tendency to behave *as if* concepts had essences, and, *as if* words had 'core' or context-independent meanings. In other words, psychological essentialism could be the key to explaining how we manage to converge on what words in context mean in the absence of fixed and predetermined word meanings.

Also, as already mentioned more than once, people are very willing to accept that they do not have defining conditions for the terms they use, that even experts can get it wrong and, perhaps in response to this, adopt a flexible strategy that allows for knowledge to be *revisable*. I mentioned then that theorising on this strategy has since led to the postulation of *psychological* essentialism, In other words, there is a *psychological* explanation for why we believe that there really are pure (Platonic) forms or essences instantiated in the objects and events around us: we are under the spell of an *inbuilt* bias. What's more, for scientists like Dawkins or psychologists interested in these issues, for our behaviour to be explained, essences do not really need to exist, we just need to *believe*, and behave *as if* they did. Psychological essentialism thus postulates an innate, and (sometimes) difficult to resist, tendency to see the world as neatly cut up along dividing lines.

A complete exploration of the ways in which psychological essentialism extends into theories of word meaning in context would be beyond the scope of this thesis; however, I hope to have at least made the point that it lends support to the radical contextualist idea that words do not need core, basic, context-independent meanings. It chips away at the idea that there is necessarily something that all instances of 'open' share, and, at the same time, offers a plausible, *psychological* explanation for the intuition that there might be, *or should be* something in common. The alternative offered is that perhaps what instances of opening share is simply falling under the class of actions that can be *described* as 'opening'. As discussed in chapter 3, however, pinning down what it is about an action that definitely makes it a type of opening action is futile, since the norm or framework of evaluation used for deciding whether something falls under 'opening' is generated in the context in which it is needed.

Finally, this discussion would not be complete without a mention of Medin and Ortony's (1989) introductory essay to a collection of papers on similarity and analogical reasoning which they entitled 'psychological

essentialism', perhaps coining the term for the first time. They begin with a warning:

> There are problems with equating concepts with undifferentiated clusters of properties and with abandoning the idea that category membership may depend on intrinsically important, even if relatively inaccessible, features (Medin and Ortony, 1989: 179).

In the alternative they propose, there is a constraining relation between surface features and deeper properties that are responsible for, or 'generate', the first. This can be illustrated with the example of a comparison between a bear and a whale. There are few if any superficial features to base similarity judgements on, but we can still say that they are similar *in that they are both mammals*. So, according to the authors, avoiding the problems of probabilistic theories of correlated features involves redefining similarity by looking beyond superficial, exclusively perceptual features.

To give credit where credit is due, Tversky's (1977) contrast model already calls for the contextualisation of similarity judgments. As mentioned earlier, he holds that the database we use when carrying out similarity judgments is particularly rich (it includes *anything* that can be deduced from our general knowledge of the world and is *relevant* to the object in question). According to Tversky, 'When faced with a particular task (e.g., identification or similarity assessment) we extract and compile from our database a limited list of relevant features on the basis of which we perform the required task' (Tversky, 1977: 329). It is this particular flexibility of similarity that allows Tversky to theorise on examples such as 'Jamaica is like Cuba, Cuba is like Russia, but Russia is nothing like Jamaica'. For Tversky, it is thus natural to consider that the collection of features that represent the objects are not necessarily surface 'perceptual' features. The similarities between Tversky's model and Kahneman and Miller's norm theory are not a coincidence since Tversky and Kahneman spent the first decades of their careers working together.

4.4 Barsalou's Comprehensive Account of Categorisation

The aim of this section is to bring the contributions of the different theorists discussed in this chapter together under a single perspective and argue for an overall interpretation of the results of categorisation research that is significantly different from the one that has usually been considered. I base most of the discussion on the ideas of Lawrence Barsalou because his work is now *the* key additional component needed to complete the picture of categorisation and to draw out its consequences for theorising on word meaning in context, which is my ultimate concern.

My starting point is Barsalou's insights on the '*instability* of graded structure', but this section also covers his proposal for 'ad hoc categories' and ends with his view of concepts, which is radical by almost any standard. By 'graded structure', Barsalou means the internal structure of a category comprising a central 'best exemplar', or 'clearest case', surrounded by less prototypical members, and grading off from less prototypical cases to nonprototypical cases at the edges of the category. The disagreement between Barsalou and prototype theorists is that they assume that this structure is stable for a given category; Barsalou, on the other hand, sees evidence of 'instability' or, in other words, *variability* and *context-dependence*. Discussion of these topics has appeared in various articles (Barsalou and Sewell, 1984; Barsalou, 1985; Barsalou, Sewell and Ballato, 1986),[111] but culminates in the seminal 1987 article. His first objective in this article is a *general* review of the main results of categorisation research; there is a marked emphasis on prototype theory probably because, as remarked earlier, prototypes were the phenomenon that attracted the most attention and their study dominated the field, for better or for worse. In the section on Rosch's influential contributions to prototype theory, I argued that a constrained *yet positive* interpretation of her findings is possible and that it would involve avoiding the idea that what she proposed could be

[111] Sewell and Ballato (1986) is given as 'under review' in Barsalou (1987) but I can find no trace of a published version. I can only assume that it remained a manuscript. He cites the results in the 1987 paper, and this is what I base my comments in this section on.

taken as, or even developed into, a full account of concepts and categories; rather, I discussed the possibility that what she brought to the attention of wide audiences was a particular *type* of categorising *effect*. Now that the other possible effects have been presented: *exemplar* effects and the effects of *holding (essentialist) beliefs about kinds*, it should be clear that prototypes were only the tip of the iceberg and that it is categorising behaviour *in general* that warrants explanation. Barsalou's great insight is to group the different categorisation effects together to reveal a deeper truth about human cognition: he calls it 'instability', but, as I will argue, it can just as well be explained as *context-dependence*. From this insight, a markedly different perspective emerges on many of the issues I have focused on so far in this thesis: namely, the ubiquity of context-dependence, meaning eliminativism, and, *particularly*, on the possibility of solving the issue of arriving at occasion-specific word meanings using not only pragmatic principles but general reasoning, common sense and their related construction processes.

Before getting into the topic of Barsalou's insights into categories and categorisation however, some words of clarification regarding his terminology are in order. Barsalou's observations centre on the fact that *different* subsets of information from long-term memory represent things-out-there-in-the-world on different occasions. In the terms of norm theory we have adopted, this means that according to Barsalou, when a subject needs to represent something out there in the world, a dog, a vacation or peace, for instance, 'different sets of representations (or information from memory) are selectively activated' depending on the needs of the particular occasion. As with most research in this field, the examples are predominantly of natural kinds, but there are some examples of artefact kinds. Barsalou's position, *uncontroversially, I should think*, is that a natural kind such as DOG is represented in a particular subject's mind *by different subsets* of information from long-term memory *at different times*. A terminological issue arises, however, because, following the tradition in categorisation research, Barsalou (1987) refers to these things-out-there-in-the-world, dogs, vacations, etc. as 'categories' (philosophers in general and some psychologists prefer to call these 'artefact' or 'natural kinds'); finally,

Barsalou sometimes again uses the term 'categories' in reference to *the representations in our minds* when he argues against the idea that there are stable, invariant representations in long-term memory. I would argue that, by keeping the above clarifications in mind, it *is* possible to follow the distinctions Barsalou does make: when he speaks of category *membership*, the issue is whether something actually *is* a member of a kind; on the contrary, any mention of structure (graded or not, variable or invariable) refers to the mental *representation*.

Furthermore, the main claim of the 1987 article is that effects such as prototypes occur because the *same* category (i.e., the same *kind*) is represented in working memory on different occasions by *different concepts*. This is Barsalou's alternative to the idea that *kinds* are represented by fixed and stable 'graded structures' or prototypes.[112] Another possible terminological difficulty is with regard to Barsalou's use of 'concept'. Initially, he follows common practice in his field: a concept is a 'knowledge structure' in the widest possible sense; a concept manages to represent something out-there-in-the-world (whatever it is a concept *of*, or whatever it is *about*) by having some 'encyclopaedic' knowledge or information as its contents (such as knowledge about what the concept refers to, accumulated experience, deductions, and so on). A concept has both a concept label (generally the word form) and associated encyclopaedic knowledge. Barsalou breaks away from the common psychological conception of 'concept' when he claims that the encyclopaedic information associated with a concept, in other words, the contents of the concept file, are *not* unchanged from occasion to occasion. This is the main insight resulting from his in depth analysis in the 1987 article. He challenges the then very prevalent idea that subjects have at their disposal stable, invariable mental representations for concepts and categories. His suggestion is that prototype theorists came to posit *stable* prototypes by mistake, on account of their disregard for the effects of context. As an alternative, he suggests

[112] I come back to Barsalou's technical notion of 'graded structure' in the text below, for now, it is only important to see that he means it to stand in stark contrast with the alleged *all-or-none* category of the classical approach.

that there is no reason to presume that we call forth the *same* representations each time that we think about (or, I would add, *communicate* about) dogs, peace, or going on vacation.

Allowing variability to permeate our notions of categories and concepts instead of assuming stability, puts the results of early categorisation research in a very different light: prototype theorists like Rosch, despite their initial efforts to only cautiously advance interpretations, suggested that categories, (i.e., the mental representations of *kinds*) *possess* a graded structure.[113] In other words, that the ordering of category representativeness that falls from the most typical member to the least by degrees on a gradient is something that belongs *as a fixture* to the category representation. How explicitly Rosch endorses this interpretation is not clear, as she only speaks of 'core meanings' and 'good examples' that 'internally structure' the instances of a category; even if she is open to the possibility that different *conditions* can cause the focal point to move, it is still nonetheless true that, in her work, there is no discussion whatsoever of the factors determining the focal point for a category *at a given time* or whether there are any *effects* of context of use on these representations. Further reduced interpretations of Rosch's data then roughly assumed that for American college students, for instance, a bird is typically a *robin* and that, given *robin* as a focal point, they invariably judge *how good an example* of BIRD another member of the category (such as a *chicken* or an *ostrich*) is depending on its similarity to this unique focal point. Barsalou does not deny that this notion of 'internal structure' represents some advantage with respect to the idea of an all-or-none judgement that makes all members

---

[113]This position can be seen in the following quote:

> Contrary to the assumption that categories are necessarily logical, bounded entities, membership in which is defined by an item's possession of a simple set of criterial features (e.g., Katz and Postal, 1964), Rosch has argued (Rosch, 1973, 1975a, 1975b, 1975c, in press) that many natural categories are continuous and possess an internal structure in which members are ordered according to the degree to which they are judged good examples (typical) of the category (Rosch, Simpson, and Miller, 1976: 491).

Notice here that an emphasis on graded structure seems motivated by the opposition it creates with what a classical account would have predicted.

equivalent because all members possess the same deciding criteria (or, using the terminology of the previous subsection, the same '*defining essence*'), but, as pointed out already, he feels an important factor – *the effects of context* – is wrongly being left out. In Rosch's prototype framework, subjects' typicality judgements and reaction times can be explained as stemming from a phenomenon of representativeness that proves very psychologically real, yet Barsalou cannot help taking issue with the unnecessary *added* assumption that for each category there is only one stable graded structure. The evidence, he notes, points markedly in the opposite direction, in the direction of 'instability'.

Barsalou's (1987) aim is therefore to dispel the idea of stable prototypes, but not because he challenges the observations of prototype theorists across the board; rather, he seeks to highlight certain findings and offer an alternative interpretation to graded structure in particular. His starting point is the idea that a particular graded structure is a category *fixture*. Instead, he suggests, graded structure should be construed as *resulting* from categorising *behaviour*; so, instead of picturing the mental representation of BIRDS as invariably representing *robins* as central, *any* ordering is theoretically possible and it becomes normal to expect different structures in different contexts. This seemingly small change suggested by Barsalou has enormous consequences: when faced with a bird related categorisation task, instead of retrieving a pre-computed, abstract representation of the category *BIRD* from memory, a *highly flexible* and context-sensitive process selectively retrieves information from memory to *construct* an ad hoc category *BIRD* in working memory. This proposal clearly belongs to a larger group of 'exemplar theories' and particularly to a group I propose to call 'exemplar-based norm theory' for short; Kahneman and Miller's (1986) and Barsalou's (1987) articles were written more or less at the same time and they cite one another as sources of evidence and support for their proposals; so, once more, the compatibilities are not

coincidental.[114] Barsalou's additional insight is that, since human beings are constantly trying to achieve goals, we should consider that among the constructions they arrive at with their highly flexible cognitive systems, there are not only ordinary taxonomic categories like BIRD, arguably well established in memory, and, incidentally, those predominantly favoured in laboratory settings, but also 'ad hoc categories' that, by definition, need to be created to serve specific needs at hand, like *THINGS TO EAT ON A DIET* or *THINGS TO TAKE ON A VACATION*. The notion of ad hoc categories is particularly important in bringing back to theorising on categories what initial interpretations of prototypes had inadvertently left out. In the remainder of this section, I address ad hoc categories in detail, I then focus on Barsalou's alternative explanation of graded structure and finish with a presentation of what Barsalou's contributions mean for the notion of concept.

One of the most important of Barsalou's multiple contributions to research on categorisation is his notion of ad hoc categories. Take, for instance *THINGS TO EAT ON A DIET*; this is not only an ad hoc category insofar as it is not, *for most people*, well established in memory, but also a 'goal-derived category', this means that the kind of information selectively retrieved from memory to create this category can include an 'ideal'; or, in other words, a particular *property* that exemplars of this category should ideally meet (e.g., *zero calories*). Zero-calorie foods are unlikely to be the most representative of the foods people eat on a diet, so *zero calories* is not a 'central tendency'; rather, it is what is associated with the goal the category is created to facilitate (i.e., lose weight) (Barsalou, 1987: 105). Although Barsalou does not explicitly mention this, it is important to add that the subject creating this ad hoc goal-derived category is clearly using not only her general reasoning and common sense but also her folk theories of dieting and nutrition. Similarly, we have seen that cultural customs are also among the considerations subjects use to create categories, as in the example of what different

[114] Kahneman and Miller (1986) also cite Hintzman and Ludlam (1980) and Hintzman (1986) which was then *in press*. The compatibilities between these proposals will be further explored in chapter 5 on memory.

populations might consider qualifies as *EDIBLE*. Finally, notice a further similarity between Barsalou's ad hoc categories and the phenomena discussed in the previous section on psychological essentialism: cows that look like horses are never very good exemplars of the category *COW*, but some of the subjects in these experiments could use their general reasoning to override superficial similarity values in order to concentrate on what they believed makes an animal the animal that it is, and as a result, certain very atypical cows, *that looked like HORSES*, were labelled as *COWS*.

Another important factor in describing the general phenomenon of ad hoc category construction is that an ad hoc category might be created *once*, for a particular purpose, and then discarded, or it might prove very useful and through repetitive instances of creation, reactivation and use become part of a subject's repertoire. For instance, once Keil's experiment is over, his subjects are unlikely to ever again categorise what looks like a horse as a *COW* on considerations of the animal's insides, its parents and offspring, or the opinion of scientists. *THINGS TO EAT ON A DIET*, on the other hand, might prove very useful, especially if the first diet is unsuccessful; the ad hoc category is not *necessarily* lost once it has been used, it can itself be stored in memory and become part of the information that is selectively retrieved from memory on a subsequent *particular* occasion, when the dieter is, say, forced to choose something from a menu in a restaurant. The main point is that cows, things to eat on a diet (or anything else out-there-in-the-world) are represented in a subject's mind by *different subsets* of information from long-term memory on *different occasions*; categorising behaviour does not depend on pre-computed category representations, rather it *generates* category knowledge 'on the fly' and the variety of information going into this process seems unbounded, it can be anything in a subject's prior experience that proves relevant to that subject's current context.[115]

---

[115] With respect to whether this conclusion contradicts or complements Rosch's contributions, I think it is important to keep in mind that it does not interfere with three of the most basic claims: (i) that subjects find rating members of a category on exemplariness a 'meaningful' task, (ii) that subjects often agree with each other, and, finally, (iii) that their

Barsalou's other major contribution is his focus on the variety of factors that make graded structure 'unstable', or, in other words, the diverse *effects of context*. For instance, he points out that linguistic contexts have an effect on which member of a category subjects will find as 'more representative'. Barsalou cites the studies by Emilie Roth and Edward Shoben (1983) showing that varying the *linguistic* contexts in which an expression appears results in different orderings of members within a category. For instance, if 'animal' is processed in the context of 'riding', then *horse* and *mule* are likely to be judged more typical. In an inventive experiment Roth and Shoben collected response times for a task in which subjects were given the phrase 'Stacey went to milk the animal' and asked to judge as fast as they could whether 'goat', 'cow', or 'bull', for instance, were members of the category *ANIMAL*. Faster response times for 'cow' than for 'goat' were taken to mean that representativeness correlated with processing making the fact that these members were more representative of their class *in these particular situations* a psychologically real and significant phenomenon (Roth and Shoben, 1983: 363-365).

Another possible factor that invites the effects of context is 'point of view'. Consider that in Rosch's experiments, robins were predominantly central, but that it must be taken into consideration that she asked a homogeneous group of people the same question in the same setting. Asian undergraduate students might disagree with American undergraduates on what is representative of BIRD, because the birds they regularly encounter are not the same. To take this further, Barsalou and Sewell (1984) conducted experiments in which university undergraduates, graduates and faculty were asked to take their own point of view when generating graded structures; then, another set of undergraduates, graduates, and faculty were asked to take the other groups' point of view when generating graded

---

ratings correlate with speed on category verification tasks (Rosch, 1973). Barsalou's contribution is to reintroduce elements that had been inadvertently left out and, taking a step back, point to a possible *global* interpretation that shows categorisation to be a dynamic and adaptive cognitive behaviour.

structures for the same classes. When taking their own point of view, substantial differences appeared between ratings given, for instance, by undergraduates and faculty. Surprisingly, however, when undergraduates were asked to generate graded structures from the faculty's point of view, typicality ratings *matched* those the faculty themselves had given. Faculty taking the undergraduates' point of view were not exact but very accurate. Graduate students also excelled at the task of matching what undergraduates and faculty had produced as graded structures (Barsalou, 1987: 106-107). Barsalou warns that this does not mean that individuals are very accurate at taking other peoples' points of view, just that when ratings are averaged, different populations are accurate at taking other populations' point of view. I would stress that this is additional evidence in favour of the idea that a very rigorous publicity constraint is not necessary to guarantee basic understanding and communication between individuals and populations; I come back to this below.

Finally, given the above results of relative between-subject reliability, Barsalou and his colleagues turned their attention to within-subjects reliability. They designed some experiments to test how stable graded structures are within particular individuals. They would ask the same individual to rate the typicality of the same members of a category on two occasions two months apart. The invariability of stable graded structure model would predict that a given subject's answers would be highly stable, especially if she was asked to take her own point of view on common taxonomic categories. The results, however, were only an agreement of .80 on average (Barsalou, Sewell and Ballato, 1986). In both Barsalou's framework and the one I am suggesting this is not at all surprising. The subjects in these experiments did not have a stable graded structure that they retrieved to accomplish the task they were given. Instead they *selectively* activated information from memory to help them order the members the first time around; and, when they were again asked to rate the same members, they did not retrieve a stable graded structure nor the graded structure used on the first trial; rather, they again relied on selectively activating relevant information from *all of* memory for the task at

hand. The fact that two months had gone by explains that they were no longer in the *same* situation as before and so unlikely to activate the *same* representations as before. A study by McCloskey and Glucksberg (1978), cited by Barsalou (1987) provides further supporting evidence: they found that for certain categorisation judgments of the type 'Is a *Y* an *X*?', subjects often changed their minds across a one-month period of time; the example given is whether YEAST is an *ANIMAL* (cited by Barsalou, 1987: 112).

To summarise where we have gotten to so far, beyond the fact that Asian undergraduates might disagree with American undergraduates on what is representative of BIRD, because the birds they regularly encounter are not the same as Americans, it is also the case that if I process 'bird' in the context of 'pet', as in 'pet bird', I construct a completely different 'category representation' than if I process 'bird' alone (as the undergraduates in Rosch's experiments so often did). Finally, although there is some reliability between and within subjects, we do not regularly arrive at the exact same representations in comparable circumstances as our peers or as other time-slices of ourselves. In Barsalou's words, 'Invariant representations of categories do not exist in human cognitive systems. Instead, invariant representations of categories are analytical fictions created by those who study them' (Barsalou, 1987: 114). This conclusion might not be appealing to those who believe that there *are* invariant cognitive structures, the cognitive equivalent of *pure forms*, and that the task of cognitive science is to identify them. In that view, finding stability in our knowledge and our concepts *that is only relative* is a meaningless pursuit. In the view I argue for, the 'instability of graded structure' is far from meaningless and the consequences it has for our theories of concepts cannot be easily dismissed because they reveal something *fundamental* about our cognition: that it is capable of generating highly flexible and adaptive representations to effectively guide our behaviour.

A full account of my eliminativist position will have to wait until after the chapter on memory, but many details, particularly on the compatibility between radical contextualism and the 'exemplar-based norm theory'

account of categories and concepts I presented and argued in favour of in this chapter, can hopefully already be made clear.

As announced early in this section, Barsalou's claims regarding categorisation lead him further away from traditional theories of concepts than most theorists in his field; at the same time, however, by these same claims, he joins the many theorists interested in context-dependence and open to eliminativism that I have presented in this chapter and the last. In a nutshell, Barsalou adopts the position that just as there are no stable pre-computed representations for categories in memory, there are no stable mental representations, as traditionally conceived, for concepts either. Rather, as described by what I have proposed to call 'exemplar-based norm theory', the cognitive systems we rely on for our capacity to interpret the world (and, I would add, *learn* from our experiences) must *generate* what Kahneman and Miller call 'norms' *to represent knowledge*. This is because, contrary to what traditional theories of concepts and knowledge assumed, long-term memory is not organised into stable invariant categorical representations or perfectly delimitated concepts; rather, information in long-term memory is largely undifferentiated and must be scanned and summarised in order to represent something-out-there-in-the-world.

A parallel description can be given of how occasion-specific word meanings are arrived at: instead of consulting or retrieving stable, context-independent word meanings (and modulating them), our language interpretation mechanisms *generate* 'norms' that represent our knowledge of what words mean; this might be in the form of Recanati's 'semantic potential' but it is too soon to tell. In accord with meaning eliminativism, this account does not assume that long-term memory stores linguistic meanings in context-independent form, or, in fact, that it stores any *specifically linguistic* information differentiated and separated in any way from general information. It rather assumes that word forms are kept in memory together with any information, *contingent or not*, that our experience associated with them, including, for instance, contexts in which forms and meanings were paired, and not only explicitly but also implicitly communicated meanings, among other things. That memory could store these reputedly contingent

associations together with word forms was mostly met with disbelief across fields interested in language processing. I turn my attention to how this disbelief vanished in the face of evidence in the following chapter.

I return now to Barsalou's contribution. His claim is that different subsets of information from long-term memory are incorporated into ad hoc concepts and the 'instability' or context-dependence uncovered in his review of categorisation research is a result of different 'concepts' of the same category (i.e., the same kind) being constructed on different occasions. For instance, Asian undergraduates do not construct the same concept for the common taxonomic category *BIRD* as American undergraduates do; and, different populations do not construct the same concept for the class of *EDIBLE THINGS*. A supposed disadvantage of following Barsalou, and psychologists in general, on this point is that there is no way for such an account of concepts to meet Fodor's publicity constraint. This point has come up before (chapter 2, section 2.8), I claimed that attempts to meet the Fodorian publicity constraint regularly failed because of Fodor's overly rigorous conditions on concept individuation. I can now add that the account of concepts emerging from the considerations in this chapter simply avoids the Fodorian publicity constraint by rejecting the underlying assumption that understanding and communication depend either on literally the same concept-type being tokened by different people at different times or on people somehow sharing the same concept through a perfect similarity of thoughts and/or experience. I contend that not only is this unattainable, it is not necessary: understanding and communication do not depend on sharing literally the same concepts; they depend on being able to converge on conceptualisations, or communicated meanings, which is facilitated by the fact that we share cognitive systems as a species, beliefs and cultural customs as populations, background knowledge and circumstances as interlocutors; in addition to all of this, we have the ability

to take another's point of view.[116] So, for instance, Waismann might not have *immediately* constructed the same concept for INTELLIGENT as his interlocutor meant to express when he said 'My dog is intelligent', but according to 'norm theory', this just means that the association of DOG + INTELLIGENT surprised Waismann. He overcame his surprise and arrived at an interpretation of what the speaker intended not by consulting pre-computed norms, but by *generating a norm* for this particular occasion. Importantly, this does not mean that Waismann now has an *INTELLIGENT BEING* category that invariably includes dogs, nor that he now necessarily, strictly, shares a concept INTELLIGENT with the man in the park; rather, the point is that the two men's differences have not hindered their relative understanding of each other. In this framework, what is important is that a hearer can *create* an ad hoc concept/category exclusively for the purposes of a particular conversational exchange. Whether he wants, or thinks it worthwhile, to hold on to this construct is another matter.

This account strongly contrasts with the standard views of cognition, according to which, when a subject needed to mentally represent something-out-there-in-the-world, she had at her disposal, stable, invariant representations of categories (i.e., invariant concepts) that only needed to be retrieved ready-made from memory. In the new action tradition/contextualist framework, retrieval is a much more dynamic and adaptive process, knowledge is not clearly differentiated into invariant concepts and it is up to construction processes benefiting from seemingly boundless types of information and previous experience, pragmatic principles, general reasoning, common sense, *and the context* to dynamically construct whatever structure the task at hand requires: whether it be

[116] I could back this claim up by pointing out that I am far from being the only theorist who finds Fodor's absolutist claims regarding concept individuation and publicity unconvincing. I would be expected however to explain which theorists finds fault with which aspect of conceptual content and publicity, something much too time consuming for this thesis. I would rather repeat that Fodor (1998) has construed his constraints in such a way that only his own theory of concepts complies with them; and that I have already cited a pair of respected philosophers, Prinz and Clark (2004), who very decidedly reject Fodor's theory of concepts as being on the wrong foundations, I am sure there are others.

Barsalou's ad hoc concepts, and ad hoc categories or, as I have repeatedly suggested, occasion-specific word meanings arrived at in a way compatible with meaning eliminativism.

4.5 Closing Remarks: Implications for Word Meaning

To close the discussion in this chapter, I would like to insist on the consequences of Barsalou's notion of ad hoc concepts. At the end of the previous chapter, the direction in which I suggested cognitive pragmatics is moving, or should be moving, is towards a more decisive rejection of the traditional semantic framework with adoption of radical contextualism and the formulation of a new framework for word meaning in context. But, ultimately, once again, I left certain issues unresolved. Now, with the addition of Barsalou's account of concepts as themselves possibly unstable and context-dependent, instead of invariable, I can come back to one important issue to clearly state my position: I have claimed that my account is partially compatible with Fodor's 'word meanings are concepts' and relevance theory's 'words encode concepts'. This claim can now be more precisely formulated as *words express concepts but the concepts words express are constructed in their contexts of use*. With regards to relevance theory's position: I claim that words cannot encode concepts (or pro-concepts or concept schemas) because there are no stable concepts (complete or incomplete) that words could map to. Even if we considered that once a partial mapping is achieved, mechanisms of completion and/or modulation allow words to express concepts they do not encode. With regards to Fodor's claim, I agree that *words express concepts*, but the concepts they express are not to be simply retrieved by the listener (that is, there is no fixed mental lexicon); there is a process of construction that involves selectively activating information from memory. But, because these concepts would violate Fodor's publicity constraint, the incompatibilities between our accounts possibly outweigh the compatibilities. For Fodor, for instance, DOG, BARK, and INTELLIGENT, are stable and fixed concepts, otherwise they would not meet his publicity constraint. In his account, memory must somehow store a fixed and context-independent mental entity DOG, and so

on. In my own account, speakers use the word 'dog' to express a concept DOG but this does not constitute a *tokening* of the concept-type DOG because there is no concept-type DOG; rather, a DOG concept must be built each time it is needed, or in other words, each instance of DOG is occasion-specific. Furthermore, in response to worries about *publicity*, I point to the fact that it is arguably more psychologically plausible that we use the vast mental resources at our disposal to *figure out* what a word means in context rather than postulate concept-types with origins that are difficult to justify. Also, arguably, the more *means* at our disposal for figuring out a word meaning in context, the less important it seems to be whether there is a context-independent, stable, linguistically-mandated word meaning; perhaps we have such things for certain of our concepts, through formal education, for instance, but this exception is of little interest to a theory of natural language and meaning.

Finally, there is an expected objection to any kind of context-dependency in concepts that simply cites the impossibility of having any sort of stable knowledge or establishing any scientific fact in the scenario I have described. This would not be the case because a distinction can be made between the concepts we use every day as ordinary people and those technical terms of specialised fields. In a slightly different form, this has already been proposed when discussing Putnam's externalist semantics. I, for example, have a layman's concept for GOLD that is not the same as that of a jeweller, investment consultant or chemistry teacher. My own brand of meaning eliminativism and contextual concepts leaves it up to science to progressively better define the natural kinds of which we have concepts and meanings - I do not think this is a very controversial position.

Chapter 5: Memory

5.1 Introduction

The aim of this chapter is to provide the last piece of the puzzle in my claim that psychological models support the kind of meaning eliminativism that I have put forward. Much has already been said about the link between memory and occasion-specific word meaning construction in previous chapters. Furthermore, since the consensus is that 'the gap' postulated by the linguistic underdeterminacy hypothesis is filled by calling on information *stored* in memory, I do not think the case for the importance of memory in general needs to be made. The case that does need to be made, however, is that recent changes have come about in theorising on memory at such a pace and have so revolutionised thinking on memory that an attentive, detailed, and fresh look at theories of memory and their implications for theorising on language, and word meaning in particular, is required.

This chapter is divided into 4 sections. Section 5.2 summarises 'memory in the cognitive era'; I focus on the assumptions that characterised this period in order to show both how they permeated thinking outside of memory research and how they were finally challenged. In section 5.3, I counter the early assumptions with my own proposal of how to construe 'memory for language'. Hintzman's model, frequently mentioned throughout this thesis, is presented in full in section 5.4. Finally, once all of this background on the way memory actually works is in place, I close this chapter with a section entitled 'Implications for a positive account of meaning eliminativism': my aim is to provide as many details as possible of how the psychological models presented fit with the kind of eliminativism I advocate.

5.2 Memory in the Cognitive Era

5.2.1 Assumptions of the Early Models

Although inquiry into memory is as old as our civilizations, already present in the writings of Plato and Aristotle, the modern era of theorizing on memory can be said to have started as late as the middle of the 20th century. The view commonly held before the arrival of the cognitive revolution was of memory as a unitary faculty. In stark contrast to this, today, a common thread to all views of memory is a certain fractionation. This modern view was surely foreshadowed by the classic works of people like Hermann Ebbinghaus, credited with the first experimental findings in memory research, and William James, credited with having popularised the 'primary' versus 'secondary' memory dichotomy. James defines secondary memory as 'the knowledge of a former state of mind after it has already dropped from consciousness' and he contrasts this with primary memory, 'the *current* state of mind' that 'endures in consciousness' for a certain length of time (2010 (1890): locations 13863 and 13791). James' work was surely further foreshadowed by those who came before him so that beginning our review to coincide with the cognitive revolution might be judged as arbitrary or simply a matter of convenience. [117] Yet, I would argue that the sheer volume of research into human faculties that came about with the cognitive revolution justifies differentiating today's *cognitive* view of memory from all those that came before, including even the relatively modern *phenomenological* view.

An important pioneer of our cognitive view of memory is Donald O. Hebb. His contribution to neuroscience is of such importance that he is sometimes simply called the father of the discipline; he was among the first to study the neural foundations of behaviour and work out a *biological* theory of learning. Hebb (1949) proposed that persistent or repeated activity 'tends to induce lasting cellular changes'; or, in other words, that *in*

[117] Among others, James cites thinkers like William H. Burnham and Sigmund Exner.

consequence of cells firing together, growth and metabolic changes further facilitate the firing of one cell by the other (Hebb, 2002 (1949): 62) This is often summarised as 'Hebb's rule'*: cells that fire together, wire together*. Hebb further proposed a consequent distinction between two types of memory: short-term memory, based on *temporary* electrical activity in the brain, and long-term memory, based on lasting or *long-term* neurochemical changes.

The concurrent appearance of computer models is also key in understanding the development of memory models in the early modern stage. In the 1950s, Donald E. Broadbent was involved in a major double innovation: he defied the associationist stimulus-response school dominating psychology and stated his memory model in terms of *information processing.*[118] Behaviourism was quickly coming to an end and so, in 1958, when Broadbent published *Perception and communication*, it was well accepted and readily adopted. His view bonded with the existing primary versus secondary memory view to produce a model of memory that would have a particularly lasting influence. From this moment on, memory would often be thought of as (i) information moving along a path that initiates with perception and ends in long-term memory; and (ii) represented with flowcharts strongly reminiscent of electric circuits.[119] Broadbent's model consists of three modules *or stores*: two primary modules, the *S-system* and the *P-system*; and a more long-term memory, the secondary memory. In this series of systems, the first store, called the *S-system*, receives information directly from the environment, it serves only to hold that information until it is passed on to the next module or lost. Only selected information arrives in the *P-system* where it forms part of the

[118] Broadbent was, of course, also in his turn influenced by earlier thinkers and researchers, particularly by his teacher at Cambridge, Frederick Bartlett, an influential pioneer of memory research and experimentation. I return to Bartlett in the discussion of verbatim memory (§ 5.3).

[119] See Broadbent, 1958: 216, Figure 5 for an example of this. The diagram shows two stages labeled *S* and *P*. While there are many arrows coming into *S*, there is only one arrow joining *S* to *P*.

subject's conscious awareness. Both of these systems together make up what William James referred to as 'primary memory'.

Broadbent's model was so influential that prominent contemporary researchers, like Ian Neath and Aimée Surprenant, among others, argue that despite decades of intervening research, subsequent models, for better or for worse, left a great number of Broadbent's key assumptions intact. Yet, if the memory model I present at the end of this chapter is to be adopted, *all* of these assumptions need to be acknowledged and most of them rejected. Neath and Surprenant (2003: 44) list three assumptions linked to Broadbent's model that have survived into the present day. In this section, I propose to look at these assumptions and how they biased research into memory well into the 1980s. Then I argue that the view of memory these assumptions reveal is not only still prominent within psychology, hence relevant in understanding discussions on memory today, but that it is almost entirely the point of view which permeated modern linguistics so that the assumptions of early memory models not only misguided research into memory, they also biased thinking in linguistics.

The first assumption listed by Neath and Surprenant is that there are distinct and separate systems in memory, each implementing a different function: *holding information*, *rehearsing information*, *filtering information*. Furthermore, the systems are lined up in a sequence, so only the end product of one system goes on to the next. A second assumption is that primary memory, the *S*- and *P-systems,* is of *limited* capacity. From the assumption that primary memory is a filter, it follows that most of the information is discarded. The information in the *S-system* has the shortest life span. If it is not immediately part of awareness, whatever was perceived is irreversibly gone. In the *P-system*, the subject's focus of attention acts as a filter blocking peripheral information from attaining long-term memory. This limitation was seen as a positive protective device that kept our minds from being overloaded with information. This second assumption limits memory in two ways. First, following assumption one, it is *the limited scope of our awareness* that functions as a filter and discards all but a small fraction of what is available to our senses. Notice that even if our awareness

can only process a small part of our environment *at the moment of perception*, it does not seem to follow that only those aspects of our environment *consciously* perceived and processed at the moment of perception are picked out and stored in memory. This, however, is the assumption behind a scope of awareness that *serves* as a filter.[120] This leads us to the second assumption contained within the first: that the time allotted to the functions of holding, rehearsing and filtering information is *contained within* the time of processing. This follows only if information is moving through memory as electricity moves through circuits, that is, in a rudimentary one-directional flow, where values are decided *once* and only end results affect subsequent stages.

The third assumption to have survived from this early model is that information in primary memory can fade to the point that it is permanently lost. To keep something in the *P-system*, for instance, it is necessary to rehearse it. If it is not rehearsed, or not sufficiently rehearsed, and is filtered out of what arrives in long-term memory *then it is as if it had never been perceived*. Not a trace of it remains. A model adopting (whether overtly or not) the above assumptions is quite limited. It not only excludes all that is available only under a certain threshold of consciousness but it devises to get rid of any information lingering just outside of awareness. The problem with this, of course, is that many aspects of our environment are not part of our awareness and *are nevertheless processed and stored in memory*. Today's cognitive science gives a very important role to the detection of underlying patterns, a feat that the mind accomplishes without awareness, without commanding any conscious effort. During the 1960s, however, the focus was on experiments testing conscious memorisation with tasks such as the intentional retention of items on a list. This was complemented with neuropsychological evidence: distinct amnesias for distinct systems

[120] The alternative is that aspects of our environment *not* available to our awareness are nonetheless somehow captured and stored in memory. That this *can* be the case was already illustrated with an example in chapter 2 (§ 2.7.3): variations in pronunciation that are imperceptible to the listener are nonetheless stored in memory; Bybee (2000), and Pierrehumbert (2001), among many others since, found robust evidence for this (see Bybee 2010, for references).

evidenced, for instance, by a patient with a preserved primary memory, but deficient secondary memory. For both of these lines of research, Broadbent's simple and clear assumptions were seen as unproblematic. The field of memory research saw a very rapid increase in activity, but not much in the way of challenges to its main assumptions. Instead, the early models were consolidated with evidence of a distinction between short-term memory and long-term memory.

As a result, at the end of the 1960s, the dominating model, Atkinson and Shiffrin's 'dual-store' model, was largely a reformulation of Broadbent's model which added interesting *methodological* developments to his insights while respecting the overall original design. For instance, the figures that describe the different components of the system in Atkinson and Shiffrin's papers are still characteristically in the form of flow charts.[121] In accordance with previous assumptions, they represent memory encoding as information being *filtered* from the senses into a first store and from the first store to a more permanent store. The dual, or 'multi-store' model's innovations include considering this first short-term store as a 'buffer' or *shield* to further filter out information that might have passed from the senses to the first store but should go no further. They also proposed adopting the term 'temporary working memory' which rightly stresses that this 'short-term store' is not necessarily a separate physiological structure in the brain: the theory is, as a result, consistent with this component also representing the temporary activation of information permanently stored in long-term memory (Shiffrin and Atkinson, 1969; Atkinson and Shiffrin, 1971). The focus of the authors' work, however, is the experimental study of processes explicitly conceived of as 'under the control of the subject' and how they affect the flow of information in and out of the short-term store.

[121] See, for instance Shiffrin and Atkinson (1969), and Atkinson and Shiffrin (1971). These two papers detail the model as it was set out in the seminal 1968 paper. In this section, I cite these two papers, the two most cited of the Atkinson-Shiffrin collaboration, rather than the harder to access 1968 text. The articles conveniently address both short and long-term memory so they provide a complete view of the model.

Prominent among the 'control processes' that the very influential Atkinson-Shiffrin model brought to the fore is *rehearsal*:

> by rehearsing one or more items the subject can keep them in the short-term store, but the number that can be maintained in this way is strictly limited' (Atkinson and Shiffrin, 1971: 83).

This is a reference to one of psychology's most cited papers, George A. Miller's (1956) 'Magical number seven, plus or minus two' already well-known at the time. Atkinson and Shiffrin further explain that 'Once an image [or trace] is lost from the short-term store it cannot thereafter be recovered from it' (1971: 83).

This is an advance on Broadbent's model in just one respect: that *not* transferring a piece of information from the short-term store into long-term memory *immediately* does not necessarily mean that it will decay and be lost forever; rather, the subject can make use of one of the processes his memory system makes available to him, namely, *rehearsal*, to maintain the information in short-term store as long as he desires (Shiffrin and Atkinson, 1969: 180). The problem with this, of course, is that it presupposes that the subject *selects* information for this operation and so it still depends on the subject's awareness. As a result, all three of Broadbent's assumptions are intact: (i) although *working* memory can also represent reactivated information held in the long-term store, it is still the subject's focus of attention which filters information coming in from the senses; (ii) the first store is of limited capacity; and, finally, (iii) whatever is not attended to is lost and *not a trace of it remains*.

It is important to stress how little these assumptions were questioned at the time. Despite the fact that the multi-store model went through various generations of development, at the end, more key assumptions had been conserved than questioned. This is due in part to the long involvement of one particularly influential researcher: Richard Shiffrin. He collaborated not only with Richard Atkinson in the model briefly described above but also in 1981 with Jeroen Raaijmakers to update the dual-store model on issues of retrieval and associations in memory; with

Gary Gillund in 1984 on notions such as recognition and recall; and again in 1997 with Mark Steyvers.[122] For decades, multi-store models of memory were firmly the received view and any observations not consistent with them were considered not so much challenges to these models but, rather, peripheral phenomena they had yet to account for. The strategy adopted by the field as a whole seemed to be one of further development of these multi-store type models. As a result, the full theoretical consequences of the biases these assumptions introduced only became apparent recently, in retrospect. In the next subsection, I present the insights that most directly challenged these assumptions.

5.2.2 Real-Life or 'Ordinary' Memory

Criticisms of the general approach to memory presented above began to appear in the late 1970s and 1980s. In a particularly critical conference presentation, Ulric Neisser stressed that, after almost a century of research, nearly *nothing* was known about the 'interesting and socially significant' aspects of memory (Neisser, 1978: 4, cited by Neisser, 1988: 1). Not only the general public's but also the researchers' notion of memory had been unfortunately restricted to what subjects could *intentionally* memorise and recollect. If the kind of experiments conducted in memory research labs all over the world were any indication, memory's primary function was to memorize strings of words, letters and other specific stimuli. Yet, even in the

---

[122] The collaboration with Mark Steyvers (Shiffrin and Steyvers, 1997) is different in that it adopts many of the insights of multiple-trace models such as MINERVA 2 which I present at the end of this chapter (see Steyvers, Griffiths and Dennis, 2006); insofar as it does, it moves away from some key features of the earlier models but it is still a dual-store model. It is also worth noting that Mark Steyvers is a key figure in bringing *probabilistic* approaches to bear on human memory models. Along with people like Joshua Tenenbaum, Thomas Griffiths, Thomas Landauer, and Susan Dumais, he developed 'latent semantic analysis', a memory model based on an analogy between information retrieval by machines (such as Google searches) and human memory. Roughly, latent semantic analysis extracts the meaning of words by ignoring the word itself and concentrating on the contexts in which it appears. It has performed well on comprehension tests but has one major flaw that keeps me from integrating it into my model of memory for language: it does not seem that the kind of operations in this model parallel *in any way* those of human subjects. For some interesting coverage of latent semantic analysis, see Kintsch and Mangalath, 2011, and references therein.

absence of a positive account, this is arguably *not* what memory is *for.*[123] A decade after his original remarks, Neisser reports that, although the question of what 'real-life' or 'ordinary' memory is for has only begun to be investigated, he can already see a fundamental shift in conceptions of memory. This achievement is not the result of work by any single individual or school of thought; there are, however, several particular proposals that, having intervened early in this new stage, are worth mentioning.

In the 1970s and early 1980s, Endel Tulving proposed to distinguish between 'semantic' and 'episodic' memory. He defined the first as 'a person's abstract, timeless knowledge of the world that he shares with others' and the second as 'concerned with unique, concrete, personal experiences dated in the rememberer's past' (Tulving, 1983: *Preface v*). Early on, the key distinction was that semantic memory involved *abstract* knowledge or *factual* information that could be used in many different situations for different purposes. For instance*, Manila is the capital of Philippines* can be considered part of someone's semantic memory. This piece of information can be used as part of any thinking process (such as making travel plans), to answer a question (such as 'What is the capital of Philippines') or to understand a sentence (such as 'Manila hasn't always been the capital of Philippines'). Before Tulving, the term 'semantic memory' had already been proposed by Quillian (1968) as part of an account of how word meanings were stored in memory. I return to the critical question of whether word meanings are stored in semantic memory below. For now, it is important to note that Tulving broadened the term to include *any* piece of information a subject might know or any concept-based knowledge he might have (Neisser, 1988). Psychologists' interest in this type of memory was great since, *insofar as it represented a subject's store of knowledge of the world*, it was clearly central to most *if not all* cognitive processes. In contrast,

[123] Hintzman (2011) summarises the point with an evolutionary remark: our hunter-gatherer ancestors hardly had need for a shopping list when they ventured out onto the savannah. Answers to the question of what memory *evolved for* are starting to appear in thoroughly thought-out forms. A particularly persuasive account is that memory is a prediction machine (see Bar, 2009a, 2009b, 2011).

episodic memory was first mostly dismissed as less interesting since it was narrowly conceived of as a subject's capacity to recollect individual events, like a particular visit made to Manila.

A further fractionation critically challenged the early assumption linking awareness *or consciousness* to memory. Also in the 1980s, Daniel Schacter revolutionised his field by introducing a distinction between 'implicit' and 'explicit' memory. He noted that studies of memory tended to require the *conscious* recollection of specific learning episodes and set out to show that information encoded during a particular learning episode could be expressed *without* deliberate recollection (Graff and Schacter, 1985; Schacter, 1987). That not all memory requires explicit learning or conscious recollection of the specific study episode would become even more important when researchers started looking into how language is learned. When this distinction was originally introduced, however, the important point was that the specific learning episode could be forgotten, while the contents were still remembered.

The joint insights of Tulving and Schacter, among others, resulted in a new framework for thinking about and labelling the different systems of memory. This new framework maintained the split between primary, or *working* memory, and long-term memory, but critically proposed that long-term memory was further subdivided into two components: 'explicit', or *declarative* memory and 'implicit', or *non-declarative* memory (Squire, 1992). In this new framework, Tulving's original dyad of semantic and episodic memory, as defined above, are subdivisions of explicit, declarative memory. This implies that their contents should*, by definition*, be declarable. If *Manila is the capital of Philippines* is an item in my semantic memory, I should be able to intentionally retrieve this item in favourable conditions. A very important second component of the new model is 'non-declarative', or 'implicit' memory, often defined as memory *without* awareness.[124] It

[124] Neath and Surprenant (2003) dedicate a whole chapter to implicit memory. As with almost all topics within memory research, there is still much disagreement about how best

assembles a host of phenomena, *unnoticed or marginalised under previous models*, and recognises them as integral parts of long-term memory. Larry Squire includes under non-declarative, implicit memory information acquired during skill learning (motor, perceptual and cognitive), habit formation, simple classical conditioning (including some emotional learning), and priming. 'Experience can cumulate in behavioral change but without affording conscious access to any previous learning episodes or to any memory content' (Squire, 1992: 233).

The possible consequences of this new framework for those studying language processing are hard to overestimate. An important clarification, however, should not be overlooked: 'semantic memory' is often defined as memory for *meanings*, *concepts*, and general (impersonal) facts. However, as stated above, *by definition*, in the new framework, *semantic memory is supposed to be a part of declarative memory* from which it follows that I should be able to make statements about its contents. If asked for the capital of Philippines, I should be able to declare 'Manila'. Yet if semantic memory is both the memory of facts *and of word meanings and* part of declarative memory, then I should be able to state the meanings of the words I know as easily as I can name the capital of Philippines. The reasons why subjects cannot and, in fact, *do not need to, store word meanings in this way* have been the topic of previous chapters. Here it suffices to say that *were* definitions available to the speaker, they would surely be stored within declarative memory and more particularly within semantic memory.[125] But *following my thesis* that they are not, the question becomes *where* that which *serves as* word meanings is actually stored. In other words, where are Recanati's 'contextualised senses' kept? Where are the memory traces of previous episodes of use stored? Assuming that what we know when we *know a language* must be stored in either *explicit* or *implicit* long-term memory, the

---

to account for this type of memory, but the main insight, that memory can be *independent* of a subject's awareness, is not in dispute.

[125] The fact that *only* definitions purposely learnt *are* available to the speaker (under favorable conditions) is actually further proof that *otherwise* they are not thus available (which does not mean that they cannot be worked out).

discussion up to this point would seem to indicate that it is to *implicit* memory that language researchers should look. Research into implicit memory, however, has only just begun, and so most questions pertaining to the specificities of implicit memory must be left to the side for now.

In the following section, I focus on the topic of *memory for language* from the point of view of linguistics. The old received assumptions regarding *what* memory records and stores and *how* memory does this have been questioned and a new outlook on memory has begun to emerge. The objective now is to shine this new light on specifically *linguistic* representations in memory. Denying that a specialised declarative memory stores standing, context-independent word meanings (as in an 'abstract mental lexicon' or 'semantic memory'), and arguing rather for an eliminativist approach to word meaning does not mean that memory is taken not to store *anything* of linguistic nature; but, I will argue, it does suggest that the task of characterising this 'memory for language', as I call it, needs to be revisited. My contributions are limited to bringing together a couple of approaches that share a certain outlook on memory for language that I advocate. To introduce them, I co-opt the terminology of chapter 3 that distinguished between the approaches of formal semantics and contextualism with regard to word meaning (i.e., 'minimalism' versus 'maximalism'); here, these terms capture two opposing traditions of what memory for language registers: accounts that suppose minimal, abstract representations, which I'll call 'representational minimalism', are confronted with accounts that suppose rich, contextualised representations, what I'll call 'representational maximalism'.

5.3 A Maximalist Model of Memory

In this section, I call attention to a number of significant overlapping assumptions in the fields of language and memory. I focus first on parallels between the breakthroughs in memory research presented in the previous section and current transformations in various branches of linguistics. Then I quickly review specific contributions that challenge the old received view

of memory and contribute to drawing an alternative rich memory model for language.

That the fields of linguistics and memory research share assumptions is in part due to the fact that the same historical contexts and forces that facilitated the specific assumptions regarding memory systems presented above (§ 5.2.1) were present as cognitive models of language perception and processing were first developed. But it is also the case that particularly influential ideas stemming from research on memory permeated thinking outside of psychology and indirectly determined certain aspects of language theories. Today, however, some of these influential ideas are being challenged, and so, I suggest, should the aspects of language theories that they suggested.

I begin this section with the case of speech perception. My objective is not only to highlight the ubiquity of shared assumptions in the areas of language and memory, it is also to suggest a certain consistency between the solutions proposed by the kind of contextualism I am advocating and those already adopted by many researchers in the varied fields of linguistics. Speech perception is a good example and point of departure because not only is it where language processing begins but also insofar as what it has to offer is *transferable* to other areas within language studies.

For decades, a significant assumption had held fast: that the 'lack of invariance' of speech signals was a *problem* that the listener had to solve (Goldinger, 1998; Fowler and Magnuson, 2012). Speech is characterised by variability: speakers differ in the shape and size of their vocal tracts, in the care with which they articulate and in their native dialects, to name but a few factors. The variability is such that all agree that there is no one-to-one mapping allowing listeners to attach language forms to the acoustic signals they perceive. So how do they identify the language forms of the messages they interpret? One of the most influential approaches to this alleged problem simply assumed that speakers were *endowed* with a way of normalising speech signals. 'Speaker normalisation', as this is called, was

thought to intervene at every level of speech perception, starting with consonants and vowels, which come together in syllables and words.[126] Roughly, it assumed clearly defined categories for vowels, consonants, *and words*; two assumptions fatally came together: that 'variable speech signals are matched to ideal templates or prototypes' and that the independent, pre-existent notion of an abstract mental lexicon *requires* normalisation (Goldinger, 1998: 252). These two assumptions strengthened the idea that perceiving speech sounds amounted to *filtering out the noise* in a process of recognition; that speech perception depended on *matching* 'noisy' signals to *canonical* representations in memory (Goldinger, 1998). The underlying assumption, of course, is that what was filtered out, labelled 'noise', had no bearing on language processing. Perhaps even more importantly, considering the conclusions of the previous chapters, *normalisation* presupposed canonical representations for vowels, consonants and words but, ultimately, could propose no satisfactory account of them. Researchers tackling this set out to find 'invariant acoustic cues' thought to underlie the supposed acoustic pattern recognition systems and ultimately language form (i.e., *word*) recognition but the results were disappointing. Very few invariant cues were found, despite the fact that marketable applications (e.g. in computer-human interaction) for any advance in this area were a considerable motivation.

An alternative construal of variability in the speech signal, which rejected the assumptions of 'normalisation' began to emerge. I propose to call this approach 'maximalism' to better highlight the oppositions with the traditional account within speech perception and the parallels with oppositions throughout language studies. Maximalism suggested that variability was not necessarily a problem that speech perception had to solve, but that, *on the contrary*, it possibly represented an important source of information *for* speech perception. Variability in speech makes it rich with information about the speaker: if we know them, we immediately

[126] Robust evidence of acoustic continua perceived as categorically distinct sounds supported normalisation early on.

recognise them by their voice; if not, we can still know quite a bit about them: information such as sex, age, even weight and height; we can know about their socioeconomic status, and, very important in an exchange, about their emotional state (Fowler and Magnuson, 2012: 13). Variability in speech has a double upside, it provides information about the speaker that can help us identify the words and expressions they are using, and, beyond that, tell us more about how to interpret these words than the words alone ever could. This proposal directly challenges a minimalist assumption at the heart of both memory and language processing: that processing, *whether it be perception or encoding,*[127] necessarily entails *reduction*. In previous chapters, similar assumptions have appeared in diverse areas of present-day cognitive theorising and I have argued for possible alternatives. I do not claim to have offered a complete list of such cases, only a few examples. There is one more area, however, that is worth a closer look: language change.

There is robust evidence from an area in linguistics receiving much attention lately that offers additional support for the maximalist view of speech perception. Empirical studies in the area of lexical diffusion of phonological change have actually confirmed and extended the results presented above. Postulating *rich* representations instead of the conventional bare representations proved necessary in explaining correlations between pronunciation variability (e.g., deletion of /t/ and /d/ endings) and frequency. Janet Pierrehumbert (2001) and Joan Bybee's (2000, 2002) model for explaining these frequency effects involves a series of arguments. First, we must recognise that pronunciation changes, such as the well-known phonetic reduction process of final /t/ and /d/, occur online *as words are being used*. It then follows that words used more frequently are exposed to reduction processes more often. This leads to the conclusion that it is *because* words like 'told' are more frequent than words like 'meant' that

[127] 'Encoding' here refers to the sense this term has in memory research, that is, roughly, registering or recording into memory.

the /d/ is deleted in 68% of utterances of 'told' and the /t/ is never deleted in utterances of 'meant'. This also supports the idea that the phonological representations used by subjects are 'gradually built up through experience with speech', or, in other words, that, *in direct contradiction* to the received 'categorical rules' of phonological theory, which rather assume *minimal* schematic representations, specific details of *the context and use* are stored in memory as part of language form representations (Pierrehumbert, 2001: 137).

In chapter 3 (section 3.2), I presented 'Two contrasting traditions' in language studies: the 'product' and the 'action' traditions. I claimed that the traditional, received view of word meaning, as in minimal or formal semantics, belongs to the former and that the emerging view of word meaning I defend, contextualism, belongs to the latter. I also mentioned that the action tradition gave rise not only to neo- and post-Gricean contemporary pragmatics, among them relevance theory, but also to the usage-based tradition in linguistics. Since the time of their shared origins, both contemporary pragmatics and usage-based approaches have significantly grown in influence. The number of publications and the range of topics covered have likewise grown. The starting point and general framework for this thesis is a post-Gricean, relevance-theoretic pragmatics; however, insofar as usage-based models can contribute to the topic of the construction of word meaning, I propose to take them into consideration. A global review of the approach would be well beyond the scope of this thesis and, given the frequent parallels between usage-based approaches and the *inferential* model I have argued for throughout, I do not feel it is necessary. I propose therefore to limit my attention to those contributions directly relevant to the picture I am drawing of maximalist theories of memory for language.

Joan Bybee, a prominent representative of the usage-based approach, has significantly contributed to our modern understanding of frequency effects in language. She holds that just as phonological representations are built up cumulatively through use, so are the representations of *all* the other elements of our languages (in the usage-based tradition these are

morphemes, words, phrases, and constructions).[128] Bybee (2010) explicitly adopts exemplar theory to model how *use* affects these representations. In stark contrast to the notion of an abstract mental lexicon, she claims that an analogue 'exemplar' representation (one that includes sub or non-categorical information) is *stored in memory* with *each* experience of a word or phrase. The exemplar representations include details such as phonetic particulars, contexts of use, and, *importantly for my purposes*, those components of speaker meaning recovered by the hearer. Finally, in agreement with multiple-trace memory models (Hintzman and Ludlam, 1980, see chapter 4, § 4.3.2.2), Bybee posits that *each and every* experience with language has an impact on cognitive representations (2010: 7-8). This last point is important for frequency effects since it cannot be the case that frequency is only registered once a certain threshold has been reached, for *how would we know that it has been reached?* Rather, it must be the case that each experience can potentially 'count' in order for accumulation to begin (Bybee, 2010: 18). Bybee is aware that the accumulation of exemplar representations in her account contradicts the received view of memory as *limited*, and she argues, as I have, that these limits are more assumption than fact.

Support for the alternative maximalist position comes from other researchers and areas in the usage-based tradition. Olga Gurevich, Matthew A. Johnson and Adele E. Goldberg, for instance, have recently revisited the long-standing assumption that when the gist of an utterance has been understood, the form is immediately forgotten. The widespread acceptance of this particular idea surely owes much to a particularly well-received notion of what it means for memory to be limited. It is because memory could not possibly hold all the language forms that we experience in our everyday conversational exchanges that we assume that, as soon as the meaning is recovered from a form, the form has accomplished its function

[128] A non-comprehensive list of Joan Bybee's publications can be found in the references section.

and can be discarded, it is no longer needed. Empirical evidence in favour of this conclusion came from the work of Frederic Bartlett (1920, 1928). Bartlett (1932) asked a panel of subjects to listen to a story and then retell it from memory some time later. Critically, the story was of an unfamiliar culture and so it would be fair to note that it tested a specific kind of memory: memory for the details of unfamiliar themes and story lines following a foreign logic. Unsurprisingly, the results indicated that what was forgotten or altered by the subjects, were those details that did not easily fit their own cultural expectations of coherence. Nonetheless, the conclusion drawn from this study was that information necessarily had to lose most, if not all, its detail in order for some relevant information to be committed to memory. Many studies followed that of Bartlett, with similar results, and rapidly a consensus was established: verbatim recall for language forms is negligible or nonexistent.

Gurevich, Johnson and Goldberg (2010) start their study with a review of classic and contemporary studies of verbatim memory and complement this with their own experiments. A first cursory look at memory studies quickly reveals that the consensus above must *at least* be further nuanced since surface form *is* remembered in certain circumstances. If, for instance, the subjects are told that they will be tested, their performance can remarkably improve. Or, if the sentences are isolated, unrelated items, memory seems facilitated. Recall is also positively influenced by emotionally loaded language. In a study cited by Gurevich and colleagues (Murphy and Shapiro, 1994), one group of subjects reads a sarcastic letter and the control group an emotionally-neutral letter. Some sentences in the letter are identical and so the test was able to demonstrate that the sarcastic context improves verbatim recall. A related revelation, very important in my opinion, is that prior studies in general were simply too drastic in writing verbatim memory off. Empirical results generally point to *limited* verbatim memory but as Gurevich and colleagues (2010) suggest, this could be read as evidence for *some* verbatim memory (in a glass half-empty or half-full fashion). In the experiment above, for instance, subjects in the neutral letter condition are reported as having correctly identified

sentences as known 71% of the time and having mistakenly labelled paraphrases as known 54% of the time. Consider that recall was even better for the subjects in the emotional condition. These results point to *imperfect* verbatim memory, not lack of verbatim memory. This distinction is important because, given the context discussed in detail at the beginning of this chapter, that is, the assumption that *memory is very limited*, it makes sense that when it came to interpreting the results of their experiments, researchers tended to focus on memory loss and that their results were interpreted as supporting not *imperfect* verbatim memory but simply as an absence of verbatim memory. The authors conclude that, for this reason, among many others to do with methodological flaws in experimental design, results from previous studies are in general unreliable in determining whether verbatim memory plays an important role in language processing. They propose to investigate this further with their own experiments.

Two initial experiments test recognition memory and two further experiments test recall. These later two experiments involve subjects listening to a story as they see accompanying pictures. The stories are carefully worded to avoid the facilitating factors discussed above and subjects are not told that a recall test will be conducted at the end. When they are asked to retell the story, the pictures are used as prompts. Responses were then transcribed and for a phrase to count as verbatim, it had to match exactly the heard phrase or vary by no more than one word. For instance, if the subject heard 'I can go places no one else can' and produced 'I can go places that no one else can', this was counted as a match. But if the subject heard 'I was like everyone else' and produced 'I used to be like everyone else', it was not counted as a match. Two versions of the story were used so as to eliminate the possibility that certain phrases were coincidental matches, that is, matches produced not through verbatim memory but accidently, simply because it is the 'natural' way of expressing a thought. The authors are very aware of the importance of formulaic language and so take this possibility very seriously. The findings are that, overall, subjects produce 14% matching verbatim phrases. Since on average they produce fewer phrases than they hear, this 14% corresponds to 11%

verbatim recall for all the phrases heard in a 300-word-long story. This 11% might not seem a very high figure, but consider that before these tests, influential researchers, like Johnson-Laird and Stevenson had found *no* evidence for verbatim memory, so the jump to 11% is noteworthy.[129] The majority of what subjects recall is not verbatim, but even if only a small percentage is, this is strong evidence that surface structures are *not* lost immediately after the sentence is understood. A final experiment confirmed and extended these results. An elaborate experimental design allowed Gurevich and colleagues to test the number of times a subject *spontaneously* used verbatim memory in constructing their own description of a video they had previously heard described for them. They conclude that reusing previously heard clauses is a 'natural tendency' even in the absence of instructions to do so. The authors are cautious not to 'pigeon-hole' their results as stemming exclusively from implicit memory since while subjects were not asked to consciously recall verbatim utterances, they probably did *consciously* search their memory for earlier content. The authors further note that the relationship between implicit and explicit memory is complex. The recall phenomena they observed surely involve interactions of the two types of memory. This, however, does not invalidate their results. If anything, it should prompt further research into implicit/explicit memory for language. Their final conclusion is that the retention of specific linguistic formulations is a natural aspect of everyday language processing (Gurevich et al, 2010: 70).

A foreseeable objection to Gurevich and colleagues' results is that they do not make it clear what function the kind of verbatim memory they discovered serves. Discovering that verbatim memory is not inexistent is an accomplishment in that it challenges a long-standing assumption regarding memory and thus opens up new paths for research but, undeniably, it leaves more questions than answers. I would argue that this is because human

[129] Researchers like Johnson-Laird, simply considered that 'Listeners do not ordinarily retain the syntax of a sentence for longer than is necessary to grasp its meaning' (Johnson-Laird et al. 1974: 704, see also Johnson-Laird and Stevenson, 1970).

memory is a relatively new field and the relation between memory and *language* is something we are only beginning to understand.

For my particular purposes, Gurevich and colleagues' results are interesting mainly because they point to a previously ignored resource speakers have at their disposal: the limited but existent access to the 'surface' form; in other words, the particular word, expression or construction used when they experienced an episode of language use. For an account that holds that memory does not store standing, context-independent, linguistically-specified word meanings that can serve as the input to processes of occasion-specific word meanings, it is critical to point out what memory *does* store. In my approach, I heavily rely on the pairings between particular uses of words or expressions and what they were, *on that occasion*, used to express (the basis for Recanati's 'contextualised senses'). So evidence of verbatim memory, *insofar as it points to memory of specific wordings paired with specific situations*, lends support to my claims.

The central aim of the following section is to address the assumption that what is stored in memory is a fixed, abstract mental lexicon. The issue at stake is the same as in chapter 3: do words have standing, context-independent meanings, or, putting it in the terms of standard relevance theory, do words *encode* or map onto concepts (while they on occasion express a concept they do not encode)? Now, however, this question is addressed from a psychological perspective: what alternative is there to the idea of fixed, pre-computed representations?

5.4 Hintzman's Memory Traces, Concepts and Word Meanings

I suggest looking to Douglas Hintzman's work on exemplar theories of concepts as a point of departure in replacing the idea of an abstract mental lexicon. More than any other approach I am aware of, Hintzman's account of memory espouses the new outlook discussed throughout this chapter: a 'maximalist' model that assumes it is unproblematic to suppose rich and contextualised rather than minimal and abstract representations. Also, Hintzman has more than enough in common with the contextualists of

previous chapters to figure among them. He is special within this group for having developed an experimental paradigm to simulate the creation of 'generic' or abstract ideas: the very well received and influential 'multiple-trace' memory model, which includes 'MINERVA 2', a computer model implementing the theory and capable of simulating human memory. Arguably, this model has allowed psychologists, for the first time, to address the question of *how abstract ideas are formed* as experience with the world accumulates. The first aim of this section is to give a detailed description of the model and of the theoretical questions with which it is concerned. My principal concern, however, is to clarify the relevance of Hintzman's model for a theory of word meaning in context.

Hintzman (1986) takes the discussion up, more or less, where 'context theory' or the 'exemplar model' (chapter 4, § 4.3.2) left off: mental representations are created as a direct consequence of a subject's exposure to different stimuli, or instances of real-world experiences. [130] Critically, however, unlike prototype theorists, there is no assumption that these representations are *fundamentally different* from the experiences themselves; rather, they are 'traces' left behind by the experience of an instance. Consequently, there are no *abstract* representations to be subsequently retrieved for categorisation tasks. In the terms of memory research, prototype theory postulates that the effect of repetition is not only the creation of traces of the individual events in *episodic* memory but also the creation of a *separate* abstract representation that would be stored in a 'functionally separate' type of memory called 'semantic' or 'generic' memory. According to Hintzman's alternative, there is no functionally separate type of memory. The most interesting possible interpretation of this idea, for my purposes, is that abstract representations (whether they be

[130] Hintzman's proposal also takes up where Pylyshyn's mental representation theorising left off. For Pylyshyn, if mental representations were images, they could conceivably only code geometrical distributions and sensory attributes, but not *information* like relations. He concluded that mental representations had to be purely symbolic. In memory trace models, this conclusion is avoided because memory traces are construed as capturing *events* with their geometrical distributions, sensory attributes, and whatever information these features made evident for cognition. Memory trace models of mental representations recover what Pylyshyn's model initially inadvertently left out.

prototypes or some other structure) are simply not playing the role prototype theorists imagined for them (that is, serving as summaries, models, schemas or *norms*) for subsequent tasks. Instead, it must be the *individual instances* that were created during experience that are 'aggregated', as in norm theory, producing a 'framework of evaluation' to serve *the function of a norm*.[131] In Hintzman's terms, his view is that 'only traces of the individual episodes are stored and that aggregates of traces acting in concert at the time of retrieval represent the category as a whole' (Hintzman, 1986: 411).

A related point of contention between prototype theorists and Hintzman is that despite allowing that prototypes are a major step forward in our understanding of concepts *compared to the traditional theory of concepts*, he finds that prototype theory is still flawed *in the same way* as many other theories of concepts: context is again a key element insufficiently taken into consideration. Hintzman illustrates this important point with an example. When subjects are asked to rate how typical a beverage is, context plays a major role in their answers (something unacknowledged by prototype theorists). In a 'neutral' context, subjects asked to rate coffee, tea and milk on a scale of prototypicality typically rate coffee the highest, with tea in second position and milk as the least good example.[132] The explanation given by prototype theory is that coffee is a prototype of the category *BEVERAGE* and that, of the two remaining beverages, *tea* resembles it more than *milk* does. But, if subjects are asked to perform this task after hearing a story about a truck driver getting a doughnut and a beverage for breakfast, *milk* and not *tea* will come in second place. So the centrality relations amongst members of a category are not fixed, as most studies in prototype theory seem to suggest. To be very clear, to the assumption that memory stores fixed, abstract representations, or

---

[131] Kahneman and Miller (1986) cite Hintzman's work: Hintzman and Ludlam (1980), and Hintzman (1986), which was then *in press*.

[132] I have put *neutral* in scare quotes since the true complexity of this issue is that there is *no* such thing as a neutral context. Consider that if this experiment had been carried out in the UK instead of the US, tea and not coffee would have probably rated higher. Any contextual factor, even those we take for granted, can influence how items are rated.

'precomputed rules' of any kind, Hintzman's radically contextualist approach counters that memory is endowed with retrieval mechanisms that *produce* the (abstract) representations, frameworks of reference, or concepts that we traditionally thought were stored, fixed, and pre-existent.

Hintzman's memory model is radically different from the 'dual-store' type models presented in the first section of this chapter and only very partially compatible with the new framework emerging around the contributions of Tulving and Schacter. A first major difference is that instead of filtering information as it flows into the memory system, Hintzman proposes a 'multiple-trace' model. Experiences produce memory traces instead of tokening abstract types stored in long-term memory. These memory traces largely correspond to what I have called *rich*, *contextualised* representations, the representations postulated by a maximalist model of memory. Hintzman (1986) describes a memory trace as a 'record' of an episode or experience. Importantly, traces, like exemplars, are supposed to *preserve* not only the features present in the experience they are a trace of, but also the configuration of these features. Furthermore, in Hintzman's model, *each* experience attended to is recorded as a trace in memory. If an experience is repeated, a pre-existing trace is not strengthened, as single-trace models assume, rather, a new memory trace is created with every experience.

A second major difference is that Hintzman's model can do without dividing memory into different stores. He adopts James's (1890) 'primary' and 'secondary memory' to differentiate between 'an active representation of a current experience' and the 'largely dormant' pool of memory traces. He also speaks of 'semantic' or generic memory in order to be able to compare his model to others and to explain it in the terms of the field he is part of (i.e., the study of memory), but his claim is that a *single* system can account for memory as a whole.

The third major difference, already mentioned above, is that, following his own ground-breaking research on categorisation (Hintzman and Ludlam, 1980), Hintzman explains the creation of abstract ideas as

occurring *not* as memories are 'encoded' or *registered* into memory, as most models before his did, but during *retrieval*.

The three differences above not only set Hintzman's *theoretical* model apart; they are part of how Hintzman *simulates* human memory with the computer model MINERVA 2. The computer model requires a couple of basic assumptions in order to simulate memory. First, it must be assumed that experiences are made up of *primitive* properties, 'some of which are abstract' and that similar experiences share the same properties (Hintzman, 1986: 412). This assumption is no different from Tversky's (1977) contrast model's 'feature matching': similarity is construed as a linear combination of the measures of common and distinctive features among objects. Hintzman explicitly states that the number of primitive properties is large and that they are *not* acquired by experience.[133] An experience can be completely described by a potentially large set of properties and the memory trace of this experience is construed as capturing some subset of the properties of the concrete experience in such a way that when the memory trace is retrieved the subject potentially accesses the properties *in the form of an ensemble*; that is, a memory trace is an *analogue*, *holistic* representation of a concrete episode of experience. Properties can be anything from simple emotional tones and modality-specific sensory features (such as colours and odours) to abstract relations (such as *before, greater than*, and so on). Mental representations of experiences are assumed to be sets of such 'primitive' properties. Despite the fact that the list of such properties can be quite large, it is assumed that it is much smaller than the variety of experiences humans can have. Finally, the model presupposes that, notwithstanding their great variety, human experiences can be described by the features or properties they share with other experiences, and that they can be individuated by their own pattern of features and by the features that they alone possess (Hintzman, 1986: 412).

[133] Hintzman further explains that on the topic of primitive schematic properties, he adopts Fodor's (1981) rationalist approach instead of an empiricist approach.

Memory traces are not perfect copies of the experiences they represent. The model uses a 'learning rate' parameter when 'encoding' an experience into memory so that the individual features of a trace can match the original event to varying degrees. The model can simulate learning in this way since a particular feature can be ignored or forgotten, its 'value' in the trace then falls to zero (Hintzman, 1988: 529). This perspective on learning highlights an important point: it has long been assumed that learning is based on extracting criterial features or defining essences from our experiences and, as mentioned earlier, selectively recording or encoding *only* these minimal, abstract representations. The idea is that what is learned from an experience is its *schematic* representation and that learning is effective insofar as it ignores irrelevant details and records only the essential properties of an experience. But this is a problematic stance on abstract representations because it disconnects experiences from one another very early on by fixing what is essential to each of them in a separate moment, i.e., the moment at which they enter memory. Once in memory, these representations can, of course, be compared to one another, but details lost during encoding would no longer be available.

As I hope previous sections in this chapter and discussions in previous chapters have shown, there is abundant evidence that suggests that details that would be lost according to this *extraction/abstraction at encoding principle* are nonetheless somehow available and can cumulatively change a mental representation and influence behaviour. Hintzman's model offers a more satisfactory account of abstract representations by adopting a couple of simple assumptions. First, details believed to have been filtered out by other models do arrive in secondary memory in the form of memory traces. In other words, in Hintzman's model, the representations in memory that serve as input to subsequent cognitive processes are richer in detail than those of any model that assumes that what is stored in memory is fundamentally different from what was experienced (i.e., more abstract, like a prototype). Memory traces are richer because, although learning is construed as imperfect (some information is lost during 'encoding') and primitive abstract properties are postulated as basic building blocks,

Hintzman's model rejects the view that memory is in danger of being overloaded and so does not construe processing as requiring the kind of type/tokens pairings Pylyshyn's early model did, or consider that filters are needed to restrict what gets into secondary memory. MINERVA 2 simply posits a *multiple* trace model:

> Every conscious experience gives rise to its own memory trace, no matter how similar it may be to an earlier one. Thus, phenomena that are repeated but nonetheless command attention will be represented in memory over and over again (Hintzman 1986: 412).

The second assumption concerns the time or stage at which an abstract representation is to be created. The upside to delaying the creation of an abstraction until there is a task at hand, or, in Hintzman's terms, 'the time of retrieval' are twofold. First, 'because information is abstracted from concrete experiences at the time of retrieval rather than during learning, no sophisticated executive routine is needed to decide when and how to tune, reorganize, or abandon memory structures' (Hintzman 1986: 423). Second, MINERVA 2 'waits until the time of retrieval to combine experiences' and thus 'has more flexibility in how the combination is to be done' (Hintzman 1986: 425).

Retrieval is a fundamental part of the model. It consists of two basic operations: a 'probe,' and an 'echo'. The probe is part of an active experience in primary memory that is communicated to secondary memory where it activates all the memory traces that shared properties allow it to find.[134] The more properties that the trace shares with the probe, the stronger the activation of this trace will be. Following the connectionist, neural model, activation can be seen as spreading from lower-level, primitive property nodes to higher-level nodes representing complex properties; activation is also horizontal. Importantly, when a particular memory trace is activated,

[134] In chapter 2 (section 2.8), I briefly introduced this idea with the notion of a 'cue' used to 'probe' memory (like typing a word can prompt a Google search) I can add now that what comes back reveals what memory contains like an 'echo' reflects the contours of the space that produces it.

the ensemble of its features is activated, so that the activation of a trace implies not only the activation of the features it shares with the probe but also those it does not share. For Hintzman, this is important in explaining how information not present in the probe can be activated.

Hintzman also suggests another image to illustrate the basic operations of retrieval: the notion of *resonance*. In this image, the probe is 'broadcast' to all memory traces and some of them resonate back. This 'echo' consists of 'content' and 'intensity'. Intensity corresponds to the total amount of activation 'triggered' by the probe. It is proportional to both the similarity of the memory traces to the probe and the number of traces activated. The content is the set of primitive properties activated by the probe. Properties not present in the probe are part of the echo content when they are present in strongly activated traces or in many of the traces a probe activates. For Hintzman, this particular way of construing the notion of how the 'echo' brings the contents of memory together is a distinguishing mark of his approach: since the probe is 'broadcast' to all memory traces, they can *all* be simultaneously activated and simultaneously contribute to the echo, '*each according to its similarity to the probe*'. Importantly, the echo can itself become an experience in the subject's awareness as *previously unrelated features come together*. This simultaneous activation and contribution of traces to the echo also means that activation is so widespread that the content of the echo primarily reflects the shared properties and 'drowns out' any sour note as in a sea of voices.

In the simulations presented by Hintzman (1986) to illustrate his model the memory traces and the probes used to activate them were made up of 23 features each. Presence and absence of a feature were represented with +1 and -1 values, respectively. In one demonstration, the first ten features were used to capture a category name and the remaining 13 for a corresponding pattern. To simulate how MINERVA 2 can find the name of a category when the input consists of only category members, for each to-be-found category, three category names were randomly generated and paired with randomly generated prototypes. Next, exemplars of the to-be-found categories were

generated by distorting the prototypes and combining these 13 features with one of the three 10-feature category names. To illustrate these purely mathematical value lists, let's replace them with an example of what they are designed to model: suppose that the to-be-found category is *BEVERAGE*, and that three sub-categories are used, *COFFEE, TEA,* and *MILK*. A prototype was generated by randomly altering, for each of the 3 sub-categories, 13 features for type-of-beverage and 10 for name-of-category. Testing consists of generating probes that lack certain features; in our example these would be either *type-of-beverage features* or *name-of-category features* in order to ascertain whether the model can produce general beverage features if given a category name or can generate a category feature list when given only one exemplar of one category.

This last task corresponds to the ambition of explaining the creation of abstract ideas mentioned earlier, something Hintzman calls the 'schema-abstraction task'. When a probe is fed into the system, an echo returns that is the result of the activation of all the stored exemplars that have similar features. Because exemplars are bundles of features, the activation of the features that the probe shares with the exemplar can also activate features in an exemplar that are not in the probe. Activation spreads from features that are strictly shared between the probe and the exemplar to features that are associated with the exemplar but not present in the probe. To give a concrete example, one of the tasks that such a system could accomplish is *associative recall*. Suppose I meet my friend's friend for the second time. The face I see before me is the probe I use to search my memory (not even purposely, it happens automatically). It will activate similar faces and a more or less distant memory of this exact face (perhaps with a different haircut). Now I hear his name again, or, for the first time, it does not much matter - this activates exemplars of *John*, for instance, and through co-activation associations are formed. A new exemplar is created where face and name features can now be automatically co-activated (in favourable conditions, of course). The next time I see this person, all of the exemplars that share features will be activated, his face will activate at least two exemplars (say, one with short and one with long hair) and, with any luck,

the activation of the face features of the second instance will spread to the name features associated with that particular exemplar thereby successfully reminding me of his name. This account of memory is not only intuitively appealing, it is also psychologically plausible and theoretically economical. In fact, it solves two tasks, remembering and associative learning, with a single process, namely, co-activation.

A full description of the model would include a formula, based on Tversky's (1977) principles, specifically devised to quantify features and compute the similarity of a particular trace to a probe, but, a description of the mathematical basis of these computations is beyond the scope of this thesis. My focus, in the remainder of this section is rather on how the model, as described in theoretical terms, applies to the issues of interest in this thesis: namely concepts and word meaning. Hintzman is aware of the interest his model has for theorising on these topics and discusses the important theoretical consequences of his proposal. From the outset, one of his objectives is to answer the question of how abstract knowledge relates to specific experience and his major claim is that 'concepts do not have unitary representations' (Hintzman, 1986; 420). When Hintzman was conducting his research, most investigators were working within the 'new' framework launched by Tulving and Schacter, and so they interpreted Hintzman's interests as concerning the relationship between 'semantic' or 'generic' memory and 'episodic' memory. Hintzman's revolution, however, consists in positing that there is perhaps only *one* memory system 'which stores episodic traces, and that abstract knowledge as such does not have to be stored but can be derived from the pool of traces of specific experiences at the time of retrieval (Hintzman, 1986: 411). Instead of fixed conceptual representations, or 'unitary representations' in Hintzman's terms, he proposes that there are only traces of episodes that contain the concept name, or label. Notice how this completely reverses the order established by the so-called 'new' framework: psychologists were particularly interested in 'semantic memory' when Tulving extended the term from word meanings to *any* conceptual knowledge a subject might have in memory because *theoretically* this would mean that all the criteria, fixed norms or conceptual

knowledge that subjects used in any task involving concepts, categories, and norms was *somewhere* in the mind, ready to be discovered. For Hintzman, however, as for Barsalou, Tversky, Kahneman, and others covered in these chapters, *there is no fixed store of concepts anywhere in our minds*.

At the same time, the theory maintains that retrieval mechanisms create *abstractions* and if these abstractions themselves become the object of conscious reflection they will be stored as traces in memory. Hintzman's model assumes that abstract ideas are created in the process of retrieval, so MINERVA 2 does not in fact make use of such previously created abstractions in simulating memory. However, Hintzman is adamant that nothing in his theory would prohibit such abstractions from entering secondary memory. In fact, he sees it as an ironic upside to his explanation of concepts as not requiring abstract representations that it in fact provides a good explanation of how such representations are generated and learned. On the force of what the model can accomplish without the need of these abstractions, however, once this clarification is given, Hintzman holds that it changes nothing to his motivation for simulating schema-abstraction tasks without abstract representations (Hintzman, 1986: 422).

Hintzman's alternative is that we look to *episodic* memory for our store of know-how. We are only beginning to see just how this works, but as an illustration, assume, as is surely the case, that no one is born knowing what the capital of the Philippines is; even those born in that city had to learn this fact. In my case, having grown up on the other side of the planet, I was perhaps supposed to memorise this from a list of world capitals at school and perhaps I was even successful in rehearsing it until the day of an exam. But subsequent years of not having any use for this information would have buried it deep in my memory so that many years later, on planning a trip, for instance, I could be said to again 'learn' that Manila is the capital in the context of planning the trip. If I never make this trip, I might again 'forget' about Manila; if I do make the trip, the number of episodes with Manila in them stored in my memory will probably be such as to greatly facilitate keeping this information available. My point is that, contrary to what the traditional approach assumed, a piece of information like 'Manila is the

capital of Philippines' is not encyclopaedic information that is static in my mind like an entry in an atlas. You know about Manila because certain episodes of your experience have, for one reason or another, accumulated in such a way as to be retrievable.

Hintzman argues that adopting his alternative allows us to avoid two major difficulties 'semantic memory' theorists face. The first concerns how knowledge that is assumed to be stored explicitly in semantic memory is first acquired and how it is revised by experience. According to Hintzman, assuming that there is an abstract schema that new input can be compared to inevitably leads to postulating a 'powerful executive routine' since it is assumed that *something* behind the scenes evaluates the effectiveness of our store of knowledge in our interactions with the world and identifies and keeps track of failures in order to '*diagnose* their causes, and *infer* the nature and amount of tinkering with existing structures that will be necessary to *insure* that the failures do not happen again' (1986: 423, *my italics*). Notice that the terms of this explanation are at the level of *intentional* explanation and that, as discussed in chapter 2 (§ 2.2.3), the consensus of scientific explanation is that such intentional explanations must be set within a physicalist framework. In other words, a full scientific explanation of how memory *revises* stored knowledge cannot leave it up to a mysterious executive routine (i.e., a homunculus) to make any decisions. Hintzman's alternative:

> The view offered here is that, to the extent that abstract knowledge as such is stored in memory, it has no special status or function. All experiences to which one attends are encoded as episodic traces, whether they violate one's expectations or not. A new experience never modifies an old memory trace. […] changes in behaviour follow automatically from the indiscriminate accumulation of new episodic traces in memory (Hintzman, 1986: 423).

The second difficulty facing semantic memory theorists identified by Hintzman is the problem of representing context-dependency in hierarchies that are assumed to be fixed. It is because what he proposes, i.e., *aggregation at retrieval*, is a possible solution to this difficulty that Hintzman is cited by

Recanati in the context of meaning eliminativism; and, it is also the reason I first turned my attention to his work. Hintzman explicitly agrees with those language theorists who question the idea that words have fixed meanings. He is aware that his memory model on its own cannot offer a complete account of how *utterances* are interpreted, but he does suggest that it represents at least a partial answer to the issue of *word meaning in context*. Hintzman's scenario assumes *no fixed conceptual representations*, only the traces of episodes in which words were used, and that knowledge about the world is *somehow* used in arriving at a specific meaning (1986: 423). Hintzman illustrates this with the common verb 'eat': since it can be used in very different circumstances, for instance, with different subjects (e.g., termite, princess, snake) and objects (e.g. crumb, pencil, melon), it is to be expected that instead of retrieving a core meaning when processing this word, what is retrieved is highly context-dependent. The word is assumed to be represented in memory by a very large number of memory traces and the key is that, while *all* these traces can potentially be activated in parallel, only a subset is 'strongly activated' in any particular encounter; critically, *the* factor determining which traces get strongly activated is the context in which the word appears.[135]

In the following final section of this chapter, I develop further my own positive account of meaning eliminativism using the insights on memory from Hintzman that I have just outlined. Throughout the latter chapters of this thesis, I have begun to provide arguments in favour of meaning eliminativism; however, as I mentioned at the end of chapter 4, a full account of my eliminativist position could not be given until Hintzman's contributions had been fully explored in a chapter dedicated to memory. The account below completes the arguments in favour of eliminativism and focuses particularly on clarifying how Hintzman's account of memory supports my claims. I also aim to bring together here the other key contributions in order to give a full picture of an eliminativist account of

[135] I return to the notion of 'context' in the following subsection; for now, I can say that context is both the 'co-text', as in the linguistic or discursive context in which a word appears, and the wider 'extra-linguistic' context of the situation.

word meaning in context and the framework I have endeavoured to draw for it.

5.5 Implications for a Positive Account of Meaning Eliminativism

My review in chapter 3 of contextualist and radically contextualist positions on word meaning ended with the acknowledgement that, to the best of my knowledge, meaning eliminativism, the account in favour of which I argue, is not explicitly endorsed by anyone in the language sciences, not even by Recanati, who is responsible for describing it and giving arguments in its favour. It seems that eliminativism appears far too radical a break with the traditional conception of word meaning to be translatable into a full-fledged account of how word meaning in context works. It has been my contention, however, that embracing this break with tradition can actually be taken as an *opportunity* for theories of word meaning, and particularly for approaches that are already progressively moving towards more radically contextualist views.

In chapter 3, section 3.5 'The second stage: abandoning the modular view', I presented some approaches that are already leaving the too restrictive framework of formal semantics behind, although not necessarily going as far as adopting eliminativism. In building a case for eliminativism, there was one major objection in particular that I needed to address: that by outright denying the existence of the *sort* of input generally assumed to serve as *the* input for the occasion-specific word meaning construction process, (i.e., either fixed linguistic meanings or something less conceptual, more schematic), the construction/modulation process that is at the heart of contemporary pragmatic accounts would be left both without a clear place to start and without sufficient constraints on its operation. In order to address this potential objection, I had to show that what we have recently learned about the workings of the mind from psychological models (such as exemplar and norm theory), and memory models (such as multiple-trace memory models), could, first of all, reveal that despite the radical changes adopting these models entail, an occasion-specific meaning construction process could still be *at the heart* of pragmatic processes of utterance

comprehension and, furthermore, that suitable alternatives are available for the tasks of triggering and constraining this construction process. To get started I followed the lead of Recanati, who cites Hintzman's multiple-trace memory model in both his 1998 encyclopaedia entry and the 2004 book; however, both times, he only does so in passing. In the 2004 book, where more details on eliminativism are provided, all Recanati does is refer his readers to Hintzman's work 'for a detailed psychological model supporting meaning eliminativism', he gives *no* description of how Hintzman's model would in fact support it (2004: 147). In order to truly overcome the objection against meaning eliminativism, my task was therefore to back Recanati's claim with a full portrayal of Hintzman's multiple-trace memory model and, with this portrayal in hand, argue that eliminativism is not only psychologically plausible but truly a better alternative than comparable accounts. The assumptions of Hintzman's model, however, are so novel, and so radically different to construals of memory within linguistics, that in order to adequately present eliminativism I needed to spell out the framework in which Hintzman's model should be understood, that is, give the background for a multiple-trace memory model, which is broadly *exemplar* based, and, crucially, *radically contextualist* in that it suspends creation of whatever structure interpretation needs until *time of retrieval*.

This chapter has added critical details to this picture: most importantly, a *maximalist* view of memory for language. I have insisted on the *details* recent research has shown memory *does* record in order to postulate that memory traces of word meaning *in use* are rich and detailed in ways we are only beginning to understand. I argued this with the case of limited but important verbatim memory: memory *does* record specific wordings paired with specific situations and can sometimes retrieve these pairings whole. I also illustrated rich, *contextualised* memory traces with the case of lexical diffusion of phonological change. Research into lexical diffusion led researchers to postulate rich phonological representations to explain the fact that the well-known reduction process of final /t/ and /d/ correlated with the frequency of the words: 'told', which is more frequent than 'meant' deleted the final /t/ 68% of the time, while it was never deleted

in utterances with 'meant'. Words are exposed to such reduction process *as they are used*, so, it was necessary to postulate the accumulation of memory traces rich with phonological detail in order to explain certain behavioural changes (e.g. deletion of some but not all final /t/s).

Limits of time and space, unfortunately, did not allow me to cover language change beyond the example of deletion; yet, perhaps a brief discussion of a particular type of *meaning* or 'semantic' change can illustrate how the point exemplified by lexical diffusion can be extended to other cases: for instance, cases where words and expressions gradually, *through the contexts in which they appear*, build up senses (or, in the terms of this thesis, 'semantic potentials') that they did not originally have. The French term 'soûl', for instance, was originally used to express SATIATION, neutrally, without a negative connotation. Then it became a euphemism for drunkenness, a way of avoiding offensive terms (such as 'ivre', 'drunk') while referring to someone who had had too much to drink. But, as the instances in which it expressed drunkenness accumulated, the term acquired its own negative connotations and eventually lost its euphemistic character. Steven Pinker (1994) refers to this as the 'euphemism treadmill': sometimes, new words chosen to replace those judged impolite are 'tainted by association' and replacements must again be found for them. The point for both phonological and semantic change is that what is recorded in memory is not a static canonical representation and there is no assumption that anything in particular is filtered out as experience flows into memory. The new assumption is that experiences produce memory traces, and the emphasis is on the fact that these representations can be described as records *preserving* the different aspects of the experience.

The first step towards implementing Recanati's insights with Hintzman's model and thereby offering *my own positive account of meaning eliminativism* is to match Recanati's notion of contextualised senses with Hintzman's notion of memory traces. Contextualised senses, according to Recanati, are the occasion-specific meanings that words and expressions have actually expressed (or have been taken to express) in their particular

contexts. *These occasions of use are captured as memory traces*. Each occasion of use of a word or expression is a distinct experience and produces its own trace. In this way, experience with language produces a vast store of rich, contextualised representations of specific episodes of words and expressions in use *and the senses to which they gave rise*. Furthermore, thanks to Hintzman's notions of cue and retrieval, the information in these traces can be selectively activated. Contextualist approaches shy of eliminativism suppose that something fixed, like a context-invariant linguistic meaning, serves as input to the occasion-specific word meaning construction process. Recall that two of these kinds of approach are what Recanati calls 'pragmatic composition' and 'wrong format'. The 'pragmatic composition view' maintains that modulation is only necessary when a literal meaning must 'cohere' with other word meanings during semantic composition, while the 'wrong format view' is the position that words have overly rich or overly abstract meanings. The key difference between them is that for the pragmatic composition view, the linguistic meaning associated with a word can stand on its own while for the wrong format view, modulation is obligatory; both, however, have linguistic meaning play a role in the construction process (Recanati, 2004: 146-147).

In eliminativism, on the other hand, there is no need for fixed stable linguistic meanings to serve as input. Rather, following Hintzman, the word form (/iːt/ 'eat' in English or /kʌməʳ/ 'comer' in Spanish, for instance), and the context, *both linguistic and situational*, in which it appears, function as a composite cue that activates memory traces according to their similarity. The linguistic form is never alone, despite the fact that it is not associated with a determinate sense, because the context-sensitive mechanism at work in Hintzman's model, and, I would argue, in my own brand of eliminativism, fashions a cue that includes both the word and its context, that is, the word form is *not* extracted from the context of the utterance or the situation in general in which it appears. And there is no reason to assume that the memory traces capture *only* the meanings (i.e., the gist) to which a certain form gave rise; rather, *anything relevant in the episode as an experience* can be captured in the memory trace, including, I would argue, a trace of social

aspects of an interpretation that led to an occasion-specific meaning. For instance, in the 'soûl' example above, there came a time when the younger generations used this word *with* a negative connotation, and the older generations had to figure out this novel use. I claim that when memory traces capture an experience with language, they also capture social aspects of use; particularly, for instance, if a hearer experiences an instance of 'soûl', or 'intelligent' in an unexpected context. Speakers, mostly unconsciously, naturally assign high values to information they gather from their conversations that ensures they 'stay in tune' with the use of their language communities, and this, I presume, is captured in the memory traces of the episodes, – this is, of course, *in addition to* the fact that memory traces can affect behaviour simply by accumulation.

Assuming that rich contextual factors are present both in the cue and in the memory traces integrates context-dependence into the very heart of the model. Sensitivity to context is ubiquitous both because the traces stored in memory are 'contextualised senses', or, in other words, *context rich senses*, rather than 'core' abstract senses; and, because the *context at hand* is captured in the cue, it is the context *in its widest sense*, with whatever word forms are being used, whatever common ground the speakers share, and whatever aims are being undertaken, that is *broadcast* to all the memory traces in secondary memory simultaneously thereby activating stored traces according to their similarity not only to the word form (e.g. /iːt/ 'eat') but also in accordance with any other aspect of the context (from emotional tones to particularities of the situation). For instance, in the 'Aqui los pobres no comen' example ('Here the poor do not eat', example 4, § 3.3.1), it is important to keep in mind that it is uttered by my friend who is visiting London. The common ground he and I share includes, for instance, when our last meal was, how much it cost, among many other things. This wealth of information, together with the speaker's specific tone of voice, can serve as a very powerful cue to activate memory traces of other utterances with similarities to this one and lead me to *recognise* the playfully plaintive quality of the tone the speaker is using. Thus, the playfully plaintive tone itself can serve as a cue and retrieve other instances of statements made in

similar circumstances with this tone of voice. All of this accompanies the actual words in the utterance and, I would argue, is at least one of the factors that steers the interpretation process away from any possibility of taking the individual elements, 'the poor' or 'eat', to refer simply, '*literally'*, to *the poor* or *the act of eating*. In this account, any similarity with previous utterances can be exploited, and the assumption is that there are myriad clues to the speaker's meaning in *this* utterance which are to be found in memory traces of previous utterances the hearer has encountered. The foreseeable objection now is that to suppose that this myriad of very rich, *contextualised* memory traces of previous experiences with language can be *selectively* activated is to suppose that there is a very powerful mechanism of selection. In the previous section, I discussed how Hintzman's model avoids the need for any kind of executive control mechanism to manage memory traces; this is also true with regard to his construal of how particularly *relevant* memory traces can be highly activated.

In fact, I would argue that Hintzman's account of the mechanisms of cue and retrieval allows a simple and elegant explanation as to why the 'echo', or, in other words, the contents of what resonates back when a probe is broadcast to all memory traces, is so particularly relevant to the task at hand. Memory traces are activated according to their *similarity* to the probe, and, importantly, as first discussed in chapter 4 (§ 4.2.3), similarity requires a framework to determine *in what respect* one thing is to be similar to another. Hintzman provides that framework: in his account, the probe that cues memory is constructed *online*, he describes it as an 'active representation' in primary memory (1986: 412) and it includes anything relevant to the subject at that moment of his experience. I take this to mean that the constraints for similarity, or, in other words, the *framework* for the similarity judgements the model depends on, are *part of* the probe or *emerge* as the probe interacts with what it activates. To get a glimpse of possible consequences for the language theorist, consider the effects of cue and retrieval on, for instance, the notion of polysemy. In the traditional account, where words are supposed to have core, fixed, literal meanings, polysemy is a sort of exception to the rule: it's a *single* form that has a

number of related senses. A classic example is the word 'newspaper' which can refer either to *the publisher*, or to *the publication*, or to *a specific copy* that someone is reading. By definition, the senses are supposed to be related and it is assumed that one sense represents the 'core' while the others are derivative. Polysemy, however, is far from unproblematic; one of the difficulties is, as Nunberg (1979) wrote, that distinguishing between core senses and derived senses, or senses a word *has* and senses it only *takes on* as a matter of the context it appears in, seems arbitrary, since we have 'no empirical grounds' to say which is which (p. 174). The distinction between semantics and pragmatics, however, *depends* on there being clear distinctions between, what in the terms of this thesis, I would call context-independent word meanings, on the one hand, and contextualised senses, on the other. The advantage of an eliminativist approach should be clear: if a context-independent or core meaning is not presupposed, then, variations in what a particular form means in a particular context are to be *expected*, and, contrary to traditional accounts, there is simply no exception to be explained with the help of the notion of polysemy; rather, the need for the notion itself seems to vanish. I presented a related claim in the section on Peter Bosch (chapter 3, § 3.5.2), who similarly claims that the traditional picture of polysemy is 'misleading' and should be replaced by his account of the 'context dependence of predicate expressions'. Like Hintzman, and in line with my brand of eliminativism, Bosch criticises the strong tendency of traditional approaches to fail to fully appreciate the role played by contextual factors in determining occasion-specific word meanings. The rejection of the notion of polysemy has much to do with Hintzman's mechanism of cue and retrieval because, even if 'polysemy' is 'deconstructed', something remains to be explained: 'indeterminacy', that is, instances in which it is not clear to the hearer whether 'the newspaper' refers to *the publisher* or *the publication*.

Hintzman's (1986) contribution is that since a word or expression functions as a probe for memory, it can activate different tightly knit subsets of traces *giving the impression* that a single word has separate but related meanings. However, when a word like 'newspaper' is actually encountered

in the context of a conversational exchange, the cue contains sufficient contextual information to only *highly* activate memory traces from the relevant subset.[136] In this account, indeterminacy (or 'unresolved ambiguity') is only encountered in cases where, for some reason, the context is insufficient to selectively activate a single tightly knit subset of traces and the echo that resonates back is 'confused' instead of 'coherent'.[137] Usually, this situation simply does not arise because words are embedded in utterances, and utterances are embedded in conversational exchanges between interlocutors who share common ground and common goals in such a way as to ensure that a single sense emerges as the most highly activated one.

Finally, as a last example to illustrate the claims of my particular meaning eliminativist account, let's think back to the example of 'Be an angel and pick up some bread on your way home'. In contrast to the relevance-theoretic account presented in chapter 3, (§ 3.3.2.4), in my account, it is not a linguistically-specified meaning that allows access to information associated with the elements that make up the utterance. Rather, the words *and the phrases in which they appear*, are used as cues to activate memory traces on the basis of similarity. Imagine the following scenario: I have overseas houseguests, among them my aunt who cooks for us in the evening. While I am out, I get the 'Be an angel and pick up some bread on your way home' message on my phone. A whole phrase can serve as a cue; in this instance, it's 'Be an angel', together with the context in which it appears. This cue selectively activates other memory traces according to similarity. As it happens, my aunt frequently uses this formula when asking for a small service. The fact that the cue contains not only the phrase she uses but the phrase *in its context* and that memory traces are also rich in contextual

---

[136] This explanation of 'polysemy' easily extends to 'homonymy', which Hintzman describes as an extreme case of polysemy (1986: 423); if the same word-form has two or more quite dissociated populations of traces then it is all the easier to avoid unresolved ambiguities in use.

[137] I come back to another application for this idea that the echo that resonates back from memory can be 'confused' or 'coherent' in the conclusion following this chapter.

details, means that the traces selected for activation will overwhelmingly have to do with doing someone, *my aunt in particular*, a small service.

This recognition might not seem a great feat accomplished by memory traces, but it is important because of what it suggests regarding the interpretive process. First, notice the amount of processing involved in the example above and compare with the example of 'angel' discussed in chapter 3. This is not because I am selecting an easy example for myself, but rather because I want to insist on one of the roles memory traces play in conversational exchanges. One of the main tenets of the exemplar-based norm theory which supports my eliminativist account is that memory traces *accumulate* and affect our behaviour. For language processing, this means that the way we interpret utterances is affected by the store of memory traces an utterance activates. In the account presented in chapter 3, the process of utterance comprehension of an utterance with 'angel' involves a core or 'literal' encoded meaning for 'angel' and the modulation of the encoded concept ANGEL in the creation of an ad hoc concept ANGEL*. Critically, however, despite the fact that this kind of ad hoc concept creation is assumed to occur frequently in modulating encoded concepts, in the particular example presented in chapter 3, the construction of the ad hoc concept proceeds as if no ad hoc creation involving ANGEL had ever occurred before. In other words, it is as if the process of forming the ad hoc concept ANGEL* from a canonical 'atomic' concept ANGEL was done from scratch, without the help of any previous occurrences in which ANGEL required modulation. In my eliminativist account, on the contrary, the assumption is rather that 'angel', *together with whatever context it appears in*, works as a cue to activate prior instances of the use of 'angel' from memory. This process has no need for a fixed input since similarity to the current instance of 'angel' activates many memory traces and highly activates only the most relevant.[138] In the case above, the cue highly activates traces of a specific

[138] Memory traces of previous utterances include the linguistic form, the contexts in which they have appeared, and the interpretations that pragmatic expectations of those occurrences legitimated. The interpretational process both creates new ad hoc concepts, or occasion-specific senses, and benefits from previous occurrences.

idiolect, again, recognising a specific expression or wording used by one individual. This results in an interpretation that, much like previous examples in this section (e.g., 'Here the poor do not eat'), steers the interpretation away from taking the elements to refer in any simple, direct or 'literal' way. In the particular instance described, the frequency of 'Be an angel', as a set expression of which there are *tens of thousands* in the English language, is so frequent in my particular database that my interpretation could almost *overlook* the occurrence of the individual word 'angel'.[139] The memory traces activated by the cue 'Be an angel' in the context given here overwhelmingly activate memory traces that aggregate to 'Please do me a service'. In this situation, I would probably have more to figure out concerning the kind of bread she expects me to buy. Nothing in the 'literal' meaning of bread, *if one existed* would be helpful; rather, I would use my episodic memory to relate my aunt and recent mentions of bread. Aided my common sense, I would conjecture that she wants fresh baguettes like the ones she's gotten used to since she arrived. Notice that if I respond 'The bakery is closed', it is equally the common ground between us, the context in which this exchange is taking place and her common sense, which would guide her to understand that 'the particular bakery in our town is closed' and not that 'there is *one and only one* bakery in existence and it is closed'. This is why, to discuss word meaning in context, utterances must come before sentences, and speaker intentions before context-independent word meanings.

The situation would, of course, be very different if the utterance were 'Sally is an angel', and I were hearing this predicated of Sally for the first time.[140] It is also an interesting exercise to imagine this utterance in a context in which nothing steers me away from a 'literal' construal of 'angel'; recall that words do not have to have standing, context-independent

---

[139] Of course the form is still registered: I encounter the English 'angel' and not the Italian 'angelo'. Eliminativism does not deny that some forms belong to some languages rather than others, it rather doubts that they can be associated with fixed, context-independent meanings.

[140] I refer to Robyn Carston's recent very interesting paper on the differences in processing of different types of metaphors (Carston 2010a).

meanings for us to *act* mostly as if they did within our language community. In this situation, I might even consider, despite knowing that Sally is a human being, that there is something fundamentally different about Sally that makes her particularly angel-like. I would argue that the kind of psychological essentialism I described in chapter 4 (§ 4.3.4) is helpful here, I can both believe that there is something angelic about Sally and leave that essence an open slot.

Parenthetically, the frequency with which this latter situation occurs, together with our natural tendency to converge, is what keeps the meanings of our words *relatively* stable. Nonetheless, my position is that this makes no difference to the fact that there are countless instances stored in memory, both individual memory and collective memory, in which a woman is referred to as an *angel* and which would, therefore, suffice for figuring out an occasion-specific meaning without the need of context-independent, fixed word meanings for 'angel'. Shakespeare has Romeo say of Juliet 'O, speak again, bright Angell: For thou art as glorious as a winged messenger of heaven' already in the 16th century, so figurative uses of 'angel' are by no means new.[141] However, turning to the dictionary for conventional (or even 'attested' senses) is not the solution. As illustrated above with arguments in favour of the rejection of the notion of polysemy, I do not think multiplying the senses that a word is taken to *have* (in some fixed, reified way) is a promising solution.

To sum up where we have gotten to before moving onto the conclusion: I now take myself to have responded to the challenge of proposing a positive account of how an eliminativist view of word meaning would work. It is only a sketch of what much more time and careful work could possibly develop into a full-fledged theory of how word meaning in context works, but I hope that, at the very least, a justification of the eliminativist approach now seems reasonably plausible. In this chapter, I have mostly done this by underlining

[141] On the dangers of minimising the cultural and historical dimensions of metaphors, see Vincent Nyckees (2000, 2007, 2008).

how psychological models of memory support eliminativism. The position presented in this last section can be summed up as follows: as soon as we give up the notion that words have context-independent, fixed meanings or encode concepts, we see that it is not that words map directly onto atomic concepts but rather that lexical forms launch retrieval processes that build ad hoc concepts to be those concepts that the lexical forms find themselves mapped to in those circumstances. So it is not that meaning is underspecified, it is rather that meaning does not work by mapping onto anything pre-established, whether specified or underspecified.

Chapter 6: Conclusion

My global aim, in this thesis, has been to present arguments in favour of a radically contextualist, i.e. eliminativist, account of word meaning. I have argued from a *theoretical* perspective, first by laying out the current situation in the discussion on context-dependence, in which I identify a general trend towards the acceptance of context-dependence as far more pervasive in language understanding than traditionally thought, even among those opposing contextualism, and, second, by bringing together theorists from diverse fields who have adopted radical contextualism, or are, I have argued, moving in that direction. From a *technical* perspective, I have argued that both radical contextualism in general, and meaning eliminativism in particular, are strongly supported by psychological models. I suggest that as a result of taking both of these perspectives, and their respective lines of argumentation together, meaning eliminativism emerges less as the marginal position it has sometimes been taken to be and, despite its radical claims, more as a strong and viable alternative among accounts of word meaning.

Eliminativism maintains that word meanings exist *only* as occasion-specific senses. These claims directly contradict the general assumption in traditional semantics, and in the product tradition more broadly, that words have stable, context-independent meanings and that sentence meaning is roughly the product of the syntactical combination of the meanings of its parts. Throughout this thesis, I have tried not to underestimate the very drastic consequences for theorising on word meaning in context that accepting meaning eliminativism would entail. With this in mind, I undertook a step-by-step presentation of arguments and counterarguments leading to my own proposal. In the introductory chapter, I focused on presenting the traditional view of concepts, which, because of its profound influence, marked the discussion of both theoretical and empirical studies for decades even after its demise. Since much of the work in the thesis would oppose the different traditions and assumptions regarding the study of concepts and word meaning that emerged after the fall of the traditional

definitional account, I proceeded, in the following chapter, with a presentation of the predominant cognitive framework for the study of concepts. In that chapter, I focused on the perspective of one particular philosopher, Jerry Fodor, and his influential (atomistic) theory of concepts; this in turn provided the background for the presentation of relevance theory's construal of concepts. I finally turned to the topic of word meaning from the perspective of pragmatics in chapter 3; my aim in the first half of that chapter was to provide background for the current discussion in *contemporary cognitive* pragmatics of the *role of context*, or, otherwise, of the *ubiquity* of context-dependence. Once this background was set, however, I identified it as the new consensus in the field and argued that pushing the logic that had led theorists thus far *further* would reveal that the new consensus is sufficiently incompatible with the traditional semantics framework to justify breaking away from it. The eliminativism I advocate is still an extreme position, but, I hope, given the arguments provided in chapter 3, it no longer appears as such an outlier.

Anticipating certain objections, I turned my attention to a careful presentation of the well-established psychological models that support eliminativism in chapters 4 and 5. Among other things, chapter 4's aim was to reveal the true extent of context dependence: it goes well beyond the kind of context dependence encountered in occasion-specific word meaning construction; it is actually ubiquitous in all kinds of interpretation. Finally, my overall objective in chapter 5 was to give a full account of Hintzman's multiple-trace memory model in order to counter one of the main objections I anticipated to an eliminativist account: that eliminativism cannot offer a full-blown account of word meaning in context because once it has rejected fixed inputs to occasion-specific meaning construction processes, it does not have a *positive* account of word meaning in context to offer; the worry, in other words, is that while eliminativism has arguments to reject the more traditional picture, it has no full-blown account to replace it. So, in the last section of the chapter on memory, I focused on drawing out the consequences of Hintzman's model for an account of word meaning in

context and integrating arguments from previous sections to give a full picture of the account I propose.

Assuming now that I have made some progress in presenting a viable eliminativist account of word meaning, I turn to a speculative presentation of how eliminativism would deal with a completely 'new' word, a topic left over from the discussion in the previous chapter. The challenge is that if eliminativism relies on cue and retrieval mechanisms that activate memory traces of a particular word-form in its previous contexts, how would it work if a word were encountered for the very first time? As the example with 'gene' below is designed to illustrate, I think the account is flexible enough to accommodate such a circumstance, which, after all, is probably more frequent than we imagine.

Before I begin, however, I must mention and put aside the much more common phenomenon of encountering 'new' words that have such a degree of transparency that they would not serve my purposes. For instance, consider two new entries recorded by the *Oxford English Dictionary* in the last ten years: 'catastrophise', and 'automagically'. From reading or hearing 'catastrophise', even on the very first encounter, the word-form can serve to activate a host of memory traces that will be relevant to the task at hand not only because they are similar to the cue, but because the act of figuring out what a speaker means by what he says is very frequent as a cognitive operation and computations involved are only slightly different for a 'new' word. The echo from the probe could then include not only memory traces with instances of the very similar 'catastrophe', but also other noun-verb pairs, such as 'drama – dramatise' that exemplify noun-verb relations, and, finally other verbs ending in '-ise'. There is an abundance of information in memory that could help make sense of 'catastrophise'. The key to the proposal I have put forth is that only the most relevant traces for the purpose at hand are highly activated, and, following norm theory, they 'aggregate' to create a 'framework of evaluation' for the task at hand, or, in other words, memory traces are selectively activated to create an occasion-specific meaning for catastrophise. The same

goes for 'automagically', here, what is interesting is the profusion of events that the word 'automatically' can be applied to in our high-tech world. This, according to multiple-trace memory, translates as an abundance of this word in our individual and collective memory stores. 'Automagically', as a cue to probe memory activates both the high-tech and the magical to express something some speaker or writer felt these domains shared. I put these type of examples aside since what I am after is a completely new word, or, at the very least, a word we can imagine appearing as a 'new' word in the English language.

With this in mind, I have chosen 'gene', which, when it first appeared, at least, would been mostly uncontaminated by any previous uses. Furthermore, 'gene' serves my particular purposes well on various other accounts. First of all, because despite coming from the very technical domain of genetics, it has already been adopted into everyday language and is often subject to quite loose uses. Finally, the account of how this word is used in everyday speech is compatible with two prominent ideas in this thesis: Putnam's externalism and psychological essentialism. There are scientists who figure as experts on matters of what the object-out-there-in-the-world which we call a *gene* is and their knowledge trickles down to the general population. As I hope my examples will leave clear, however, it is not necessary to *know* very much about the scientific technical term 'gene' to competently use or understand certain expressions with the word 'gene'. Furthermore, as suggested in the section on psychological essentialism, people are generally willing to admit both that they themselves are not experts on what a gene *is* and that there is *probably something* that scientists have discovered, or are looking for, that singularly defines or determines what a gene is. The word has been adopted into everyday speech and the metaphysical reality of genes is presumed.

Very briefly, just as background for what follows, the word was coined to refer to the material entities of heredity discovered by Mendel. Before the word 'gene' was coined, the word 'genetic' already existed, but was not a word belonging to everyday speech. It was used to speak of origin, development and common evolutionary origin *but not heredity*. The

scientific discovery of genes forever changed this and consequently today, the notion of heredity is also attached to 'genetic'.[142]

A final reason this word serves my particular purposes is that it balances a relatively recent introduction into everyday speech with both a high familiarity and, except for experts, little technical knowledge of what the actual things-out-there-in-the-world *genes* are. I hope this will facilitate the reader's task of imagining someone hearing the word *for the first time* in a conversational exchange and having to *figure out* what it could mean.

Imagine that after spending Christmas with her partner Peter's extended family Mary asks him 'Why does everyone in your family know how to cook so well?' and Peter answers 'It's in our genes'. Suppose Mary does not know anything about genes, she has never heard the word used before and so has no memory traces with the particular form /dʒiːn/ 'gene', or any other sufficiently similar form, like /dʒɨnɛtɪk/ 'genetic', in her store of memory traces. In my account, she still has quite a bit to go on. She expects an answer to her question concerning why a particular skill is so well represented in Peter's family, so she can include this when she fashions the cue that serves to probe her memory. She also has the sentence form that Peter has used. To reflect that the word is utterly unknown to her, let's suppose the linguistic form represented is something like this: 'It's in our ––––––(s)'. Notice that, although Mary has no idea what the word 'gene' means (in this or *any* context), she can still tentatively interpret the 's' at the end as a sign of a plural.

In my eliminativist account, the partial form (with its open slot) and contextual clues are 'broadcast' to the entirety of Mary's memory store and, based on similarities, some memory traces resonate back. Of course, it is mere speculation to say *which* traces would be activated in such a scenario, and describing them succinctly when they are holistic experiences rich in detail is complex, but, the objective is to give an idea of the process, so I put these caveats aside for now. Let us suppose that Mary is attempting to build

---

[142] For instance, before this change in the meaning of genetic, today's 'developmental psychology' was termed 'genetic psychology' (*Oxford English Dictionary* –online version).

an occasion-specific meaning for the unknown word (and an interpretation of the utterance as a whole) on the basis of what she does know: the topic of conversation: Peter's sense of 'our' for things his family shares. This notion of *sharing* is well represented already in the cue Mary can construct to probe her memory since the response she is expecting is something like *what* it is that Peter's family shares, and the form Peter uses suggests that 'there is something that is in their ––––(s)'; it would be reasonable to suppose it is something they share. The echo that resonates back could include memory traces of utterances that have included expressions such as 'our blood', 'our ancestors', 'our roots', 'our traditions', 'our history'; or, using not only the notion of sharing, but also the linguistic form used, traces of utterances such as 'It's in our upbringing'. 'It's in our traditions', 'It's in our interests' (these suggestions are, of course, not exhaustive). The probe might not activate a very tight-knit subset of traces to do with families and sharing but this is unproblematic because Mary can suspend the creation of an actual occasion-specific word meaning for 'genes' on this occasion. If she feels that nothing too important depends on a fuller understanding of Peter's intended meaning, she can still keep the whole episode as a trace in memory *with whatever little sense she has made of the utterance*. Perhaps she can ask Peter for clarification, like Waismann about the utterance predicting 'intelligence' of a dog, or she can simply register the episode. If she encounters the word again soon, this initial trace will serve a purpose; if the word is never again encountered, after some time the trace will fade, ultimately becoming irretrievable (like all those words we thought we had learned while on vacation abroad…).

Notice that the traces that are activated do not *stand in* for an occasion-specific meaning; rather, they are more or less strongly activated in order to be subsumed, or 'scanned and summarised' in the terms of norm theory, in order to give Mary a framework for her interpretation. Notice also that there are *other things* that families share such as quirks, pets, hereditary good or bad looks, diseases. I do not mention these above because, given the context, Mary is unlikely to find them relevant.

Now I propose to imagine that the above initial conversation took place in the 1970s, a circumstance that would explain why it was the first time Mary heard this word, and focus on some specifics to do with the fact that memory traces involving the word 'gene' accumulated not only in her memory but also in the collective memory of her language community, with particular consequences. With the help of Google's Ngram project, a massive online database containing approximately 4% of all books ever printed, the appearance of the word *gene* can be visualised and its frequency monitored up to the year 2008.[143]

On Google Ngrams, we can visualise the exponential growth of the word 'gene' (see figure 3 in the annexe). Notice that it appears around the 1920s and remains relatively low in frequency until around the year 1970 when we can speculate that its contexts of use widened as it was adopted into everyday speech. This rise in frequency culminates around the year 2000, and the upward trend evens out in recent years. 'Ngrams', as these graphs are known, also allow us to see the word 'gene' paired with, for instance, 'good' and 'bad' (see figure 4). Occurrences of the phrases 'good genes' and 'bad genes' first appeared, like the word 'gene' on its own, around the 1920s, their rise in frequency follows that of 'gene' and 'genes' until the end of the 1980s. From 1990 to just after the year 2000, 'bad genes' *doubles* in frequency, 'good genes' shows exponential growth from 1990 to just after the year 2000. This pattern of explosive frequency is quite common among expressions with 'gene'. To illustrate this, I have also included Ngrams for the expressions 'gene pool' (figure 5), 'faulty genes', 'genetic disorder', 'genetic mutation', and 'cancer gene' (figure 6) and, finally, 'gay gene' (figure 7), which I do not discuss in detail here (but see Urquiza, 2011).

[143] Google Ngrams (http://books.google.com/ngrams) is a convenient way of visualising general trends in word use. It must be used with precaution because it contains two types of errors: those linked to the accuracy of character recognition (the data base is linked to Google's scanning project) and, more important to us, the year attributed to the source is sometimes mistaken. For the very general trends that interest us here, however, these errors are not an issue. For a detailed presentation of Google Ngrams, see Michel et al (2011).

My point is that with these graphs in hand, we can easily imagine how instances of the word 'gene' appeared and became relatively frequent in a short space of time. Each encounter with 'gene', alone or paired with another word, in a construction already encountered, or in a new construction, is recorded in memory as a memory trace. We can also suppose that initial encounters will resemble Mary's in that a probe including the information at hand returns an echo that is informative, although, since the word is not well represented in memory, it cannot activate a tight-knit subset of traces. In Hintzman's terms, this is a 'confused' instead of 'coherent' echo (see chapter 5, section 5.5). However, arguably, what resonates back does not need to be perfectly 'coherent' to serve in the next encounter. Each encounter activates some traces on the basis of similarity to the situation at hand and, in the terms of norm theory, 'generates' a norm, or a framework of evaluation to serve in the interpretation. Even in situations where this framework of evaluation fails to produce an interpretation, and the utterance comprehension process is abandoned (because cognitive efforts are outweighing cognitive effects) *this episode is recorded in memory*. This might frequently be the case at the beginning of the frequency trend; yet, as encounters with the word add up, attempts at interpretation, *and their results*, also add up. All of this is captured in memory and can be 'scanned and summarised' by retrieval mechanisms that selectively activate only those memory traces that are relevant to the context at hand. Eventually, successful interpretations accumulate in the form of Recanati's contextualised senses. As contextualised senses accumulate, a certain semantic potential appears, the hearer/reader feels he knows what the word means; even if he still has relatively little scientific knowledge of genes.

A point of contention might be whether it is justified to use Google's Ngrams to hypothesise about memory traces accumulating in a language community. I can only offer some suggestive thoughts in favour of this new tool but I stress that I use it only as a way of visualising trends. The raw data that the Ngram viewer uses to illustrate the frequency of words over time is the

relative frequency of these words in printed books. To arrive at their numbers, the team at Google counted how often a word appeared in the publications of a particular year and then divided it by the total number of words also appearing in the works of that year (Michel, 2011). How well does this reveal the underlying frequency of use by people in conversations? And, even more difficult, how well does this reveal the kind of populations of memory traces with regards to a particular word that my account assumes? Perhaps frequency levels on Ngrams do not relate to memory trace population sizes directly, but, I would argue, the sheer size of the corpus employed (5 million books and a corpus of 11 billion words for the year 2000, for instance) means that the trends are extremely robust and that trajectories as clear as the ones pictured in figures 1 and 2 deserve an explanation. One of the directions in which I would like to take my future research is into more corpus-based studies. The Ngrams provided here are only an illustration, but I feel that corpora hold great potential for the work of the language theorist who takes seriously the role of actual instances of use in determining word meanings.

In this last section, I have attempted to illustrate meaning eliminativism from the perspective of a pragmatic utterance comprehension process. I have focused on the fact that, since in my account, speakers and hearers rely heavily on memory traces of past utterances to build the occasion-specific word meanings their current situation requires, an interesting way to put the account under pressure is to imagine a scenario where there are no memory traces of the particular language form (e.g. 'gene') available. I hope that a meaning eliminativist account that includes memory traces now appears sufficiently robust to cope with such a situation.

Annexe

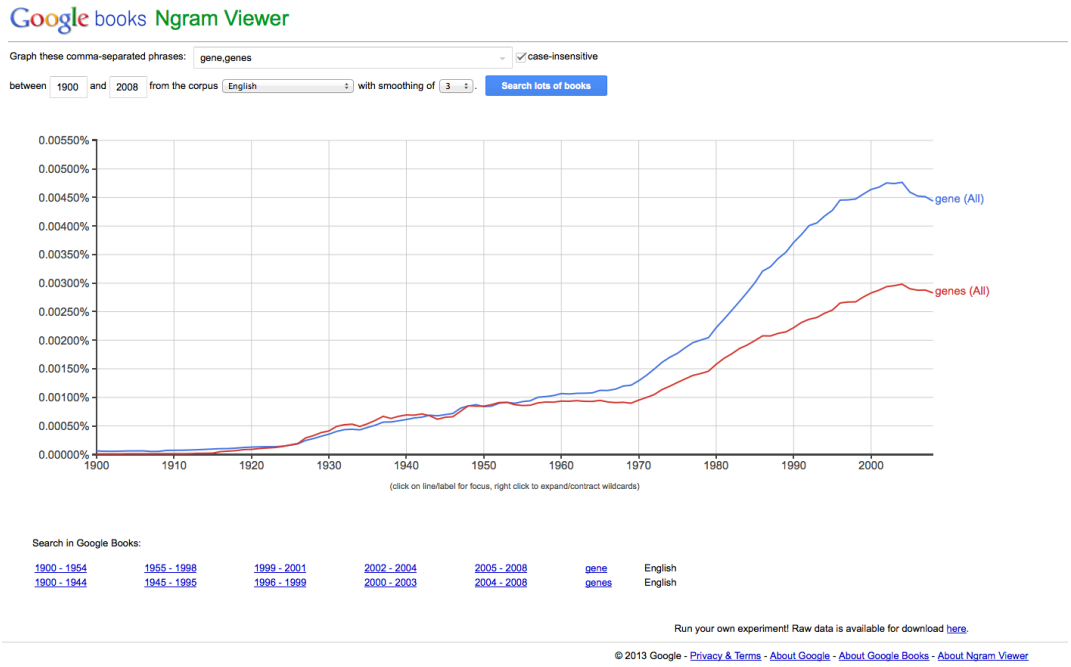

Figure 3: Ngram for ‘gene’ and ‘genes’

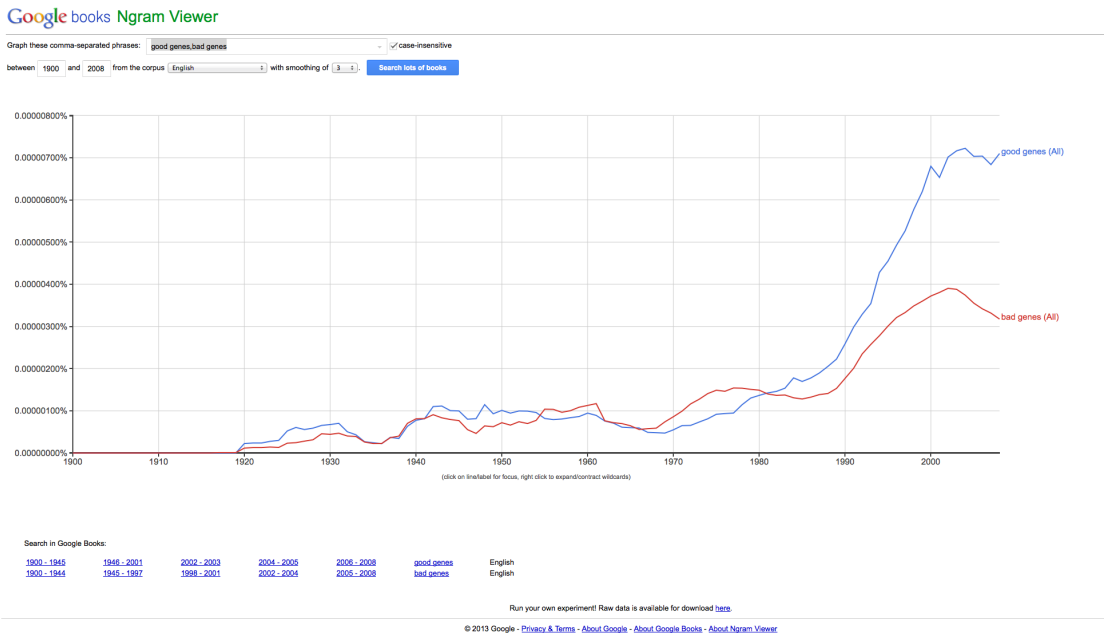

Figure 4: Ngram for ‘good genes’ ‘bad genes’

Figure 5: Ngram for ‘gene pool’

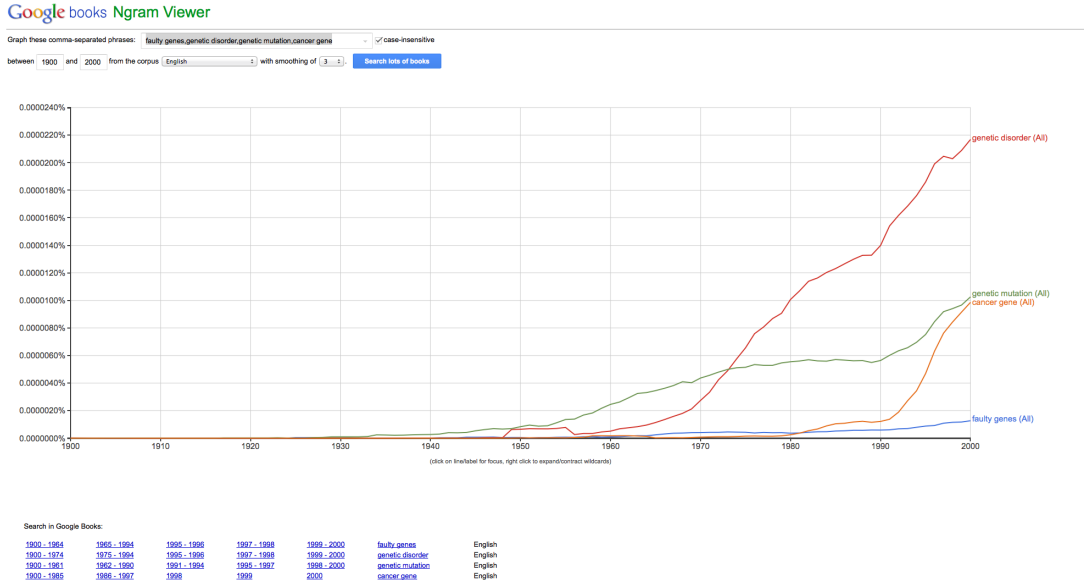

Figure 6: Ngram for ‘faulty genes’, ‘genetic disorder’, ‘genetic mutation’, ‘cancer gene’

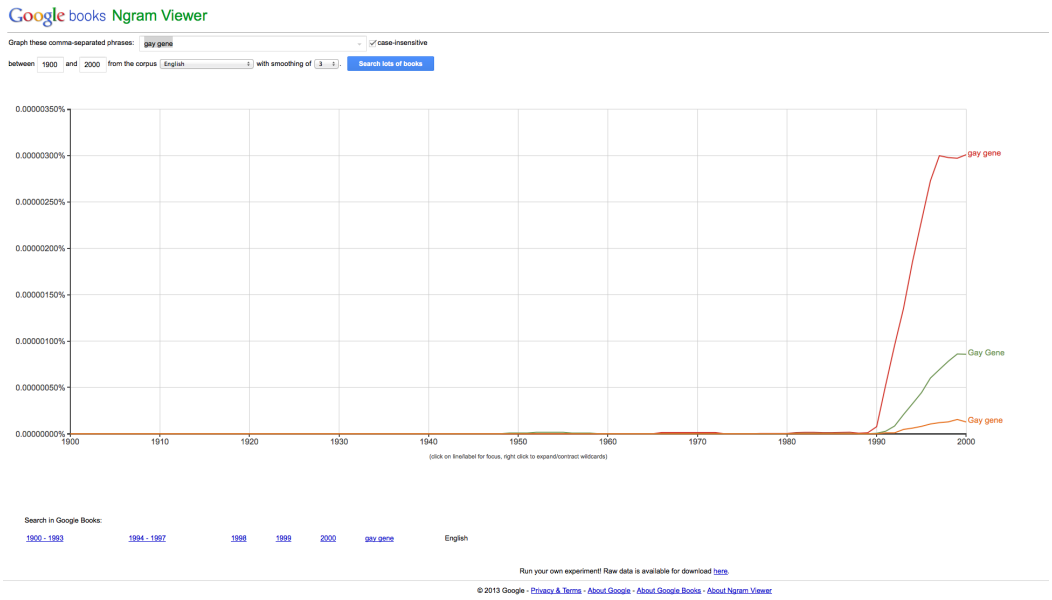

Figure 7: Ngram for ‘gay gene’

References

Armstrong, Sharon L., Gleitman, Lila R., & Gleitman, Henry. (1983). What some concepts might not be. *Cognition, 13*(3), 263-308.

Asher, Nicholas. (2011). *Lexical Meaning in Context: A Web of Words*. Cambridge: Cambridge University Press.

Assimakopoulos, Stavros. (2012). On encoded lexical meaning: philosophical and psychological perspectives. *Humana.Mente Journal of Philosophical Studies, 23*, 17-35.

Atkinson, Richard C., & Shiffrin, Richard M. (1971). The control processes of short-term memory: Report for The Institute for Mathematical Studies in the Social Sciences. Stanford University.

Bach, Kent. (1994). Conversational impliciture. *Mind & Language, 9*(2), 124-162.

Bach, Kent. (2003). Minding the gap. In C. Bianchi (ed.), *The Semantics/Pragmatics Distinction* (pp. 27-43). Stanford: CSLI Publications.

Bach, Kent. (2005). Context *ex Machina*. In Z. G. Szabó (ed.), *Semantics Versus Pragmatics* (pp. 15-44). Oxford: Oxford University Press.

Bar, Moshe. (2009a). The proactive brain: memory for predictions. *Philosophical Transactions of the Royal Society B-Biological Sciences, 364*(1521), 1235-1243.

Bar, Moshe. (2009b). Predictions: a universal principle in the operation of the human brain. *Philosophical Transactions of the Royal Society B-Biological Sciences, 364*(1521), 1181-1182.

Bar, Moshe. (2011). *Predictions in The Brain: Using Our Past to Generate a Future*. Oxford: Oxford University Press.

Barsalou, Lawrence W. (1982). Context-dependent and context-independent information in concepts. *Memory and Cognition, 10*(1), 82-93.

Barsalou, Lawrence W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology-Learning Memory and Cognition, 11*(4), 629-654.

Barsalou, Lawrence W. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser (ed.), *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization* (pp. 101-140). Cambridge: Cambridge University Press.

Barsalou, Lawrence W. (1993). Flexibility, structure, and linguistic vagary in concepts: Manifestations of a compositional system of perceptual symbols. In A. F. Collins, S. E. Gathercole, M. A. Conway & P. E. Morris (eds.), *Theories of Memory* (pp. 29-101). Hove: Lawrence Erlbaum.

Barsalou, Lawrence W. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society B: Biological Sciences, 358*(1435), 1177-1187.

Barsalou, Lawrence W. (2005). Abstraction as dynamic interpretation in perceptual symbol systems. In L. Gershkoff-Stowe & D. H. Rakison (eds.), *Building Object Categories in Developmental Time* (pp. 389-431). London: Lawrence Erlbaum.

Barsalou, Lawrence W. (2008). Grounded cognition. *Annual Review of Psychology, 59*(1), 617-645.

Barsalou, Lawrence W., & Sewell, Daniel R. (1984). Constructing representations of categories from different points of view *Emory Cognition Report # 2*: Emory University, Atlanta, Georgia.

Barsalou, Lawrence W., Sewell, Daniel R., & Ballato, Susan R. (1986 (under review)). The instability of categories as measured by graded structure: Emory University, Atlanta, Georgia.

Barsalou, Lawrence W., Yeh, Wenchi, Luka, Barbara J., Olseth, Karen L., Mix, Kelly S., & Wu, Ling-Ling. (1993). Concepts and meaning. In K. Beals, G. Cooke, D. Kathman, K. E. McCullough, S. Kita & D. Testen (eds.), *Chicago Linguistics Society 29: Papers from the Parasession on Conceptual Representations* (pp. 23-61). University of Chicago: Chicago Linguistics Society.

Bartlett, Frederic Charles. (1920). Some experiments on the reproduction of folk stories. *Folk-Lore, 31*, 30-47.

Bartlett, Frederic Charles. (1928). An experiment upon repeated reproduction. *Journal of General Psychology, 1*, 54-63.

Bartlett, Frederic Charles. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge: Cambridge University Press.

Beckner, Clay, Ellis, Nick C., Blythe, Richard, Holland, John, Bybee, Joan, Ke, Jinyun, …Five Graces, Group. (2009). Language is a complex adaptive system: position paper. *Language Learning, 59*, 1-26.

Berg, Jonathan. (2002). Is semantics still possible? *Journal of Pragmatics, 34*(4), 349-359.

Bianchi, Claudia (ed.). (2003). *The Semantics/Pragmatics Distinction: Proceedings from WOC 2002*: CSLI Publications.

Blakemore, Diane. (1987). *Semantic Constraints on Relevance*. Oxford: Blackwell.

Bloom, Paul. (1996). Intention, history, and artifact concepts. *Cognition, 60*(1), 1-29.

Bloom, Paul. (2000). *How Children Learn the Meanings of Words*. London: MIT Press.

Bloom, Paul. (2002). Mindreading, communication and the learning of names for things. *Mind & Language, 17*(1-2), 37-54.

Bloom, Paul. (2010). *How Pleasure Works : The New Science of Why We Like What We Like*. London: Bodley Head.

Borg, Emma. (2004). *Minimal Semantics*. Oxford: Clarendon.

Borg, Emma. (2012). *Pursuing Meaning*. Oxford: Oxford University Press.

Bosch, Peter. (2007). Productivity, polysemy, and predicate indexicality. In B. D. T. Cate & H. W. Zeevat (eds.), *Logic, Language, and Computation* (pp. 58-71). Berlin: Springer-Verlag Berlin.

Bosch, Peter. (2009). Predicate indexicality and context dependence. In P. DeBrabanter & M. Kissine (eds.), *Utterance Interpretation and Cognitive Models. Current Research in the Semantics/Pragmatics Interface* (Vol. 20, pp. 99-126). Oxford: Emerald Group Publishing Limited.

Brennan, Susan E., & Clark, Herbert H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology-Learning Memory and Cognition, 22*(6), 1482-1493.

Broadbent, Donald E. (1958). *Perception and Communication*. New York: Pergamon Press.

Burton-Roberts, Noel. (2005). Robyn Carston on semantics, pragmatics and encoding. *Journal of Linguistics, 41*(2), 389-407.

Burton-Roberts, Noel (ed.). (2007). *Pragmatics*. Basingstoke: Palgrave Macmillan.

Bybee, Joan. (2000). The phonology of the lexicon. Evidence from lexical diffusion. In M. Barlow & S. Kemmer (eds.), *Usage-Based Models of Language* (pp. 65-85). Stanford: CSLI Publications.

Bybee, Joan. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change, 14*, 261-290.

Bybee, Joan. (2003). Cognitive processes in grammaticalization. In M. Tomasello (ed.), *New Psychology of Language, Vol. 2* (pp. 145-167).

Bybee, Joan. (2007). Constructions at work: The nature of generalization in language. *Journal of Child Language, 34*(3), 692-697.

Bybee, Joan. (2010). *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Cain, Mark J. (2002). *Fodor: Language, Mind, and Philosophy*. Cambridge: Polity.

Cappelen, Herman, & Lepore, Ernest. (2005). *Insensitive Semantics: A Defense of Semantic Minimalism and Speech Act Pluralism*. Oxford: Blackwell.

Cappelen, Herman, & Lepore, Ernest. (2006). Précis of *Insensitive Semantics*. *Philosophy and Phenomenological Research, 73*(2), 425-434.

Cappelen, Herman, & Lepore, Ernest. (2007). Relevance Theory and shared content. In N. Burton-Roberts (ed.), *Pragmatics* (pp. 115-135). Basingstoke: Palgrave Macmillan.

Carnap. (1956). *Meaning and Necessity: a Study in Semantics and Modal Logic*. Chicago: University of Chicago Press.

Carruthers, Peter, & Boucher, Jill (eds.). (1998). *Language and Thought: Interdisciplinary Themes*. Cambridge: Cambridge University Press.

Carston, Robyn. (1988). Implicature, explicature and truth-theoretic semantics. In R. M. Kempson (ed.), *Mental Representations: The Interface Between Language and Reality* (pp. 155-181). Cambridge: Cambridge University Press.

Carston, Robyn. (2002a). *Thoughts and Utterances: the Pragmatics of Explicit Communication*. Oxford: Blackwell.

Carston, Robyn. (2002b). Linguistic meaning, communicated meaning and cognitive pragmatics. *Mind & Language, 17*(1-2), 127-148.

Carston, Robyn. (2002c). Metaphor, ad hoc concepts and word meaning - more questions than answers. *UCL Working Papers in Linguistics, 14*, 83-105.

Carston, Robyn. (2004). Relevance Theory and the Saying/Implicating Distinction. In L. R. Horn & G. L. Ward (eds.), *The Handbook of Pragmatics* (pp. 633-656). Oxford: Blackwell.

Carston, Robyn. (2007). How many pragmatic systems are there? In M. J. Frápoli (ed.), *Saying, Meaning and Referring: Essays on François Recanati's Philosophy of Language* (pp. 18-48). Basingstoke: Palgrave Macmillan.

Carston, Robyn. (2008a). Linguistic communication and the semantics/pragmatics distinction. *Synthese, 165* (Special Issue on the Semantics/Pragmatics Distinction), 321-345.

Carston, Robyn. (2008b). *Minimal Semantics* -by Emma Borg. *Mind & Language, 23*(3), 359-367.

Carston, Robyn. (2009a). The explicit/implicit distinction in pragmatics and the limits of explicit communication. *International Review of Pragmatics, 1*(1), 35-62.

Carston, Robyn. (2009b). Relevance Theory: contextualism or pragmaticism? *UCL Working Papers in Linguistics, 21*, 19-26.

Carston, Robyn. (2010a). Metaphor: Ad hoc concepts, literal meaning and mental images. *Proceedings of the Aristotelian Society, 110*(3), 295-321.

Carston, Robyn. (2010b). Explicit communication and 'free' pragmatic enrichment. In Soria Casaverde & Romero (eds.), *Explicit Communication: Robyn Carston's Pragmatics* (pp. 217-287): Palgrave.

Carston, Robyn. (2010c). Lexical pragmatics, ad hoc concepts and metaphor: from a relevance theory perspective. *Italian Journal of Linguistics, 22*(1), 153 - 180.

Carston, Robyn. (2012). Word meaning and concept expressed. *The Linguistic Review, 29*(4), 607-623.

Carston, Robyn. (2013). Word meaning, what is said, and explicature. In C. Penco & F. Domaneschi (eds.), *What Is Said and What Is Not*. Stanford: CSLI Publications.

Carston, Robyn, & Hall, Alison. (2012). Implicature and explicature. In H.-J. Schmid (ed.), *Cognitive Pragmatics* (Vol. 4). Berlin: De Gruyter Mouton.

Chomsky, Noam. (1959). A review of BF Skinner's *Verbal Behavior*. *Language, 35*(1), 26-58.

Clark, Billy. (2013). *Relevance Theory*. Cambridge: Cambridge University Press.

Clark, Herbert H. (1992). *Arenas of Language Use*. London: University of Chicago Press & Center for the Study of Language and Information.

Clark, Herbert H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Collins, Alan F., Gathercole, Susan E., Conway, Martin A., & Morris, Peter E. (1993). *Theories of Memory*. Hove: Lawrence Erlbaum.

Dawkins, Richard. (2011a). *The Magic of Reality: How We Know What's Really True*. London: Bantam Press.

Dawkins, Richard. (2011b). The tyranny of the discontinuous mind. *New Statesman, 19*, 54-57.

Dawkins, Richard. (2014). Essentialism. *Edge*. http://edge.org/response-detail/25366

Dupoux, Emmanuel, & Mehler, Jacques (eds.). (2001). *Language, Brain, and Cognitive Development; Essays in Honor of Jacques Mehler*. London: MIT Press.

Edelman, Shimon, & Shahbazi, Reza. (2012). Renewing the respect for similarity. *Frontiers in Computational Neuroscience, 6*, 1-19.

Escandell Vidal, M. Victoria, Leonetti, Manuel, & Ahern, Aoife (eds.). (2011). *Procedural Meaning: Problems and Perspectives*. Bingley: Emerald.

Fodor, Janet Dean, Fodor, Jerry A., & Garrett, Merrill F. (1975). Psychological unreality of semantic representations. *Linguistic Inquiry, 6*(4), 515-531.

Fodor, Jerry A. (1975). *The Language of Thought*. New York: Crowell.

Fodor, Jerry A. (1980). The mind-body problem. *Scientific American, 244*(1), 114-123.

Fodor, Jerry A. (1981a). *Representations : Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, Massachusetts: MIT Press.

Fodor, Jerry A. (1981b). The present status of the innateness controversy *Representations. Philosophical Essays on the Foundations of Cognitive Science* (pp. 257-316). Cambridge, Massachusetts: MIT Press.

Fodor, Jerry A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. London: MIT Press.

Fodor, Jerry A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, Massachusetts: MIT Press.

Fodor, Jerry A. (1990a). Information and representation. In P. P. Hanson (ed.), *Information, Language, and Cognition* (pp. 175-190). Vancouver: Vancouver Studies in Cognitive Science.

Fodor, Jerry A. (1990b). *A Theory of Content And Other Essays*. Cambridge, Massachusetts: MIT Press.

Fodor, Jerry A. (1998). *Concepts: Where Cognitive Science Went Wrong*. Oxford: Clarendon Press.

Fodor, Jerry A. (1999). Information and representation. In E. Margolis & S. Laurence (eds.), *Concepts: Core Readings* (pp. 513-524). London: MIT Press.

Fodor, Jerry A. (2000). *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. London: MIT Press.

Fodor, Jerry A. (2001). Language, thought and compositionality. *Mind & Language, 16*(1), 1-15.

Fodor, Jerry A. (2003). *Hume Variations*. Oxford: Clarendon Press.

Fodor, Jerry A. (2004). Having concepts: a brief refutation of the twentieth century. *Mind & Language, 19*(1), 29-47.

Fodor, Jerry A. (2008). *LOT 2: The Language of Thought Revisited*. Oxford: Clarendon Press.

Fodor, Jerry A. (2009). So what's so good about Pylyshyn? In D. Dedrick & L. Trick (eds.), *Computation, Cognition and Pylyshyn* (pp. ix-xvii). Cambridge, Massachusetts: MIT Press.

Fodor, Jerry A., Garrett, Merrill F., Walker, Edward C.T., & Parkes, Cornelia H. (1980). Against definitions. *Cognition, 8*(3), 263-367.

Fodor, Jerry A., & Lepore, Ernest. (1992). *Holism: A Shopper's Guide*. Oxford: Blackwell.

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

Fodor, Jerry A., & Lepore, Ernest. (1998). The emptiness of the lexicon: Reflections on James Pustejovsky's *The Generative Lexicon*. *Linguistic Inquiry, 29*(2), 269-288.

Fodor, Jerry A., & Lepore, Ernest. (2002). *The Compositionality Papers*. Oxford: Oxford University Press.

Fowler, Carol A., & Magnuson, James S. (2012). Speech perception. In M. Spivey, K. McRae & M. Joanisse (eds.), *The Cambridge Handbook of Psycholinguistics* (pp. 3-25). New York: Cambridge University Press.

Frápolli, María José. (2007). *Saying, Meaning and Referring: Essays on François Recanati's Philosophy of Language*. Basingstoke: Palgrave Macmillan.

Geeraerts, Dirk. (1997). *Diachronic Prototype Semantics: A Contribution to Historical Lexicology*. Oxford: Oxford University Press.

Geeraerts, Dirk. (2010). *Theories of Lexical Semantics*. Oxford: Oxford University Press.

Geeraerts, Dirk, & Cuyckens, Hubert (eds.). (2007). *The Oxford Handbook of Cognitive Linguistics*. Oxford: Oxford University Press.

Gillund, Gary, & Shiffrin, Richard M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*(1), 1-67.

Giora, Rachel. (1999). On the priority of salient meanings: Studies of literal and figurative language. *Journal of Pragmatics, 31*(7), 919-929.

Goldinger, Stephen D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251-279.

Grice, Herbert Paul. (1957). Meaning. *Philosophical Review, 66*, 377-388.

Grice, Herbert Paul. (1967). *Logic and conversation*. Paper presented at the William James Lectures.

Grice, Herbert Paul. (1989). *Studies in the Way of Words*. Cambridge, Massachusetts: Harvard University Press.

Groefsema, Marjolein. (2007). Concepts and word meaning in relevance theory. In N. Burton-Roberts (ed.), *Pragmatics* (pp. 136-157). Basingstoke: Palgrave McMillan.

Gurevich, Olga, Johnson, Matthew A., & Goldberg, Adele E. (2010). Incidental verbatim memory for language. *Language and Cognition, 2*(1), 45-78.

Hall, Alison. (2008). Free enrichment or hidden indexicals? *Mind & Language, 23*(4), 426-456.

Hall, Alison. (2009). Subsentential utterances, ellipsis, and pragmatic enrichment. *Pragmatics & Cognition, 17*(2), 222-250.

Hall, Alison. (2011). Ad hoc concepts: Atomic or decompositional?. *UCL Working Papers in Linguistics*, 23, 1-10.

Hampton, James A. (2006). Concepts as prototypes. In B. H. Ross (ed.), *Psychology of Learning and Motivation: Advances in Research and Theory, Vol 46* (Vol. 46, pp. 79-113). San Diego: Elsevier Academic Press Inc.

Hampton, James A. (*forthcoming*). Categories, prototypes and exemplars. In N. Reimer (ed.), *Routledge Handbook of Semantics*: Routledge.

Hanson, Norwood Russell. (1958). *Patterns of Discovery. An Inquiry into the Conceptual Foundations of Science*. Cambridge: Cambridge University Press.

Hebb, Donald Olding. (2002 (1949)). *The Organization of Behavior. A Neuropsychological Theory*. London: Lawrence Erlbaum.

Hintzman, Douglas L. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review, 93*(4), 411-428.

Hintzman, Douglas L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*(4), 528-551.

Hintzman, Douglas L. (2011). Research strategy in the study of memory: fads, fallacies, and the search for the "coordinates of truth". *Perspectives on Psychological Science, 6*(3), 253-271.

Hintzman, Douglas L., & Ludlam, Genevieve. (1980). Differential forgetting of prototypes and old instances - simulation by an exemplar-based classification model. *Memory & Cognition, 8*(4), 378-382.

Homa, Donald, Sterling, Sharon, & Trepel, Lawrence. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology-Human Learning and Memory, 7*(6), 418-439.

Homa, Donald, & Vosburg, Richard. (1976). Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Human Learning and Memory, 2*(3), 322-330.

Hopper, Paul J., & Traugott, Elizabeth Closs. (2003). *Grammaticalization* (2nd ed.). Cambridge: Cambridge University Press.

Horn, Laurence R., & Ward, Gregory L. (eds.). (2004). *The Handbook of Pragmatics*. Oxford: Blackwell.

Jackendoff, Ray. (1992). *Languages of the Mind: Essays on Mental Representation*. London: MIT Press.

James, William. (2010 (1890)). *The Principles of Psychology* eBooks@Adelaide.

Johnson-Laird, Philip N, Robins, C, & Velicogna, Lucy. (1974). Memory for words. *Nature, 251*(5477), 704-705.

Johnson-Laird, Philip N, & Stevenson, Rosemary. (1970). Memory for syntax. *Nature, 227*(5256), 412-412.

Kahneman, Daniel. (2011). *Thinking, Fast and Slow*. London: Allen Lane.

Kahneman, Daniel, & Miller, Dale T. (1986). Norm theory - comparing reality to its alternatives. *Psychological Review, 93*(2), 136-153.

Kant, Immanuel. (1977 (1783)). *Prolegomena To Any Future Metaphysics That Will Be Able To Come Forward As Science: The Paul Carus Translation. Extensively Revised by James W. Ellington.* (2nd ed.). Indianapolis: Hackett.

Kaplan, David. (1989 (1977)). Demonstratives: an essay on the semantics, logic, metaphysics, and epistemology of demonstratives and other indexicals. In J. Almog, J. Perry, H. K. Wettstein & D. Kaplan (eds.), *Themes from Kaplan* (pp. 481-563).

Keil, Frank C. (1989). *Concepts, Kinds, and Cognitive Development*. Cambridge, Massachusetts: MIT Press.

King, Jeffrey C., & Stanley, Jason. (2005). Semantics, pragmatics and the role of semantic content. In Z. G. Szabo (ed.), Semantics Versus Pragmatics (pp. 111-164). Oxford: Oxford University Press.

Kintsch, Walter. (1974). *The Representation of Meaning in Memory*. Hillsdale, New Jersey: Lawrence Erlbaum.

Kintsch, Walter, & Mangalath, Praful. (2011). The construction of meaning. *Topics in Cognitive Science, 3*(2), 346-370.

Koriat, Asher, & Goldsmith, Morris. (1996). Memory metaphors and the real-life/laboratory controversy: Correspondence versus storehouse conceptions of memory. *Behavioral and Brain Sciences, 19*(2), 167-188.

Kripke, Saul. (1977). Speaker's reference and semantic reference. *Midwest Studies In Philosophy, 2*(1), 255-276.

Laurence, Stephen, & Margolis, Eric. (1999). Concepts and cognitive science. In E. Margolis & S. Laurence (eds.), *Concepts: Core Readings* (pp. 3-81). London: MIT Press.

Levy, Elena, & Nelson, Katherine. (1994). Words in discourse - a dialectal approach to the acquisition of meaning in use. *Journal of Child Language, 21*(2), 367-389.

Machery, Edouard. (2010). Précis of *Doing Without Concepts*. *Behavioral and Brain Sciences, 33*, 195-244.

Margolis, Eric, & Laurence, Stephen (eds.). (1999). *Concepts: Core Readings*. London: MIT Press.

Margolis, Eric, & Laurence, Stephen. (2007). The ontology of concepts - abstract objects or mental representations? *Noûs, 41*(4), 561-593.

McClelland, James L., & Rumelhart, David E. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Vol. 2. Psychological and Biological Models.* Cambridge: MIT Press.

McCloskey, Michael E., & Glucksberg, Sam. (1978). Natural categories - well defined or fuzzy sets. *Memory & Cognition, 6*(4), 462-472.

Martinich, Aloysius P. (2008). *The Philosophy of Language* (5th ed.). Oxford: Oxford University Press.

Medin, Douglas L., Goldstone, Robert L., & Gentner, Dedre. (1993). Respects for similarity. *Psychological Review, 100*(2), 254-278.

Medin, Douglas L., & Schaffer, Marguerite M. (1978). Context theory of classification learning. *Psychological Review, 85*(3), 207-238.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., . . . Google Books Team. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176-182.

Miller, George A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-97.

Miller, George A. (1978). Semantic relations among words *Linguistic Theory and Psychological Reality* (pp. 60-118).

Monk, Ray. (1991). *Ludwig Wittgenstein: The Duty of Genius*. London: Vintage.

Moravcsik, Julius. (1994). Is snow white? In P. Humphreys (ed.), *Patrick Suppes: Scientific Philosopher* (pp. 71-85): Kluwer.

Murphy, Gregory L. (2004). *The Big Book of Concepts* (2nd ed.). London: MIT Press.

Murphy, Gregory L., & Medin, Douglas. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*(3), 289-316.

Murphy, Gregory L., & Shapiro, A. M. (1994). Forgetting of verbatim information in discourse. *Memory & Cognition, 22*(1), 85-94.

Neath, Ian, & Surprenant, Aimée M. (2003). *Human Memory: An Introduction to Research, Data, and Theory* (2nd ed.). London: Wadsworth.

Neisser, Ulric. (1978). *Memory: What are the important questions*. Paper presented at the International Conference on Practical Aspects of Memory, Cardiff, Wales.

Neisser, Ulric (ed.). (1987). *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*. Cambridge: Cambridge University Press.

Neisser, Ulric. (1988). What is ordinary memory the memory of? In U. Neisser & E. Winograd (eds.), *Remembering Reconsidered: Ecological and Traditional Approaches to the Study of Memory* (pp. 356-373). Cambridge: Cambridge University Press.

Neisser, Ulric. (1988). New vistas in the study of memory. In U. Neisser & E. Winograd (eds.), *Remembering Reconsidered: Ecological and Traditional Approaches to the Study of Memory : 2nd Emory Cognition Project Conference : Papers* (pp. 1-). Cambridge: Cambridge University Press.

Neisser, Ulric, & Winograd, Eugene (eds.). (1988). *Remembering Reconsidered: Ecological and Traditional Approaches to the Study of Memory: 2nd Emory Cognition Project Conference: Papers*. Cambridge: Cambridge University Press.

Nosofsky, Robert M., Little, Daniel R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review, 118*(2), 280-315.

Nunberg, Geoffrey. (1979). The non-uniqueness of semantic solutions: polysemy. *Linguistics and Philosophy, 3*(2), 143-184.

Nyckees, Vincent. (2000). Quelle est la langue des métaphores? In C. Détrie (ed.), *Cahiers de praxématique* (Vol. 35, pp. 115-139). Presses Universitaires de la Méditerranée: Publications de l'Université Paul Valéry, Montpellier 3, Presses Universitaires de la Méditerranée.

Nyckees, Vincent. (2007). La cognition humaine saisie par le langage : De la sémantique cognitive au médiationnisme. *Corela, Numéros spéciaux, Cognition, discours, contextes*. http://edel.univ-poitiers.fr/corela/document.php?id=1659

Nyckees, Vincent. (2008). Le sens figuré en langue et en discours: les sources linguistiques de l'énonciation métaphorique. In C. Sakai & D. Struve (eds.), *Regards sur la métaphore, entre Orient et Occident*: Éditions P. Piquier.

Olson, David R. (1970). Language and thought - aspects of a cognitive theory of semantics. *Psychological Review, 77*(4), 257-273.

Perry, John. (1998). *Indexicals, contexts and unarticulated constituents.* Paper presented at the Proceedings of the 1995 CSLI-Amsterdam Logic, Language and Computation Conference.

Perry, John, & Blackburn, Simon. (1986). Thought without representation. *Proceedings of the Aristotelian Society, Supplementary Volumes, 60*, 137-151+153-166.

Pierrehumbert, Janet B. (1990). Phonological and phonetic representation. *Journal of Phonetics, 18*(3), 375-394.

Pierrehumbert, Janet B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. J. Hopper (eds.), *Frequency and the Emergence of Linguistic Structure* (pp. 137-157).

Pinker, Steven. (1994). *The Language Instinct: The New Science of Language and Mind*. London: Allen Lane.

Pinker, Steven. (1994). The game of the name. *New York Times*.

Pinker, Steven. (2008). *The Stuff of Thought: Language as a Window into Human Nature*. London: Penguin.

Posner, Michael I., & Keele, Steven W. (1968). On genesis of abstract ideas. *Journal of Experimental Psychology, 77*(3P1), 353-&.

Posner, Michael I., & Keele, Steven W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology, 83*(2), 304-&.

Prinz, Jesse, & Clark, Andy. (2004). Putting concepts to work: Some thoughts for the twenty-first century. *Mind & Language, 19*(1), 57-69.

Prinz, Jesse J. (2002). *Furnishing the Mind: Concepts and their Perceptual Basis*. London: MIT Press.

Prinz, Jesse J. (2012). Regaining composure: A defense of prototype compositionality. In M. Werning, W. Hinzen & E. Machery (eds.), *The Oxford Handbook of Compositionality* (pp. 437-453). Oxford: Oxford University Press.

Putnam, Hilary. (1975). The meaning of "meaning". *Mind, Language and Reality*, 215-271.

Putnam, Hilary. (1981). *Reason, Truth and History*. Cambridge: Cambridge University Press.

Putnam, Hilary. (1999 (1970)). Is semantics possible? In E. Margolis & S. Laurence (eds.), *Concepts: Core Readings* (pp. 177-187). London: MIT Press

Putnam, Hilary. (2008). Meaning and reference. In A. P. Martinich (ed.), *The Philosophy of Language* (pp.306-313) Oxford: Oxford University Press.

Putnam, Hilary. (2013). The development of externalist semantics. *Theoria, 79*(3), 192-203.

Pylyshyn, Zenon W. (1973). What mind's eye tells mind's brain: a critique of mental imagery. *Psychological Bulletin, 80*(1), 1-24.

Pylyshyn, Zenon W. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge, Massachusetts: MIT Press.

Pylyshyn, Zenon W. (1999). Concepts: Where cognitive science went wrong. *Trends in Cognitive Sciences, 3*(2), 81-82.

Pylyshyn, Zenon W. (2003a). Return of the mental image: are there really pictures in the brain? *Trends in Cognitive Sciences, 7*(3), 113-118.

Pylyshyn, Zenon W. (2003b). *Seeing and Visualizing: It's Not What You Think*. London: MIT Press.

Pylyshyn, Zenon W. (2009). Perception, representation, and the world: The FINST that binds. In D. Dedrick & L. Trick (eds.), *Computation, Cognition, and Pylyshyn* (pp. 3-48). London: MIT Press.

Quillian, M. Ross. (1968). Semantic Memory. Cambridge Massachusetts: Air Force Cambridge Research Laboratories.

Raaijmakers, Jeroen G. W., & Shiffrin, Richard M. (1981). Search of associative memory. *Psychological Review, 88*(2), 93-134.

Ratcliff, Roger. (1978). Theory of memory retrieval. *Psychological Review, 85*(2), 59-108.

Ratcliff, Roger, & Murdock, Bennet B. (1976). Retrieval processes in recognition memory. *Psychological Review, 83*(3), 190-214.

Rayo, Agustin. (2013). A plea for semantic localism. *Noûs, 47*(4), 647-679.

Recanati, Francois. (1993). *Direct Reference: from Language to Thought*. Oxford: Blackwell.

Recanati, Francois. (1994). Contextualism and anti-contextualism in the philosophy of language. In S. Tsohatzidis (ed.), *Foundations of Speech Act Theory* (pp. 156-166). London: Routledge.

Recanati, François. (1998). Pragmatics. In E. Craig (ed.), Routledge Encyclopedia of Philosophy. London: Routledge. Retrieved August 19, 2014, from http://www.rep.routledge.com/article/UO28SECT7.

Recanati, François. (2001a). Déstabiliser le sens. *Revue internationale de philosophie*(2), 197-208.

Recanati, Francois. (2001b). What is said. *Synthese, 128*(1-2), 75-91.

Recanati, Francois. (2004). *Literal Meaning*. Cambridge: Cambridge University Press.

Recanati, Francois. (2010a). *Truth-Conditional Pragmatics*. Oxford: Oxford University Press.

Rey, Georges. (1983). Concepts and stereotypes. *Cognition, 15*(1–3), 237-262.

Rey, Georges. (1985). Concepts and conceptions - a reply. *Cognition, 19*(3), 297-303.

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

Rosch, Eleanor. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (ed.), *Cognitive Development and the Acquisition of Language* (pp. 111-144). Oxford: Academic Press.

Rosch, Eleanor. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology-General, 104*(3), 192-233.

Rosch, Eleanor, & Mervis, Carolyn B. (1975). Family resemblances - studies in internal structure of categories. *Cognitive Psychology, 7*(4), 573-605.

Rosch, Eleanor, Simpson, Carol, & Miller, R. Scott. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology-Human Perception and Performance, 2*(4), 491-502.

Roth, Emilie M., & Shoben, Edward J. (1983). The effect of context on the structure of categories. *Cognitive Psychology, 15*(3), 346-378.

Schacter, Daniel L. (1987). Implicit memory - history and current status. *Journal of Experimental Psychology-Learning Memory and Cognition, 13*(3), 501-518.

Schacter, Daniel L. (1992). Understanding implicit memory - a cognitive neuroscience approach. *American Psychologist, 47*(4), 559-569.

Schacter, Daniel L., Chiu, C. Y. Peter, & Ochsner, Kevin N. (1993). Implicit memory - a selective review. *Annual Review of Neuroscience, 16*, 159-182.

Schank, Roger C. (1982). *Dynamic Memory : a Theory of Reminding and Learning in Computers and People*. Cambridge: Cambridge University Press.

Schank, Roger C. (1999). *Dynamic Memory Revisited*. Cambridge: Cambridge University Press.

Schank, Roger C., & Abelson, Robert P. (1977). *Scripts, Plans, Goals and Understanding : an Inquiry into Human Knowledge Structures*. Hillsdale, New Jersey: Lawrence Erlbaum.

Schank, Roger C., Collins, Gregg C., & Hunter, Lawrence E. (1986). Transcending inductive category formation in learning. *Behavioral and Brain Sciences, 9*(4), 639-651.

Schmid, Hans-Jörg. (2008). New words in the mind: Concept-formation and entrenchment of neologisms. *Anglia-Zeitschrift Fur Englische Philologie, 126*(1), 1-36.

Schmid, Hans-Jörg (ed.). (2012). *Cognitive Pragmatics*. Berlin: De Gruyter Mouton.

Schneider, Susan. (2011). *The Language of Thought: A New Philosophical Direction*. London: MIT Press.

Scott, Kate. (2013). This and that: a procedural analysis. *Lingua, 131*, 49-65.

Searle, John R. (1978). Literal meaning. *Erkenntnis, 13*(1), 207-224.

Searle, John R. (1980). The background of meaning *Speech Act Theory and Pragmatics* (pp. 221-232). London: Reidel.

Searle, John R. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.

Searle, John R., Kiefer, Ferenc, & Bierwisch, Manfred (eds.). (1980). *Speech Act Theory and Pragmatics*. London: Reidel.

Shapiro, Lawrence. (2009). Things and places: how the mind connects with the world. *Mind, 118*(487), 1168-1174.

Shiffrin, Richard M., & Atkinson, Richard C. (1969). Storage and retrieval processes in long-term memory. *Psychological Review, 76*(2), 179-193.

Shiffrin, Richard M., & Steyvers, Mark. (1997). A model for recognition memory: REM-retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*(2), 145-166.

Smith, Edward, & Medin, Douglas. (1981). *Categories and Concepts*. Cambridge, Massachusetts: Harvard University Press.

Soria Casaverde, Maria Belen, & Romero, Esther (eds.). (2010). *Explicit Communication: Robyn Carston's Pragmatics*. Basingstoke: Palgrave Macmillan.

Sperber, Dan. (2001). In defense of massive modularity. In E. Dupoux & J. Mehler (eds.), *Language, Brain and Cognitive Development: Essays in Honor of Jacques Mehler* (Vol. 47).

Sperber, Dan. (2005). Modularity and relevance: How can a massively modular mind be flexible and context-sensitive? In P. Carruthers, S. Laurence & S. P. Stich (eds.), *The Innate Mind: Structure and Content* (Vol. 1, pp. 53). Oxford: Oxford University Press.

Sperber, Dan, Clement, Fabrice, Heintz, Christophe, Mascaro, Olivier, Mercier, Hugo, Origgi, Gloria, & Wilson, Deirdre. (2010). Epistemic vigilance. *Mind & Language, 25*(4), 359-393.

Sperber, Dan, & Wilson, Deirdre. (1986/95). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Sperber, Dan, & Wilson, Deirdre. (1995). Postface. In D. Sperber & D. Wilson, *Relevance: Communication and Cognition* (2nd ed., pp. 255-279). Oxford: Blackwell.

Sperber, Dan, & Wilson, Deirdre. (1998). The mapping between the mental and the public lexicon. In P. Carruthers & J. Boucher (eds.), *Language and Thought: Interdisciplinary Themes* (pp. 184-200). Cambridge: Cambridge University Press.

Sperber, Dan, & Wilson, Deirdre. (2002). Pragmatics, modularity and mind-reading. *Mind & Language, 17*(1-2), 3-23.

Squire, Larry R. (1992). Declarative and nondeclarative memory - multiple brain systems supporting learning and memory. *Journal of Cognitive Neuroscience, 4*(3), 232-243.

Stanley, Jason. (2000). Context and logical form. *Linguistics and Philosophy, 23*(4), 391-434.

Steyvers, Mark, Griffiths, Thomas L., & Dennis, Simon. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences, 10*(7), 327-334.

Tomasello, Michael (ed.). (2003a). *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure volume 2*. London: Lawrence Erlbaum.

Tomasello, Michael. (2003b). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. London: Harvard University Press.

Tomasello, Michael. (2009). *Why We Cooperate*. Cambridge, Massachusetts: MIT Press.

Traugott, Elizabeth Closs. (1989). On the rise of epistemic meanings: an example of subjectification in semantic change. *Language, 65*(1), 31-55.

Traugott, Elizabeth Closs. (2004). Historical pragmatics. In L. R. Horn & G. Ward (eds.), *The Handbook of Pragmatics* (pp. 538-561). Oxford: Blackwell.

Traugott, Elizabeth Closs. (2010). Grammaticalization. In A. H. Jucker & I. Taavitsainen (eds.), *Historical Pragmatics* (pp. 97-126). Berlin: Walter de Gruyter.

Travis, Charles. (1985). On what is strictly speaking true. *Canadian Journal of Philosophy, 15*(2), 187-229.

Tulving, Endel. (1983). *Elements of Episodic Memory*. Oxford: Clarendon.

Tversky, Amos. (1977). Features of similarity. *Psychological Review, 84*(4), 327-352.

Urquiza, Catalina. (2011). Lexical pragmatics and memory traces. *UCL Working Papers in Linguistics, 23*, 47-68.

Waismann, Friedrich. (1951). Verifiability. In A. Flew (ed.), *Logic and language* (Vol. 1, pp. 117-144).

Werning, Markus, Hinzen, Wolfram, & Machery, Edouard (eds.). (2012). *The Oxford Handbook of Compositionality*. Oxford: Oxford University Press.

Wharton, Tim. (2003). Interjections, language and the 'showing-saying' continuum. *Pragmatics and Cognition, 11*(1), 39-91.

Wilson, Deirdre. (2003). Relevance and lexical pragmatics. *Italian Journal of Linguistics/Rivista di Linguistica, 15*(2), 273-286. (Special issue on pragmatics and the lexicon).

Wilson, Deirdre. (2005). New directions for research on pragmatics and modularity. *Lingua, 115*(8), 1129-1146.

Wilson, Deirdre. (2011). The conceptual-procedural distinction: Past, present and future. In M. V. Escandell Vidal, M. Leonetti & A. Ahern (eds.), *Procedural Meaning: Problems and Perspectives* (pp. 3-31). Bingley: Emerald.

Wilson, Deirdre, & Carston, Robyn. (2007). A unitary approach to lexical pragmatics: Relevance, inference and ad hoc concepts. In Burton-Roberts (ed.), *Pragmatics* (pp. 230-260). Basingstoke: Palgrave Macmillan.

Wilson, Deirdre, & Sperber, Dan. (1993). Linguistic form and relevance. *Lingua, 90*(1-2), 1-25.

Wilson, Deirdre, & Sperber, Dan. (2002). Truthfulness and relevance. *Mind, 111*(443), 583-632.

Wilson, Deirdre, & Sperber, Dan. (2004). Relevance Theory. In L. R. Horn & G. Ward (eds.), *The Handbook of Pragmatics* (pp. 607-632). Oxford: Blackwell.

Wilson, Deirdre, & Sperber, Dan. (2012). *Meaning and Relevance*. Cambridge: Cambridge University Press.

Wittgenstein, Ludwig. (1953). *Philosophical Investigations. Translated by G. E. M. Anscombe*. Oxford: Basil Blackwell.

Wray, Alison. (2013). Formulaic language. *Language Teaching, 46*, 316-319.