Word meaning, concepts and the representation of abstract entities from the perspective of radical pragmatics and semantic externalism

Georg Kjøll

Ph.D. thesis

Centre for the Study of Mind in Nature,
Department of Philosophy, Classics, History of Art and Ideas

Ph.D. programme in Linguistics
Faculty of Humanities
University of Oslo

November 2010

Acknowledgments

In the writing of this thesis, I have benefitted from the continuous support and encouragement of many good colleagues, friends and family.

First and foremost, I am extremely thankful for the vast amount of work Deirdre Wilson has put into the supervision process. Her careful reading and commenting on every new draft, and her always being willing to discuss and offer critical insight on the issues I have been thinking about, has made this project much better and more enjoyable than it otherwise would have been. With her patience, and her never failing to offer help, even with matters on which we disagree considerably, she has consistently surpassed all reasonable expectations anyone could have of a supervisor.

I gratefully acknowledge the encouragement of my co-supervisor, Jan Terje Faarlund, to whom I have always been able to turn for advise and a critical, linguistic perspective. Along with him, my colleagues at the Centre for the Study of Mind in Nature (CSMN) at the University of Oslo have been invaluable company and a source of inspiration throughout these last three years. I regard myself as lucky to have been part of such a vibrant research community, and its high level of activity and pleasant social environment have made the inherently lonely endeavour of writing a thesis feel a lot less solitary.

I want to thank the CSMN core group for believing in my project and, together with the Faculty of Humanities and the Department of Philosophy, Classics, History of Arts and Ideas at the University of Oslo and the Research Council of Norway for funding the work in this thesis.

A lot of people have read through previous drafts and very generously provided me with comments and advice: Endre Begby, Taina Bucher, Robyn Carston, Ingrid Lossius Falkum, Luan Huang, Terje Lohndal, Bjørg Nesje Nybø, Jonas Pfister, Mihaela Popa and Ivo Spira. I would like to thank Nicholas Allott in particular, who, in having offered his help as a reader and discussion partner from day 1, deserves acknowledgment and credit for his positive influence on my work.

I have relied on many more people for discussions and counsel, especially in my ventures into philosophy. I am grateful to Trine Antonsen, Julian Fink, Thiago Galery, Thomas Hodgson, Torfinn Huvenes, Anders Nes, Georges Rey, Anders Strand and Andreas Sveen for giving of their time to discuss word meaning with me. I extend a special thanks to Heine Holmen for always being willing to engage in discussion and thereby exposing my philosophical prejudices.

Much of the work in the upcoming chapters have been presented at various conferences and academic forums, and I thank the participants at the International Conference for the Philosophy of Language and Linguistics 2009 at the University of Łodz, the 2009 workshop on the Emergence of Intersubjectivity at the Center for Subjectivity Research, University of Copenhagen, the 2009 and 2010 Annual Meetings for the European Society for Philosophy and Psychology at the Central European University in Budapest and the Ruhr Universität Bochum/Essen and the Pragmatics Reading Group at the University College London for comments and interesting discussions.

Getting feedback from good colleagues on presentations at a series of internal events at the CSMN, among them the Language and Rationality Seminar, the CSMN Colloquium and the joint workshop on semantics with the University of Tromsø's Center for Advanced Study in Theoretical Linguistics, has served as a vital part of my academic training, contributing greatly to my getting the core ideas behind this thesis into shape. I also thank the students at the course on Pragmatics and Relevance Theory for stimulating discussions about linguistic underdeterminacy.

Finally I wish to thank my family, Camilla, Ivar and Lillian Kjøll and Taina Bucher, for support and encouragement throughout my education, and for the unfailing belief they have in me.

I thank Taina for love, companionship and for providing purpose and real meaning to my life. I am also grateful for her persuading me to come along on her research stay in New York City, where a substantial part of this thesis was written.

Introduction

This thesis is about mental and linguistic representation. The work in the upcoming chapters will attempt to answer the questions of how people understand words, how they are used to convey *meaning* and how this meaning is acquired.

Of course, this topic is in itself potentially bottomless. The idea that *words express meaning* is so central to the way we pre-theoretically take language to work, and its implications for theories of language and behaviour so great, that any study which attempts to address this question risks ending up in a linguistic and philosophical morass. So much has been said about so many aspects of word meaning, that it would be easy to spend the better part of the time and space available to do a PhD just trying to get an overview of the topic.

The strategy I have used in trying to avoid this becoming a Sisyphean task is to take a somewhat narrow focus in investigating the notion of word meaning. I will concentrate almost exclusively on two approaches to the issue, that of Relevance Theory (Sperber and Wilson 1995; Wilson and Sperber 2004) and informational semantics/conceptual atomism (Fodor 1998a; 2008). The two approaches have distinct goals in the study of word meaning: Relevance Theory is interested in how meaning is communicated in linguistic interaction, while informational semantics is more concerned with the nature of meaning independent of communicative contexts.

My aim in this thesis is to lay the foundation for a unitary account of how meaning is represented and communicated, drawing on the strengths and insights of each of the two accounts.

An unfortunate side effect of combining the strengths and insights of two independently developed theories is that one inevitably ends up with double the weaknesses as well. I will therefore spend quite a lot of time and effort defending some of the positive claims made by both Relevance Theory and informational semantics, discussing some of the main objections they face and considering potential obstacles to merging them. But even though such a merger means that I will end up with a wider array of counter-arguments to deal with, I think the pay-off will be comparatively greater.

The reason for this is that I think the two accounts complement each other in important ways. Relevance theoretic pragmatics takes as its object the process of communication, trying to explain how people interpret what others intend to communicate in a given communicative situation. But although meaning in communicative interaction is the pragmatician's central concern, Relevance Theory also relies on an explanation of how words have meaning

independently of any communicative situation. Intentional communication often (though not always) depends on words somehow providing the input to pragmatic processes, and it therefore seems to me that the story one can tell about interpretation is somewhat constrained by what one's theory of the semantic input is.

In their writings on the topic, Sperber and Wilson have relied on the idea that words encode *concepts* in describing both the input to pragmatic processes and the nature of word meaning outside any communicative situation. Sperber and Wilson see concepts as mental items which can occur independently of language and are not necessarily lexicalized, In their view, a word used in a specific communicative act will not necessarily convey the very concept it encodes, but may convey an inferentially related concept that is more general or more specific than the encoded one.

The idea that words used in communication somehow convey concepts has a long tradition in the philosophical literature, but Relevance Theory endorses a view most closely associated with Jerry Fodor, which treats these concepts as *atomic* and analyses their content in *externalist* terms, seeing them as (mainly) getting their content from standing in a relation to something in the world. Especially in recent relevance theoretic writings on *lexical pragmatics*, concepts have come to play a prominent role, although not much work has been done from the relevance theoretic perspective on investigating what concepts actually are and what cognitive work they are required to do other than being the meanings of words.

But given that conceptual semantics is ontologically prior to pragmatics, in the sense that there would be little need for pragmatics if there were no conceptual representations to communicate, one's views on the semantics of conceptual representations are likely to constrain one's pragmatics. In this thesis I aim to contribute, then, to a better understanding of the pragmatics of linguistic communication by considering what concepts have to be like in order to do the work required of them by the lexical pragmatic theory.

In addressing this question, I hope to fill what some see as a gap in Relevance Theory, and address worries that have been raised about the semantic foundation on which lexical pragmatics is built. It has been claimed, for instance, that there is a tension "between the adoption of an (atomistic and externalist) view of concepts (such as Fodor's) and the description that Relevance Theory in fact gives to concepts" (Reboul 2008: 524). Reboul claims that "This is an issue that has to be addressed if Relevance Theory is to continue to be of central importance in contemporary pragmatics" (*ibid*). A critical investigation of the informational semantic view of concepts and how well it squares with the central tenets of Relevance Theory will hopefully go some way towards dealing with these concerns.

The fact that concepts and mental representations are seen as ontologically prior to natural language makes investigations of informational semantics, on the other hand, independent of pragmatic theories. As Fodor is prone to point out, "there is no reason to suppose that 'how you think' or 'what you can think about' depends on what language you speak" (2008: 218). But Fodor also holds that informational semantics cannot be completely cut off from the communicative domain, since it needs to interact in some way with "Grice's communication theory; or, anyhow something like it." (Fodor 1998b: 68). Fodor claims that "An acceptable account of thought ought to say something about how thoughts are expressed; and if thoughts are what have content in the first instance, then it is natural to suppose that what communication communicates is the content of thoughts" (*ibid*).

But nowhere does Fodor offer a story about how a communicator goes from a thought to a natural language utterance, implementing some sort of procedure that the speakers/hearers of a language can use to translate between concepts and words (see Fodor 2008: 216, n26). This lack has provoked Fodor's critics to call for rules or procedures "by means of which [expressions] are mapped to a relevant region in this system of 'thought', acquiring its semantics there" (Hinzen 2007: 51, echoing Chomsky 2000: 176-177). What I will try to show is that pragmatics may help to address these concerns, thus complementing the Fodorian theory of concepts with important insights about how content can be communicated from thinker to thinker.

I will also try to show how pragmatics can help to explain how concepts expressing certain types of content are acquired. Though many concepts are formed on the basis of perceptual input (someone sees a dog and thereby acquires the concept DOG), according to Fodor's brand of externalism, much conceptual content is not plausibly sustained via perceptual mechanisms. All humans are capable of talking and thinking not only about things they can see and touch, but also about a host of other entities with which they have never had perceptual contact. I have in mind here so-called *abstract entities*, a category which includes such things as democracy, feminism, inflation, happiness, justice, love, etc. and could potentially be expanded to include many, many more.

Though I will show how Fodor's approach to semantics is neutral with respect to how a concept gets its content from mind-external input, Fodor (1998a: 75) claims that informational semantics is "untenable" unless one can show how concepts are formed or tokened as the result of a causal process. A considerable amount of the work in this thesis is therefore motivated by the challenge of explaining how concepts for a range of ontologically problematic entities are acquired – if not by direct perception. I will argue that some key

insights from the relevance theoretic view of communication as an exercise in *Theory of Mind* will help to shed light on how people rely on each other to sustain *semantic access* to entities that they have never perceived.

The approach I take to word meaning in this thesis is wholly constructive. Since I am acutely aware of the massive dimensions of the topic I have taken on, I will limit myself to giving a positive account, leaving rival theories (mostly) outside of the scope of my discussions. I do acknowledge, though, that it would have been interesting to look at some alternative approaches to mental and linguistic representation in the light of the suggestions I will make, thereby contextualizing my constructive story a bit more. I am thinking here in particular of other externalist approaches to semantics, such as Dretske's (1981) version of informational semantics, Millikan's (1993; 2000) teleological semantics and the so-called anti-individualist theories of content, of which Burge (1979; 1986) is the central proponent. Readers familiar with these views will recognize quite a few affinities with the account I will be defending, but a discussion of the overlaps and disparities between all these will unfortunately fall outside my scope here.

As regards the representation of abstract entities, it would also have been valuable to compare the line I take with what is perhaps the dominant linguistic approach to this topic: the conceptual metaphor/embodied cognition theories of Lakoff and Johnson (1980; 1999; Johnson 1987; see also Fauconnier and Turner 2002; Kövecses 1986; for a critique, see Murphy 1996; 1997). Lakoff and Johnson's claim that the meaning of terms designating non-perceivable objects is grounded in concrete, bodily experiences has become something of a standard treatment of the representation of entities like love, happiness and ideas. In ignoring this important literature, I am also disregarding a vast amount of theoretical and empirical research on the topic of mental representation.

This may be unfortunate, but I believe there are good reasons for limiting my scope in this way. Most importantly, a full treatment and critique of the conceptual metaphor approach to cognition, a research program that now spans thirty years and includes thousands of articles and monographs, would require much more time and space than I have in my thesis. And given that there are not many alternatives to the Lakoff and Johnson approach to the representation of abstract objects in the cognitive science/philosophical literature (Barsalou 1999; Prinz 2002 are notable externalist exceptions; but see Dove 2009: 418-423 for a critique), I think my energy is better spent on developing a constructive alternative from a different perspective.

The thesis is divided into two main parts. The first, which comprises chapters 1, 2 and 3, focuses on the nature of semantic content and its relation to words. Here I am interested in how words are i) used and understood in a context and ii) represented independently of communicative interaction. The second part, which comprises chapter 4, 5 and 6, looks at how content can be seen as individuated via a mind-world relation, and considers the nature of this relation. I will focus on two main questions: how concepts are acquired, and what in the external world supplies input to this acquisition process. I will be particularly concerned with how the mechanisms connecting the mind and the world are sustained in cases where the entities that are represented are not directly perceived or perceivable.

The structure of the individual chapters is as follows:

Chapter 1 introduces the idea of radical pragmatics as advocated by Relevance Theory (Sperber and Wilson 1995). I start by discussing a range of examples which I take to show that there is a significant gap between what is explicitly expressed by an utterance and the meaning of the sentence uttered. On the basis of these examples, which are often taken to demonstrate what has been called the *problem of linguistic underdeterminacy*, I will argue for the view that the goal of communication is an enlargement of "mutual cognitive environments", and that this does not necessarily entail that speaker and hearer should end up entertaining identical thoughts.

I defend Sperber and Wilson's view of communication against charges raised by Cappelen and Lepore (2007), who claim that Relevance Theory is committed to the thesis that content can never be shared between interlocutors. Discussing a particular case of communication where no exact thought sharing takes place, I show how Sperber and Wilson distinguish successful from unsuccessful communication in terms of the potential sharing/non-sharing of *contextual implications*.

In chapter 2, I investigate what approach to semantics might lay a suitable foundation for Relevance Theory. Starting from Sperber and Wilson's view that *concepts* (or conceptual representations) provide input to the pragmatic interpretation process, I ask what mechanisms might provide the concepts with content independently of communication. Fodor's (1998a; 2008) account of concepts serves as the starting point for answering this question, and I go through Fodor's treatment of the metaphysics of meaning as it is situated in the broader framework of Computational Theory of Mind. Outlining the explanatory labour concepts are supposed to do, I defend Fodor's *informational semantics/conceptual atomism*, which claims that the content of concepts is constituted solely by an appropriate meaning-making lawful connection between concepts and properties, against some of the main objections it faces. I

argue that these objections are all in principle answerable, but that many questions about the nature of semantic content still remain.

Some of these questions concern the relationship between words and concepts, and I argue towards the end of the chapter that data which shows substantial variation between lexical items across languages is hard to square with what Fodor labels *the publicity constraint* if one accepts the relevance theoretic assumption that words encode concepts.

In chapter 3, I discuss some ways in which the pragmatic theorist could deal with these worries. I argue against giving up the view of concepts as the context-independent meanings of words, and instead propose an alternative to the idea that words *encode* concepts. I suggest that the relationship between words and concepts may be better seen as one of *potential activation*, where any given word may give access to a range of concepts, or lead to the construction of a new one, depending on the context in which the word is tokened. The mechanism by which an appropriate concept is selected by hearers is seen, on this approach, as a semantically constrained pragmatic process driven by the search for relevance.

I defend this alternative conception of word meaning against some expected objections and discuss what is lost and what is gained by giving up on semantic encoding. I also compare this view with some other related accounts of word meaning, and outline some theoretical and methodological advantages that a theory which does not rely on a relationship of default mechanisms or semantic encoding in specifying word meaning would have over traditional approaches.

In chapter 4, I move away from questions about the metaphysics of meaning towards a consideration of how semantic content is sustained and concepts acquired. I start by looking at what I call *the argument from ontology*, which holds that objects in the world are too unruly and/or inaccessible to people's perceptual systems and therefore unsuitable to provide content to thoughts and utterances. Though some theorists (Chomsky 2000; Jackendoff 2002) take the argument to show that informational semantics is empirically and explanatorily inadequate, I defend the view that it does not apply to Fodor's theory of meaning once the scope and aim of informational semantics has been clarified.

Nevertheless, I take up the argument from ontology as a genuine challenge to informational semantics, and hold that it prompts the theorist to give a positive account of how people acquire concepts and sustain what Fodor calls *semantic access* to entities they cannot perceive. I go on to show how Fodor's theory of concepts provides a basic framework for understanding how concept acquisition can be mediated by either *deference* to other people or explicit theory construction. Using the representation of the entity inflation as a case

study, I show how Relevance Theory can contribute to fleshing out the idea of deference and thereby contribute to a fuller theory of concept acquisition.

In chapter 5 I take up some challenges faced by the account of concept acquisition I have outlined, concentrating on two problem cases. The first is the representation of fictional and/or metaphysically impossible entities such as Sherlock Holmes, elves and ghosts. The second is the representation of entities I call medium-abstract, in that the concepts representing them are not plausibly thought of as acquired on the basis of either direct perception or deference or theory-construction alone. Examples include emotion concepts, such as LOVE and HAPPINESS, and normative and moral concepts, such as JUSTICE and MORALLY RIGHT.

In contrast with Rey (2005a), who claims that externalism about concepts expressing unrealisable properties like *elfhood* is untenable for naturalistic theories of content, I suggest an informational semantic treatment of metaphysically impossible entities according to which concepts like ELF get their content from *mind-dependent* properties. In dealing with the problem of medium-abstract entities, I use the account of the word-concept relation outlined in chapter 3. My claim is that a number of concepts are potentially activated by a lexical compound such as 'morally right', some of which are acquired on the basis of deference or explicit theories, while others have an innate or broadly perceptual basis.

In chapter 6, I outline some implications of the story of concept acquisition I have proposed, and suggest that different types of acquisition processes lead to the formation of different types of concept, characterised by different properties. I suggest some empirical predictions that follow from this, and discuss the account of concepts proposed by Fodor and presupposed by Relevance Theory within a broader cognitive science perspective. I conclude that even though a consequence of the Fodorian notion of meaning is that concepts are not directly accessible to empirical research, having something in one's theoretical vocabulary which fills the crucial role that Fodorian concepts play is a necessary prerequisite for any type of psychological generalisation to take place.

Some notes on vocabulary, abbreviations and typographical conventions

RT – Relevance Theory
IS – informational semantics
CTM – Computational Theory of Mind
LoT – the Language of Thought (also referred to as *Mentalese*)
MoP – Modes of Presentation (also referred to as a concept's *syntactic form*)
ToM – Theory of Mind (also referred to as *mindreading*)
IA – Isomorphism Assumption (about the word-concept relation)

In this thesis, *radical pragmatics* will refer to theories which hold that most sentence tokens need to go through a range of pragmatic processes in order for the meaning it explicitly conveys to be recovered. *Relevance Theory* is one instance of a radical pragmatic theory.

I take *semantic externalist* theories to be those that claim content to be primarily constituted by a link between the mind and the world. *Informational semantics/conceptual atomism*, which is a strong construal of semantic externalism, holds that the mind-world relation *exhausts* semantic content.

Even though my starting point in this thesis is words and communication, a lot will be said on the topic of *concepts* in what follows. Concepts will be taken to be mental particulars, symbols in a mind-internal language that has both a content and a form (MoP).

I will follow the conventions of writing concepts in small capitals (DOG). Ad hoc concepts will be marked by an asterisk (DOG*). Properties will be annotated in italics (*being a dog*) while lexical items are in single quotes ('dog').

Part I:
Word meaning, concepts and communication

1. Relevance Theory, content similarity and communicative success[1]

...as exact as a resemblance
- Gertrude Stein

1.1. Introduction

This chapter approaches the analysis of word and sentence meaning from the perspective of radical pragmatics. It has two main aims: to present the relevance-theoretic account of communication, and to defend the view that successful linguistic interaction does not necessarily have to result in two interlocutors sharing thoughts.

I start by outlining the linguistic underdeterminacy thesis, which holds that in general, what is explicitly conveyed by an utterance goes well beyond the meaning of the sentence uttered. I present a few cases of what have been seen as instances of this phenomenon in the recent linguistic and philosophical literature, and briefly discuss two different strategies, one semantic and one pragmatic, that a theorist can pursue in analysing them.

Focusing on the pragmatic strategy, I show how Relevance Theory (Sperber and Wilson 1995) treats communication as a process governed by the search for relevance. According to Sperber and Wilson, linguistic meaning alone radically underdetermines what a speaker can explicitly communicate by uttering a sentence in a wide range of different situations. They analyse utterance interpretation as a process of developing a fragmentary sentence meaning into a hypothesis about the speaker's meaning, guided by a relevance-based comprehension heuristic.

In this chapter, I show how Sperber and Wilson's account suggests that human verbal communication is risky and potentially prone to errors, so there can be no guarantee that speaker and hearer will end up sharing identical thoughts as a result of their interaction. On this account, successful communication need not involve the exact reproduction of thoughts: all that matters is that the thoughts of speaker and hearer are similar enough for the purposes of that particular interaction. I discuss some objections raised by Cappelen and Lepore (2005; 2007) to similarity-based accounts of successful linguistic communication, and respond to charges of incoherence levelled at the radical pragmatic approach.

---

[1] A shorter version of this chapter is published as 'Content similarity and communicative success', in *International Review of Pragmatics* 2(1): 2010. Some of the work in sections 1.3 and 1.4 is drawn from the unpublished work in Kjoll (2007).

Towards the end of the chapter, I ask what distinguishes the relevance theoretic view of successful communication from other, more traditional accounts of communication as thought sharing. I conclude by suggesting that some conceptual ground-clearing remains to be done on both sides of the debate, and raise some questions about the view of sentence meaning as fragmentary and incomplete which remain to be answered by relevance theorists.

1.2. Meaning and communication

1.2.1. The problem of linguistic underdeterminacy

The words people utter in communication have a curious feature. There is something which makes them, when used and combined in the proper manner, particularly apt to convey a *speaker's meaning*. When I utter sentences such as the following:

1. I was here yesterday
2. The book is on the table

I can generally trust my hearer to understand what I wanted to put across by uttering these words. I can assume that, through some reliable process, my hearer will be able to interpret my utterance so as to create some kind of understanding between us.

But what are the features of words which allow this to take place? What is the reliable process underlying the generation of meaning and understanding? And how can a speaker know that the hearer has indeed grasped whatever it was that was being conveyed? A standard answer in the philosophical and linguistic literature is that (most of) the words which feature in utterances have meanings, and that when a hearer knows these meanings and the rules by which they are combined, it follows that he will also know what is being conveyed.

But how direct is the relationship between the meanings of words and what is conveyed by uttering them in concrete situations? It seems that, in many cases, knowing the meaning of the sentence uttered is not enough in itself to enable a hearer to recognise what the speaker intended to convey. Sometimes, factors that seem to be external to the linguistics of the utterance need to be taken into account. For instance, in interpreting some utterances, in addition to understanding the meanings of the words uttered, a hearer will need to take into account who his interlocutor is, where the utterance is taking place, and probably what day it is. This follows from the fact that the very same words 'I was here yesterday' can convey quite different meanings in different situations.

In uttering 'I was here yesterday' when standing in the music venue Le Poisson Rouge at 158 Bleecker Street, New York City the 15th of June 2010, what I convey is that Georg

Kjoll was at Le Poisson Rouge on the 14th of June 2010. If Keith Jarrett utters these words when standing on the stage of the Salle Pleyel, 252 Rue du Faubourg Saint-Honoré, Paris on the 27th of November 2008, what he conveys will be that Keith Jarrett was on the stage of the Salle Pleyel on the 26th of November 2008. If I, or Keith Jarrett, or someone else, utters the very same sentence while travelling the Peruvian Amazon jungle next year, it will mean something else again. Words like 'I', 'here' and 'yesterday' are so-called *indexical expressions*, a class of words whose meanings are taken to vary greatly depending on the situation in which they are uttered (see Kaplan 1989; Perry 1997).

The fact that a word or a sentence can seemingly convey a number of different contents, depending on the occasion of use, is often referred to in the literature as *context-sensitivity*. As shown above, some words, such as the indexicals 'I', 'now' and 'here', are uncontroversially context-sensitive. There are also expressions like 'this' and 'that' (demonstratives), 'today', 'tomorrow', 'current', 'actual' and 'present', which clearly display the same characteristic, so that 3 will be understood as referring to different people if uttered in 2008 and 2010:

3. The current president of the United States is doing a good job

Nouns like 'outsider', 'enemy', and adjectives like 'right', 'left', 'local' and 'imported', are also generally regarded as context sensitive. In addition, there are cases of *lexical ambiguity*, where, because of historical accidents, different lexical items share phonological and/or orthographic form, as in 'bank', 'fly', 'case' or 'race'. Most people would agree that in interpreting an utterance containing these words, the situation of utterance needs to be taken into account in deciding what or who is being talked about .

Some further, time-worn examples of (alleged) context-sensitivity come from the case of metaphors (example 4), category extensions (example 5) and metonymy (example 6):

4. Bob is a rolling stone
5. Loudon is the new Dylan
6. The ham sandwich is waiting for the check

Utterances such as these also illustrate the point that there is often a gap between the linguistic meaning of a word or phrase and what it can be used to convey on a particular occasion. This gap needs to be bridged if a hearer is to understand the speaker's meaning. Bob is obviously not a stone, rolling or otherwise, although he might be a bit like one in certain respects. Similarly, Loudon cannot actually *be* Dylan, although he is taken by many to share some affinities with Dylan. And while the knowledge that a ham sandwich was served to a

particular customer might be a helpful way for a waiter to identify him, the sandwich itself will not be expected to settle the check.

These are, of course, different from the examples above containing indexicals and the like, since metaphors intuitively express something false if interpreted "literally", while utterances featuring context-sensitive expressions will simply fail to express anything at all if reference is never assigned. However, for anyone interested in how utterances are interpreted, metaphors and indexicals can be seen as different facets of the same issue, in that they both show how knowing what is conveyed by uttering a sentence is often not reducible to knowing the literal meaning of the sentence uttered. And to the extent that both indexicals and metaphors contribute to the explicit content of utterances, they both raise what has been called the *problem of linguistic underdeterminacy*: the problem of how to account for the fact that what a speaker *explicitly* conveys by an utterance often goes well beyond what the words themselves mean.

These two types of underdeterminacy, which are striking illustrations of this problem, have been treated as exceptions to a general rule of linguistic interaction proposed by traditional models of communication: that what is explicitly communicated by an utterance should stay as close to the literal meaning as possible. In classical rhetoric, for instance, metaphors and metonymies are described as departures from a norm of literalness. These departures had to be recognized as such by the speaker's audiences, serving somehow to enhance and ornament the orator's style of speech. There are clear echoes of this in Grice's treatment of figurative language (Grice 1989).

But apart from these apparent exceptions, the explicit content conveyed by the use of words and sentences in general was traditionally seen as remaining constant even across fundamentally different situations. So when Keith Jarrett utters 'Snow is white' in situation A, what this explicitly communicates will be the exact same thing as when Charlie Haden utters the same sentence in a distinct situation B. And traditionally, the type of content conveyed by 'Snow is white', rather than what is conveyed indexical or metaphorical uses is seen as more typical.

However, as the linguistic underdeterminacy issue has been studied in more detail, it has become increasingly unlikely that the gap between sentence meaning and utterance meaning can be bridged by treating indexicals, reference assignment and ambiguity, on the one hand, and metaphor and metonymy, on the other, as the exceptions that confirm the rule. In addition to the examples above, there is a wide range of cases where the linguistic meaning

appears to be fragmentary or incomplete. Some of these can be grouped together and analysed as a restricted linguistic category, while others seem to defy classification.

Take the *possessive relation* as an example:

7. Maybelle's shoes are new

'Maybelle's shoes' can be applied to (a variable number of) shoes standing in any number of relations to Maybelle. These include, but are not confined to, the shoes Maybelle owns, the shoes she does not own but usually wears, the shoes she used to own, the shoes she designed, the shoes she told a friend to buy, and so on indefinitely. Moreover, 'new' can convey a number of different meanings depending on the relation between Maybelle and the shoes, and on other extra-linguistic factors.

A further range of cases where the communicated meaning does not stay constant across different situations involves *definite descriptions*:

8. The book is on the table
9. The idea is working!

Here, something more than the linguistic meaning alone will be required to pick out the particular book or idea the speaker has in mind, and thus to interpret the utterances in 8 and 9. Many utterances also display *illocutionary indeterminacy*, in that the intended illocutionary force of the utterance depends on more than its linguistic meaning. An example is 10, where three distinct utterances of the same sentence type can be interpreted as an assertion in one case, a question in another[2] and a command in yet a third:

10. The homework will be handed in by next Friday

The explicit content of other utterances is linguistically underdetermined in that they contain lexical items which need to be specified to a specific domain or degree, cf. the quantification in 11-12 and the comparison class in 13-14 (the bracketed material is what I assume is non-lexicalized material that needs to be added in the course of the interpretation process, given an appropriate context):

11. Everything [in the world] is connected
12. Everything [that needs to be connected for the computer to work] is connected

[2] It might not be so easy to get the question interpretation in the English version of this sentence, but translating it into another language might make the intuition clearer. Compare e.g. Spanish 'La tarea se entrega el proximo viernes' with the appropriate rising intonation. In English, the combination of declarative word order and interrogative intonation is notably found in so-called "echo questions" (see Blakemore 1994; Noh 2000: chapter 4).

13. June is tall [for a 5 year old]
14. Johnny is tall [for a basketball player]

The fact that an utterance of the same sentence type will give rise to different interpretations depending on how 'everything' and 'tall' are specified can be seen from differences in their truth-conditions. If 'tall' is specified according to a comparison class for basketball players when the speaker is in fact talking about children, an utterance of 5 ('June is tall') will come out as false. If the situation is set up the other way around, and the basketball player Johnny is understood as tall compared to pre-schoolers, the utterance in 6 will be trivially true.

Another type of lexical underdeterminacy can be found in the following utterances, where in order to yield a pragmatically plausible interpretation, 'drink' may have to be specified as involving a particular type of consumption of a particular type of substance, and 'tired' may have to be specified as involving a particular degree of tiredness:

15. Make sure to buy enough beer for the party. Elvis drinks.
16. Carl's not coming to the movies. He's tired.

If 'Elvis drinks' is understood literally, to mean simply that Elvis drinks something sometime, it will not convey anything pragmatically plausible, since it is obvious that all living humans take in liquid on a regular basis throughout their life. 'Drink', here, will have to be understood as conveying something like 'drinks large quantities of alcohol'. In example 16, the content of 'tired' needs to be specified to a particular type and degree of tiredness in order to provide any kind of explanation for why Carl cannot go to the movies. Presumably, there are many ways of being (mentally and/or physically) tired, with a variety of gradations in each type of tiredness. Not all of those will be sufficient to explain absence from a scheduled trip to the cinema.

It is also possible to find the opposite type of case, where an utterance is taken to convey something less specific than the meaning of the sentence uttered:

17. After Jerry Lee's speech, the room was silent.
18. Denmark is flat. We could go there for our cycling holiday.

A room is very seldom completely silent: there is generally likely to be an electrical hiss or background noise present, although these would probably be ignored in interpreting 17. And if 'flat' in 18 is interpreted literally, as meaning completely or geometrically flat, the utterance will be strictly speaking false. In both cases, some kind of departure from what is often regarded as the literal meaning of the word would have to be made in order for the utterance to be interpreted as contributing some true information about actual states of affairs.

It seems clear that in interpreting all sorts of utterances, figuring out what is being conveyed depends on the hearer taking into account a range of contextual factors external to the linguistic (speaker-independent) meaning of the sentence uttered: i.e. its *linguistic semantics*. The following are further examples familiar from the recent literature on meaning and context (as before, the material in brackets is assumed not to be part of the linguistic meaning of the sentence uttered, but somehow recovered in context):

19. Paracetamol is better [than generic drugs of the same type]
20. The roof is strong enough [to support a crew of workmen]
21. Rosanne got sick and [then, as a consequence] stayed home from work
22. Kathy's eaten [breakfast, this morning]
23. Cindy's eaten [using her hands]
24. Tara Joan is new [in her job as a jewellery designer]
25. John is too young [to go to the movies with his parents]
26. [I'd like one ticket to] Oxford, please!
27. There's [enough] milk [to use in your cereal] in the fridge

Many who are interested in the process of linguistic interpretation and how meaning is recovered during an act of communication will see the above data as problematic. How do people reliably and often effortlessly bridge the gap between what seems like the linguistic meaning of the sentence uttered and what is conveyed by uttering it in context?

1.2.2. The problem of mental access

There seem to be two main types of possible strategy for a theorist who agrees that (all or most of) the examples above show a gap between the apparent linguistic meaning of the utterances and what they explicitly convey[3]. First, she could argue that the apparent linguistic underdeterminacy is merely superficial, and that once the appropriate empirical and theoretical work is done, we will see that there are *unpronounced elements* in the linguistic structure which, although not phonetically realised, trigger insertion of the bracketed material. Alternatively, she could maintain that in at least some cases, the inserted material is not linguistically required, but is added in the course of comprehension for pragmatic rather than strictly linguistic reasons.

The first option has been pursued by many philosophers and linguists in the formal semantics tradition, who have proposed a number of ingenious ways of cashing out the idea of

[3] In this chapter I am using the theoretically neutral term 'explicitly convey', instead of more familiar notions like 'what is said', in order to avoid venturing into the many debates on the explicit/implicit distinction in Gricean pragmatics. See Carston (2002: chapter 2) and Recanati (2001) for discussion and review of some positions in the debate.

unpronounced elements. Most recently, and perhaps most famously[4], the philosopher Jason Stanley (2000; 2002; 2007) has claimed that "All truth-conditional effects of extra-linguistic content can be traced to logical form" (2000: 391). As he puts it,

> The intuitive truth-conditions of [an] utterance are due to the assignment of values to the parts of the sentence uttered, and combination in accord with syntactic structure. In each case, the belief that this cannot be the case is due either to an impoverished conception of syntactic structure, or an impoverished conception of available semantic resources (Stanley 2007: 21)

If he is right, the problem of determining the explicitly communicated content of an utterance is reducible to finding hidden syntactic variables, which have values assigned to them on the basis of some formally-specified function.

Though Stanley and his peers have been met with an impressive battery of counter-arguments[5], it is worth noting that there is some widely acknowledged independent evidence for the presence of unpronounced syntactic elements in at least some sentence types, so the proposal is not without some initial plausibility. Examples include *covert pronouns* in so-called PRO-drop languages like Spanish (28), as well as the fairly uncontroversial case of *syntactic ellipsis* (29):

28. He llegado [have arrived] → [PRO] he llegado [I have arrived]
29. Roy is crying, but Johnny is not → Roy is crying, but Johnny is not crying

That said, there is a big leap from this evidence to the claim that *all* effects on truth-conditional content can be traced to the linguistic form of an utterance[6]. The defender of the semantic view therefore has a big task in hand in explaining how all the types of variation illustrated in just the few examples above can be analysed in terms of syntactico-semantic variables hidden in the linguistic structure.

The alternative to the semantic solution is the proposal that *pragmatic* mechanisms are responsible for bridging the gaps between sentence meaning and explicitly communicated

---

[4] There are also many theorists, with a history of publication on the topic that is significantly longer than Stanley's, who address the underdeterminacy problem, devising formal machinery for dealing with particular cases. I use Stanley as an example since I take him to be unique among semanticists in approaching the underdeterminacy issue as a whole, dealing explicitly with the overarching theoretical questions.

[5] See e.g. Cappelen and Lepore (2005: chapter 6), Cappelen and Hawthorne (2007), Carston (2002: 197-205; 2008b), Collins (2007), Hall (2008; 2009), Neale (2007), Ostertag (2008). For positive arguments for the semantic solution, see the collection of essays in Stanley (2007) and Marti's (2006) article.

[6] Readers well versed in the history of modern linguistics will notice certain affinities between Stanley's position and claims made by proponents of generative semantics, who argued that elements accounting for truth-conditional variability could be traced back to the "deep structure" of sentences (see e.g. Lakoff 1971 for a classical defence of this position). In the end, the generative semantic position was abandoned due to descriptive inadequacies and inconsistencies with the Chomskian foundation on which it was supposed to be built (see Newmeyer 1996: chapter 8 and 9 for historical overview).

content in the examples above. According to this view, the semantics and syntax of the sentences 1-27 do not provide enough information for the resulting utterances to be truth-evaluable, but merely act as starting points for the interpretation process. People use utterances as input to interpretive mechanisms that have as their output hypotheses about what the speaker meant. This output may depart to a greater or lesser extent from the semantics of the sentence uttered, and one's views on the exact scope of the underdeterminacy problem will determine how powerful the pragmatic mechanisms have to be to deal with it.

The proponents of what has become known as radical contextualism, or radical pragmatics[7] (the term I'll prefer from here on), are happy to acknowledge the inherent malleability of the meanings conveyed by a whole range of natural language expressions. Radical pragmaticians[8] hold that most utterances need some form of pragmatic elaboration for the explicitly communicated content to be grasped, since the semantics is very seldom enough on its own.

Relevance Theory (Sperber and Wilson 1995; Wilson and Sperber 2004; Carston 2002), the radical pragmatic theory I will be concerned with in this thesis, treats all of the examples in 1-27 as too linguistically "impoverished" to determine the explicitly communicated content by semantic means alone. Sperber and Wilson (building on insights from Grice 1989) hold that first and foremost, successful human linguistic communication involves the expression and recognition of intentions. In their framework, the speaker must have both an informative intention and a communicative intention, where an informative intention is the intention to inform the audience of something, and a communicative intention is the intention to have the informative intention recognised (or, more precisely, made 'mutually manifest'). For communication to succeed, the informative intention must be recognised, but does not have to be fulfilled, whereas the communicative intention must be fulfilled, but does not have to be recognised (see Sperber and Wilson 1995, chapter 1; I will return to these notions in section 1.3.2 of this chapter).

When addressed with an utterance such as 8 ('The book is on the table'), a hearer will automatically form a hypothesis (or set of hypotheses) about which book the speaker must

---

[7] Carston (2009: 25) argues that there are a number of theoretical issues which distinguish radical pragmatics from radical contextualism: for instance radical pragmatics, unlike radical contextualism, holds that "only a few words in the language are inherently context-sensitive" the vast majority being, rather, "susceptible to the pragmatics of the speaker-hearer interaction such that they can be used to communicate an indefinite range of different concepts" (*ibid*). I wish to follow her in this, but allow myself to make some broad generalisations for expository purposes in introducing the problem of linguistic underdeterminacy.

[8] I will prefer the term *pragmatician* over Carston (2009)'s suggested *pragmaticist* for distinguishing the linguistic view from the philosophical position of *pragmatism*, which has very little to do with pragmatics as discussed here.

have had in mind. If there are several tables and books in the near vicinity, he may potentially use all the information available to him in coming up with what he takes to be the best hypothesis about what 'the book' and 'the table' are intended to refer to. This information can be of any type: it can be provided by visual input, or derived from the discourse directly preceding the utterance, memories of previous conversations, encyclopaedic knowledge and so on. In example 21 above ('Rosanne got sick and stayed home from work'), assumptions about what happens when people are sick, among other things, will affect the way the relation between the two phrases is construed.

In this case, the logical conjunction of the two phrases 'Rosanne got sick' and 'Rosanne stayed home from work' will have to be elaborated on in order for the hearer to reconstruct the relation between them as both causal and temporal. On this approach, recognition of the speaker's informative and communicative intentions involves the integration of linguistic information with contextual information, and the goal of pragmatics is to explain how this is done. The explanation offered by Relevance Theory will be discussed below.

The upshot of claiming that a powerful and heavily context-dependent pragmatic process is ultimately responsible for constructing a hypothesis about what is explicitly communicated, is that the gap between the linguistic structure of utterances and what they are used to communicate becomes considerable. For instance, if the semantics of 'and' does not determine any specific temporal or causal relation between the conjuncts, the possibilities for interpreting a conjoined utterance will be almost open-ended. The same point will apply to the other utterances above, such as those where the semantics of 'tired' is assumed to pick out a very general feeling of *tiredness* which has to be specified or narrowed down in different ways from situation to situation.

If it is granted that the problem of linguistic underdeterminacy shows that what is explicitly communicated by utterances varies from situation to situation, then the appeal to a powerful pragmatic process that takes impoverished semantic information as input and enriches it using contextual information may explain the data above nicely. But critics of the radical pragmatic approach to communication have raised worries about how well this explanation squares with some important facts about human linguistic interaction.

Intra-linguistic communication generally happens effortlessly and reliably, and most of the time people have no problem in interpreting what other people mean, even when they have different starting points. If what Relevance Theory and other radical pragmaticians claim is true, and there is therefore a huge gap between sentence meaning and what is explicitly

communicated, what is it that explains how people so often and so predictably understand each other? If an utterance of any one sentence can be interpreted in countless different ways, what explains how people so often and predictably arrive at interpretations which are at least broadly similar, if not identical?

Words, sentences and the utterances in which they figure are public representations, which are publicly accessible in the same way to all participants in a linguistic exchange. The thoughts a speaker wants to convey, and the intentions she has in producing an utterance, in contrast, are private and not accessible to her interlocutor in the same way as they are to her. If the gap between the public and the private representations is as big as Relevance Theory claims, how does a hearer ever reach a certain conclusion about the mental states of his interlocutor?

This is what I will call the *problem of mental access*. It is one of the reasons why many people who are worried about the linguistic underdeterminacy problem opt for a semantic rather than a pragmatic account. If it could be shown that the impoverished nature of the semantic input is merely superficial, it would be possible to reduce the scope of the mental access problem by postulating variables that are assigned their content via formal functions. If there are unpronounced and hidden syntactico-semantic elements present in the linguistic structure of sentences like 'Kathy's eaten', which can explain why it is interpreted as 'Kathy's eaten [breakfast this morning]' in one situation and 'Kathy's eaten [using her fingers]' in another, the problem of how we know other's people's mental states might be reducible to the allegedly more tractable problem of how we know the syntax and semantics of natural languages[9].

The same desire to reduce the impact of the problem of mental access seems to underlie attempts by semantic minimalists (Borg 2004; Cappelen and Lepore 2005) to explain away intuitions about the semantic incompleteness of some of the examples discussed by radical pragmaticians. In their view, sentences like 'Kathy's eaten' are truth-evaluable as they stand, without any need for a pragmatic mechanism to flesh out this minimal meaning into what can be intuitively called its "communicated content". *Contra* semanticists like Stanley, they claim that it is indeed the job of pragmatic processes to recover the "communicated

[9] Or so it is argued, by Stanley and his peers. But as Neale (2007) points out, the fact that there is nothing in Stanley's account to explain the resolution of indexicals (which semanticists often refer to as a "pre-semantic" process) speaks strongly against the account being immune to the problem of mental access. And the existence of implicatures, which by definition are not traceable back to the form of the uttered words, makes it even more obvious that the problem cannot be merely explained away – no matter how far one tries to push it into the background.

content", but that the processes involved are too unruly and unsystematic for their outputs to form part of what is reliably shared between interlocutors.

Still, minimalists claim that unless utterances have some kind of stable, truth-evaluable content, there is nothing to explain how understanding is achieved between interlocutors in communication. If the semantic input is as gappy, and the possible ways of fleshing it out are as numerous, as is claimed by the proponents of radical pragmatics, it would be impossible to explain how linguistic communication can indeed be so fast and reliable. Hence the appeal to 'minimal' truth-evaluable semantic contents such as *Kathy has eaten.*

Though I will not engage any further with the minimalists' or semanticists' analysis of the problem of linguistic underdeterminacy[10], the lessons to be learned from their rejection of the pragmatic approach is clear. The greater the scope of the underdeterminacy problem according to pragmaticians, the bigger the problem of mental access will turn out to be. In what follows, then, I will outline the relevance theoretic approach to communicated meaning, and discuss and evaluate how Sperber and Wilson address the problem of epistemic access.

In the remainder of this chapter, I will look at how Relevance Theory analyses the process by which a hearer forms a hypothesis about the speaker's informative and communicative intentions based on what they see as impoverished semantic input. In the light of some objections to the radical pragmatic solution to the problem of linguistic underdeterminacy (where I will focus on the arguments in Cappelen and Lepore 2005; 2007), I will examine how Relevance Theory treats the sharing of thoughts and meaning between linguistic interlocutors, and ask what the benchmark of successful communication might be if the pragmatic view of communicated meaning is assumed.

In the next chapter, I will look more closely at the idea of word meaning used in Relevance Theory, and examine the implications of this account for the notion of content in general, thereby returning to the general question about word meaning this thesis started with.

[10] See the works already cited for positive accounts by semanticists and minimalists of some of the above examples. For critique of these accounts from a pragmatic perspective, see Carston (2002; 2004; 2006; 2008a; 2008b), Recanati (2004; 2007), Hall (2008; 2009), Begby (under review) and the references cited therein. The collections edited by Preyer and Peter (2005, 2007) and Szabó (2005) are also highly useful for getting an idea of what different theorists hold to be at stake in the debate.

1.3. Relevance Theory

1.3.1. The principles of relevance

Above, I have hinted at the treatment Relevance Theory suggests for cases of linguistic underdeterminacy. In this section, I will look in more detail the pragmatic mechanism Sperber and Wilson propose to account for how hearers bridge the gap between sentence meaning and speaker's meaning.

As already explained, Relevance Theory builds on some central insights from Grice: in particular, that people communicate first and foremost by expressing and recognizing *intentions*. Contra Grice, who held that a speaker's intentions can be recognised on the assumption that he follows a Cooperative Principle and generally adheres to a set of maxims of conversation (Grice 1989), Sperber and Wilson maintain that there is a single property of *relevance* which guides people's interactions with the world and with each other.

They formulate this firstly in their Cognitive Principle of Relevance, which states that "Human cognition tends to be geared to the maximisation of relevance" (Wilson and Sperber 2004: 610). According to Relevance Theory, the cognitive system is in a constant (and largely automatic and sub-attentive) search for information that can be productively processed, and for ways to improve its representation of its environment. All sorts of things, any input to cognitive processes, can be relevant: sights, sounds, utterances, thoughts, memories, conclusions of inferences, as long as they connect with background information the person has available to "yield conclusions that matter to him" (Wilson and Sperber 2004: 608).

Wilson and Sperber label these conclusions *positive cognitive effects*, which they define as effects that make "a worthwhile difference to the individual's representation of the world – a true conclusion, for example" (ibid). An input to cognitive processes can achieve relevance in a number of ways: it can serve to strengthen an already held assumption, it can weaken or revise an existing assumption, or it can combine with an already held assumption to yield a warranted *contextual implication*, "a conclusion deducible from the input and the context together, but from neither input nor context alone" (ibid)[11].

However, positive cognitive effects are only one of two factors affecting the relevance of an input. The other is the amount of mental effort required to construct a mental representation of the input, access an appropriate set of contextual assumptions, and derive some positive cognitive effects. According to Relevance Theory, other things being equal, the

[11] The list should not be taken to be exhaustive, as there are other worthy candidates for types of positive cognitive effects, such as changing one's preference system or improving or re-organizing memories (Wilson and Sperber 2002: 601).

greater the positive cognitive effects derived from an input, and the smaller the mental effort needed to derive them, the greater the relevance of the input to the individual who processes it, at that time. To claim that human cognition tends to be geared to the maximisation of relevance is to claim that it tends in the direction of increasing cognitive effects and reducing processing effort when an opportunity presents itself.

Of all the potential inputs we find in our surroundings, the richest source of positive cognitive effects by far are other individuals. However, there is an important difference between information we retrieve from general observation of the environment and what we get when communicating with other people, so Relevance Theory claims that utterances occupy a privileged position compared to other types of inputs.

According to Wilson and Sperber, utterance comprehension relies not only on a general ability to attribute goals, beliefs and desires to other individuals, but on a more specific ability to treat utterances as a type of *ostensive stimulus* that automatically creates an expectation of relevance. They formulate this in their Communicative Principle of Relevance: "Every act of ostensive communication creates a presumption of its own optimal relevance" (2004: 612). An optimally relevant utterance is defined as one that is relevant enough to be worth the hearer's processing effort, and, moreover, the most relevant one compatible with the speaker's abilities and preferences.

The Communicative Principle of Relevance justifies the addressee of an utterance or other piece of ostensive behaviour in searching for cognitive effects that would contribute to optimal relevance in a way the speaker might have foreseen. Wilson and Sperber argue that this search is guided by a comprehension heuristic which may be seen as implementing the following procedure: "a) Follow a path of least effort in computing cognitive effects: Test interpretive hypotheses (disambiguations, reference resolutions, implicatures, etc.) in order of accessibility, and b) Stop when your expectations of relevance are satisfied" (2004: 613).

To take a concrete example, consider a situation in which a friend and I are discussing new fiction and he tells me about Don DeLillo's latest work. He encourages me to give it a read, and I ask where to buy it. He answers by uttering 8 (repeated here as 30) while pointing to a table covered with books:

30. The book is on the table

According to Relevance Theory, the linguistic structure of 30 is not enough on its own to enable me to find the particular item my friend had in mind, since I have to pick it out from a pile of other salient books. However, by drawing on information made available by my

memory and perceptual systems - for instance, the name of the author just mentioned, the titles of the books in front of me and the way they are positioned on the table, I should be able to pick out a book entitled 'Falling man'.

Relevance Theory suggests that this is the best hypothesis about my friend's intentions because having my attention drawn to one of the other potential candidates will not lead to greater cognitive effects for less effort in a way the speaker could manifestly have foreseen. For instance, my friend's diary is harder to pick out from the description given, and information about its whereabouts does not combine with any easily accessible information I have to yield conclusions that matter to me. By interpreting 'the book' as referring to Don deLillo's *Falling Man,* I can combine the information derived from the utterance with the assumption that my friend has encouraged me to give DeLillo's book a read, and conclude that my friend is offering to lend it to me (for more on reference assignment and accessibility from a relevance theoretic perspective, see Scott 2008).

Now look at example 21, reproduced as 31 below:

31. Rosanne got sick and stayed home from work

In this case, a hearer looking for optimal relevance may feel encouraged to elaborate on the causal relation between the events described by the two conjuncts, since the first suggests a possible motivation for the second and the resulting interpretation would explain why the speaker chose to mention both. The hearer is also likely to construe the second event as temporally succeeding the first, since this is a fairly stereotypical sequence of events, and it will therefore require little effort to use it to compute sufficient cognitive effects. If one imagines a context in which someone is organizing a lunch meeting that Rosanne was supposed to attend, the information that Rosanne stayed at home from work after getting sick will combine with the information about the lunch and yield the contextual implication that Rosanne will not attend the meeting because she got sick and had to stay at home.

1.3.2. Communicating without infallible epistemic access

The relevance-theoretic machinery outlined in the previous section is meant to address the first of the two problems I raised in section 1.2 – that of linguistic underdeterminacy. Sperber and Wilson hypothesize that a hearer, motivated by the presumption of relevance created by the ostensive stimulus, sets off on a search for cognitive effects that might make the utterance relevant in the expected way – following a path of least effort in enriching a fragmentary

semantic representation ('I have eaten') into something fully propositional (S says that she has eaten breakfast today).

But as already mentioned, postulating a powerful pragmatic process to bridge the gap between sentence meaning and speaker's meaning magnifies the epistemic problem. If there is no direct, one-to-one correlation between the information encoded by a given utterance and the thought it can be used to express, many different thoughts may in principle be assigned as the interpretation of any given utterance. How can the hearer then know that the chosen interpretation is correct?

The answer is that one cannot know for certain. Sperber and Wilson argue that people have no sure-fire way of predicting others' mental states. In situations where two individuals are looking at the same object, or being jointly presented with the same piece of information (1995: 18), there is simply no way of knowing *with certainty* that they will construct the same mental representations, use the same contextual assumptions and draw the same conclusions, or indeed whether they are paying attention to the same stimulus. And even when the two individuals *do* arrive at the same interpretation of an utterance, or happen to construct the same mental representation of an object and draw the same conclusions, they have no way to verify for sure whether this is in fact the case.

This is linked to what in psychology is often referred to as the problem of mutual knowledge, where a requirement of absolute certainty leads to an infinite regress. If, in order to communicate, a speaker and hearer need to know not only that they entertain the same piece of background information, but that they knows that they do, and so on, an infinite regress ensues. "By the very definition of mutual knowledge, people who share mutual knowledge *know* that they do" (Sperber and Wilson 1995: 19), so in order to establish that they have mutual knowledge, the speaker needs to know that the hearer knows, the hearer needs to know that the speaker knows, and that she knows that he knows, that she knows that he knows that she knows and so on indefinitely[12].

Sperber and Wilson therefore argue that mutual knowledge cannot be a prerequisite to successful communication, and propose to replace it with a notion of mutual manifestness which they claim is weaker in just the right way. According to their definition, an assumption is *manifest* to an individual at a given time iff he is capable of representing mentally at that time and accepting the resulting representation as true or probably true (Sperber and Wilson 1995: 39). An individual's *cognitive environment* is a set of assumptions that are manifest to

[12] For discussions about mutual knowledge and the implications of the infinite regress problem for studies of communication, see the collection of essays in Smith (1982).

him, i.e. all the assumptions that he can perceive or infer at a given time, including some that he has never entertained before. These assumptions have to be evidenced, but they need not be true, and the notion of manifestness is therefore weaker than the notion of *knowledge*: something can be manifest without being known, or even mentally represented (1995: 40). This captures the idea that beliefs one has never entertained before, although one is in some sense "disposed" to entertain them (e.g. that there are no kangaroos on the moon, that Napoleon and Moses never met etc.,) may come to play a part in the cognitive economy.

The move from manifestness to mutual manifestness involves the further assumption that two individuals may share a cognitive environment in which the same assumptions are manifest. As Sperber and Wilson point out, "One thing that can be manifest in a given cognitive environment is a characterisation of the people who have access to it" (1995: 41). If two people share a cognitive environment in which it is manifest that they both have access to it, then according to Sperber and Wilson, every manifest assumption is also *mutually manifest.* Thus, if the lights go out during a conversation between a friend and myself, the assumption that the lights have gone out is mutually manifest to us, and in Sperber and Wilson's terms, this assumption forms part of our *mutual cognitive environment*.

Relevance Theory presents a slightly different picture of communication than the one based on 'thought sharing' presupposed in many contemporary linguistic and philosophical accounts. Sperber and Wilson see communication as "a matter of enlarging mutual cognitive environments, not of duplicating thoughts" (1995: 193)[13]. On their model, two people, when they begin communicating, start out with a set of mutually manifest assumptions that they gradually expand on as their interaction proceeds. Any utterance by one of them may potentially add to their common stock of mutually manifest assumptions, which can in turn be extended by subsequent utterances.

What happens when we communicate is that a speaker utters a string of sounds (or signs something, or makes marks on paper) which a hearer recognizes as an ostensive stimulus. This sets him off on a path of least effort searching for cognitive effects that will make the utterance relevant in the expected way. Whenever it becomes manifest to both him and his interlocutor that they share an environment in which certain effects are manifest to

[13] It is worth keeping in mind, when critical points about the relevance theoretic view of communication are addressed later, that Sperber and Wilson do not deny that people share information when communicating. On the contrary, they are explicit on the point that "any account of human communication must (…) incorporate some notion of shared information" (1995: 38).

both of them, their mutual cognitive environment grows. This, in effect, is what the essence of communication turns out to be on the relevance-theoretic picture.

Recall from section 1.2.2 that on the relevance-theoretic picture, communication involves both an informative intention and a communicative intention, which I presented there in rather simplified form. I can now present the fuller versions, which appeal to the notions of manifestness and mutual manifestness:

> *Informative intention:* to make a certain set of assumptions manifest to the audience
> *Communicative intention:* to make the informative intention mutually manifest.

Here, the informative intention is to provide evidence for a certain set of assumptions which the speaker wants the hearer to believe, and the communicative intention is to add to the mutual cognitive environment of speaker and hearer the assumption that the speaker had this informative intention. According to Sperber and Wilson, communication is successful as long as the informative intention becomes mutually manifest, whether or not the hearer ends up believing what the speaker intended to convey.

Sperber and Wilson's appeal to the weaker notions of *heuristics*, *manifest assumptions* and *mutual cognitive environment* instead of their stronger counterparts; *algorithms*, *known facts* and *mutual knowledge* reflects the fact that agents' interaction with the world and each other is generally based on non-demonstrative inference, which may lead to conclusions that turn out to be false. Everybody misrepresents perceptual stimuli from time to time: our visual system is always subject to illusions, our hearing or memory also prone to occasional failure. The same point applies to communication: "Since it is obvious that the communication process takes place at a risk, why assume that it is governed by a failsafe procedure?" ask Sperber and Wilson rhetorically (1995: 44-45).

Humans have no way of peeking inside each other's heads, but they do have more or less reliable procedures for inferring the intentions, beliefs and emotional states of others, even if they can never be 100% sure of the conclusions they draw. If one takes this seriously, and follows Sperber and Wilson in rejecting the possibility of arriving at certain knowledge about other people's cognitive states, "failures in communication are to be expected: what is mysterious and requires explanation is not failure but success" (1995: 45).

I have tried to sketch Sperber and Wilson's solution to this mystery by giving relevance theoretic analyses of examples 30 ('The book is on the table') and 31 ('Rosanne got sick and stayed home from work') above. As could be seen from the way these are treated, such natural language sentences are seen as semantically incomplete, in that linguistic

information alone does not determine the intended meaning, but has to be contextually enriched in order to satisfy the expectations of relevance raised by the utterance.

The Cognitive Principle of Relevance, the Communicative Principle of Relevance and the relevance-theoretic comprehension heuristic are meant to explain how the hearer is able to bridge the gap between an incomplete semantic representation and a hypothesis about the speaker's meaning.

1.3.3. Assumption schemas, word meaning and concepts

The *ostensive-inferential* picture of communication Relevance Theory advocates marks a significant break from a traditional view of utterance comprehension as a process primarily driven by the meanings and composition of words. According to this traditional view, which Sperber and Wilson (1995: chapter 1) calls *the code model of communication*, there is a direct correspondence between the linguistic meaning of a sentence and the proposition expressed by uttering that sentence on a given occasion. Knowing the meaning of the sentence uttered, plus the values of a few parameters for speaker, hearer, location of utterance, time of utterance etc., will on this model determine the explicit content of an utterance.

Sperber and Wilson take the problem of linguistic underdeterminacy to show that this cannot be the true story about how communication works. In their view, semantic representations, which they see as "incomplete logical forms, i.e. at best fragmentary representations of thoughts" (1995: 193) provide no more than a starting point for constructing a hypothesis about the explicit content of an utterance. Semantic representations themselves are "mental objects that never surface to consciousness" (*ibid*), and, as Sperber and Wilson put it, merely serve as *assumption schemas* which have to be contextually enriched in order to yield a truth-evaluable explicit content.

What exactly is an assumption schema or starting point for constructing a full-fledged hypothesis about the *speaker's meaning*? Sperber and Wilson (1995) take most words to stand in an encoding relation to mental items called *concepts*, which provide the linguistically specified meanings of words like 'tired', 'sick', 'new', 'get', 'eat' 'Rosanne', 'Oxford' and so on. The utterance of a word like 'sick', Sperber and Wilson suggest, will systematically activate a concept SICK, this correlation being what provides the word with its meaning. A concept such as SICK is assumed to play two formally distinct functions in mental life: as a constituent of the logical form of thoughts, and "as an address in memory, a heading under which various types of information can be stored and retrieved" (1995: 86).

Sperber and Wilson assume that the information a concept may give access to "falls into three distinct types: logical and encyclopaedic and lexical" (*ibid*). The *lexical* information is about "the natural-language counterpart of the concept: the word or phrase of natural language which expresses it" (*ibid*). The encyclopaedic entry attached to a concept "contains information about its extension and/or denotation: the objects, events and/or properties which instantiate it" (1995: 87), while the logical entry consists of deductive rules which apply to logical forms containing the concept. The encyclopaedic entry is not seen as contributing to the content of the associated concept, but as providing a potential source of contextual assumptions; by contrast, the logical entry is seen as constraining the content of assumptions in which the concept figures (1995: 89).

Many words are taken by Relevance Theory to express concepts with rather general meanings, denoting a wide range of objects, events or states of affairs. For instance, Wilson and Carston (2007: 246) treat the verb 'rest' as encoding a concept "which covers any degree of inactivity (physical or mental), from sleeping to staying awake but not moving much to performing a range of not very strenuous tasks (with many more possibilities in between)". According to Wilson and Carston (2007), when a hearer comes across an utterance featuring the word 'rest', he may use this general concept REST as a starting point for constructing a more specific concept, with a more restricted denotation. In doing so, he will draw on contextual assumptions in the encyclopaedic entry of REST, and use them to create a narrower *ad hoc concept*, REST*, which shares these encyclopaedic properties, and which yields enough implications to satisfy his expectation of relevance. This ad hoc concept is constructed for the purposes of the specific linguistic exchange, and is the outcome of a process triggered by the search for relevance.

Sperber and Wilson (1998) consider the case of 'tired', which they take to encode a concept denoting a wide range of degrees of tiredness, including very minimal ones. As they point out, an utterance expressing such a very general concept would not yield enough implications to be deemed relevant on most occasions: for instance, if someone utters 'I'm tired' intending to provide a reason for not going to the cinema, it is hard to see what reason that very general concept would provide. Relevance Theory therefore argues that someone faced with this utterance would have to assume that the speaker intended to convey a more specific ad hoc concept TIRED*, which would carry enough implications to satisfy his expectations of relevance.

How tired does one have to be in order to be *tired enough* for it to serve as a valid excuse for not going out? "Well, there is no absolute scale of tiredness (and if there were, no

specific value would be indicated here)", according to Sperber and Wilson (1998: 194). The concept constructed expresses merely "an ad hoc, circumstantial notion of tiredness" (*ibid*) – whatever degree of tiredness would have the implication that the speaker is unable to muster the energy to take a trip to the cinema.

A hearer faced with any utterance, then, will use whatever information the linguistic system provides him with to construct a hypothesis about the speaker's meaning, following a path of least effort in computing cognitive effects. Typically, for words like 'rest' or 'tired' with very general meanings, he will have to construct *narrower* meanings in order to satisfy his expectation of relevance. For other words, he may have to construct meanings broader than the one encoded: for instance, a word like 'silent' (encoding the concept SILENT, taken to denote a complete absence of sound) may be used to convey a range of broader ad hoc concepts denoting various degrees of approximation to silence.

I will return to the issue of narrowing and broadening and the relevance theoretic view of word meaning in the next chapter, but propose to analyse a previously introduced example in order to get a clearer grip on how all this plays out within a framework where utterances have "assumption schemas" or "incomplete logical forms" as their semantic representations. Take another look at utterance 21 above (repeated here as 32):

32. Rosanne got sick and stayed home from work

The logical form of this utterance might be something like the following assumption schema once it has been syntactically parsed[14] and reaches the interpretive system:

33. [ROSANNE$_X$ GET$_{tense1}$ SICK] & [PRO$_Y$ STAY$_{tense2}$ HOME-FROM-WORK]

Here, the pragmatic mechanisms have to fill out the intended reference of 'Rosanne' (and the empty pronominal slot PRO), since there is nothing in the semantics of Rosanne which specifies who it refers to. It will also have to work out whether the two conjuncts should be seen as linked merely by the logical conjunction '&', taken to be the meaning of the word 'and', or whether the events they describe should be taken to stand in a more specific temporal and/or causal relation (see Carston 2002: chapter 3 for discussion).

In order to derive enough implications to satisfy his expectations of relevance, the hearer of 33 may also have to narrow (what I here assume are) the general conceptual

[14] This is somewhat simplified for expository purposes. Wilson and Sperber (2004) suggest that syntactic parsing, development of logical form, recovery of implicatures etc. involve *mutual parallel processing*. In actual utterance interpretation, then, the workings of the cognitive system is not assumed to be strictly linear.

meanings of SICK and HOME-FROM-WORK into something more specific. Within the constraints imposed by the grammar (including what in Relevance Theory is labelled 'procedural meaning', see Wilson and Sperber 1993) he will furthermore have to figure out the relevant time-frame within which the events in question took place.

As a result of this (sub-attentive) pragmatic process, the hearer may end up with something like the explicature[15] in 34:

34. ROSANNE$_{REF}$ GOT SICK* at T_1 and [as a consequence] ROSANNE$_{REF}$ STAYED* HOME-FROM-WORK* at T_2

According to Relevance Theory. this explicature will act as a premise for the derivation of contextual and logical implications, which may in turn lead to further effects and so on. Since the assumption schema (in 33) used in constructing the explicature (in 34) is so minimal, Sperber and Wilson claim that there can be no guarantee that the speaker and hearer will end up entertaining the exact same thoughts. As long as sentence meaning only provides *a clue* to the speaker's meaning (rather than incontrovertible evidence for that meaning), there will be a chance that the thoughts the hearer ended up with diverge somewhat from the ones the speaker had in mind.

For instance, the speaker and hearer of 32 may assign reference to the same person called Rosanne, but individuate her differently, or they may construct slightly different ad hoc concepts SICK*, with the speaker arriving at a more specific conception of the type of illness Rosanne suffers from (migraine?) than the hearer, who constructs a more general SICK** concept which encompasses, say, migraine, the flu and a severe cold.

However, there is no need to assume that these divergences will automatically lead to failure of communication, since Sperber and Wilson explicitly deny that speaker and hearer have to entertain identical thoughts for communication to succeed[16]. In many circumstances, people may understand each other perfectly well despite there being some differences in form or content between the thoughts they end up with. What matters is that those thoughts share

[15] In Relevance Theory, the notion of 'explicature' covers the explicitly communicated content of an utterance. The term was chosen to contrast with Grice's term 'implicature', and to take account of possible pragmatic contributions to explicit content. See Sperber and Wilson (1995: 172-193) and Carston (2002: chapter 2) for discussion.

[16] I am, here and throughout this chapter, working with a pre-theoretical notion of communicative success, which treats communication as successful if the interlocutors, as a result of an utterance, manage to coordinate their behaviour appropriately (an idea I have in common with other approaches to communicative success, e.g. that of Pagin (2006; 2008) to whom I refer for a more detailed discussion of this criterion). This has a certain element of vagueness, which is meant to reflect the fact that standards of what counts as successful communication will differ from situation to situation, depending on the degree and type of co-ordination the speaker was aiming to achieve. My main concern in this chapter is to show how Relevance Theory provides the tools for developing more detailed explanations. See section 1.4.3 for further discussion.

enough logical and contextual implications to bring about the degree of co-ordination that the speaker hoped to achieve.

In the case above, the implications which enable the speaker and hearer to co-ordinate may be that Rosanne is not at work today, that Rosanne will not attend the lunch meeting, that Rosanne may not meet tomorrow's deadline, and so on. Since all of these are derivable from the assumption that Rosanne is SICK*, or that she is SICK**, the divergences between the speaker's and hearer's interpretations are immaterial in this case. Every utterance will lead to an indefinite number of cognitive effects, not all of which will have been foreseen and intended by the speaker or actually entertained by the hearer (as opposed to merely being made manifest). Thus, speakers and hearers may well achieve the required degree of co-ordination by sharing only some of the cognitive effects they derive.

Since sentence meanings serve only as *clues* to the contextual implications (and other cognitive effects) the speaker intended to achieve, it is the implications, rather than the sentence meanings themselves, which are important to the theorist who wishes to distinguish situations where communication is successful from those where it is not. When two people succeed in communicating, they end up having access to at least some of the same contextual implications, thus enlarging their mutual cognitive environments and enabling co-ordination to take place. Misunderstandings arise when the mutual cognitive environment is either not extended at all, or not extended to the degree necessary for co-ordination to take place. With 33, for instance, this could happen through a wrong assignment of tense, so that the hearer took the speaker to be describing some event in Rosanne's distant past, or a larger divergence between different ad hoc construals of SICK (I will return with a discussion of this in section 1.4.3).

Thus, the relevance theoretic notion of shared logical and contextual implications gives the theorist a possible means of analysing communicative success and failure. It also suggests a possible theoretical explanation of how interlocutors can arrive at a degree of coordination that is "good enough" for their purposes – without relying on the exact duplication of thoughts.

As mentioned in section 1.2.2. giving up on the traditional assumption of a direct relation between sentence meaning and explicit content, does lead to some objections from theorists who see success in communication as inevitably involving such a relation. And, as I will show in the next section, abandoning the view that exact duplication of thoughts is a prerequisite to successful linguistic interaction adds further fuel to the fire. Here I will address a recent and very influential critique of the relevance theoretic approach to communication,

and consider how we might explain what counts as successful communication if identity of thoughts is not the answer.

1.4. The 'non-shared content' critique of Relevance Theory

1.4.1. The role of 'similarity'

The relevance theoretic picture of communication outlined above has, through its break with a dominant and traditional way of seeing linguistic interaction, met with some resistance in the linguistic and philosophical literature. Recently, it has been criticised by Cappelen and Lepore (2007), whose objections I turn to in this section of the thesis.

Cappelen and Lepore (2005; 2006; 2007) who themselves favour an approach on which grasping a minimal, truth-conditional content is a necessary condition on utterance comprehension, argue that the relevance theoretic view of successful communication entails a commitment to what they label the Non-Shared Content Principle (NSC). According to NSC, "When a speaker utters a sentence, S, thereby intending to communicate the proposition that p, the audience will not grasp p [but instead] some proposition (or set of propositions) R-related to P" (Cappelen and Lepore 2007: 117). The R-relation between the thought a speaker wants to communicate and what a hearer ends up as his interpretation of the sentence intended to convey this thought is one of *similarity*, according to Cappelen and Lepore.

In their view, the position that a proponent of NSC is committed to is that a speaker's meaning and a hearer's interpretation will *never* strictly coincide as a result of linguistic interaction. It follows that ordinary communicators are subject to an illusion when they think they communicate successfully, since they really never grasp *exactly* what the other says and means. This, Cappelen and Lepore argue, runs counter to all sorts of intuitions everybody has about language and communication, and is an undesirable outcome since people do not merely *think* they understand each other when they report speech, attribute beliefs and assess each other's assertions, they *do* understand each other

What Cappelen and Lepore see as committing Relevance Theory to the NSC principle is Sperber and Wilson's reliance on the notion of sentence semantics as a radically incomplete input to the comprehension process. When a hearer develops a sub-propositional assumption schema (such as the one in 33 above) into a fully-fledged explicature, the way she does this will depend on which interpretation would be optimally relevant, according to RT. But, Cappelen and Lepore claim, which interpretation is optimally relevant will depend on the belief base of the interpreter, which is shifting and variable from person to person and situation to situation. There is therefore no way for someone uttering a sentence S to predict in

advance how it will be interpreted by her interlocutor. By extension, there is no way for the pragmatic theorist to explain how different people, acting upon different beliefs and coming from different perspectives, can settle on interpretations that are similar enough for understanding to take place.

Here, I take Cappelen and Lepore to be making two objections which, though they are related, need to be separated for their worries to be appropriately addressed. The first is that Relevance Theory does not have the resources to explain how communication can succeed across different (physical) *contexts*, given that the gap between sentence and utterance meaning is so big. The second objection is that, even if Sperber and Wilson had a way to deal with cross-contextual communication, their claim that the goal of communication is not necessarily to achieve identity of thoughts between interlocutors leads to a reliance on *similarity* of thoughts. But similarity is a notoriously unstable notion, since there is no principled account of how two thoughts can be similar enough for the purpose of linguistic interaction to be achieved.

There are thus several points worth remarking on in Cappelen and Lepore's critique, and some have already been discussed in print. Wedgwood (2007) has argued that Cappelen and Lepore misunderstand fundamental aspects of the goal and scope of Relevance Theory, leading to a divergence of views on what 'successful communication' amounts to. According to Wedgwood, Cappelen and Lepore overlook the fact that RT's view of communication as enlarging mutual cognitive environments implies that communication is successful "iff through it two people can tell, on the basis of the evidence available to them, that they have some more assumptions in common than they had before" (Wedgwood 2007: 656).

According to Wedgwood, Cappelen and Lepore's claim that "similarity constitutes a crucial part of the relevance-theoretic explanation of communication" confuses Relevance Theory, "an explanation of what interlocutors do, with the possible effects of applying RT (as viewed by an omniscient third party)" (2007: 657). When faced with an ostensive stimulus, an addressee does not evaluate possible interpretations by thinking "I reckon that interpretation q is reasonably similar to the intended interpretation p and therefore communication has been successful'" (2007: 656). Rather, "he says to himself 'Having employed the best means available, I calculate that the speaker must have intended to communicate q' – even if an omniscient third party might be able to identify that the speaker really intended to communicate the similar proposition p" (*ibid*).

Wedgwood argues that assessments of similarity therefore form no part of the pragmatic process of utterance interpretation which Sperber and Wilson are trying to explain,

which is important "because any approach that actually took similarity to be an intrinsic part of the process of interpretation would quickly lapse into incoherence". (2007: 657). In Wedgwood's view, Cappelen and Lepore's criticisms do not succeed in showing that RT is implicitly committed to the Non-Shared Content principle, since the distinction between explaining content sharing from the point of view of an omniscient third party and explaining "the subjective assessment of intended meaning that drives RT as an explanatory framework" (Wedgwood 2007: 655) is never made.

Though I agree with Wedgwood's claim that a difference in perspectives between Cappelen and Lepore and Sperber and Wilson may result in some of the NSC criticisms missing the mark (I return to this issue in the next section), I do think Wedgwood is too quick to dismiss Cappelen and Lepore's 'similarity' objections as inconsequential for Relevance Theory. Even though one of the main aims of Relevance Theory may be to explain how hearers arrive at a subjective assessment of the speaker's intended meaning, another, equally important matter still needs to be addressed by any pragmatic theory. If Relevance Theory does give an accurate account of how speakers and hearer end up understanding each other, it should also be able to explain in more objective terms how and why this happened.

Recall the example above, of a speaker who utters 32 ('Rosanne got sick and stayed home from work'). Here, Relevance Theory suggests that communication may succeed even if the speaker has in mind an ad hoc concept SICK* (≈ flu or migraine) that denotes a more specific type of *sickness* than the broader concept SICK** (≈ general queasiness) constructed by the hearer. What this means is that the hearer will form a hypothesis about the speaker's intended meaning that diverges somewhat from the speaker's actual meaning, i.e. the two thoughts are only *similar*.

The important question is, what makes this a case of successful communication, as opposed to one where speaker and hearer form two concepts SICK* and SICK*** which are also similar, but where the latter denotes graver illnesses (cancer, a chronic heart condition) and the divergence results in a genuine misunderstanding? If the theory does not aim to explain why the two thoughts are similar enough in the first case but not in the second, this rules it out as an analytical tool in accounting for failures and successes in communication. But surely, explaining how and in which cases people manage to communicate - and by extension, how and in which cases they do not - should be one of the goals of any adequate theory of communication.

I believe that Cappelen and Lepore, in raising their objections to radical pragmatics, point to an important problem that cannot be merely brushed aside. In what follows, I show

how Relevance Theory may go about solving what I take to be a problem for all theories of communication which hold that identity of thoughts and/or content is too strict a requirement for communication to be successful. Even though this point does not form an explicit part of Cappelen and Lepore's (2007) argument, I will argue that any appeal to similarity ultimately relies on a notion of identity at a fundamental level. By looking more closely at some of the core notions of relevance theoretic pragmatics, I will show how the sharing of logical and contextual implications does provide a measure of successful communication, and how the degree to which they overlap provides a benchmark for whether the relevant content has been shared or not.

1.4.2. Context and flexibility

As already mentioned, Cappelen and Lepore (2007) argue that Sperber and Wilson's proposed comprehension heuristic, and their notion of cognitive effects, are not enough to guarantee a workable standard of similarity of content across contexts. Cappelen and Lepore claim that any logical form (Sperber and Wilson's *assumption schemas*) "can be developed into an indefinite number of explicatures, and ... it is impossible to predict in advance which development various readers will end up with" (2007: 131). In their view, there is "not even a guarantee that these [developments] will be similar", since any two developments will in principle be similar in some respects (2007: 131). "The relevance theorist could try to suggest that the interpretive results in [a context] C will be similar by the similarity standards of C. But again, given the radical variability in standards between contexts, what could possibly guarantee this claim?" (ibid), they ask rhetorically.

Cappelen and Lepore seem to be particularly worried about what happens when two communicators do not share a physical environment, or when the addressee of an utterance is to some extent unknown, as is the case with academic writing. They use their own text as an example and observe rightly that

> "The sentences of this chapter have certain logical forms. The readers of this chapter will develop these until they satisfy the Principle of Optimal Relevance. Which development satisfies that principle for a particular reader R will depend on the contextual effects these logical forms have on R." (2007: 130)

They take a wrong turn, though, when they claim that there is "no way to predict in advance which development of these logical forms various readers will end up with. There are infinitely many such developments and common sense dictates that readers will all end up in different places" (2007: 130-131).

As Wedgwood (2007: 661) points out, it is slightly odd to criticise Relevance Theory for providing no way of explaining how communicated content is determined, since providing such an explanation is in fact the theory's "raison d'être". However, I choose here to interpret C&L's critique as directed more at what they think is the weak heuristics of Relevance theory, rather than as committing them to the unjustified claim that RT offers no heuristics at all. The main source of their concerns, I believe, stems from their view of "context". It seems clear, in reading their criticisms of Radical Contextualism in general, and RT in particular, that they have a very specific idea of what context has to be for a relevance theorist.

Though they do not discuss this point explicitly, judging from some extracts in Cappelen and Lepore (2005; 2007), it seems that for them, context amounts to something like the intuitive notion of *situation of utterance*:

> "Often, people in different contexts are asked to do the same thing, e.g. pay taxes. They receive the same instructions, are bound by the same rules, the same laws and conventions (…) When people over a period of time, across a variety of contexts, try to find out whether something is so, they typically assume content stability across those contexts" (2007: 122)

> "Sometimes the audience of an utterance doesn't share a context with the speaker. This can happen in any of several ways, the most salient of which being the reproduction of a speech act, as in published articles. Writers often have no idea who their reader is; they know next to nothing about her beliefs; or about her perceptual environment; all they know is that it is not shared (2005: 213)

What they seem to overlook in their exposition of Relevance Theory, though, is the fact that context is defined as a technical notion in Sperber and Wilson (1995). Relevance Theory treats context as "a psychological construct, a subset of the hearer's assumptions about the world" (1995: 15). Moreover, "the selection of a particular context is determined by the search for relevance" (1995: 141), and takes place within a relevance-oriented account of communication and cognition.

The difference in outlook between the two approaches reveals itself in small, but far from insignificant details. For instance, Cappelen and Lepore write that "[t]he cognitive effects of an utterance on a person at a time *t* will depend, essentially, on the beliefs the interpreter has at *t*. These vary between interpreters" (2007: 130)". Though this is correct at one level, it is important to acknowledge that not *all* the beliefs an individual is capable of entertaining contribute to the cognitive effects of an utterance, and, moreover, that not all the cognitive effects derived from an utterance will be recognised by the audience as *intended by the speaker*.

Sperber and Wilson emphasise that a context is a *subset* – not the totality – of the assumptions manifest to an individual in a given cognitive environment, and they explicitly reject the idea that every utterance needs to be evaluated against the hearer's whole belief base: "If the context included the whole of the hearer's encyclopaedia, virtually any new information that a speaker could express would have some contextual effects […]. This line is clearly not worth pursing" (1995: 137), they conclude.

Contexts are treated in Relevance Theory as dynamic entities that may start off comprising very little information, and are gradually expanded during the course of a conversation. There is no default setting on what has to be included in a given context: for instance, when someone reads a book or an academic paper, the initial context often has no information drawn from the immediate physical environment, but consists rather of background knowledge activated by the title and abstract of the paper. And despite the fact that information about an individual's physical surroundings may be very relevant in many situations, they are not as likely to enter into the picture when he talks over the phone, or writes e-mails or reads books.

Here, information from the (short- and long-term) memory system is likely to be deemed capable of making a greater potential contribution to relevance, and will therefore be selected as part of the subset of mentally-represented assumptions brought to bear in utterance interpretation. Moreover, Sperber and Wilson reject the "traditional" pragmatic picture on which context is selected in advance of the interpretation process, so that "relevance is seen as a variable to be assessed in function of a pre-determined context" (1995: 141). In their view, "The assessment of relevance is not the goal of the comprehension process, but only a means to an end, the end being to maximise the relevance of any information being processed" (1995: 142).

What Relevance Theory postulates, then, is a system in which context is constructed from assumptions *manifest* in the individual's cognitive environment, depending on their relative accessibility and expected contribution to relevance[17]. Two individuals who have access to the same manifest assumptions $P_1 \ldots P_n$) will share a cognitive environment, and if it is manifest to both of them that they share this environment, the assumptions $P_1 \ldots P_n$ will be not only manifest to each of them, but *mutually manifest* to both. Just as two individuals may start out with a very limited range of mutually manifest assumptions, constituting a mutual cognitive environment that is gradually expanded in conversation, the initial context may be

[17] Any one of a range of manifest assumptions could, in principle, contribute to relevance, but they will not form part of the context unless they are actually entertained.

quite minimal, and further contextual assumptions may have to be accessed at the cost of some processing effort. This picture, with its use of minimal contexts and a limited range of mutually manifest assumptions, is clearly not as unwieldy as Cappelen and Lepore suspect. Moreover, the fact that cognition in general is seen as relevance-oriented will make it possible for speakers to predict, in at least some situations, what contextual assumptions hearers are likely to access, and what conclusions to draw.

1.4.3. Content similarity

But how about the second part of Cappelen and Lepore's objection, the claim that there is no way to ensure an adequate similarity relation between two thoughts? As already mentioned, I will not follow Wedgwood (2007) in discounting the role of similarity in pragmatic theory, since I see it as an important task for Relevance Theory (as for any pragmatic theory) to explain what makes communication fail and what makes it succeed[18]. If this is to be done, it is necessary to provide some account of what makes the thoughts of speaker and hearer "similar enough", even though this will raise a puzzling but familiar problem.

The problem, discussed in the philosophical literature going back to Plato, in a nutshell, is this: Similarity presupposes identity. If two objects are similar, they are similar because they share certain features or have things in common. These features can be either literally shared (i.e. identical) or they can be similar. But if they are merely similar, it would have to be as a result of having some of *their* features in common. There is nothing that stands in the way of the features of the features of the object themselves being merely similar rather than identical, but in that case, the similarity relation must again depend on their having features in common, and so on until one reaches a level where strict identity obtains. So if the thoughts that speaker and hearer end up entertaining are not identical but merely similar, this similarity relation has to be explicated in some way or other.

I have already given away parts of Sperber and Wilson's solution to this problem in going through example 32 ('Rosanne got sick and stayed home from work') above. What I claimed was that two interlocutors may represent a piece of information differently – and as a result entertain different thoughts – yet still communicate successfully[19]. The speaker of 'Rosanne got sick' could have had in mind a very specific notion of *sickness*, but the hearer

[18] However one wants to cash out the idea of communicative success/failure, cf. footnote 16.
[19] In fact, given the notion of 'thought' used in Relevance Theory, there are two ways in which thoughts can differ while still enabling the individuals who entertain them to coordinate well enough. They can either have different contents which are similar enough for the purpose of the interaction, or they can have identical content but differ in their *mode of presentation*, an idea I will get back to in section 1.4.4.

may not have picked up on this, and instead constructed a broader ad hoc concept SICK**, with the result that the thoughts of speaker and hearer are only similar rather than identical. The question is, how is this similarity relation to be construed?

As suggested above, a relevance theorist could here appeal to the notion of contextual implications[20]. Relevance Theory holds that thoughts are entertained not in isolation but in a context of mentally represented background assumptions. According to Wilson (1995: 208) "two propositions resemble each other in a given context to the extent that they share logical and contextual implications in that context" (see also Sperber and Wilson 1985/1986; Sperber and Wilson 1995: chapter 4.7 for more on the notion of interpretive resemblance). Two thoughts which individuate a certain person differently and diverge in their respective 'ad hoc' construals of the concept SICK, can still be similar enough for communication to be successful as long as they share the cognitive effects that contribute most to making the utterance satisfy the specific presumption of relevance raised by the utterance.

In the scenario I set up above, the context included information about an impending job meeting, where Rosanne hasn't yet showed up despite being expected. I proposed that, in this case, the implications that contribute most to satisfying the presumption of relevance raised by the utterance will be that Rosanne is not at work today, that Rosanne will not attend the lunch meeting and that there's a chance Rosanne will not meet tomorrow's deadline; if these are recovered by the hearer, communication will be successful.

Now, the speaker's or hearer's particular construal of SICK may lead to a great many more implications, depending on the available assumptions about migraines, Rosanne's supply of migraine medicines and the effectiveness of these etc. Moreover, all sorts of other implications will be derivable from the utterance construed in this way (e.g. that no one is sitting at Rosanne's desk today, that nobody whose first name begins with 'R' is at work, that Rosanne will not have needed to take the bus this morning and so on). If these seem relevant enough to either the speaker or the hearer, they may be entertained as thoughts and lead on to further implications, which will have to be taken into account in the theorist deciding whether communication has been successful or not. My claim is that as long as it does not become mutually manifest that these implications have to be derived in order to make the utterance

[20] There are, of course, other ways of explicating the similarity explanation between thoughts. The solution I present here may be particular to the relevance theoretic view of communication and the (Fodorian) conception of thought they rely on, and might therefore be seen as having limited transfer value. For an attempt to cash out similarity of communicated content in terms of possible world semantics, see Pagin (2006).

relevant in the expected way, they will not be treated as part of the speaker's meaning, and should not affect the interaction between speaker and hearer[21].

But consider the slightly modified version of the 'Rosanne' exchange above, where the hearer constructs an ad hoc concept SICK*** (based on mistaken assumptions about Rosanne's medical history, which he thought he shared with the speaker) denoting very grave and life-threatening illnesses such as cancer or heart conditions. The thought that the hearer constructs here will still lead to the three contextual implications I suggested above were the ones that contributed most to satisfying the presumption of relevance raised by the utterance, but will also lead to several other highly relevant implications which, as it happens, were not intended by the speaker. A thought such as 'ROSANNE GOT SICK***[from cancer]' may lead to the implication that Rosanne will be gone for a long time, that she has to undergo hospital treatment, that she may never recover fully, that the hearer should make arrangements to visit her in hospital and so on.

This is a case where the thoughts of speaker and hearer share some implications, but the hearer mistakenly attributes to the speaker the intention to convey a number of highly relevant implications that were not part of her intended meaning, The result is certainly a misunderstanding, and one that is likely to affect their future interactions (the hearer may propose that they go and see Rosanne in hospital, come up with potential work replacements etc.). From a theorist's perspective, this is a misunderstanding because the speaker's informative intention did not become mutually manifest, and it did not become mutually manifest because of a mismatch in the cognitive environments of speaker and hearer.

As Cappelen and Lepore (2007) quite rightly point out, there is no fixed standard of similarity that one can appeal to in an account of comprehension, and therefore no point in trying to stipulate how many shared implications would be enough for communicative success. However, this does not present a problem for Relevance Theory, which is not in the business of making predictions about sentence types, as is emphasized by Hall (2009: section 5). Given "the context-sensitive nature of pragmatic processes", one has to consider "the details of the particular context of utterance, and the context-specific processing" of a *token*

[21] Whether an implication is attributed as part of the speaker's meaning or merely entertained as a further thought of the hearer's does not depend on the similarity metric used by the theorist, but only on the working of the relevance-theoretic comprehension heuristic. It is important to note that this heuristic provides no justification for going beyond the *most accessible* set of implications that make the utterance relevant enough to satisfy the presumption of relevance. Any further implications are derived on the hearer's sole responsibility. Moreover, implications of the sort that nobody whose first name begins with and 'R' is at work today are hardly ever likely to contribute to relevance, and are therefore unlikely to be derived at all.

utterance (Hall 2009: 118) before considering how a term is interpreted and the utterance understood.

The fact that relevance itself is defined as a comparative rather than a quantitative, concept by Sperber and Wilson (1995: 79, Wilson and Sperber 2004: 610, using the terminology of Carnap 1950) reinforces this conclusion. Since there is no absolute standard for measuring relevance across individuals and times, there is no manner in which one can decide how many implications will have to be shared between interlocutors for communicative success to be achieved by a given utterance type. This has also been a reason for my reluctance to give a definition of communicative success that would apply across individuals and times (see footnote 16). But since communication always happens in a context, and since the success/failure of communication has to be assessed from the theorist's omniscient third-party perspective, I believe that the lack of such type level predictions does not count against the theory.

1.4.4. The sharing of thoughts vs. the sharing of implications

Even though I have argued in this chapter that its reliance on similarity of thoughts is no obstacle to Relevance Theory's ability to explain successful communication, it is important to emphasise that pushing the problem of similarity into the domain of implications does not make it disappear. Since similarity presupposes identity, it follows that shared implications themselves cannot be merely similar, in that any appeal to similarity must rest on strict identity at another level. So if the claim that thoughts are similar if and only if they share implications is to have any force, the implications themselves must ultimately be literally the same.

The moral of the story then becomes: if there is any kind of sharing going on in cognition and communication, there has to be something, somewhere, that is *literally* shared, i.e. identical. For Cappelen and Lepore (2005; 2006; 2007), it is the thoughts themselves. For relevance theorists, it is a subset of logical and contextual implications[22]. But the question remains: what makes the relevance theoretic story a more plausible option than the traditional Fregean idea of thought-sharing advocated by Cappelen and Lepore? Why is it better that the locus of identity be at the level of logical and contextual implications rather than thoughts?

[22] And as several people have pointed out to me; the notion of mutual cognitive environment also presupposes literal sharing of assumptions. Some subset of assumptions manifest in my cognitive environment will have to be identical to some subset manifest in yours, or as Sperber and Wilson (1995: 41) put it: "the total shared cognitive environment of two people is the intersection of their two total cognitive environments, i.e. the set of all facts that are manifest to them both."

This, of course, depends. It depends not only on the nature of contextual implications, but also on what notion of 'thought' one is operating with. Relevance Theory is concerned with an "individualist psychological notion of thought" (Carston 2002: 33), on which a thought and the proposition it expresses come apart (Carston 2002: 30). In this framework, contextual implications - which may be manifest to an individual without actually being entertained - come a lot cheaper (in the sense that they are easier to share) than thoughts.

The idea behind this "psychological notion", which Relevance Theory has taken from Jerry Fodor's seminal work on the Language of Thought (Fodor 1975), is that the content of a thought can be distinguished from the vehicle which carries it. A thought, according to Fodor's account, involves both a Mode of Presentation and a (referential) semantics, i.e. a form and a content. Committed as he is to physicalism, Fodor (see e.g. his 1998a: 15-22, 2008: chapter 7) requires a thought to have some sort of material realization, and therefore involve some type of activity in the brain. Thus, the thought 'Rosanne is not at work today' will have both a content (or express the proposition that) ROSANNE IS NOT AT WORK TODAY and (in the case of a human; physical-neuronal) a Mode of Presentation. Though Fodor wants contents to be type identical across speakers (as a result of a "non-negotiable" publicity constraint, see Fodor 1998a: 28-34 and chapter 2 of this thesis), it is not clear that this is a reasonable demand on the vehicles of thought.

In fact, Fodor explicitly rejects a type identity condition on the physical realizations of thoughts, commenting that "it surely can't be taken for granted that a certain basic concept is realized by the very same neurological property in *different* minds. What about Martians? Or infants? Or dogs and cats? Or, for all I know, you and me?" (2008: 89). Even if two people have thoughts with exactly the same content, it cannot be taken for granted that these thoughts are physically realized in exactly the same way (involving the same type of neurons in the same brain region), since this rules out a priori that creatures with different types of brains can think "the same things".

Sperber and Wilson share Fodor's doubts about the possibility of "typing tokens of primitive mental representations (…) by their neurology" (Fodor 2008: 90) and take this to show that the vehicles of at least some types of thought may be inherently private[23]. They comment

[23] This is a view somewhat stronger than the one Fodor (2008) advocates. He does not suggest giving up entirely on the publicity requirement on the MoP's of concepts and thoughts, but instead endorses functionalism "about the relation between computational psychology and its various physical implementations" (2008: 90). I do not think the position one takes on the publicity of MoPs has a bearing on the issues discussed in this chapter, but will return to the topic in chapter 3, section 3.3.5.

> "It seems plausible that in our internal language we often fix time and space references not in terms of universal co-ordinates, but in terms of a private logbook and an ego-centred map; furthermore, most kinds of reference – to people or events for instance – can be fixed in terms of these private time and space co-ordinates" (Sperber and Wilson 1995: 192)

Sperber and Wilson conclude not only that natural language is too weak to encode all humanly thinkable thoughts, but that it seems "neither paradoxical nor counterintuitive to say that there are thoughts that we cannot exactly share, and that communication can be successful without resulting in an exact duplication of thoughts in communicator and audience" (1995: 193). Thus, Modes of Presentation cannot be straightforwardly assumed to be of the same type across thinkers. However, contrary to Cappelen and Lepore's claims, this does not entail giving up on the sharing of *content*.

For contextual implications are still perfectly sharable on the Relevance Theoretic account. Implications in general, and contextual implications in particular, are not necessarily physically instantiated, and do not always need to be entertained by a thinker. They therefore do not involve the subjective mode of presentation inherent to thoughts. They can, of course, come to be represented and thereby physically realized in a speaker or hearer, but for purposes of pragmatic theory, they need only exist as abstract contents – tools by which a theorist can assess the import of a given utterance or thought.

The fact that implications need not be entertained as fully-fledged, physically realized thoughts has an obvious consequence. All thoughts and utterances have a huge range of logical and contextual implications, many of which are not relevant enough to be computed and represented by a relevance-oriented cognitive system. I tried to show this in discussing the misconstrual of SICK above, where I claimed that the thought that 'Rosanne is not at work today' may contextually imply that no one logged in to Rosanne's computer this morning, or that no one is sitting in Rosanne's chair right now, or that no one whose first name started with 'R' is at work today, and so on. These are clear cases of contextual implications that in normal circumstances would not make a worthwhile difference to the individual's representation of the world, and would therefore not be computed.

Literally sharing thoughts, then, comes out as involving more than sharing implications, since in order to have the same thought as another individual, one would have to share not only content but also a physical form, the (neuronal) process individuating the thought's mode of presentation. Though it may well be that the publicity of thoughts can be cashed out by other means than type identity, arguing that the locus of identity is to be found at the level of contextual implications is still theoretically preferable. Sharing an implication

involves only identity of content, whereas the traditional account involves sharing both the content and form of a thought.

1.5. Conclusion

In this chapter I have presented a radical pragmatic account of communicated meaning and content sharing. I have discussed Relevance Theory's claim that communication involves a speaker engaging in ostensive behaviour as evidence of her intention to convey a certain thought (or set of thoughts). I have shown how, according to Sperber and Wilson, if communication is successful, the thought that the hearer ends up attributing to the speaker will share logical and contextual implications with the speaker's own thought, regardless of whether it is identical or merely similar to that thought[24]. I have tried to show how this addresses Cappelen and Lepore's worry about Relevance Theory being committed to a Non-Shared Content principle.

By subjecting the notion of shared contextual implications to scrutiny, I also hope to have shown how Sperber and Wilson's view of linguistic interaction is easier to defend than a more traditional view of successful communication as necessarily involving literally shared thoughts. Of course, whether Relevance Theory really is theoretically preferable to the traditional position will depend on much more careful examination of the details of alternative approaches than I am able to offer here.

Since an evaluation of other approaches to communication is outside the scope of my thesis, I will not pursue this point further, but it seems to me that more work needs to be put into examining how key theoretical terms in this debate are being interpreted by both sides. It may be, for instance, that the premises of the debate between minimalists and radical pragmaticians would alter substantially if it turned out that Cappelen and Lepore were relying on some other conception of thought than the "individualist, psychological notion" (Carston 2002: 33) that Relevance Theory makes use of. It if turned out that Cappelen and Lepore read Relevance Theory as denying that the sharing of thoughts *in a Fregean sense* (where these are abstract objects, what Frege: 1918 terms "gedanke") plays any role in successful communication, this reading is clearly false.

---

[24] Though the hearer has to end up with a thought that is relevantly similar to the one the speaker had in mind for communication to be successful, it is not crucial (in most situations) that he entertains it as a belief. He may merely *metarepresent* the assumption as a belief of the speaker or somebody else, the only thing the account is committed to is the relation between the objects of whatever attitude the interlocutors have. I take Cappelen and Lepore to agree with this, so the issue of believing and accepting is not really at stake in the debate.

In fact, Frege's notion of *gedanke*, is closer to what Sperber and Wilson call contextual implications, which I argued were best seen as public and sharable in Relevance Theory. At least, nothing in their position entails a denial of the sharability of *gedanke*. What Sperber and Wilson deny is that psychologised thoughts (what I guess Frege would label "vorstellung") are sharable in the same way that Frege took *gedanke* (but not *vorstellung*) to be. As long as Cappelen and Lepore (2007) fail to specify what they mean by "thought", it is still an open possibility that there is not much disagreement between minimalism and radical pragmatics on the notion of what is required for successful communication to take place after all[25].

Underlying my contribution to this debate, there are also unanswered question about the notion of "content" and how this is individuated. In the final sections of this chapter, I have argued that for thoughts to be similar in content, they must have at least some logical or contextual implications which are identical in content. But what is this content, and to what extent can it plausibly be seen as public? If Relevance Theory claims that the Mode of Presentation of many thoughts is inherently private, what guarantees that the content of these thoughts is shareable?

These are issues that I will take up again in the next two chapters. There, I will look at what input lexical semantics provides to the pragmatic interpretation process and what Relevance Theory takes to be the context-independent (encoded) meaning of a word. I will subject to closer scrutiny the view of conceptual content which underpins recent research in lexical pragmatics, and ask how well the relevance theoretic program squares with theoretical constraints on concept possession.

In so doing, I will return to some of the questions with which I introduced this chapter, trying to flesh out and re-examine the relevance-theoretic view of word meaning that I have made use of and defended so far.

[25] This might be a surprising outcome of the debate, given the way Relevance Theory is taken by Cappelen and Lepore to be in stark opposition to minimalism. But, then again, some of the key misunderstandings in Cappelen and Lepore's reading of Relevance Theory, identified by Carston (2006) and Wedgwood (2007), may mask the fact that their alternative account shares quite a lot of central assumptions with Sperber and Wilson's.

2. The publicity of meaning and the problem of translation: Merging Relevance Theory and the Computational Theory of Mind

...as the American television star Miss Piggy once remarked, it is inadvisable to attempt to lift more than you can eat. Semantics commits suicide when it tries to swallow the world.
- Jerry Fodor

2.1. Introduction

In the previous chapter, I introduced the notions of word and sentence meaning from the perspective of Relevance Theory. I showed how Sperber and Wilson (1995) see the linguistically encoded inputs to the interpretive mechanism as incomplete "assumption schemas", with the pragmatic processes given the task of developing them into something fully propositional. Part of this task was taken to involve the broadening/narrowing of concepts that were seen as too specific or too general to yield enough cognitive effects in a given context.

These concepts are assumed to be the context-independent meanings of words, the hypothesis being that they get their content from somehow standing in an appropriate relationship to something in the world. In this chapter, I will look more closely at this hypothesis, and at the notion of concepts which underlies work in relevance theoretic pragmatics. The aim of doing this is to get a better insight into the questions about word meaning with which I started out in this thesis. The hope is that an inquiry into the nature of concepts as described in the radical pragmatic account I have been defending so far will shed light on how concepts can have content in the first place, and how this content is related to the meanings lexical items can express.

I will start by looking in more detail at the machinery of lexical pragmatics which I touched on in the previous chapter. In section 2.2, I discuss the relevance theoretic approach to *lexical modulation* and *ad hoc concept construction* and show how this approach can help to analyse a variety of cases of the linguistic underdeterminacy problem. In section 2.3, I go on to discuss the view of concepts put forward and defended by Jerry Fodor. His treatment of lexical concepts as atomic entities with no internal structure is largely endorsed by Relevance Theory, and I show how concepts which lack internal structure are seen by Fodor (focusing on his 1998a) as getting their content from standing in a lawful mind-world relation.

In 2.3.1 I introduce the broader framework of Computational Theory of Mind, in which the notion of *concepts* plays an important part, and outline some of the cognitive roles concepts are supposed to play in this theory. In section 2.3.2 I look in some detail at Fodor's motivations and reasons for treating concepts as atomic, showing how he argues that the alternatives to what he calls *informational semantics* are theoretically inadequate. In 2.3.3 I discuss some objections to Fodor's theory, and conclude that even though these are serious, they are answerable in principle when the resources and scope of Fodor's semantic theory are properly clarified.

In section 2.3.4, I show how Fodor's version of informational semantics, despite being thoroughly externalist, can be seen as metaphysically and epistemologically neutral. I argue that this makes it resistant to counter-examples from ontologically problematic entities, leaving questions about what is out there in the world up to empirical investigation. I suggest, though, that this conception of semantics is quite thin and needs to be supplemented by other theories in order to answer some of the overarching questions about meaning. In particular, it would benefit from being supplemented by a story about how concepts are acquired, an issue I return to in part II of the thesis.

In section 2.4, I take a step back and look at how Fodor's version of informational semantics, and the constraints he sets on concept possession, square with Relevance Theory's view of concepts and word meanings. Section 2.4.1 argues that to a certain extent, radical pragmatics can serve to fill in some of the gaps in Fodor's account, thereby helping it respond to charges of explanatory inadequacy. Despite this desirable outcome of the RT/IS coupling, I show, in section 2.4.2, that one constraint Fodor sets on concepts may prove to be an obstacle to the merger. I argue that maintaining a principled, a priori view of the publicity of concepts is not automatically compatible with the claim that there are words which encode general concepts with very broad denotations.

In particular, I suggest that the issue of how meanings are related across languages is problematic for Relevance Theory and argue, in section 2.4.3, that accounting for the content of concepts encoded by (near-synonymous) lexical items in different languages without violating the publicity constraint is no easy task. In section 2.4.4 I consider whether an appeal to *properties* can help negotiate the publicity constraint, before concluding that an account which can explain cross-linguistic meaning relations without relying on a particular metaphysics would be methodologically preferable. Section 2.5 provides some further motivations for developing such an account, a task to which chapter 3 is be devoted.

2.2. Word meaning and lexical pragmatics

As highlighted in chapter 1, Relevance Theory sees the gap between the material encoded by natural language utterances and the meanings they are used to convey as so great as to give rise to a radical proliferation of pragmatic processes in interpreting utterances.

Although a unitary pragmatic process driven by expectations of relevance is ultimately responsible for bridging the gap between the meaning of a linguistic string and the explicitly communicated content, there are several pragmatic sub-tasks which the hearer has to carry out in linguistic interpretation. There is *disambiguation*, where the hearer has to identify which of two (or more) homophonic and/or homographic lexical items ('bank', 'race', 'fly') the speaker was using on a particular occasion.

There is *saturation*, which is involved in the interpretation of context-sensitive expressions such as indexicals and demonstratives (e.g. 'I', 'you', 'him', 'this', 'there', 'here', 'now', 'yesterday'). According to Carston (2004; who adopts the vocabulary of Recanati 1993), saturation is a linguistically mandated pragmatic process, in that no proposition will be expressed by an utterance containing an unsaturated expression. For instance, an utterance of 'I am here now' cannot be regarded as true or false until the lexical items 'I', 'here' and 'now' have had their referents pragmatically assigned.

In addition to the linguistically motivated processes of disambiguation and saturation (which are generally agreed to contribute to the explicit content of utterances[26]), defenders of radical pragmatics claim that *free enrichment* processes are responsible for the recovery of various linguistically unencoded elements. Free enrichment is pragmatically motivated, in that it takes as input material which is truth-evaluable as it stands, and enriches it in order to satisfy expectations of relevance. The following are examples (taken from Carston 2004, who herself has taken them from various sources; the bracketed material is unpronounced but assumed to be recovered given the appropriate discourse context):

35. Something [dramatic] has happened
36. Jack and Jill went up the hill [together]
37. Louise has always been a great lecturer [since she's been a lecturer]
38. Sue got a PhD and [then] became a lecturer.
39. Mary left Paul and [as a consequence] he became clinically depressed

[26] Though it of course varies from theory to theory which expression types are seen as constituting this "genuinely" context-sensitive category.

A separate pragmatic process, though one that is also motivated by pragmatic concerns, is *lexical modulation*, or *ad hoc concept construction*. Some examples are given in 40-46:

40. The crowd was SILENT* [≈'there were hardly any sounds to be heard from the crowd']
41. Is there more beer? Cathy's glass is EMPTY* [≈'the glass contains an insignificant amount of liquid']
42. Ralph DRINKS* [≈'drinks more alcohol than he strictly speaking should']
43. John is TIRED* [≈'tired from having run a marathon']
44. John is TIRED** [≈'tired from having stayed up all night']
45. The apple is RED* [≈'red on the peel']
46. The watermelon is RED** [≈'red on the flesh']

The idea behind the notion of an ad hoc concept is that a word (e.g. 'silent', 'drinks', 'tired', 'red' in examples 40-46) is used to communicate an occasion-specific concept which is different from the one it linguistically encodes, but shares some encyclopaedic properties with the encoded concept, and hence gives rise to some of the same contextual implications. The word 'silent', for instance, is taken to encode a concept SILENT that picks out states of affairs where there is no sound whatsoever. It goes without saying that there will be very few situations which can be described as literally SILENT, and the encoded meaning will have to be *broadened* in order to yield appropriate (and true) contextual implications on the particular occasion of use. The same goes for EMPTY, DRY, BALD, FLAT and other concepts which are taken by Relevance Theory to have highly restricted denotations.

For the words 'drinks', 'tired' and 'red' in examples 47-51, the situation is a little different. Here, the concepts assumed to be encoded will have to be *narrowed* (i.e. made more specific) to yield appropriate cognitive effects in a given discourse context. Everybody (literally) drinks something every now and then, so the claim that 'Ralph DRINKS' is not relevant unless 'drink' is appropriately enriched to something like DRINKS* [≈'drinks more alcohol than he should'] in a context where there is, for instance, a question about Ralph's ways with wine. Similarly, someone can be tired in different ways, or to different degrees, and in interpreting 43 or 44, the general concept TIRED may have to be pragmatically narrowed in different ways depending on the context. And in order for the use of the word 'red' to achieve relevance, it may have to be understood as picking out a particular shade of red, distributed over the object in a particular way. The process by which this *lexical modulation* happens is fuelled by the search for cognitive effects, and guided by the relevance theoretic comprehension heuristic, as outlined in chapter 1.

The last few years have seen great advances in the field of *lexical pragmatics* (with important relevance-theoretic contributions by Carston 1997, 2002; Sperber and Wilson 1998; Sperber and Wilson 2008; Vega Moreno 2007; Wilson 2003; Wilson and Carston 2006; Wilson and Carston 2007, in addition to interesting work by theorists using other frameworks, e.g. Blutner 1998; 2004) where a range of what are traditionally seen as distinct tropes or figures have been reanalysed as outcomes of a unitary, relevance-driven process. For instance, according to Sperber and Wilson (2008) and Wilson and Carston (2007), approximation, hyperbole and metaphor can be placed on a continuum of gradual broadening, where an approximate use lies closest to the literal meaning and hyperbole and metaphor involve greater departures.

On this approach, examples like 40 (The crowd was SILENT*) or 41 (Cathy's glass is EMPTY*) are seen as involving marginal broadenings of the encoded meaning, while the following lie further out on the continuum (examples adapted from Sperber and Wilson 2008; Wilson and Carston 2007):

47. The bathwater's BOILING*
48. *Falling Man* PUT-ME-TO-SLEEP*
49. Joan is the KINDEST-PERSON-ON-EARTH*
50. Joan is an ANGEL*
51. Archie's a MAGICIAN*

If 47 is understood as conveying that a bath is uncomfortably hot, without necessarily reaching 100 degrees Celsius, the case would traditionally be classified as a hyperbole. The utterance in 48 can be used felicitously to describe a situation where the reader of the book in question didn't actually fall asleep, but was merely in a state of extreme drowsiness. Similarly, the utterer of 49 presumably does not have grounds to claim that Joan is the kindest person walking the face of the planet, and so uses the expression loosely to convey a particularly high degree of benevolence. In each case, a subset of the encyclopaedic properties of the linguistically encoded concept is added to the context and used to derive an appropriate set of cognitive effects.

According to Sperber, Wilson and Carston, the metaphors in 50 and 51 are extreme cases of broadening, in that they involve a fairly radical departure from the original denotation of ANGEL and MAGICIAN – a set of divine creatures and people with supernatural powers, respectively. Many metaphorical cases are taken to involve narrowing as well as broadening, in that the input to broadening is only a particular subset of the original denotation. Thus, in the case of ANGEL, the interpretation conveyed would be based on the properties of good

angels, rather than of dark angels, or avenging angels, who also fall within the linguistically-specified denotation of 'angel'. Many metaphors, then, are special in that they involve both narrowing and broadening in parallel, and as such illustrate the complementary nature of the two processes.

Other linguistic phenomena that seem to be treatable in terms of this lexical pragmatic approach include idioms ('Bill kicked the bucket last night', see e.g. Vega Moreno 2007), category extensions ('do you mind Xeroxing these for me?', 'for suitcases; pink is the new black'), neologisms ('he Houdinied his way out of the closet', see Wilson 2003) and slang ('Emma came to my house last night, it was so random', see Kjoll 2007). The upshot is that analysing the meaning conveyed by the use of words in terms of narrowing and broadening proves to be a very potent linguistic tool.

But as with all new theories, the lexical pragmatic enterprise raises some questions and leaves some issues unresolved. One of these concerns the reliance of Relevance Theory on the idea that most words in natural language encode concepts[27] – seen as mental items that are stored in long-term memory and figure as the constituents of thoughts. Sperber and Wilson (1995), following the writings of Jerry Fodor, hold that many of these concepts are *atomic*, in the sense that "the meaning of most words cannot be defined in terms of, or decomposed into, more primitive concepts" (Sperber and Wilson 1995: 91).

Carston also bases her lexical pragmatic work on the assumption that morphologically simple words encode *atomic* concepts: "they don't have definitions (sets of necessary and sufficient component features) and they are not structured around prototypes or bundles of stereotypical features" (Carston 2002: 321). In Relevance Theory, the encoding relation is seen as semantic and direct, in that the word 'dog' will always activate and be activated by the concept DOG, and the word 'tired' will always activate and be activated by the concept TIRED, even though these may not be communicated by uttering the phrase 'tired dog' and are "replaced" by the more relevant ad hoc concepts TIRED* and DOG*. In other words, *linguistic semantics* links a word to an encoded concept, and a *pragmatic* process governed by the search for relevance leads to adjustment of the communicated meaning.

Aside from this, there is little explicit discussion in the RT literature on what concepts are and what role they play. Perhaps as a consequence, some critical questions about the nature of word meaning in Relevance Theory have been raised in the last few years (starting

[27] These are the so called *content words*, to be contrasted with *function words*, such as the articles 'the', 'a', pronouns, particles and connectives, which are taken to encode not concepts but rules or *procedures* (see Wilson and Sperber 1993, Wilson 2009).

with Carston 2002: chapter 5; followed by Burton-Roberts 2007; Fretheim 2008; Groefsema 2007; Vicente 2005; Young 2006; Vicente and Martinez Manrique 2010; see also Reboul 2008: 523, who claims that "there's a longstanding tension (...) between the adoption of an (atomistic and externalist) view of concepts (such as Fodor's) and the description that Relevance Theory in fact gives to concepts"). Both the idea that the relation between words and concepts is one of encoding and the idea that concepts really play a role at the foundations of cognition and communication have been subjected to scrutiny, with various other proposals about alternative ways to see word meaning put forward.

What is at stake in a discussion of the relation between words and concepts is the nature of word meaning and representation. Asking how one gets from lexical items like 'tired' and 'dog' to the mental items they express amounts to subjecting a whole theory of thought and language to scrutiny. In order to get a clearer idea about whether concepts are plausible candidates for being the meanings of words, and whether the relation between them can be seen as one of encoding, I propose to take a step back from pragmatic theory and venture into the philosophy of mind.

Within Relevance Theory, the view of words as encoding unstructured atomic concepts goes back to Jerry Fodor's work within the Representational/Computational Theory of Mind (CTM for short, see Fodor 1975; 1981b; 1990; 1994; 1998a; 2003; 2004; 2008; for an overview see Cain 2002: chapters 3-5). The CTM is a theory of the workings of the mind, where concepts play an important role in explaining some of what are taken to be the central features of human thinking. Concepts then, despite the fact that they are not directly observable, are supposed to behave in ways that allow the theorist to explain human cognition. From this, one can derive certain requirements on concept possession, the nature of concepts and the role they are expected to play in a theory.

A natural way to start considering whether and in what way concepts can serve as the meanings of words is to look at the theoretical requirements on something's being a concept. Once one knows what work they have to do in the cognitive domain, one may get a clearer picture of how they can also plausibly be seen as encoded by words. What I will do in what follows, then, is to outline the theory of concepts put forward in Fodor's recent writings (in particular his 1998a; 2004 and 2008) and situate it in the larger context of Computational Theory of Mind. I will go through the central tenets of CTM, outline the motivation behind the theory of concepts and also address some of the criticisms the theory has met. As well as illuminating the topic of this thesis – the meaning, use and representation of words – I hope

that once this is done, I will be in a better position to give some answers to the question of how mental and linguistic items are related.

2.3. Computational Theory of Mind and Conceptual Atomism

2.3.1. The Language of Thought hypothesis

Thinking, according to Jerry Fodor, is a mental process. It is a mental process that is essentially *productive* and *systematic*, which means that one can think things one has never thought before and one can do so with considerable reliability. For instance, if I can think the thought JOHN LOVES JILL, I can also think the thought JILL LOVES JOHN. Also, if I can think the thoughts BROWN COWS ARE BEAUTIFUL and HAIRY DOGS ARE DIRTY, I can also, quite plausibly, think HAIRY COWS ARE DIRTY and BROWN DOGS ARE BEAUTIFUL etc.

Throughout his career, Fodor has argued that the most plausible way to account for the productivity and systematicity of thought is to claim that the system used in thinking consists of *discrete* (separate) and *fully compositional* units, along with rules for combining them. The constituents of the thought JOHN LOVES JILL are three discrete units JOHN, JILL and LOVES, rather than one long unit JOHN-LOVES-JILL, since otherwise the theory would fail to capture the generalisations above. The mantra is that the meaning of a thought, or sub-part of a thought (such as HAIRY COW), is exhausted by the meanings of its constituents (in this case HAIRY and COW) and the rules by which they combine.

Thinking, more specifically, is a type of computation. Fodor sees thought as involving truth-preserving formal operations over token mental representations (symbols), very much like the system first hypothesised and built by the computer scientist Alan Turing in the nineteen-forties. Fodor regards this as a largely innocuous view, and he attributes some form or other of it to Plato, Aristotle, Descartes, Hume and the British Empiricists. He labels it the Computational Theory of Mind, and formulates its key idea in the following way: "[t]okens of symbols are physical objects with semantic properties [and] computations are those causal relations among symbols which reliably respect semantic properties of the relata" (Fodor 1998a: 10). The symbols computed over are, according to Fodor, primitive, unanalysable concepts, mental particulars which work as *the atoms of meaning*.

Fodor's particular version of the Computational Theory of Mind is known as the Language of Thought hypothesis, since the system over which computations take place is held to be language-like, with its own syntax and semantics. This, for Fodor, is the only plausible basis for explaining the defining properties of human thoughts: productivity and

systematicity[28]. Sentences in the Language of Thought, like sentences in natural languages, consist of strings of discrete units which can be combined to yield new meaningful strings: "Just as the semantics of sentences are constructs out of the semantics of words, so the semantics of thoughts are constructions out of the semantics of the concepts that are their constituents" (Fodor 2008: 19).

But crucially, the medium of thought is not natural language (English, Japanese, Wolof etc.) as there are a number of properties that set LoT apart from spoken languages. Most apparent is the lack of phonological features; on Fodor's account, LoT symbols are also *amodal*, which allows the system to integrate and compute over information from all sensory modalities. Also, the medium of thought is free of the underdeterminacy of natural languages, and thoughts are taken to express only full, truth-evaluable propositions. The reason for this is that "whereas it's thoughts that equivocal sentences equivocate between (…), there doesn't seem to be anything comparable around that could serve to disequivocate thoughts" (Fodor 2003: 156; Fodor 1998b: chapter 6; but see Carston 2002: 74-83 for discussion).

Making LoT independent in principle from natural languages has several explanatory virtues. One is that it allows for an explanation of the intelligent behaviour of non-linguistic agents (such as animals and babies), whose cognitive systems also possess the features of productivity and systematicity (see e.g. Fodor 1987: 152-153). It also makes it possible to explain the dissociation between intelligence and linguistic skills in humans (Allott 2008: 112; Smith and Tsimpli 1995). And importantly, the claim that intelligent thought is conducted in an amodal system of symbols which is in principle independent of natural language preserves neutrality towards the question of exactly what content and rules a thinker (be it a human or a dog or a Martian) employs to do her thinking. Fodor sees this as a wholly empirical question, and the LoT hypothesis therefore does not entail a commitment to any particular view on how language affects thought[29].

---

[28] Originally, the argument was that the Language of Thought hypothesis is the *only one* available for scientists interested in explaining the productivity of cognition, but the emergence of *connectionism* some time after Fodor's 1975 book *Language of Thought* has changed that. Fodor has engaged in the debate with connectionists at various points since, and consistently accuses them of being unable to capture the *systematicity* of thought. See Fodor (1997), Fodor and McLaughlin (1990) and Fodor and Pylyshyn (1988) for the original arguments, and Cain (2002: 102-111) for a review.

[29] This is the thorny issue of linguistic relativity, one of the most debated in a century's worth of work in linguistics and cognitive science. Though Fodor himself has expressed strong feelings on the matter (see Pinker 1994: 404-405, Pinker also, in chapter 3, gives an entertaining historical overview of the debate and provides some very good arguments against the Sapir-Whorf hypothesis), I take the LoT's neutrality on the matter to be a great virtue, and an excellent argument for using the Computational Theory of Mind as a working assumption if one is interested in the relationship between language and thought.

2.3.2. Mental content

The Language of Thought hypothesis provides the theoretical backdrop for Fodor's favoured view of content, and thereby places some clear constraints on what concepts can be. But noticeably, much of the Computational Theory of Mind's explanatory potential also *rests* on what story one tells about concepts. Traditionally, concepts have been thought of as mental images, definitions, bundles of semantic features or prototypes, but Fodor insists that none of these squares with the central tenets of CTM (or else that they are theoretically incoherent or explanatorily inadequate). He therefore makes a radical departure from many of his CTM allies on the content issue – proposing what he calls *conceptual atomism*.

Fodor's arguments for his atomistic account are essentially negative, and almost all are derived from the initial requirements on the theory to explain compositionality and systematicity[30]. He also emphasises the importance for a theory of content of a principled *stability* (the fact that it should be possible to maintain the same thought over a period of time) and *publicity* (the fact that it should be possible to share the same thought across individual thinkers). The strategy he has adopted is to pick apart the tenets of alternative accounts, claiming that none of them satisfies all the non-negotiable requirements on a theory of content, only one option thereby remaining in logical space.

The position he ends up with is a merger of two philosophical theses, *informational semantics* and *conceptual atomism*. Atomism, as explained in the Sperber and Wilson (1995) and Carston (2002) quotes above, claims that concepts have no internal structure. Informational semantics, on the other hand, is the hypothesis that the content of primitive conceptual items is determined by some sort of constitutive link between the mind and the world. For Fodor, this is a *nomic* (lawful) relation that reliably locks a concept to a mind-external property. The two theses are naturally compatible, according to Fodor, since externalism claims that "the content of a thought depends on its external relations; on the way that the thought is related to the world, not on the way that it is related to other thoughts" (Fodor 1994: 4)[31]. So the contents of mind-internal concepts like DOG, COFFEE, WATER, TIRED,

[30] And, probably, from the fact that he works within a naturalistic programme. Fodor is therefore heavily committed to *physicalism*, the thesis that "only matter has causal powers" (Fodor 2008: 196), so that abstract objects (for instance, mere propositions) are not enough to cause behaviours. His first *non-negotiable* condition on what concepts have to be, that "they satisfy whatever ontological conditions have to be met by things that functions as mental causes and effects" (1998a: 23), essentially follows from this.

[31] There is no entailment relation between them, though. Despite Fodor's (1998a) claim that arguments for atomism are also reasons to favour externalism, "the issue of whether lexical concepts are internally structured is entirely independent of the issue of whether content is determined by informational relations or conceptual [*sic*], or inferential role" (Rives 2009: 215).

BREAK, FAST, BREAKFAST and so on are constituted by their standing in a certain relation to the world, and in Fodor's view, this rules out the possibility that there are content-constitutive *internal* connections between concepts. Fodor ends up with the view that having a concept DOG is not (even partly) determined by acquired beliefs about animals, barking or tail-wagging, but rather by being in a mind-world relation in virtue of which the DOG concept has the content that it does.

Fodor describes the psycho-causal process that mediates between the mind and the world in this way a "mechanism of 'semantic access'", which is "what sustains our ability to think *about* things" (1998a: 75). The content of the concept DOG is exhausted by its being 'locked' to the property *being a dog*, and acquiring the concept DOG reduces to a process of "getting *nomologically locked* to the property that the concept expresses" (Fodor 1998a: 125). Coming across the property *being a dog* (as instantiated in an exemplar of *canis lupus familiaris*) for the first time, the individual will typically open up a conceptual address DOG which will function as a constituent in his reasoning/imagining/wondering/thinking about dogs in the future.

This conceptual address may be thought of as a *mental file* which provides a gateway in memory to all sorts of stored information about dogs. All beliefs about dogs (whether idiosyncratic or factual) are stored as assumptions in a mind-internal file (in Relevance Theory known as an encyclopaedic entry) accessible through the file name DOG. My (generic) beliefs that DOGS ARE ANIMALS, SOME DOGS BITE, MOST DOGS BARK, DOGS ARE MAN'S BEST FRIEND are stored together with thoughts about particular dogs I have encountered or owned, and may all be potentially accessed upon the tokening of a DOG symbol.

Together, the Computational Theory of Mind and informational semantics/ conceptual atomism lead to a theory of thought according to which we think in file names: "tokens of file names serve both as the constituents of our thoughts and as the Mentalese expressions that we use to refer to the things we think about" (Fodor 2008: 94-95)[32]. According to Fodor, "[i]t's only *the name* of the M(house) file (not the file itself) that serves as a constituent of one's thoughts when one thinks about houses [i.e. when one thinks about houses 'as such'= de dicto] (2008: 97)[33].

---

[32] Compare Sperber and Wilson (1995: 86), who claim that "Formally, we assume that each concept consists of a label, or address, which performs two different and complementary functions. First, it appears as an address in memory, a heading under which various types of information can be stored and retrieved. Second, it may appear as a constituent of a logical form, to whose presence the deductive rules may be sensitive".

[33] It is important to maintain a principled distinction between the name of a file and whatever it contains, since only in this way is the directedness of thought explained. When one thinks about dogs or houses, one generally

Though the beliefs stored in the filing cabinet under a given file name undoubtedly differ in strength and degree of accessibility (see Sperber and Wilson 1995: chapter 2 for more on this point), none of them holds a special status in virtue of being *constitutive* of the concept in question. That is to say, if the belief DOGS ARE ANIMALS does not figure in my DOG file, I will nevertheless be in possession of a functioning DOG concept, according to Fodor. Clearly, this goes against many people's intuitions about the semantics of 'dog', and most theorists working on concepts and word meaning would treat the belief that DOGS ARE ANIMALS as *analytic*. A thinker who uses a DOG concept that does not in any way activate the concept ANIMAL does not possess a proper, functioning DOG concept, the story goes, and consequently, most lexical semanticists see it as a task for any theory of content to account for why we have intuitions about such inferences.

Fodor is reluctant to grant a role to analytic, content-constitutive beliefs in his theory for a number of reasons, the most striking being his disapproval (to put it mildly) of two philosophical theses; *holism* and *molecularism*. Holism holds that smaller units of meaning are derivative from the bigger systems of which they form a part, and is described as "the doctrine that only whole languages or whole theories or whole belief systems *really* have meanings" (Fodor and Lepore 1992: x).

According to a holistic account of concept possession, all the beliefs an individual has about an object are constitutive of his thoughts about that object, so that all sorts of idiosyncratic assumptions about dogs that everybody has determine their particular DOG concept. This in turn makes an individual's particular DOG concept dependent not only on on his changing, peculiar beliefs about dogs, but also on everything else he believes, since the individual beliefs themselves depend on beliefs about the constituents of the beliefs, and so on indefinitely. The logical consequence is that if your DOG concept is partly constituted by your beliefs that LABRADORS ARE THE MOST BEAUTIFUL DOGS and that THE SMELL OF WET DOGS REMINDS ME OF MY CHILDHOOD SUMMER HOLIDAYS, your DOG concept will depend on idiosyncratic beliefs about SMELL, WET, BEAUTIFUL etc. etc.

This goes against several of Fodor's central assumptions, most notably that concepts should be sharable and stable. Since "one's beliefs change by the millisecond" (Fodor 2004: 35),it follows from holism that nobody shares any concept with anyone, not even with different time-slices of themselves; nobody can repeat a thought or even remember what she used to think (the concepts she used to think with are gone). The fact that holism is

---

does not think everything one knows about (or associates with) the object in question. Fodor (2003) accuses traditional associationist accounts of cognition of not being able to explain reliable directedness.

incompatible with the principle of compositionality (see Fodor and Lepore 1992; Fodor and Lepore 2002: part III for more on this and further arguments) adds further weight to the scales and rules out any use for holistic theories of meaning as a component in CTM.

For someone who accepts Fodor's arguments against holism but still wants to allow for the possibility that there are content-constitutive beliefs or inferences linked to concepts, the only remaining option is *molecularism*. Molecularism, which in its various shapes and forms is prevalent in contemporary philosophy, linguistics and cognitive science, is the idea that "acquiescence in some, but *not* all, of the inferences that a concept licenses is constitutive of having the concept" (Fodor, 2004: 35). This view, on the face of it, would seem acceptable to a CTM proponent (and indeed was defended in Fodor 1975[34]), but Fodor now rejects it on the ground that there is no principled way to answer the question of "*which* C[oncept]-containing inferences are possession conditions for C" (*ibid*).

In Fodor's view, answering the "which-question" amounts to overcoming the challenge of finding a workable analytic/synthetic distinction, which does not seem possible given Quine's (1963) famous arguments against the feasibility of maintaining a clear distinction between analytic and synthetic inferences. Fodor concedes that Quine's arguments are not enough by themselves to dismiss the possibility of content-constitutive inferences, but claims that an independent truth-maker of analyticity is needed for the distinction to do any proper work. Since "nobody knows what analyticity *is*, nobody can give a clear account of what might make ascriptions of analyticity true (/false)", he concludes (2004: 35).

The arguments against analyticity are a recurring ingredient in Fodor's writings on content atomism, but since the publication of his (1998a) *Concepts*, he has started giving more emphasis to what he calls the circularity argument. This argument plays a central role in his *Having concepts* (Fodor 2004) and *LOT 2* (Fodor 2008: chapter 2), and can be summarized with the slogan "having a concept is prior to applying it in inference". The idea, in short, is this: the inference FROM X IS A DOG to X IS AN ANIMAL presupposes the prior possession of the concepts DOG and ANIMAL. The inference cannot therefore be what (solely)

[34] Fodor's views on this topic has changed over the years, from his 'Language of Thought' (1975), where he appealed to so-called *meaning postulates* to constrain the content of concepts, to an intermediate stage where he endorsed an appeal to postulates only for sub-parts of the mental vocabulary. From his 'Having concepts' and on, though, he has given up on the idea that there are content-constraining inferences at play even for logical items like 'and', 'or', 'not' etc. Sperber and Wilson (1995) adopted the idea of meaning postulates used in Fodor's work around and before they published the first edition of Relevance, and still endorse the view that there are content constraining elimination rules stored in the logical entries for concepts (see Sperber and Wilson 1995: 86ff). For a defence of the appeal to meaning postulates against the arguments in Fodor (2004) see Horsey (2006), who argues that the logical entries used in Relevance Theory can help to address what he sees as inadequacies in informational semantics.

determines the content of the concepts; since in order to avoid vicious circularity, it must be possible to specify the content of the inference without reference to the concepts it is constitutive of.

Fodor directs these arguments against what is known as Inferential Role Semantics (IRS), the thesis that the content of a concept is constituted by (some or all of) the computations it enters into in a cognitive system. Fodor (2004; 2008) maintains that any account that takes this as its starting point is wrong-headed, and claims that there is no way to specify conceptual content functionally without succumbing to the above objection. This holds for all types of inferentialist theories, including those that try to get around the problem by claiming that concepts are abilities or sorting procedures.

To make matters worse, inferences, procedures or abilities do not compose systematically, which means that purely inferential approaches are unable to explain productivity and systematicity, according to Fodor, who takes this to rule out all inferentialist theories of content from a CTM explanation. Other theories of content, such as so-called prototype and exemplar theories, fall prey to the same compositionality objection (see Fodor 1998a: chapter 5; Fodor and Lepore 2002: chapter 2; Connolly et al 2007 for experimental evidence and Jönsson and Hampton 2008 and Prinz 2011 for critical discussion). There are some alternative theories which might, initially, seem compatible with the Computational Theory of Mind, among them various decompositional approaches (e.g. Jackendoff 1992; Pustejovsky 1995), but Fodor dismisses them as explanatorily inadequate (Fodor 1998a: chapter 3; Fodor and Lepore 2002: chapter 5)[35]. His conclusion is that the only theory left standing, once the dust has settled, is conceptual atomism.

Though I find his arguments compelling, this is not the place to evaluate Fodor's negative claims about the range of possible accounts of concept possession. My goal here has only been to outline the motivations behind his positive account, laying out some of the initial requirements a theory of content has to meet. It is important for a radical pragmatician to keep these in mind, since they will be relevant if one should decide later that an account of word meaning is better off without conceptual atomism, and thus pursue other options than Fodor's for explaining the content of words (see my chapter 3, section 3.2.1). Considering the amount of criticism the atomistic theory of concepts has received over the years, this may very well be tempting.

---

[35] Other ways of individuating content *internalistically*, for example via grammar (Hinzen 2007), are also seen as succumbing to the charge of explanatory inadequacy, see Kjoll (2009).

In what follows, I will therefore look more closely at Fodor's theory of concepts, and try to defend it against some of the criticisms it faces. I will go through some of the best-known objections and try to pick apart what I see as the substantial issues from those that are based on misunderstandings of the theory and its scope. I do not have the space to discuss all the charges levelled against Fodor here, but will counter a select few, referring readers to the discussions in Cain (2002), Horsey (2006) and Laurence and Margolis (1999a; Margolis and Laurence 2003) as well as the multiple reviews of *Concepts* in *Mind and Language* (Hampton 2000; Keil and Wilson 2000; Landau 2000; Peacocke 2000; Pietroski 2000) and Fodor's reply to the critique in the same issue (Fodor 2000b) for a fuller picture. See also the reviews of *Concepts* by Bach 2000; Gross 2001; Laurence and Margolis 1999b; Levine and Bickhard 1999. Newer discussions of Fodor's work on concepts can be found in a *Synthese* special issue featuring interesting contributions by Edwards (2009), McLaughlin (2009), Schneider (2009a) and Segal (2009).

After discussing some of the main problems with Fodor's theory, I will focus on two major explanatory challenges facing the proponent of informational semantics/conceptual atomism. One comes from considering the relations among lexical meanings across languages, and the other involves the acquisition of concepts for non-perceptual entities. I will argue that, rather than revealing fatal flaws in informational semantics itself, these are points at which other theories might step in and fill the explanatory gaps left open by Fodor. The subsequent chapters of this thesis will therefore be devoted to just such a task.

2.3.3. Objections to informational semantics

The first issue to arise when discussing conceptual atomism (especially among linguists, it seems) is that the account appears to entail *radical nativism*. The argument, here presented by Pinker (2007: 94), is that "If concepts are undefinable, that means they aren't built out of more elementary concepts, which means that they must themselves be elementary concepts, which means they must be innate". The hidden premise here is that elementary(/primitive) concepts are unlearnable, which, combined with Fodor's contention that many concepts are unanalysable yields the unappealing conclusion that such representations as those of CARBURETTOR, TROMBONE, SOFTWARE, WICKET and CAULIFLOWER are innate. Despite the unattractiveness of its outcome, Fodor has at various points endorsed the premises of this argument, thereby prompting his critics to rule out conceptual atomism as a plausible account of thought and semantic content.

But a subtlety missed by these critics is that even though atomic concepts are primitive and therefore unlearnable, "learned' and 'innate' don't exhaust the options" (Fodor 2008: 130). Though it is easy to read Fodor's arguments (dating back to his 1975 book) as defending the innateness of such unlikely concepts as DUNGAREE and BUREAUCRAT, I take Fodor to have been arguing for a methodological point from the beginning. He may be read as holding that the cognitive sciences are in need of a new account of *concept acquisition* if the ontogeny of thought and language is to be explained.

In his recent works, this point comes across more clearly. In *Concepts*, for instance, Fodor argues that "it would be nice if a theory of concepts were to provide a principled account of what's in the primitive conceptual basis, and it would be nice if the principles it appealed to were to draw the distinction at some independently plausible place" (1998a: 28). In other parts of his (1998a) book he outright rejects radical nativism (see his chapter 6), and in his (2008) book he tentatively proposes a model of concept acquisition that does not depend on the innateness of individual concepts, making use instead of the idea that concepts "lock" to mind-dependent properties.

Fodor's theory of concepts, then, is intended to be empirically neutral on the issue of nativism, and leaves it up to actual investigation to determine which concepts are parts of an innate mental repertoire. Even though it might be pointed out that Fodor himself has so far not succeeded in coming up with an adequate account of concept acquisition[36], this is no reason to exclude a priori the possibility that there could be such a theory, thus ruling out a potentially fruitful alternative to both concept learning and radical nativism.

If the nativism problem is the first reason linguists give for steering clear of informational semantics, philosophers are prone to cite the so-called "Frege problem" as a major motivation for not endorsing the theory. This problem relates to how concepts which denote the same object (and lock to the same property) can come to behave differently in a person's mental life. Frege's (2010 [orig. 1892]) original example involved THE MORNING STAR and THE EVENING STAR, which, despite having the same celestial body as a referent (the planet Venus), can nevertheless be used in thoughts and utterances of the type THE MORNING STAR IS THE EVENING STAR.

---

[36] Or, as Horsey (2006: 29, following Laurence and Margolis 1999b) does, claim that Fodor's new notion of "locking" is not much different than his earlier talk of "triggering" of innate concepts, so that it is not "terribly clear that Fodor's new position is less radically nativist than his old position". Horsey uses this as an incentive to "look more closely at how the process of concept acquisition that Fodor proposes might work in practice" (2006: 31), investigating specifically how locking via perceptual mechanism might be analysed. I will return to the issue of concept acquisition in Part II of this thesis.

If "reference" really were all that mattered, this thought should be a tautology of the form a=a, but its potential informativeness shows that this cannot be true. Think of other co-extensive concepts such as BOB DYLAN/ ROBERT ALLEN ZIMMERMAN, RAMBLIN' JACK ELLIOTT/ ELLIOT ADNOPOZ and BYZANTIUM/ CONSTANTINOPLE/ ISTANBUL, where it is perfectly possible for someone to have a thought about Bob Dylan being the greatest poet of the 20th century, while the next moment thinking that Robert Zimmerman (who, let's say, was in the same high school literature class) was a terrible writer.

The original solution proposed by Frege to these types of cases was to claim that content is determined not only by reference, but also by *sense*, leaving him with a two-tiered theory of meaning. Fodor, who does not want senses to be a part of his story about concepts[37] does indeed find these cases worrying, and spends a chapter of his (2008) *LOT 2* on working out a possible solution. The account he gives here is one that was touched on in chapter 1 (section 1.4.4) and involves the Mode of Presentation of the concepts in question. THE MORNING STAR and THE EVENING STAR do indeed have the same referent (the planet *Venus*) and therefore the same content, but in a thinker ignorant of the astronomical facts, they are realized as two different concepts with different Mentalese orthography.

Returning to the notion of *mental files*, Fodor sees each of these concepts as giving access to a different file, with a number of associated *memos* (containing beliefs, assumptions etc.), even if they happen to have the same referent. The crucial point here is that the concepts are realized differently syntactically, and thus play different roles in the mental life of an individual – despite having the same content. This would continue until the ignorant person discovered that both stars are Venus at different times of the day, or that Bob Dylan is a pseudonym for Robert Zimmerman – in which case the respective files might merge and start to behave as coextensive in both content and form (see Fodor 2008: chapter 3 on this point).

Granted, this does not deal with the full range of problems raised by the Frege cases, and it remains to be seen whether the file name story can help to explain why co-extensive concepts fail to substitute in belief reports (Fodor 2009 argues that one has to appeal to pragmatics to find a solution), but I think the account in terms of mental files goes a long way towards showing that one does not necessarily have to appeal to a two-tiered theory of meaning to solve the problem of co-referential concepts (see my chapter 6, section 6.2.2 for further discussion)

---

[37] This is mainly because he thinks his commitment to naturalism rules out senses, which were held by Frege to be abstract objects, see Fodor (2008: 18, n. 34). For discussion see Margolis and Laurence (2007) and Beck (2010).

Another problem many see with Fodor's purely referential theory of meaning is that it lacks the resources to explain intuitions about the *analyticity* of certain terms. So, for instance, the concept KILL does not just pick out acts of killing: many people have the intuition that it *really means* (or at least entails) CAUSE TO DIE. Similarly, one does not really count as having a concept like DOG if one is not able to make the inference DOG ├ ANIMAL, the story goes. For centuries, philosophers have claimed that one of the central tasks of a theory of meaning is to capture such meaning relations.

Some linguists also see it as a central task for a theory of meaning to account for people's intuitions not only about definability, but also synonymy, antonymy, hyponymy, etc., which they take to show that concepts are complex, composite entities (see e.g. Jackendoff 2002; Pinker 1999). By contrast, Fodor, in his recent work does not want to appeal either to decomposition or to an extra level of meaning to account for these intuitions. His failure to say anything instructive about the analytic data does not bother him, since he does not take for granted that the intuitions in question are *semantic* in nature. He offers instead to explain them away[38];

> "Informants, oneself included, can be quite awful at saying what it is that drives their intuitions; sometimes it's just a fragment of underdone potato. This holds all the way from chicken sexing to judgements of grammaticality and modality. Good Quinean that I am, I think that it is always up for grabs what an intuition is an intuition of" (1998a: 86-87).

It may very well be, then, that the intuitions people have about KILL (entailing causing somebody to die) are really about the property expressed by the concept KILL - in which case the issue of analyticity is a concern for metaphysics, not semantics. "It is perfectly consistent", Fodor argues: "to claim that concepts are individuated by the properties they denote, and that *the properties* are individuated by their necessary relations to one another, but to deny that knowing about the necessary relations among the properties is a condition for having the concept" (1998a: 74).

And even though a Fodorian conceptual atomist denies that the inference DOG ├ ANIMAL is constitutive of the concept DOG, this does not mean that it does not play an important role in the mental life of an individual thinking about dogs. As Margolis and

---

[38] Fodor (1998a: 80-86) also tentatively suggests that some of the cases most often advanced in favour of analyticity intuitions might be "one-criterion" concepts, where there are privileged ways of telling whether they apply (see Rey 2009a for critical evaluation). Horsey (2006) suggests that the appeal to meaning postulates stored in the logical entries for concepts (see footnote 34) can help account for what he calls *psycho-semantic* analyticity. Though I think Fodor, in appealing to "Quinean" methodological principles, succeeds in relieving himself of (at least some of) the burden of accounting for intuitions about analyticity, I acknowledge the weight of this objection.

Laurence (2003: 205) have pointed out, "explanatory roles that are often accounted for by a concept's structure needn't actually be explained directly in terms of the concept's nature". In their view, the belief that DOGS ARE ANIMALS figures in the mental file associated with DOG, albeit very strongly and very highly accessible in most cases, which may be the reason why people tend to grant this assumption special status in their metalinguistic judgments.

True, appealing to information merely *associated* with a concept to explain intuitions and behaviour "may seem like a drastic step, but virtually any theory of concepts will do the same in order to explain at least some inferences in which concepts participate" (Margolis and Laurence 2003: 205). If someone has the firm belief that dogs are dangerous, this will cause them to engage in dog-avoiding behaviour, without this strong belief being seen as analytic, as Laurence and Margolis point out.

A purely referential semantics has also come under fire for being too thin to provide a genuine psychological and linguistic explanation. As many authors are quick to point out (e.g. Hinzen 2007; Wilks 2001) it is not hugely informative for a semantic theory to say that 'dog' means DOG, which means *dog,* or that 'house' means HOUSE, which expresses *house*, and this account is absolutely no help to people interested in modelling the workings of the human mind. It is not clear, though, whether this is a genuine objection to informational semantics.

It is an entirely reasonable hypothesis that the contents of people's primitive concepts are constituted solely by a lawful mind-world relation, even though this approach to semantics may be hard to build into an artificial cognitive system. The fact that Fodor's account of concepts is not much use to someone writing a dictionary containing the word 'dog', or does not explain all the workings of the incredibly complex central thought system, simply means that whatever remains unexplained is left to theories of other mechanisms[39].

2.3.4. A metaphysically and epistemologically neutral semantics

It is something of a common theme to many of the charges raised against Fodor that his theory, though it may be logically coherent and lets concepts perform all the tasks generally expected of them (they satisfy the requirements of shareability, stability and compositionality), explains too little of what philosophers, linguists and cognitive scientists are interested in when studying the mind. In this vein, it is often claimed that Fodor's informational semantics, by relying purely on a link between the mind and the world to

[39] Cf. Fodor (1994: 110): "About the best you can do in lexical semantics is: 'chair' means *chair*, 'cat' means *cat*, and, likewise, 'Cicero' means *Cicero*. If these platitudes strike you as unsatisfactory, perhaps you should stop asking the questions of which they are the answers."

individuate content, may provide us with an explanation of how people come to think about chairs, tables and various natural kinds that are locatable in the world, but that the explanation stops here[40].

In ordinary conversation, we speak of houses, trees, water, other people and so on, but we are also able to refer successfully to value, reputation, embarrassment, morality, betrayal, inflation, profits, perfection, achievements, yesterday's news, Tuesday, Morris Halle's Ph.D degree, Mahler's Second Symphony and other types of what the internalist Ray Jackendoff (2002) has labelled "ontologically curious beasts". These are entities which, if they have physical realizations in the world, are not easily categorized or characterized. And if these entities are not physically realized, it has been claimed that there is no way in which informational semantics can explain how we lock to the properties they instantiate.

Though I think this *argument from ontology* poses interesting problems for informational semantics, there is nothing in the theory itself that favours concrete over abstract objects. Contrary to what many critics suggest (e.g. Hampton 2000), a Fodorian is not committed to claiming that entities such as grumpy, kitsch and God must have mind-independent physical realizations somewhere in the world for this content to be grasped by an agent. Fodor's informational semantic account makes an important distinction between semantics and epistemology: that is, between the fact of concepts having content and the manner in which they have and acquire it.

This means that what is important for semantics is not that there are actual causal mechanisms locking a concept to a corresponding property (for instance, via someone perceiving the property's instantiations), but that there *could be* such a locking relation in actual or counterfactual circumstances. "What a thought represents is largely independent of its *actual* causal history if the informational version of externalism is true" argues Fodor (1994: 90), adding that "Thoughts of cats are thoughts *of cats* not because cats *do* cause them but because cats *would* cause them under circumstances that may be largely or entirely counterfactual" (*ibid*).

---

[40] It has been claimed (by e.g. Prinz 2005) that Fodorian concepts are not the right kinds of things to explain how mental representations serve as tools for categorizing, which is the cognitive function most psychologists are interested in when studying the mind. There are a couple of things one could remark to this, the first being that Fodor indeed does assume that "Concepts are categories and are routinely employed as such" (1998a: 24). He does not elaborate, though, on the way he thinks atomistic accounts of concepts can help understand exactly how concepts are in fact applied in categorization, but there is, as far as I can see, no principled argument to the effect that the atomistic story *could not* explain this in the end (see Edwards 2009: especially section 5.2 for discussion).

Even though it might be important for epistemology and a theory of concept acquisition (among other things) to describe the circumstances in which a concept expressing a property would sustain a counterfactual, it is their "*availability* to sustain these counterfactuals, and not the actual history of their operation, that the metaphysics of content cares about", according to Fodor (1994: 119). Fodor therefore emphasizes that "it's *that* your mental structures contrive to resonate to *doghood*, not *how* your mental structures contrive to resonate to *doghood*, that is constitutive of concept possession" (Fodor 1998a: 76), and ends up defending an *epistemologically neutral* semantic position.

This position also ensures that Fodor's theory is *metaphysically neutral*, in the sense that all that matters for semantics is that there is a nomic connection between a concept and a property for semantic access to be sustained. This lawful relation need not be between an actual entity and a concept, since according to informational semantics, the existence of a nomic relation between a concept and a property, however this is mediated, suffices for concept possession. Fodor insists that "'however mediated' should be understood to include, in principle, nomic relations that are not "'mediated at all" (1998a 79). But as Rey (2005a; 2005b; 2009; 2011) points out, and as Fodor himself (1998a) admits, for an approach to semantics which treats the relation between a concept and a corresponding property as lawful (whether it is counter-factually or actually sustained), there must be "laws about everything that we have concepts of" (Fodor 1998a: 146). This may be quite plausible for natural kinds, but is more problematic for a range of other cases.

Rey (2009a: 190), for instance, is sceptical about whether there are "genuine laws relating states of the brain to *unicornhood*, *being a ghost*, or even serious *triangularity*", raising the notorious problem of so-called "empty concepts" for informational semantics. In the mental life of any individual there are concepts of *unrealized* entities, like Sherlock Holmes, Eugene Onegin and unicorns, and even some for *nomologically impossible* entities, like ghost or elf. How does the proponent of informational semantics account for the lawful relation between mental items and entities which do not exist, or for those that cannot exist in any possible world?

I will return in chapter 5 to a fuller investigation of this problem and a proposal about how fictional and nomologically impossible entities are best handled on an informational semantic account, but want to underline here that the prospects for dealing with abstracta on Fodor's approach are not as bleak as is often thought. A serious examination of some of the points in his (1998a) book shows that Fodor, unlike many of his externalist peers, advocates a view of concepts that is genuinely neutral both epistemologically and ontologically.

Fodor maintains that "there can be no primitive concept without a corresponding property for it to lock to" (1998a: 165), and acknowledges the initial implausibility of the idea that there are natural laws which pick out even such rather mundane categories as *doorknobhood* or *being Tuesday*. But according to Fodor, the semantics of concepts for such non-natural kinds as doorknobs is not determined by laws directly about doorknobs and the like: it is determined by laws about *us as humans* and *how our minds resonate* to the property instantiated by the kind.

In a sense, then, many of our concepts express *mind-dependent* properties: for instance, whether an entity is fast, audible, blue, a doorknob or a giraffe "depends, inter alia, on how it affects our minds" (Fodor 1998a: 149). Someone who wants to endorse Fodor's version of informational semantics, then, is not committed to any particular metaphysics about, say, *blueness*, *being a giraffe* or *being a doorknob*. I will argue in chapter 5 that this point extends to the issue of metaphysically impossible entities, and allows for an externalist treatment of the representation of entities that are not realized or even realizable in the world.

The fact that Fodor favours a *nomic-informational* over a *causal-informational* account makes for a fairly thin view of semantics. Many of the things a theorist interested in word meaning might want to study, such as how semantic content is acquired and what sustains it, is to be left to epistemology and "the pragmatics of linguistic communication" (Fodor 2008: 88, n57) [41]. This does not imply, of course, that metaphysics, epistemology or pragmatics is in any way uninteresting or irrelevant to the broader context of word meaning. It just means that explanation of how content is mediated will not be part of the semantics, but rather belongs to a semantic theory's "foreign affairs" (Fodor 1994: 81).

To many, kicking the whole explanation of how concepts are acquired (and by extension, how words are learned) and how semantic access is sustained may make for a more impoverished semantic theory. But while this in one sense is true, the division between semantics on the one hand and acquisition processes on the other, has the desirable consequence that the theory of concepts is not committed to any particular view on ontology or epistemology. I regard this as methodologically sound, since it does not force a proponent of informational semantics into defending the metaphysical realism of a given entity a priori. The theorist who relies on Fodor's thin notion of semantics can safely leave questions about

[41] Here, Fodor appeals to pragmatics even though it falls under his "first law of the nonexistence of cognitive science" (1983: 107ff), since he regards it as too unconstrained to be scientifically investigated. See Carston (2002: introduction) and Sperber and Wilson (1995: 65-67; 2002) for critical discussion of Fodor's law.

what is out there in the world and what counts as a reliable mechanism of semantic access to actual empirical investigation.

That said, I agree with the claim of Horsey (2006: 31) that "Fodor's account would be more convincing if he had a detailed and plausible story to tell about the psychology underlying concept acquisition: what the mechanisms are and how exactly they work". And as Horsey points out, a number of theorists (Cowie 1998; Landau 2000; Laurence and Margolis 1999; Levine and Bickhard 1999) do see this lack of a positive story about the mechanisms by which concepts *actually* lock to properties as a significant weakness in Fodor's approach to concepts.

Horsey (2006: chapter 5) tries to remedy this by giving a detailed account of how locking can be seen as sustained by *perceptual mechanisms* on an informational semantic approach, thus contributing a valuable insight into how concepts are formed and acquired. However, the acquisition of concepts as the result of direct perceptual contact between a mind and something in the world can only be part of the story. According to the argument from ontology, many of the things people talk and think about are not perceivable or tangible, so the acquisition of concepts for these things cannot plausibly be explained by an appeal to the workings of the perceptual system.

In fact, Fodor argues that even though "perceptual access heads the list of the ones that mediate our semantic access" (1998a: 77) to such things as *doghood*, there are many ways other than perception to sustain the meaning-making connection between a mental representation and a property. Indeed, he maintains that the list is open-ended and argues that theoretical inference, technological extensions (radar, heat-detectors), other proxies (sheep bells etc.) or mere talk or *deference* to other individuals (Fodor 1998a: 77-78) can serve to mediate between a concept and a property.

In chapter 4, I will return to the issue of concept acquisition and try to supplement Horsey's positive story with an account of how concepts representing properties without perceivable instances are acquired. In doing so, I will take as my starting point Fodor's suggestions about content mediation via deference and theory construction, and look more closely at how these ideas can be developed into a fuller theory of concept acquisition.

2.4. The translation problem and the publicity of meaning

2.4.1. Polysemy, and the relationship between semantics and pragmatics

Though many theorists may see Fodor's thin conception of semantics as counting against using it to answer the "big questions" in linguistics and philosophy of language, I think informational semantics makes for a very good theoretical starting point for anybody

interested in the content of thought, word meaning or lexical acquisition – and in how all of these are related.

I started this chapter by asking whether concepts could plausibly be seen as word meanings, as is claimed by Relevance Theory. I also wanted to know whether and how well the Fodorian informational semantic account squares with Relevance Theory's approach to lexical pragmatics. As it turns out, the merger of Relevance Theoretic pragmatics with Fodorian informational atomism is a happy coupling, in that Sperber and Wilson (1995)'s pragmatic machinery needs a lot of space to navigate – incidentally the space that is left wide open by Fodor's account of concepts.

To illustrate my point, I will take the case of *polysemy*, where one lexical item can mean or express several related things ('newspaper' as in a type of reading material, the management of a publishing company, the management's editorial policy etc., or 'run' as in 'run a marathon', 'run on gasoline', 'run a shop'; see Murphy 2002; Falkum 2010). While Fodor (1998a: chapter 3) has denied that polysemy is a genuine linguistic phenomenon, and Fodor and Lepore claim that alleged cases of polysemy do not reveal any "deep facts about lexical semantics", but rather arise "from *how things are in the world* (or anyhow, how we take them to be)" (2002: 117)[42], a pragmatic theory can explain how different aspects of a word's meaning get picked out depending on the circumstances of utterance (Wilson and Carston 2007)[43].

The relevance-theoretic idea pursued by Wilson and Carston is, in short, that some words, like 'tired', 'open', 'run', 'cut' and so on, which all have a very general application or a polysemous quality, encode concepts that denote a wide range of events and states of affairs: for instance, OPEN points to a property instantiated by all and only *opening* actions. Wilson and Carston (2007: 233) hold that "There is no standard or stereotypical method for *cutting*, *opening* or *leaving* tout court, but there are standard methods for *cutting hair*, *cutting a lawn*, *opening curtains*, and so on, each of which involves a narrowing of the more general concepts CUT, OPEN AND LEAVE". Wilson and Carston (2007) claim that the pragmatic process takes a lexical input such as 'open' or 'tired', which encodes the general concepts OPEN or

[42] This has led Pustejovsky (1998: 289), among others, to object that Fodor and Lepore's position on this issue makes them not "even recognize the relevance of polysemy as a key aspect of linguistic creativity and a window into the generative nature of thought".

[43] Whether the phenomenon of polysemy is a natural kind, and whether all cases allegedly falling in under it can be treated by the lexical machinery proposed by Wilson and Carston (2007) is not clear, though (Falkum 2010). I will return to this issue at various points throughout this thesis (see especially chapter 5, section 5.3.2).

TIRED, and narrows their denotation to yield more precise concepts constructed ad hoc for the purposes of the specific linguistic exchange.

Like all the pragmatic enrichment processes discussed in this chapter and the previous one, "narrowing is driven by the search for relevance, which involves the derivation of cognitive effects, and in particular of contextual implications" (Wilson and Sperber 2004: 617). A particular type of *opening* action may have relevant implications in a given discourse context while other types of activities will not. The interpretation may be narrowed in different ways depending on whether 'open the washing machine' is uttered in the context of washing clothes or a plumbing job. Assumptions about the activities leading up to someone's being 'tired', as well as information about the speaker, previous utterances and so on, may be added to the context, which helps determine how the general concept TIRED is narrowed down to a relevant ad hoc concept.

From the start, Sperber and Wilson have argued that the majority of what goes on in communication is a type of pragmatic inference, not the mere coding-decoding of linguistic signals as tradition has had it (see chapter 1, section 1.3.3). If words encode atomic concepts, and these in many cases express highly abstract properties such as *opening* or *tiredness* which would not make the utterance relevant in the expected way, then pragmatics *has to* remedy this by way of inference. I believe Relevance Theory gives a very plausible and coherent theoretical account of how this might happen.

But the question of what the contents of these general concepts are, and in what way these contents are plausible candidates for being the meaning of the word 'tired', is so far left unanswered. In what follows, then, I will look closer at what theoretical requirements concepts have to meet, in order to play the role they are assigned in the Computational Theory of Mind. The claim I will make is that the pragmatic theory has to reconcile what it takes to be the meaning of general lexical items like 'open', 'cut' and 'tired' with the requirements Fodor places on concept possession. The main obstacle will be the so-called *publicity constraint*, which demands that concepts are in principle shareable across time and space.

2.4.2. Conceptual constraints and cross-linguistic variation

In his (1998a) discussion of concepts, Fodor introduces a number of "non-negotiable" constraints that he believes have to be met in order for concepts to play a proper explanatory role in cognition. I have already mentioned the compositionality requirement, which holds that "Mental representations inherit their contents from the contents of their constituents"

(Fodor 1998a: 25). In Fodor's view, this is the only way to explain productivity and systematicity, which he takes to be defining features of human thought.

Another theoretical constraint he sets on mentally represented concepts is that they must "satisfy whatever ontological conditions have to be met by things that function as mental causes and effects" (1998a: 23), and that they should name categories, i.e. have things in the world that "fall under them" (1998a: 24, see footnote 40). He also maintains that quite a lot of them will have to be "learnable, and that "it would be nice if a theory of concepts were to provide a principled account of what's in the primitive conceptual basis", drawing a distinction between innate and learned concepts in "some independently plausible place" (1998a: 28).

Fodor's final constraint is the publicity (or shareability, or generality) constraint[44], which claims that "Concepts are public; they're the sorts of things that lots of people can, and do, *share*" (*ibid*). This publicity is assumed to hold across time and space for people with "the same types of minds": for instance, "Barring very pressing considerations to the contrary, it should turn out that people who live in very different cultures and/or very different times both have the concept FOOD (1998a: 29). He emphasises that "If a theory or an experimental procedure distinguishes between my concept DOG and Aristotle's (…) that is a very strong prima facie reason to doubt that the theory has got it right about concept individuation" (*ibid*).

Now, Fodor should not here be seen as arguing for a radical concept *universalism*. I assume it is uncontroversial that not everybody shares every concept with everybody else. His claim, rather, seems to be that the theory has no right to decide in advance what concepts people with different mother tongues and cultural backgrounds have. If someone makes a case for children having wildly different concepts from adults, "so that children can't think the same thought that we and grown up Hopi do" (2008: 55, n7), then so be it. But surely, Fodor comments, you will want "such claims to be *empirical*; you don't want to be forced to make them by a priori assumptions about the individuation of mental(/linguistic) representations" (*ibid*).

In Fodor's theory of concepts, the publicity requirement is supposed to make sure that such issues are left open, at least until someone comes up with solid empirical evidence to suggest that a given concept is not shared among specific thinkers. The claim I will make in this section is that the lexical pragmatic appeal to general concepts like TIRED, OPEN and

[44] Rey (2009a:187, n6) points out the unfortunate epistemic associations of using the term 'publicity', preferring instead to talk about 'sharable' concepts. I will use the terms interchangeably in what follows, but since I quote the relevant passages from Fodor (1998a) quite a lot, talk of publicity will still feature prominently.

HAPPY stands in danger of violating the publicity requirement, depending on some specifics as to how the theory should be understood.

The conflict between the appeal to general concepts and the publicity requirement becomes apparent when one starts to consider lexical items across languages[45]. As is well known, different languages lexicalize different meanings in different ways. Somewhat strikingly, many of the words which are seen in lexical pragmatics as encoding general concepts, such as 'open', 'cut' and 'break', are particularly susceptible to linguistic variation. Bowerman and Choi (2003), for instance, discuss six possible translations of 'open' from English to Korean. Majid et al (2008) find that although "there is considerable agreement across languages in the dimensions along which cutting and breaking events are distinguished (...) there is variation in the number of categories and the placement of their boundaries". Even within closely related languages there are significant variations, with Majid et al (2007a) reporting experimental tests which show that informants label events described in English as 'cut' and 'break' with one of three words[46] in German ('brechen', 'schneiden' or 'reissen'), four in Dutch ('scheuren', 'snijden', 'knippen' and 'hakken') and five in Swedish ('hugga', 'bryta', 'skära', 'klippa' and 'slita').

Why is this a problem for the theory? According to Relevance Theory, lexical items in a given natural language encode corresponding concepts. These may or may not be deployed without modification, since they are narrowed or broadened in many cognitive or communicative situations; but postulating that there are concepts corresponding to different words across languages entails a claim about the cognitive repertoire of a thinker. Depending on one's precise view of concepts, this may not accord with the publicity constraint. It will also, as Horsey (2006: 66-67) points out, force one to say something about meaning relations across languages.

To take just one example, consider 'tired', which is treated by relevance theorists as encoding a general concept TIRED, which denotes all and only instances of *tiredness*. In Norwegian, this can be translated into one of three words, 'trett', 'sliten' or 'lei'. A quick search in a corpus of literary translations[47] confirms this:

---

[45] A version of this objection has also been used against Fodor (1998a) by Hampton (2000). For discussion, see my chapter 3, section 3.2.2.

[46] The experimental paradigms makes use of 'clusters' rather than individual words, since some words in each language show a higher degree of specificity, such as the verbs 'slice', 'chop', 'snap', 'smash'. For a review of a batch of experimental literature on the typology of 'break' and 'cut', see Majid et al (2007b) and the special issue of *Cognitive Linguistics* (volume 18, issue 2) on the topic.

[47] Extracted from the Oslo Multilingual Corpus: http://www.hf.uio.no/ilos/OMC/. The translations are marked with a 'T' after the file name. All results were downloaded October 15, 2008.

52. Theresa's too **tired** to cook at conference times (DL2)
53. Theresa er alltid for **sliten** til å lage mat etter at hun har vært på møter (DL2T)
54. "You are **tired**, girl" she said (DL1)
55. "Du er **trett**, jenta mi," sa hun. (DL1T)
56. Then he announced he was **tired** of the road, and even of music (JSM1)
57. Men så sa han at han var **lei** av å farte omkring, til og med lei av musikken (JSM1T)

Taking at face value the relevance theoretic view that words encode concepts, this would presumably mean that while English speakers have one concept TIRED, Norwegians possess three different ones, SLITEN, TRETT and LEI. If so, the challenge is to say something about how these three concepts encoded by the Norwegian lexical items relate to the English item.

2.4.3. Publicity across languages

For a pragmatic theory which aims to say something about meaning relations across a given pair of languages, there seem to be three possibilities as to how this relationship can be cashed out. One could claim that 1) the combination of the three concepts encoded by the Norwegian lexical items, SLITEN, LEI and TRETT, express sub-parts of the same concept as the one encoded by 'tired'. When these smaller conceptual units are considered together, they have *the same* extension and express the same property as TIRED.

Alternatively, one could 2) hold that the three concepts encoded by the Norwegian lexical items express the properties *sliten*, *lei* and *trett,* respectively. The concept TIRED, on the other hand, expresses the property *tired*. These could then be seen as *different* properties, with different, non-overlapping instantiations. The concepts encoded by the Norwegian words would have a content distinct from the English 'tired' (and possibly also from each other). Another way of explicating the relations between the concepts encoded by the English and Norwegian words would be 3) to claim that even though the content of the concepts encoded are different in some respects, they are the same in others. That is, one might claim that SLITEN, LEI and TRETT have *similar* contents to TIRED (and perhaps to each other). In what follows, I will consider these three different solutions in turn.

Opting for the first explanatory route, where the conceptual items encoded by the Norwegian words are seen as expressing contents that can be subsumed under the more general concept TIRED, may initially seem to be an acceptable theoretical move. According to this way of explicating the relation, the only difference between the concepts encoded by the Norwegian lexical items and the one encoded by the English word is that the Norwegian conceptual counterparts individuate the world more finely.

A solution on this lines might be seen as explaining the translations between English and Norwegian quite well. A word such as ‘sliten’ will be reliably translated back to ‘tired’, and so will ‘trett’ and ‘lei’, since their extension is taken to fall squarely within the more general one expressed by TIRED. ‘Tired’, on the other hand, will be translated into any of the three Norwegian words depending on the discourse context. This solution is intuitively appealing in that it fits a common-sense understanding of lexical mapping across languages. When one consults dictionaries, one finds that a word which may have a single entry in the source language will be listed with two or more corresponding entries in the target language.

Formally, such meaning relations could be represented in the manner proposed by Hjelmslev, who divided lexical representation into “meaning zones” showing the distribution of different words across languages. Here is the meaning zone for a cluster of words for tree/wood in German, Danish and French (from Hjelmslev 1966: 50, quoted and revised in von Fintel and Matthewson 2008, who added further lexical counterparts from English):

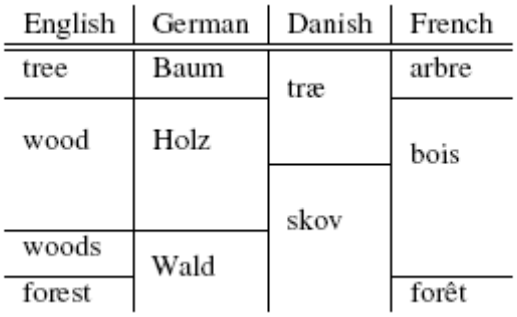

Figure 1 – Hjelmslev’s ”meaning zone” for tree/forest, with additions in English on left.

But despite the intuitive appeal of this solution, it seems that it does not really hold up when faced with the actual data from language use and lexical representation. It is well known that most words map imperfectly across languages (von Fintel and Matthewson 2008), and the neatly ordered ‘tree’ schema above, if one disregards the fact that it captures only a partial picture (‘holz’ can also mean ‘timber’ and so on, in addition one could wonder how well it extends to the rest of the world’s 6000 languages), is not really applicable to the way most other words work, This is revealed by a quick look at the behaviour of ‘tired’.

A search for the word ‘tired’ in the Oslo Multilingual Corpus yields 38 hits. A count reveals that 26 uses are translated as ‘trett’, 7 are given as ‘lei’ in Norwegian, while 5 uses are translated into ‘sliten’. If these words only expressed sub-parts of the general concept TIRED, one would expect uses of ‘sliten’, ‘lei’ or ‘trett’ to be translated back into English only as ‘tired’. But a reverse search, on Norwegian texts translated into English, reveals that this is far

from the case. Many uses of 'sliten' or 'trett' are not given in English as 'tired', but rather translated as 'weary', 'sleepy' or 'exhausted'. 'Lei' is also translated in some cases as 'fed up' or 'bored'. The following are some examples:

58. "Klokken ett," sa kirketjeneren med en tørr, **trett** stemme (LSC2)
59. "One o'clock," said the usher with a dry, **weary** voice (LSC2T)
60. Han sov mye, tok piller og ble **trett** (KA1)
61. He slept a lot, took pills and grew **sleepy** (KA1T)

Though this data does not unequivocally show that word meanings cannot be categorized and represented as standing in part-whole relations to each other, since one might appeal to lexical modulation to explain why translations are imperfect[48], I think the disorderly behaviour of the data raises an empirical problem for anyone who wants to opt for this solution to the translation problem. Bilingual speakers sometimes report the intuition that, when translating, a word in the target language expresses the meaning of a word in the source language, *but not quite accurately*. Drawing on introspective data, I think 'lei' is a good example, in that it is not really captured by 'tired', but rather conveys a different sensation, which is not really captured by 'bored' or 'fed up' either.

Faced with the different behaviours of these linguistic entries, the pragmatic theorist might opt for the second solution to the translation problem. She could claim that the contents of the concepts encoded by the Norwegian lexical items simply *differ* from the one encoded by 'tired'. But this, of course, leaves the theorist with no way of explaining how translations across languages are possible at all. How is it that 'tired' gets translated into 'sliten' and not some other Norwegian lexical item if there is nothing to relate the two? How can we know how to get from *any* lexical item in English to one in another language if there is no correspondence in content?

With the first two solutions unable to account for the data, the third option suggested above, the claim that lexical items across languages are merely *similar in content*, might seem tempting. One might argue that TIRED has relevantly similar content to the three concepts encoded by 'sliten', 'trett' and 'lei' and/or that the similarity is a matter of overlapping extensions. Visually, this "meaning overlap" could be presented in a diagram such as the following:

---

[48] A relevance theorist may claim, for instance, that the concept expressed by the Norwegian word 'trett' has as its content a sub-part of what is expressed by TIRED independently of any context, but that TRETT is broadened in use in examples 58 and 60, and therefore has to be given a different English translation if the content is to be captured. However, this raises some methodological problems which l will return to discuss briefly in the conclusion to this chapter.

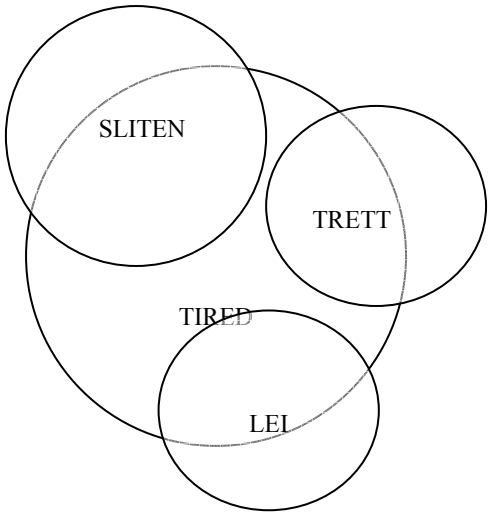

Figure 2 – Concepts expressed by 'tired' and the Norwegian translations.

Here, TIRED is presented as *partially overlapping* with the Norwegian concepts SLITEN, TRETT and LEI. Following the conclusions drawn from the corpus data, we can see how the concepts seen as encoded by 'sliten', 'trett' and 'lei' also "cover ground" which falls outside the meaning of TIRED. The following diagram is an expansion of Figure 2 with the English translations of the Norwegian lexical items added:

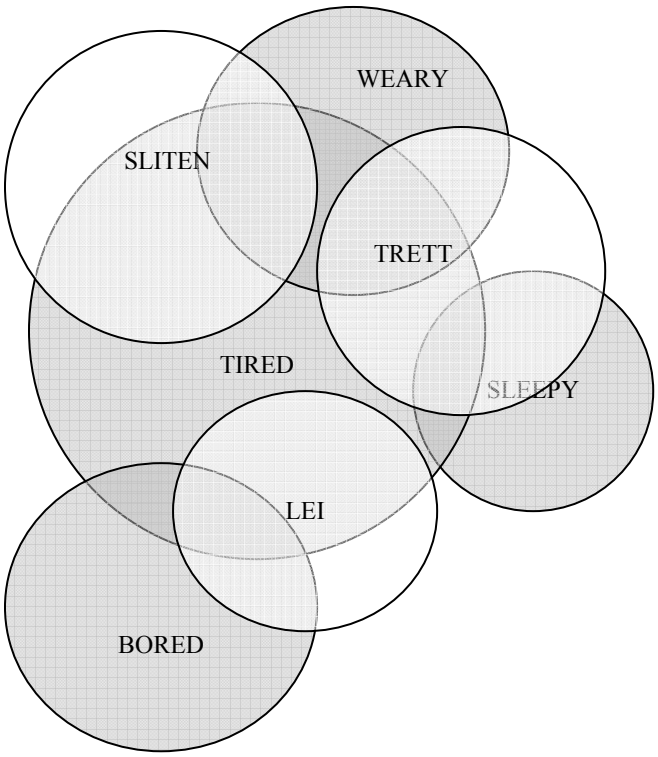

Figure 3 – Concepts expressed by 'tired', the Norwegian translations and the "semantically related" words in English. Grey colouring indicates concepts expressed by English.

Another quick look at the corpus data may help to see how this picture may be a fruitful one for cross-linguistic analysis. The sentences

62. Theresa's too **tired** to cook at conference times. (DL2)
63. "You are **tired**, girl," she said. (DL1)

are translated as 64 and 65 respectively:

64. Theresa er alltid for **sliten** til å lage mat etter at hun har vært på møter. (DL2T)
65. "Du er **trett**, jenta mi," sa hun. (DL1T)

Here, the two uses of the word 'tired' can be analysed in Relevance Theory as conveying narrower concepts whose denotations overlap with the denotations of SLITEN in 62 and TRETT in 63.

If we go to the Norwegian data, the following sentences may illustrate how uses of 'trett' sometimes convey concepts that fall outside the denotation of TIRED:

66. "Klokken ett," sa kirketjeneren med en tørr, **trett** stemme. (LSC2)
67. Han sov mye, tok piller og ble **trett**. (KA1)
68. "One o'clock," said the usher with a dry, **weary** voice. (LSC2T)
69. He slept a lot, took pills and grew **sleepy**. (KA1T)

If one takes these uses of 'trett' to involve narrowings of the Norwegian concept TRETT, the reason why they are not translated back as 'tired' might be that they convey ad hoc concepts that fall outside the denotation of TIRED. Figure 4 illustrates a possible analysis of this process:

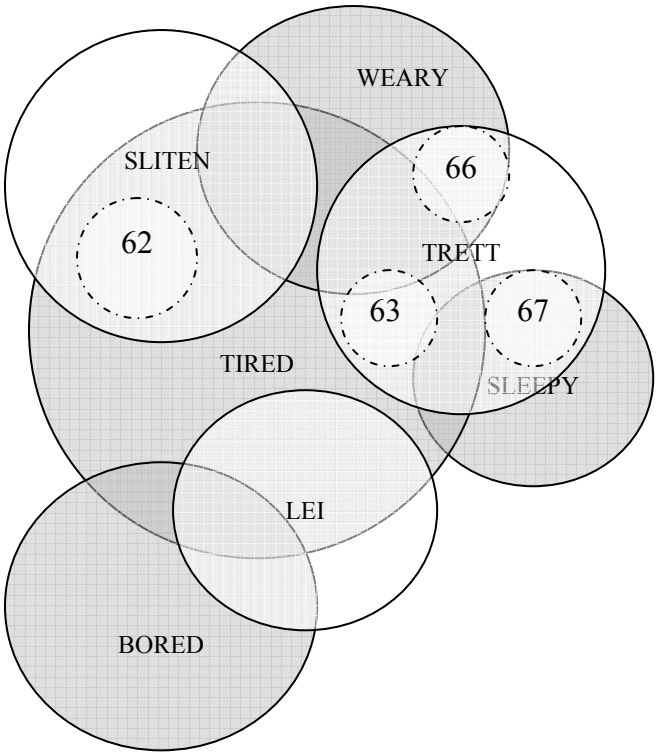

Figure 4 – Ad hoc concepts expressed by TIRED (62 and 63) and TRETT (66 and 67).

What is nice about this appeal to concept similarity is that a schema like the one above could be expanded to include whole vocabularies (for words at the same level of categorization, at least), and thus neatly and concretely explain lexical meaning relations. Using large amounts of translation data from corpora and the like, one can get accurate description of meanings and the mapping of lexical items across and within languages. This could, in principle, be used to model actual human conceptual and linguistic representation, contributing to a better understanding of word meaning, as suggested by, among others Dyvik (2005)[49].

Unfortunately, though, the similarity view of meaning relations comes with its own problems, some of which are familiar from the discussions in the previous chapter. There, I considered the *problem of similarity* for a theory of content, which holds that all similarity relations between two items (be they words, thoughts, implications etc.) have to be explicated by an appeal to identity at some level. Applied to the case in question, the problem of similarity becomes one of specifying how the content of concepts expressed by the Norwegian and English translations are the same and how they differ. Or in other words: how

[49] Thanks to Janne Bondi Johannessen for bringing this point to my attention.

can one cash out the metaphors I used above, "meaning overlap" "concepts covering ground", without explaining exactly what is literally shared between the concepts?

This problem is a serious one, I think, since some immediately accessible responses turn out on closer inspection to be theoretically unviable. One might claim that the contents of words like 'tired' and 'sliten' are similar in virtue of sharing some but not all of their structure. In many semantic theories, such appeals to structure assume that concepts have sets of features (or microfeatures, or properties, or attributes, see e.g. Moss et al. 2009), which together make up its content. But Relevance Theory, being committed to conceptual atomism, which claims that conceptual content is exhausted by the mind-world link[50], has no such bits of internal structure to appeal to in explicating similarity.

Equally unavailable is a solution based on appeals to a concept's encyclopaedic information in explicating similarities in conceptual content[51]. Even though the encyclopaedic entries for concepts such as TIRED and SLITEN should overlap to some extent (while differing in other respects), this information is regarded as merely collateral by the conceptual atomist. Allowing this encylopaedic entry to play a role in content individuation effectively leads to meaning holism (see this chapter, section 2.3.2), which is regarded by Fodor (2004; Fodor and Lepore 1992) as fundamentally in conflict with the publicity requirement. If the move to make encylopaedic information part of content was supposed to address the translation problem, and with it the publicity requirement, this solution defeats the purpose with which it was introduced[52].

2.4.4. An appeal to metaphysics

It should be clear that, despite the attractions of the similarity view of conceptual content, the conceptual atomist has somewhat limited resources at her disposal in accounting for the relations between the concepts encoded by words within and across languages. This results

---

[50] And possibly the logical entry (see footnote 34). But the logical entry does not help in determining the differences between the contents of concepts like TIRED, WEARY, BORED and SLITEN, since these presumably contain the same meaning postulate, possibly glossed as something like PHYSICAL SENSATION OF A CERTAIN KIND.

[51] In this vein, though in a slightly different context, Groefsema (2007) suggests that Relevance Theory should make the notion of conceptual content dependent on both encylopaedic and logical information. The inadequacy of this proposal is pointed out by Reboul (2008).

[52] With an eye to the spatial diagrams above, a theorist may be tempted to argue that content similarity is analysable in terms of the concepts simply being "geographically" adjacent, or similar in that they occupy overlapping regions in "conceptual space". This way of thinking about the matter forms part of, among others, Churchland's (1986; 1998) "state-space semantics", but is problematic as a solution to the similarity problem since it presupposes "what the location of a concept along that dimension is to *mean*" (Fodor 1998a: 34; see also Fodor and Lepore 2002: chapter 8; Calvo Garzon 2000)

from her insistence that concepts have no internal structure and do not decompose into smaller units, relying instead solely (or mainly) on a mind-world link to individuate content.

But such externalist conception of semantics also yield a potential solution to the problem of translation, namely an appeal to metaphysics. If the similarity relation between the concepts encoded by the English word 'tired' and the Norwegian word 'sliten' cannot be explained in terms of an overlap between the concepts themselves, overlapping *properties* might do the job. The idea would be that a pair of concepts judged as similar, such as TIRED and SLITEN, or TIRED and WEARY, are similar in virtue of having overlapping extensions, being locked to *similar properties*.

With this solution, the problem of similarity is kicked out of the cognitive domain into the metaphysical one, where the job of accounting for relations of sameness, difference and (thereby) similarity is performed by properties and their overlapping *instances*. If property P1 has *n* (actual or possible) instances which also instantiate property P2, and *m* (actual or possible) instances which do not instantiate P2, the two properties are similar. If all objects, events and states of affairs which instantiate P1 in the actual or all possible worlds also instantiate P2, the two properties are identical. If no instances of P1 are instances of P2 in any possible world, they are different.

Though Sperber and Wilson do not explicitly discuss their metaphysical position, or present a particular view of properties, this might be an acceptable solution for the relevance theorist. And should Sperber and Wilson wish to endorse the appeal to properties in explaining relations between concepts, I have no detailed counter-argument to offer (though see Goodman 1966 for some classical arguments against resemblance nominalism; Paseau forthcoming gives a recent review). I will just briefly, before concluding this chapter, suggest some reservations that a theorist wishing to square Fodor's informational semantics with relevance theoretic pragmatics might have about letting overlapping properties individuate similarity relations.

In going through the arguments against informational semantics in section 2.3.4 above, I suggested that a methodological advantage of Fodor's account was his advocating an ontologically neutral semantics. In particular, I claimed that Fodor's view of properties as *mind-dependent*, from which it follows that whether something is taken to instantiate a given property depends on how it strikes minds like ours, ensured that a proponent of informational semantics need not be committed to any particular story about the metaphysics of conceptual contents. On this approach, what is out there in the world is not an issue for the semanticist to decide, and questions of existence are better left to empirical investigations.

By depending on instances to explicate similarities in content, there is a risk that this desirable ontological neutrality will be breached, since it potentially forces the theorist to postulate instances of all sorts of properties without ever leaving the armchair. This is not, of course, an argument to the effect that a commitment to instances will prove to be too metaphysically strenuous for the relevance theorist. It may be that relying on instances of properties to explicate similarity is ultimately harmless, and that the theorist may get by appealing to merely *possible*, rather than *actual* events, objects and states-of-affairs as instantiating the properties one wishes to compare. What I am claiming, rather, is that by appealing to similarities at the level of properties, the relevance theorist ends up with a heavier ontological commitment than someone who maintains that conceptual content is public in principle.

It is also worth noting that there are some problem cases that the theorist who relies on similarity of properties will have to deal with. There are, for instance, many concepts which, if they correspond to properties at all, have no realized or even realizable instances in the world. Many concepts represent so-called metaphysically or nomologically impossible properties, such as *ghosthood*, *spectrehood*, *being clairvoyant* or *ogrehood*, which (depending on how the metaphysics turn out in the end) are taken to be *unrealisable* in the actual or any possible world. But what is there then to explain the similarity relations between concepts like GHOST, SPECTRE and OGRE? Neither of the corresponding properties can have any instances, which entails that they have all their instances in common. According to the explanation above, this should make them the same property. But surely, the properties *being a ghost* and *being an ogre* are not the same?

In the case of impossible entities, it seems that the appeal to metaphysics will not help in explaining meaning relations. True, these cases are problematic for all externalist theories of content, and it should not be taken for granted that a proponent of informational semantics who wants to preserve strict ontological neutrality is any better off than anyone else here. I will return to the problem of metaphysically impossible entities and the like in chapter 5, where I will advocate a view of the acquisition and representation of concepts for these entities that aims to preserve Fodor's ontologically neutral position.

Even though appealing to metaphysics to explain at least some similarities in conceptual content does indeed look like a viable solution to the problem of translation, the ontological commitment that comes with it may be a reason to prefer an account which does not need to rely on overlapping instances to explain content. If there are ways to explain overlap in meaning between words within and across languages which can preserve the

publicity of meaning without entailing any assumptions about what is out there in the world, such an account would be theoretically leaner, given its weaker ontological commitments.

Furthermore, a case could be made on methodological grounds against any attempt to stipulate similarity across groups of speakers a priori. If the claim is that differences in lexical repertoires across languages and speakers lead to different conceptualisation, this could be seen as running counter to one of the motivations behind the publicity constraint. Fodor claims that "Barring very pressing considerations to the contrary, it should turn out that people who live in very different cultures and/or very different times both have the concept FOOD" (1998a: 29), and he wants all claims about the concepts possessed by a given speaker to be empirical.

Though the matter of linguistic relativity is an empirical one for anybody interested in how language and thought map onto each other, the importance of the issue is particularly clear for someone working with the Language of Thought hypothesis. As I pointed out earlier in this chapter, section 2.4.2, from the assumption that LoT is an amodal system ontologically prior to natural language, it follows that this system should also be treated as *theoretically* prior to natural language. According to the LoT proponent, concepts come before language for both the child who wants to build a lexicon and the theorist who wants to say something about how the child (and adult) acquires and uses word meanings.

Fodor claims that

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

Concepts, as constituents of LoT sentences, should therefore in principle be independent of what natural language a given thinker speaks (Fodor 2008: 218)[53]. Having a theory which leads to claims about the conceptual repertoire of whole groups of thinkers a priori, solely by looking at which language(s) they speak, therefore faces the charge of reversing the order of explanation. Unless one makes empirical predictions about particular classes of conceptual items, postulating differences in mental inventories across whole groups of speakers seems to go against the theoretical priority that concepts are supposed to benefit from[54].

[53] Again, this is not a meant to be a case for concept universalism, and the publicity constraint does not rule out linguistic effects on the Language of Thought *tout court*.
[54] However small the practical import of such a postulation would be, cf. the discussion below.

I propose, then, that if a solution to the translation problem can be found in an account which explicates the relations between lexical items without appealing to similarity at the level of conceptual content, this would be methodologically preferable. In the next chapter I will therefore suggest a modified account which attempts to explain meaning relations between words within and across languages by appealing not to similar concepts, *but tooverlapping lexical entries*.

The idea is that, in order to preserve the methodological virtues of the Computational Theory of Mind alongside the explanatory power of Relevance Theory, the *encoding relation* between words and concepts will have to be rethought. I will outline what such an approach entails in terms of the lexical pragmatics of Relevance Theory and address what I see as the main objections to my alternative to encoding. The conclusion will be that while some things are lost, others are gained, and with the new picture of the mapping between language and thought in place, the merger of Computational Theory of Mind and Relevance Theory is a happy one.

2.5. Conclusion

In this chapter I have discussed the notion of word meaning from the perspective of lexical pragmatics. Following Relevance Theory, I have taken the meanings of lexical items to be mentally represented concepts, and considered what role concepts are supposed to play in cognition and communication. I introduced the account of concepts developed in the Computational Theory of Mind/informational semantics frameworks, and outlined the explanatory work a theory of thought and content is supposed to do according to the recent writings of Jerry Fodor.

I discussed several objections to Fodor's approach to concepts and informational semantics, and concluded that because of Fodor's "thin" construal of semantics, the objections were in principle answerable. Despite this, I argued that there is a problem when it comes to combining Fodor's views with some of the tenets of relevance theoretic lexical pragmatics – in particular the idea that some words such as 'tired' and 'break' encode general concepts. I suggested that considerations from cross-linguistic lexical representation forced the pragmatic theorist to treat the concepts encoded by words such as 'tired' and 'break' as merely similar across groups of speakers.

I looked at some ways in which this similarity relation could be explicated, and decided that an appeal to metaphysics was the best candidate for the job. I showed, however, that his leads to some ontological commitments which the theorist might be better off without.

I also suggested that a similarity view of conceptual content has some consequences for the issue of linguistic relativity, potentially violating the insistence in the publicity constraint that all claims about divergence in conceptual repertoire arising from differences in people's lexicons should be strictly empirical.

Towards the end of the chapter, I argued that the "translation problems" raised by the 'tired' type of cases therefore provides a reason to look for alternative ways of construing the word-concept relation. But conceivably, there are a number of strategies someone who wants to avoid such an outcome could pursue. I have discussed a couple of theoretical moves one could make to get away with a similarity story about conceptual content, but one could also opt to downplay the whole translation problem and its implications for Relevance Theory. One option would be to deny that the data I have discussed are very representative, and argue that they only hold for a small class of lexical items, like 'tired' and 'happy'. The rest, one could claim, are pretty isomorphic and map onto cross-linguistic synonyms a lot better, thus presenting no problem for the theory.

The trouble with this claim is that, even if the translation problem may stand out a lot more clearly with the general lexical items discussed in Relevance Theory, the data shows that the lexical discrepancies across languages are huge. As von Fintel and Matthewson (2008) show, there are probably no meanings that are lexicalized across all languages in the same way. Certainly, some semantic universals exist, but even the best candidates (WATER, COLD, WHITE) do not map perfectly across languages, since they correspond to one lexical item in one language and two or more in others.

Alternatively, one could try to argue that even though there may be a genuine translation problem, and even though it may be impossible in principle to construct an account of content similarity across speakers, in practice this does not matter. Sperber and Wilson claim (see e.g. their 2008: 97ff) that speakers of a given language may well not assign the same encoded concept to a given word, and that with millions more concepts than words, the fact that some words (in the same or different languages) encode different concepts makes no difference to the possibilities of shared cognition or communication.

According to Relevance Theory, people in linguistic interaction will be able to converge on the same interpretation through via lexical narrowing or broadening even if they come from different conceptual starting points. What is important for pragmatics to explain is

not what concepts individuals or groups of people have independently of any discourse context, but rather what happens when they are engaged in linguistic interaction[55].

I think this polemical strategy in one sense is right, but I maintain that no matter what one takes the job description of pragmatics to be, problems arising from underlying assumptions in the domain of representation might have repercussions for other aspects of the theory. Even though Relevance Theory has made increasing use of the notion of ad hoc concepts in the writings of recent years, there have to be, somewhere, stable concepts that can serve as the starting point for narrowing and broadening. Concepts cannot be ever-changing and in constant flux, since a communicator has to have a repertoire of contentful mental items prior to engaging in linguistic interaction (as is indeed pointed out by Sperber and Wilson 1998: 184).

Also, the theory appeals to the stability of content to support the claim that speaker and hearer do not literally have to share thoughts in order for communication to succeed. As highlighted in the previous chapter, the whole relevance theoretic idea that thoughts are similar in virtue of sharing contextual implications presupposes that the implications are in fact literally shared. I argued that this made the similarity problem impossible to escape from, since any construal of a "good enough communicative understanding" presupposes literal sharing of content somewhere.

Whether the type of content similarity I have examined and discussed in this chapter is indeed harmless will depend on whether it carries over to contextual implications, then. Not having a good story to tell about the metaphysics of implications and their content, I will leave this question unanswered, but will simply note that the mere risk of a breakdown in the neat explanation of how people do not need to end up with identical thoughts in order for communication to be successful is a good reason for abiding by the publicity constraint. If she wishes to preserve the explanation of when communication is successful and when it breaks down, the pragmatic theorist who downplays the importance of this constraint therefore has the burden of proof on her to show how the explanation of sharing contextual implications is not affected by the idea of content being merely similar outside of a context.

Because of the problems with appeals to similarity of content, Fodor regards the publicity constraint on concept possession as "non-negotiable" (1998a: 34). Even though one might take the above responses to publicity worries to show that this might be too strong, the methodological, metaphysical and theoretical advantages that come with respecting this

[55] Thanks to Deirdre Wilson and Mark Jary for pressing me on this point.

constraint may weigh in favour of an account which need not rely on content similarity in explicating meaning relations. The efforts of the next chapter will therefore be devoted to developing an account which does just that.

3. Concept activation and the mapping between language and thought

3.1. Introduction

In this chapter, I outline a view of the relationship between words and concepts designed to address the worries about publicity raised in the previous chapter. According to the view I will present, there is no encoding relation between words and concepts. Rather, lexical items potentially activate a number of corresponding mental items, without any one concept occupying a privileged semantic role. Though I will claim that this view is compatible with the central tenets of Relevance Theory, it departs from the standard RT account in a number of ways, raising some questions about the nature of lexical pragmatic processes as discussed by e.g. Wilson and Carston (2007).

My aim in what follows is to discuss some positive and negative consequences of abandoning the idea of an automatic, direct or otherwise isomorphic connection between words and concepts. The argument will be that although doing without encoding is problematic in that one loses an explanation of the normativity of meaning, it does not accord with "folk linguistic" intuitions, and it leads to a proliferation of concepts and properties, these weaknesses can be remedied by altering the theory's account of lexical entries. On the positive side, I will argue that the account I propose more easily respects the publicity constraint and theoretical priority of concepts, making the study of mental vocabulary independent of lexical semantics and typology.

In the first section, 3.2.1, I consider whether the problems raised in the previous chapter could be solved by abandoning the idea that *concepts* are word meanings. I discuss what the alternatives to concepts might be, and suggest that whatever alternative notion is chosen, it will have to help account for the productivity, systematicity, stability and shareability of thought. Having dismissed the prospects of finding an alternative that is equally well suited to the cognitive roles that concepts play, I go on in section 3.2.2 to consider the *relationship* between words and concepts in Relevance Theory and Computational Theory of Mind. Although Fodor relies on word-concept isomorphism in his writings on meaning, I show that nothing in his account hangs on how the relationship between lexical and conceptual items is specified. I also foreshadow some problems that Relevance Theory, and especially its approach to lexical pragmatics, might face in abandoning the view that words encode concepts, which I take up again in section 3.3.

In section 3.3.1, I outline my conception of words as *potentially activating* concepts, and show how words may give access to cognitive domains in which a number of atomic concepts are stored or created ad hoc. Which of these concepts will be activated or formed by the tokening of a corresponding word is a matter for pragmatics. I argue, in section 3.3.2, that the pragmatic processes involved are guided by retrieval constraints which form part of the lexical entries of words, and which specify in what cognitive domains the pragmatic mechanism is allowed to search for or construct the appropriate concept for that particular occasion. I outline some consequences of abandoning an encoding account, one of which is that words have no "literal meaning". Though I concede that this clashes with linguistic intuitions, I show that this is not such a serious matter if the framework of Relevance Theory is presupposed, since "literal meaning" does not play the same important role here as in many contemporary linguistic and philosophical approaches.

In section 3.3.3 I outline some consequences of my account for the lexical pragmatic analysis of some cases of linguistic underdeterminacy, and suggest that even though some aspects of the Relevance Theory view of *broadening* are incompatible with my account, most of the analytic power is preserved even when encoding is lost. In 3.3.4 and 3.3.5 I consider some possible objections to my account of potential activation, among them the accusation that it may lead to massive ambiguity and a reliance on type physicalism. I respond by showing that while it is true that concepts and properties are proliferated, the way lexical items are seen as univocal and the underlying concepts formally distinct makes the account immune to ambiguity objections. It is also unproblematic, I argue, in that concepts and properties are cognitively and ontologically "cheap". I dismiss the type physicalism accusation on the grounds that it is fully possible to cash out the vocabulary I use in functional terms.

In 3.4 I take a step back and consider some methodological advantages of treating words and concepts as theoretically separable (section 3.4.1), as well as comparing my approach to other theories which do not rely on isomorphism between words and concepts (section 3.4.2). My central claim is that even though there might be explanatory and theoretical advantages to abandoning a direct link between words and meanings, the heavy load that word-concept isomorphism carries in many accounts of word meaning will have to be relieved at one point or another in whatever is proposed to replace it.

3.2. A choice between two paths

3.2.1. Giving up on concepts

In the previous chapter, I examined and questioned the view that atomic concepts are the meanings of words. I concluded that the informational semantic account of conceptual content advocated by Fodor (1998a; 2008) provides a solid foundation for the study of meaning, with concepts playing the explanatory roles expected of them in a cognitive system that displays features of productivity, systematicity and shareability. I argued that even though the limited scope of the theory means that it will have to be supplemented by work from other domains in order to explain some of things that cognitive science is interested in, informational semantics provides a plausible explanation of how thoughts have their content.

However, there were some problems with the attempt to merge the lexical pragmatic account of word meaning with Fodor's conceptual semantics, most notably, how to explain meaning relations across languages. The problem arose from the conceptual atomist's limited resources for explaining similarity of word meanings across languages in a way that satisfies Fodor's publicity constraint on concept possession. As I see it, a theorist who agrees with my claim that the proposed solutions to this problem are (methodologically) unsatisfactory has two possible options to pursue. She could 1) argue that this shows the inadequacy of the view that atomic concepts are the meanings of (at least some) content words or 2) retain this view of concepts, but claim that there is something wrong with the way the *relationship* between words and concepts is specified.

One version of the first option has been suggested by Carston (2002: chapter 5). She asks whether there may be some words that do not encode *concepts*, but rather function as pointers to regions in memory where one finds "certain bundles of information from which the relevance-constrained processes of pragmatic inference extract or construct the conceptual unit which features in the speaker's thought (Carston 2002: 361). Instead of concepts, she suggests that there "might be a sizeable class of words" which express concept schemas, or pointers, or addresses in memory.

Carston stops well short of arguing that we should reject the notion of concepts as a basis for word meaning *tout court*. Instead, her proposal amounts to a suggestion that "there are different kinds of lexical meanings, with some words encoding full-fledged concepts, others encoding a schema or a pro-concept (…) and others a procedure or inferential constraint" (Carston 2002: 363). But the kinds of cases discussed in the previous chapter, 'open', 'happy' and 'tired', are plausible candidates for having their content not in virtue of

encoding a concept, but rather mapping "to an address (or node, or gateway, or whatever) in memory", according to Carston (2002: 360).

This alternative view of word meaning is not pursued further in Carston's other writings, but her conceptions of words as 'pointers' is intriguing nonetheless. And it seems clear that a type of representational pluralism which treats different types of words as having qualitatively different types of meanings could potentially be a fruitful research program to pursue[56]. It is important to remember, though, that if a theorist wants memory addresses, traces, nodes, bundles of information or other cognitive mechanisms to serve as word meanings, these will have to fill the place currently occupied by concepts. That is, if what underlies the meaning of 'tired' is not a concept but something else, this something else will have to be compositional, learnable and shareable in the same way that concepts are required to be, granted that one can have productive, systematic, stable and sharable thoughts that feature the meaning of the word 'tired' as a constituent.

But as Fodor has argued throughout his career, accounting for systematicity, productivity etc. is no easy task for a theory of cognitive content. In the previous chapter, I showed how Fodor believes that other alternatives to informational semantics fail to account for these features of intelligent thought. Whether or not he is right, and whether or not one sees informational semantics as explanatorily successful, it should at least be clear that someone who dispenses with conceptual atomism in constructing a new account of word meaning within Relevance Theory must account for systematicity and compositionality in some other way – or attempt to explain the need for these away.

To take just one example, Young (2006) has tentatively suggested that an alternative, "radical view" of cognition on which it "consists in clustering primitive conceptual features into conceptual 'bundles', thus creating new 'concepts' from moment to moment to serve a particular cognitive purpose" (2006: 276) might provide the basis for word meaning in Relevance Theory. But in doing so, he owes us an explanation of how these bundled features, which "need not be discrete or permanent entities, but rather an amorphous continuum of conceptual properties" (*ibid*), can productively and systematically combine, be learned and shared across thinkers. Eschewing the explanation of thought as involving computations over symbols that can combine in discrete infinity therefore takes the theorist back to the drawing

[56] Already, one distinction between different types of word meaning, namely the one between function and content words, is part of all linguistic theories. In Relevance Theory, this is captured in the distinction between procedural and conceptual meaning (see Wilson and Sperber 1993; Wilson 2009).

board, leaving her with no way of accounting for what seem to be the most essential features of thought.

Because I share Fodor's (1975; 2008) pessimism about the possibility of coming up with viable alternatives to the Language of Thought hypothesis, I prefer to hang on to the idea of concepts as the building blocks of cognition for as long as possible. Although representational pluralism is very much in fashion in cognitive science and philosophy (see Dove 2009; Weiskopf 2009a and 2009b for some interesting accounts) I would also be cautious about the kind of proposal Carston (2002: chapter 5) considers. Distinguishing a class of (content) words which does not encode concepts from a class that does, stands in danger of leading the theory down a slippery slope, unless some principled way of carving up the categories can be found.

Certainly, if one takes seriously the translation problem raised by lexical variability, there will be a lot of concepts other than the ones discussed by Carston which threaten the publicity constraint. With no immediately available alternative in the relevance theoretic literature, it is not clear that getting rid of concepts is a fruitful way of answering the questions raised in the previous chapter.

3.2.2. Questioning 'encoding'

A relevance theorist who wishes to stick with Fodor's account of concepts as the basis of word meaning but is troubled by the objections raised in the previous chapter still has the option of questioning the *relationship* between words and concepts[57]. As previously shown, in Relevance Theory this relationship is seen as one of *encoding*, which is a notion based on and reinforcing a traditional view of the relationship between language and thought.

Though implicit in much work in linguistics, psychology and philosophy, the connection between the lexical and conceptual systems of the mind is assumed by most to be one of close parallelism (Rives 2009: 203; Vigliocco and Vinson 2007: 195). Most share the "Cartesian" perspective on which the function of language is to convey thoughts, and words must therefore "express" concepts in some way. Fodor too, on the rare occasions that he talks

[57] The relevance theorist is also of course free to pursue any of the many other theories of content apart from informational semantics out there (inferential role semantics, holism, functionalism, prototype theory, decompositionalism etc) but it is not clear 1) how these would solve the problems raised in the previous chapter or 2) how they meet the challenges raised against them by Fodor (1998a, 2004, 2008). Another option, not previously mentioned in this thesis, would be to abandon the whole idea of thought as the basis of meaning, as does Gauker (2003) and investigators in a new "Un-Cartesian linguistics" project: http://www.dur.ac.uk/philosophy/uncartesianlinguistics/. But I take this to be at odds with the whole Gricean program on which Relevance Theory is founded, and wholly irreconcilable with the view of natural language communication as linguistically underdetermined. See Fodor (1998b: chapter 6; 2001) for discussion.

about it explicitly, seems to see the relationship between words and concepts as fairly direct. In a review of Peacocke's (1992) book *A study of concepts* he says that they share the supposition that "Concepts are word meanings. The concept DOG is what the word "dog" and its synonyms and translations express. This ties theories of concepts to theories of language" (1998b: 28).

In the introduction to his (1998a) book, Fodor mentions that he will "move back and forth pretty freely between concepts and word meanings; however it may turn out in the long run, for purposes of the present investigation word meanings just are concepts" (1998a: 2). In other places (most explicitly in his 1975) Fodor seems to suggest that the vocabulary of natural languages and thought correlate fairly directly, so that most words map onto a corresponding concept.

Johnson (2004: 335) argues that Fodor accepts what he dubs the "Isomorphism Assumption", which states that "the structure of [most of] our words is mirrored in the concepts they express". From this it follows that concepts can be studied by examining words, with evidence about aspects of lexical items being brought to bear on the issue of conceptual content. If so, the relationship between words and concepts will have to be fairly direct on Fodor's account, but Johnson also acknowledges that Fodor is not directly *committed* to an isomorphism between the conceptual and the lexical.

In his (1998b) review article, Fodor qualifies his earlier claims, arguing that word meanings cannot *ipso facto* be concepts, since there are many words which do not express concepts (e.g. demonstratives) and quite plausibly; a range of concepts without corresponding words to express them. He admits that "Getting clear on the word-concept relation is no small matter" (Fodor 1998b: 34 n1), but emphasises that whatever one takes to constitute the relation can be seen as independent of a theory of concepts. As argued in the previous chapter, an underlying conjecture of Fodor's Language of Thought hypothesis is that "the semantics of thought is prior to the semantics of language. So, for example, what an English sentence means is determined, pretty much exhaustively, by the content of the thought it is used to express" (2008: 198).

In fact, the assumption is even stronger, since

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

The ontological and theoretical priority of thought content to language may explain why Fodor is relatively agnostic about objections to his account based on facts about natural languages. So when Hampton (2000: 301) raises a version of the translation problem discussed in the previous chapter, complaining that Fodor's referential semantics has problems accounting for the fact that there are indefinitely "many cases where a concept is lexicalized in one language but not in another", Fodor merely counters that "there are notoriously terrible problems about deciding when (if ever) translations preserve meaning" (2000b: 354). He claims that though it may turn out that lexical mappings across languages are imperfect, "The pertinent question is not whether some languages lexicalize things that others don't; it's whether there are cases where what is lexicalized (expressed by an unstructured lexical item) in language A is synonymous with what's expressed by *a phrase* in language B" (2000b: 353-354).

Accounting for lexical variation by postulating that complex descriptions are the contents of words across languages raises both practical problems (how to come up with adequate descriptions) and methodological problems (how to decide which language has the unstructured concept as the meaning of a word) and Fodor concedes that "Since it's going to be part of my story that most words are undefinable (…) I'm committed to claiming that this sort of case can't arise too often" (1998a: 42 n 2).

But like Hampton (2000; and von Fintel and Matthewson 2008), I take the cross-linguistic data to show that lexical variation is massive and the mapping between words unruly, and will therefore propose getting rid of the isomorphism between words and concepts assumed by Fodor. For Fodor, nothing much hangs on the Isomorphism Assumption, since he holds that the content of thought (and therefore of concepts) is ontologically and theoretically prior to natural language. It is the relatively weak notion of words *expressing* concepts that is needed by many of the theorists who endorse the IA, with stronger assumptions about isomorphism being "relatively unevidenced" as long as no one has provided any good reasons to think that it could not "be false in both directions", according to Johnson (2004: 354).

For the lexical pragmatics approach, things might be a little different though, since Relevance Theory relies more explicitly on some specifics of the proposed relationship between words and concepts. Sperber and Wilson (1995) treat words as semantically encoding concepts. The relation is semantic in that concepts serve as the (typically literal) meanings of the words which encode them. The encoding relation also provides an automatic starting point for the pragmatic processes which ultimately yield the main output of utterance interpretation.

So 'tired' encodes TIRED, which is locked to an abstract notion of *tiredness* which may be too general to be deemed relevant enough in many cases in which the word 'tired' is uttered. Nevertheless, this is the content of the word outside any context. As explained in the previous chapter, what happens in a specific communicative situation is that a hearer who comes across an utterance of 'John is tired after having run a marathon', will use the abstract TIRED concept as a point of departure for constructing the narrower TIRED* concept, denoting a more specific property of *tiredness,* for the purposes of the linguistic interaction.

While some words are taken by Relevance Theory to express very general concepts, RT lexical pragmatics assumes that other words may encode concepts that are too specific to be satisfied in most contexts and may need to be broadened to yield an overall interpretation that is relevant enough. So 'silent' is taken to encode SILENT, which is locked to a property *silent* instantiated where there is absolutely no sound. This is the literal linguistic meaning of 'silent', but since this property will almost never be instantiated (and therefore an overall interpretation based on the literal meaning of the word will be false in most contexts), a hearer will use it as a mere starting point for constructing a broader ad hoc concept SILENT*.

What happens if one attempts to make do without word-concept encoding in Relevance Theory, then? Clearly this depends on what it is replaced with, but at the very least, the alternative account stands in danger of weakening some of the explanatory potential of lexical pragmatics. Without an explanation of how one goes from a word to a given concept, the process of narrowing and broadening has no initial starting point, and one is thus left without the neat lexical pragmatic explanation of how one can go from a too general or too specific concept to a more relevant one constructed "on the fly", which need not necessarily be lexicalized in a given natural language.

Moreover, something needs to be said about the *normativity of meaning*, and the fact a word or expression has conditions for correct application[58] if these are not specified semantically. One of the nice things about postulating an isomorphic (or otherwise) direct link between a word and a concept is that it explains how lexical items can be used correctly or incorrectly. Even though the norms which determine the meanings of words outside of a communicative situation might be "bent" in communication due to considerations of

[58] Whether claiming that "for an expression to have a meaning is for it to possess conditions of correct application" (Whiting 2007: 134) constitutes normativity in a substantial sense is not a given (see Boghossian 2005; Buleandra 2008; Gluer and Wikforss 2009; Hattiangadi 2006; 2009; Speaks 2009; Whiting 2007; 2009 for a recent debate on the normativity of meaning), but I believe it is relatively uncontroversial that some kind of socio-culturally determined normative connection has to hold between the lexical and conceptual levels – even if these are seen as theoretically independent of each other.

relevance (or informativeness, or whatever one's favourite pragmatic theory appeals to), there has to be *some* normative connection between a word and a concept if one wants to explain why 'dog' can mean DOG but not, e.g., KITTEN, BASEMENT or CHEESEBURGER. Severing the connection between words and concepts therefore runs the risk of undermining the whole story about how words communicate thoughts and the normative predictability of how this happens.

After introducing what I see as a plausible alternative to the theory of encoding, then, I will aim to address the problem of normativity and argue for a way to preserve the explanatory potential of lexical pragmatics in the absence of a direct, semantic connection between words and individual concepts. I will also address other potential problems with this theoretical move, among them the risk of sacrificing theoretical economy and succumbing to type physicalism. Towards the end of the chapter, I will also highlight the benefits of this alternative view of the mapping between language and thought, and argue that these outweigh the theoretical costs of the suggested move.

3.3. The alternative to encoding

3.3.1. Concept activation

The idea I will pursue and defend in the remainder of this chapter is the following: Words do not semantically encode concepts, they merely *potentially activate* them. There is no isomorphic or otherwise direct link between a lexical and a mental item, and no way to tell outside of a context which particular concept a word can activate. Lexical items, therefore, have no meaning in and of themselves; they function solely as *mediators* to a separate level of content.

I suggest that the best way to think of the relationship between a person's lexical and conceptual inventories is by analogy to a multi-layered map. At the bottom, one finds a conceptual level, where a finite set of mental items are organized in a (functional) grid-system. Each concept has its content in virtue of being locked to a property external to the mind, while its Mode of Presentation (its form/syntax) is individuated by its (functional) address in a given domain. "Hovering above" the conceptual levels one finds layers of lexical information aligned to fit sets of corresponding concepts. On this picture, every speaker of a language has at least two levels of "semantics" (see section 3.4.1 of this chapter), where the bottom level gets gradually populated by symbols in the course of concept acquisition. As new concepts are acquired, the layer above this gets woven in parallel, aligning itself to the concepts at the level below.

Learning a natural language, then, is learning how to associate words with their corresponding meanings, as Fodor (1998a: 9) claims. A child growing up to speak a language is hypothesised to acquire new atomic concepts from infancy, which gradually fill up the bottom layer of the cognitive "map". From the age at which she starts to learn a public vocabulary she will (necessarily later than the age at which she starts to acquire concepts) form a second layer which is developed and shaped in accordance with the already formed concepts below, in a process of trial and error[59]. If the child is a bilingual learner, she will form two layers shaped individually according to the languages learned (if she learns Chinese and English, the top layers will not look very much alike, if she learns Flemish and Dutch, there will be greater correspondence between the two lexical levels).

I wish to follow Wilson and Carston (2007: 238) in their claim that "lexical adjustment may be a one-off process, used once and then forgotten, creating an ad hoc concept tied to a particular context that may never occur again" and endorse the line Sperber and Wilson (1998: 197) take when they hold that

> "inferred senses may be ephemeral notions or stable concepts; they may be shared by few or many speakers, or by whole communities; the inference pattern may be a first-time affair or a routine pattern -- and it may be a first-time affair for one interlocutor and a routine affair for another, who, despite these differences, manage to communicate successfully".

But according to the view I am suggesting, there are no general concepts TIRED, CUT, OPEN, HAPPY etc., like the ones discussed in the previous chapter. Instead, individual concepts locked to distinct properties of different types of *opening*, *cutting*, *being tired* and *being happy* are stored or formed ad hoc.

In a communicative situation, the fact that someone has uttered the lexical item 'tired' can potentially lead to the activation of any one of the underlying TIRED concepts. What determines which particular concept is activated in a given context is the result of a pragmatic process of searching for cognitive effects. Should the property of *tiredness* most relevant for the communicative exchange not already have a corresponding, pre-formed concept, the hearer will be able to create one drawing on the contextual resources available to him in the

---

[59] I am here simplifying to an embarrassing extent, ignoring a whole array of thorny issues raised and discussed by a vast literature on lexical acquisition. But a proper discussion of the exact nature of the word-concept relation is unfortunately outside the scope of my work here, since I am merely trying to argue for the initial plausibility of an alternative view of word meaning. In the extension of this, and of the work in the chapters that follow, there will be a range of interesting questions about the cognitive, semantic and pragmatic mechanisms by which children manage to map words onto the right concepts and the individuation of the mapping relation which I hope to return to in future work.

situation of utterance, in the same way envisaged by Relevance Theory and explained in chapter 1.

What I propose is very much in accordance with the circle schemas representing meaning relations I suggested could account for the translation/polysemy data in the previous chapter. Here is the proposed English-only diagram of the relations between 'tired', 'weary', 'sleepy' and 'bored' again, without the translations into Norwegian:

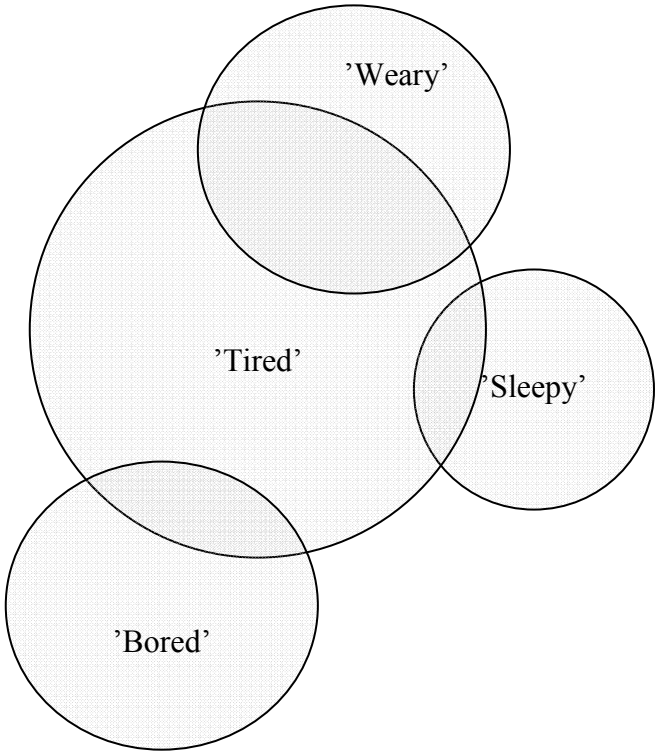

Figure 5 – Schematic representations of 'tired' and some related adjectives

However, there's a crucial difference between the idea proposed in chapter 2 and the one I am endorsing here, though. On the picture I suggest to replace the word-concept encoding relation, the large circles here are *not concepts*. They are merely lexical entries without content[60], hovering above the concepts themselves. The concepts are stored at a separate level below the lexical entries, thus making the amended, two-layered picture look more like this:

[60] The second layer is made up purely of vocabulary items which are therefore not plausibly seen as representational. This makes my account immune to some principled arguments against dual theories of content offered by, among others, Martinez-Manrique (2010).

:

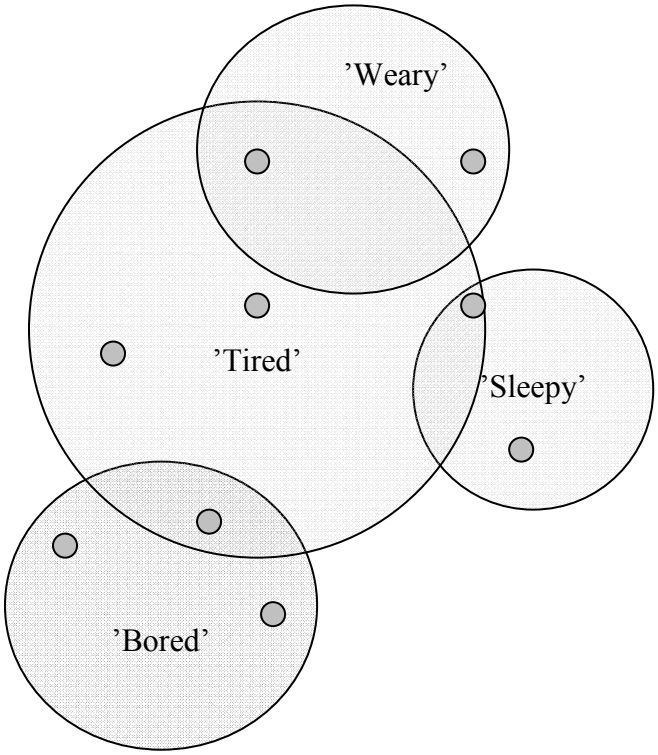

Figure 6 – Schematic representation of the lexical entries 'tired', 'sleepy', 'weary', 'bored', with correlating concepts stored "below".

Here, the atomic TIRED, SLEEPY etc. concepts are represented as minor grey circles, aligned with, but separate from, the lexical entries themselves. The polysemy of TIRED is explained by the fact that it potentially activates (say) five different concepts, all expressing distinct properties. The overlap in meaning between lexical items in and across languages is explained in terms of their potentially activating some, but not all, of the same concepts.

The move from talking about overlapping concepts to talking about overlapping lexical items provides the theory with the resources to find a way out of the similarity problem discussed in the previous chapter. While in that framework concepts could not be seen as similar to one another because there was nothing in the current relevance theoretic literature to individuate the relation between them[61], on the alternative account the similarity relation between word meanings can be explicated by appeal to their potentially activating identical/different *concepts*. So the lexical item 'tired' is similar to the lexical item 'weary' in terms of potentially activating one concept that is shared between the two, and a number of

[61] I did concede, though, that there were other candidates (such as properties) that conceivably could be up for the job.

other concepts that are not shared. Including the lexical levels of other languages makes the explanation generalize cross-linguistically, so one can imagine that 'tired' is similar to 'sliten' (Norw.) in terms of potentially activating four of the same concepts, while two concepts are better captured by 'trett' and one by 'lei'[62].

My prime motivation in moving away from the idea of *semantic encoding* and talking instead about *potential activiation* has been the prospect that this alternative conception will more easily satisfy Fodor's publicity constraint on concept possession. With this constraint, which I introduced in chapter 2, section 2.4.2, Fodor insists that an adequate semantic theory should assume that concepts are sharable across thinkers and times, barring empirical evidence to the contrary. Even though there are bound to be huge discrepancies in the conceptual repertoires of individuals, the semantic theory should not make a priori assumptions about which concepts are present, instead letting claims about conceptual differences between people be empirical.

And I think the alternative notion of word meaning I have presented so far does indeed preserve the neutrality Fodor wants. This can be seen in the way data from natural language cannot be taken to imply anything about what is contained in a given speaker's conceptual repertoire, once assumptions about word-concept isomorphism or other types of default activation between lexical and conceptual items are done away with. From typological claims such as that a given language A has three lexical items expressing a notion which a language B expresses with one word, there will be no implications about what concepts speakers of A and B possess.

In contrast with theories which rely on a direct correspondence between a word and a concept, the view of words as only potentially activating mental items makes no suppositions (however weak, see chapter 2 section 2.5) about speakers' mental representations a priori. This gives it a methodological advantage over other ways to conceptualise word meaning within the Fodorian programme, in that it is in keeping with the spirit of the Language of Thought as a system ontologically and theoretically prior to natural languages.

3.3.2. Giving up on "word meaning"

In spite of the advantages outlined above , it should be fairly clear that an account such as the one I have proposed faces some explanatory and theoretical challenges. I have already

[62] These numbers are all just meant by way of examples, of course. Deciding to what extent and in what way words are similar in and across languages is an empirical question, to be settled by the best available methods (if one ever becomes available).

mentioned a couple in the discussion above, in particular that of explaining the normativity of meaning. Without a direct link between a word and a concept, some other way of specifying why 'tired' can be used to convey a range TIRED concepts (as opposed to a range of WEARY or HAPPY concepts) is needed.

What I suggest instead of what Relevance Theory sees as a semantic link between words and concepts is the idea that each lexical address (lexical entry) in the map contains *retrieval constraints*, which limit the domains in which the cognitive system is allowed to search for a relevant concept. Just as on the original lexical pragmatic picture, the *search for relevance* determines which concept gets activated (or formed) in a specific linguistic exchange, and I am assuming that all words (except function words) provide access to something like geographical coordinates which specify in what areas of conceptual space it can look for a relevant concept. There is no default activation on this picture, and though there will be contextual factors (frequency, recency and so on) which affect the relative ease of concept retrieval (cognitive effort): no concept in a given domain occupies a privileged position relative to other concepts stored in that domain.

As in the original relevance theoretic proposal (Sperber and Wilson 1995: 86), my version of the lexical entry will have to contain phonological and orthographic information about the natural-language counterpart of concepts. With Sperber and Wilson, I assume that the lexical entry will also include "information about its syntactic category membership and co-occurrence possibilities, phonological structure, and so on" (1995: 90). In addition, I would be responsive (though not committed) to the proposal that other syntactico-semantic phenomena, such as argument structure (Hale and Keyser 1993; 2002), might form part of the separate lexical entry, as suggested by Johnson (2004: 354).

However, my version of the lexical address will need something further which is not required on Sperber and Wilson's account. To make up for the loss of the semantic link between word and concept, my lexical level needs to contain clear specifications of which conceptual regions can be activated. In order for the lexical item 'dog' to give access to distinct concepts locked to different properties of *doghood*, the procedures contained in the lexical entry have to specify that the pragmatic mechanism can look in *all and only* the precise region containing these concepts. An utterance of the word 'tired', then, will lead to the potential activation of any concept within an enclosed conceptual region, but will not be able to give access to concepts like WEARY or BORED. The only way, I suggest, that the tokening of a lexical item can give access to a concept outside the specified conceptual range is via metalinguistic correction or a metaphorical mapping (more of which later).

Moving away from the "traditional" view of words as encoding concepts to an account on which words merely potentially activate concepts also leads to the consequence that one loses the idea of words having *literal meanings*. Since a given lexical item potentially activates a range of concepts and none of these occupies a privileged (default) position compared to the others, words have no real meaning on my account. This goes starkly against the intuition shared by many people that words have some kind of content by default, an intuition that has served as the basis of most semantic and pragmatic accounts of language and communication.

In the pragmatic approach of Grice (1989), the semantic content of sentences, viz. the syntactic combination of the intuitive literal/conventional meanings of the words, plus disambiguation and reference resolution, amounts to "what is said" by someone uttering that sentence. If "what is said" is false, irrelevant, uninformative and/or un-perspicuous, a hearer who assumes that the speaker is observing the Co-operative Principle will have sufficient motivation to look for what the speaker might have pragmatically implicated, thus recovering the intended meaning behind the utterance. Though the general framework has been amended, adjusted and attuned in many contemporary neo-Gricean pragmatic theories, the role of literal word meaning as contributing to a psychologically important level of "what is said" is maintained in most accounts (e.g. Bach 1994; Horn 1984; Levinson 1983). In more recent Grice-inspired minimalist approaches to semantics, it also serves as a "minimal defense against confusion, misunderstanding, mistakes and it is that which guarantees communication across contexts of utterance" (Cappelen and Lepore 2005: 185; Borg 2004: 58 seems to have some of the same motivations in proposing her literal sentence-meaning "as a non-cancellable level of content in a linguistic exchange").

But not so in Relevance Theory, where semantic representations of sentences are incomplete logical forms, non-truth evaluable entities that function merely as evidence that points a hearer towards the speaker's intended meaning. Though Sperber and Wilson (1995) suggest that most words have a corresponding concept which serves as the semantic content of that word, they argue that this semantic content never actually surfaces to consciousness. "Semantic representations become mentally represented as a result of an automatic and unconscious process of linguistic decoding" (1995: 193), Sperber and Wilson contend.

Literal word meaning therefore does not serve the same important purpose in RT as in other broadly Gricean approaches to communication[63]. As discussed in chapter 1 (section 1.3.3), Relevance Theory takes a radical approach to the problem of linguistic under-determinacy, and claims that the meanings of sentences are mere "assumption schemas" which (almost) always have to be pragmatically developed into a fully propositional form to satisfy a hearer's expectation of relevance. These schemas are used "to identify first the propositional form and then the explicatures of an utterance. It is these explicatures alone that have contextual effects, and are therefore worthy of conscious attention", claim Sperber and Wilson (1995: 193).

If the new account of lexical items can serve the purpose of helping a hearer reconstruct the speaker's intended meaning, it should not matter for the relevance-theoretic account that words have no meaning outside of a context[64]. It may be unfortunate in terms of violating "folk linguistic" intuitions, but as Fodor claims, there is nothing in people's intuitions about semantics which identify them as *semantic intuitions*. "Informants, oneself included, can be quite awful at saying what it is that drives their intuitions", according to Fodor (1998a: 86). I share the belief that it is "*always* up for grabs what an intuition is an intuition of" (Fodor 1998a: 87) and think that it is just as likely that people's intuitions about word meaning are *meta-linguistic* as that they are purely semantic.

The process of reflecting on word meaning is, after all, not the same as applying these meanings automatically in linguistic interaction. If Sperber and Wilson (1995) are right, and semantic representations never surface to consciousness, there should be no reason to think

---

[63] The fact that words can productively and systematically combine to yield new meanings is taken by many semantic theorists to entail that natural languages must be *compositional*. Claiming, as I have done, that words have no inherent meaning makes it hard to see how this can be explained. It could be argued, then, that giving up on encoding has an additional unfortunate consequence not discussed so far. However, this depends on what sense of compositionality is required for the productivity and systematicity of natural language to be adequately accounted for, and theorists are far from unified on this issue (see Pagin and Westerståhl 2010a; 2010b for a recent review). If what is required is a very strong sense of compositionality, in which the truth-conditions of sentences need to be determined exclusively by its constituent parts and their modes of composition, it is clear that my account cannot provide this. But in this respect, I do not think I am much worse off than Sperber and Wilson or other radical pragmaticians who claim that semantic representations stop well short of supplying truth-conditional content (see Carston 2002: 70-74 and Fodor 2001; 2008: 219 on this point). If a "weaker" version of compositionality is required, according to which natural language words should be able to recursively combine to yield new structures, I do not see an a priori argument to the effect that my lexical entries cannot be compositional in this sense. Whatever lexical-syntactic information is needed to do this can be built into the model independent of the conceptual level, much in the manner that Pietroski (2009; 2010) envisages for his notion of lexical items as "instructions to fetch concepts".

[64] The fact that I am giving up on encoding and other forms of automatic or default activation of concepts also makes my account unable to yield out-of-context predictions about which of a range of potential concepts gets activated from occasion to occasion. However, I take this not to be a problem since the property of relevance, which guides the search for the best candidate, is seen as *comparative* rather than *absolute* (Sperber and Wilson 1995: 79, see also the discussion in chapter 1, section 1.4.3).

that a conscious process of introspection provides the theorist with anything informative on what the word-concept relation amounts to. Concepts, and their relation to public items like words, are at the current stage of technological (and theoretical) development inaccessible and inscrutable for cognitive researchers and philosophers alike, so accounts of concepts will have to proceed first and foremost from theoretical considerations. I will return to the issue of concepts and empirical scrutability in chapter 6.

3.3.3. Preserving lexical pragmatics

In chapter 1, I argued that the lexical pragmatic machinery proposed by Relevance Theory provides a very potent analytic tool for understanding language. It is desirable, then, that whatever one chooses to replace current relevance theoretic notions does not undermine the explanatory potential of lexical pragmatics, maintaining the central ideas responsible for its accomplishments.

I see it as fortunate that the alternative account I have suggested preserves the idea of ad hoc concepts found in Relevance Theory, and it is fully compatible with the idea that "lexical adjustment may be a one-off process, used once and then forgotten, creating an ad hoc concept tied to a particular context that may never occur again" (Wilson and Carston 2007: 238). It also preserves the analyses of different uses of the same word as giving rise to distinct, but related meanings. With the uttering of the word 'tired', a specific concept of TIRED locking to a property instantiated by a certain person, say, someone who has just run a marathon, will either be retrieved from long-term memory or formed ad hoc. On another occasion, 'tired' may activate a TIRED concept locking to a property of *tiredness* instantiated by someone who has been up all night reading fantasy novels. The process by which this happens will still be pragmatically fuelled, taking words as clues to the speaker's meaning, yielding an output that is truth-evaluable and conceptual.

But there is a noteworthy difference between my alternative proposal and the relevance theoretic one in that while the lexical pragmatics advocated by Sperber, Wilson and Carston talks of a general concept like TIRED being narrowed to particular degrees from occasion to occasion, the potential activation account claims that there is no narrowing *of the concept itself,* or its denotation. Rather, it is the area potentially activated by a given lexical item that gets specified, limiting the domain in which the cognitive system will search for the relevant concept. Upon coming across a lexical item, the whole conceptual area covered by the lexical entry is a potential place to search, but only a narrower sub-section of the large

area in which one finds TIRED concepts will be an actual candidate for the pragmatic system to search[65].

Some further differences between the standard RT lexical pragmatics and my proposal concern the issue of broadening. Recall that for e.g. Wilson and Carston (2007) there are some words which encode strictly defined concepts. Examples include RAW, which applies to things that are not cooked, SILENT, which denotes states of affairs where there is no sound whatsoever, and EMPTY, which mean that the item in question has strictly no contents. What makes these concepts convey more general meanings is a process of broadening in order to meet expectations of relevance. Examples include the use of SILENT* to talk about rooms with only a slight electric hiss in the background, or a SILENT** that denotes the sound of the crowd at a football stadium immediately after an away team has scored a goal.

My suggestion that there are clear boundaries around the domain in which concepts can be searched for effectively precludes the same type of treatment as the one argued for by Wilson and Carston (2007). By not allowing 'tired' to activate concepts that fall outside its domain (e.g. those potentially activated by 'bored'), I have also ruled out there being a broadening of the 'silent' region responsible for the generations of concepts like SILENT* and SILENT**. What I will have to claim instead is that these types of concepts already fall under a bigger lexical entry, 'silent', potentially comprising lots of meanings not normally categorized as part of the literal 'silent'.

For other, more radical types of what Wilson and Carston (2007) analyse as broadenings (see also Sperber and Wilson 2008), I can follow their lead and adapt aspects of the relevance theoretic account of metaphors and category extensions. Wilson and Carston (2007) argue that in getting from the literal conceptual content of a word like 'chameleon', a hearer constructs a new ad hoc concept that comprises not only actual chameleons but also people who share with chameleons certain properties relevant in the context. So an utterance such as the following:

70. Sally is a chameleon

may be used to communicate that Sally "has a capacity to change her appearance to fit in with her surroundings, remaining unnoticed by her enemies and escaping attack etc." (Wilson and Carston 2007: 235). To arrive at the ad hoc concept CHAMELEON* conveyed here, the hearer is

[65] It follows that the treatment of polysemy will have to rely on the activation of discrete conceptual entities, instead of the narrowing of one underspecified concept in utterances like 'The *meal* was delicious but lasted three hours'. However, this does not amount to a reduction of polysemy to homonymy, since the two tokens of MEAL concepts will be formally distinct in the Language of Thought. More on this below.

hypothesised to draw on the encyclopaedic properties of chameleons and use them to construct a new category of which both chameleons and Sally are members.

For the chameleon case, I would suppose that there is a lexical entry picking out a conceptual domain containing one or more CHAMELEON concepts (a lizard expert presumably has several of them while laymen possess but one or two). One of these concepts (if there is more than one) is activated to yield access to the encyclopaedic entry containing beliefs about actual *chameleons*. The hearer then draws on the relevant belief(s) and uses them to construct an ad hoc concept in another conceptual domain.

In the case of category extension, such as in the following example (from Wilson and Carston 2007: 236), the process should function in more or less the same way:

71. Iraq is this generation's Vietnam

Here, one particular concept out of a range of possible VIETNAM concepts will be activated, and encyclopaedic information about the United States warfare in the country collected from the mental filing cabinet. This, in turn, is used to construct the relevant ad hoc concept VIETNAM***** denoting "the category of disastrous military interventions" (Wilson and Carston 2007: 236) of which both the Iraq and Vietnam wars are claimed to be members. In a different context, 'Vietnam' may come to activate another VIETNAM concept, giving e.g. access to beliefs about cheap, warm and beautiful places to go on holiday.

As far as I can see, there is nothing in Wilson and Carston's account of metaphorical broadening that is incompatible with the alternative view of the word-concept relation that I have advocated so far in this section. Also, my account is congruent with Sperber and Wilson's (2008) contention that some very standard metaphors become conventionalised and lexicalized as ambiguous lexical entries. With a case of the tokening of a word 'pigsty' in 72 and 73, I will claim, like Sperber and Wilson, that it may have to be disambiguated before the right lexical entry can serve to activate the relevant concept for that context:

72. Your room is a pigsty
73. We built a store room where the pigsty used to be

It seems, then, that much of the explanatory potential of the lexical pragmatic machinery is preserved on the alternative conception of word meaning I have defended, little being lost in terms of preserving Relevance Theory's analytic qualities.

Though my run-through of consequences for lexical pragmatics is unduly elliptical, and no doubt leaves most questions unanswered, all I have wanted to do here is show that

most of the interesting features of the lexical pragmatics machinery can be preserved even with an alternative view of the mapping between words and concepts in place. Only time and further research will show whether a view of words as potentially activating concepts can be as explanatorily potent and theoretically consistent as the account proposed by Wilson and Carston (2007). My modest claim for now is that the alternative to encoding can also handle the cases analysed by Wilson and Carston's (2007) version of lexical pragmatics, without going against any of the central tenets of Relevance Theory.

3.3.4. Sense "ambiguity" and memory load

Another striking outcome of giving up the idea of the mind containing a stock of general concepts that are specified from occasion to occasion is a radical proliferation of mental items and properties. If words like 'tired' do not express one general concept that can be narrowed to yield specific senses of *tiredness* appropriate to the context, but rather is seen as potentially activating any number of separate concepts, this seems like a clear violation of principles of theoretical economy, most notably the Modified Occam's Razor (MOR). In his *Logic and Conversation* lectures, Grice (1989: 47 [orig. 1967]) proposed the MOR as stating that "Senses are not to be multiplied beyond necessity". This, in effect, is a methodological principle to the effect that where the theorist has a choice, she should not postulate more meanings than she has independent reason to believe are needed to explain a certain linguistic phenomenon.

Grice used the principle to argue against semantic ambiguity treatments of, among other things, truth-functional connectives, favouring instead a pragmatic account which he held to be more economical. His point was that instead of assuming a complex semantics for words such as 'and' in order to capture the various temporal and causal relations it can express, a pragmatic account that appeals to one sense plus implicatures is preferable pursuant to MOR. Grice's principle has often been extended as an argument against all types of linguistic ambiguity accounts, and it seems clear that any theory which relies on massive ambiguity in explaining the apparent flexibility of word meanings quickly runs into problems.

Consider, for instance, an approach according to which 'tired' is not lexicalized as one word in the mental vocabulary, but rather figures as a number of different items expressing different meanings. One may find a word 'tired_1' meaning *mentally tired*, 'tired_2' meaning *physically tired*, 'tired_3' meaning *drained of all bodily energy*, 'tired_4' meaning *drained of most of one's bodily energy* etc. Considering all the different manners in which a person can be tired and all the different physical and mental hallmarks associated with these quickly

makes one realize the impracticality of construing a lexical entry for each and every one of them. Furthermore, if there are massively ambiguous lexicons, it would be near impossible explain linguistic/communicative coordination. Two people with a stock of 20 different entries for 'tired' (indexed $tired_1$…$tired_{20}$) in their minds would not only have to be able to retrieve the right TIRED meaning for themselves, they would also have to somehow convey to their interlocutor that they mean, say, $tired_7$ and not $tired_{14}$ or $tired_1$.

If massive lexical ambiguity were the upshot of the potential activation account I have proposed, that would surely render it a non-starter. But the alternative proposal entails only an increase in the number of concepts, not words, as there is only one postulated lexical entry 'tired' which gives potential access to a number of underlying concepts. The lexical entry 'tired', though potentially activating a number of different concepts, is still univocal, according to my account. The concepts potentially activated, in turn, are *not* ambiguous in the way a range of lexical items $tired_1$…$tired_N$ would be, as they occupy distinct addresses in conceptual space and/or are formally realized in different Modes of Presentation from concept to concept. Recall that according to the Computational Theory of Mind, the Language of Thought is an amodal system independent of natural language. Just because two word tokens are formally univocal in a given natural language, nothing is thereby entailed about how the corresponding concepts are "orthographically" realized in the mind.

Even though greater theoretical economy is always to be preferred other things being equal, I do not see an independent reason (as Grice demands) to restrict people's mental vocabulary to the point where it corresponds numerically with their public lexicon. In fact, given the relative ease with which one can create and store new meanings ad hoc (Barsalou 1982; 1983; 1987), compared to the slow process of novel word formation and stabilisation, it would be surprising if a person's conceptual repertoire weren't vastly bigger than his natural language counter-part (Sperber and Wilson 1998). Taking into account the aforementioned points on the theoretical independence of the Language of Thought should also lead one away from using public languages as a guide to a calculation of the stock of mental representations.

Neither does there seem to be any good empirical reason for putting a low ceiling on the size of the human mental vocabulary, as it is generally agreed that the capacity of the human long-term memory easily exceeds what it is possible to test for experimentally. Memory is cognitively cheap, and people can remember and differentiate between thousands upon thousands of individual percepts and events, retain information about impressions and episodes in great detail years after they took place. A famous study in the 1970s showed how people could determine, after viewing 10.000 scenes a few seconds at the time, which of two

images had been seen with 83% accuracy (Standing 1973). The same impressive memory capacity has been showed for other input types, such as audio (Miller and Tanis 1971).

Though largely a toy exercise, some psychologists and mathematicians have tried to calculate, on the basis of these and more recent experiments, what the upper bound of the human capacity for storing information is (for a review, see Dudai 1997). Landauer (1986) suggested that the brain of a normal adult may contain "a billion bits" of information, while Furman et al (2007) speculate that on the assumption that an adult may record and store for later reproduction approximately one event per minute (a number which they retrieve from a long-term memory experiment with audio-visual narratives), that yields "a potential capacity of 10^3 items per day, or an order of magnitude of 10^7 per lifetime" (Furman et al 2007: 464)[66].

Interestingly, humans are not the only species which record impressive results in the memory domain. Other creatures, such as rats, horses, sea lions, scrub jays, pigeons and chimps also score very highly on memory tests in experimental settings. Despite the obvious limitations on possible research paradigms, it has been showed that not only do these creatures have a capacity to store a wealth of information in long-term memory, they are also likely to retain the information over a period of several years. Non-human mammals, in particular, also display faculties which surpass the limits of testability. Fagot and Cook found, in a recent investigation, that baboons "memorized a minimum of 3,500–5,000 items in our task and could have even retained thousands more with continued testing" (Fagot and Cook 2006: 17566)[67].

Other studies which have tested for robustness and longevity of memory have discovered that "rhesus monkeys (Macaca mulatta) performed nearly perfectly on oddity learning sets 7 years after original testing" (Hanggi and Ingersoll 2009: 452). Statistically significant results have been found from 2 to 5 years after first testing on squirrel monkeys and gorillas, while a sea-lion "demonstrated the use of an identity concept to familiar and novel sets of stimuli in a 10-year memory test", according to Hanggi and Ingersoll (2009),

---

[66] Good numbers and estimates of the human memory capacity are hard to find, since, as Standing discovered in his first set of extensive experiments, the ability to memorize far outlasts the ability to concentrate and stay motivated during an experimental task: "The use of large learning sets precludes the use of a truly immediate test of memory, due to the considerable time needed to view the stimuli even once. This also leads to considerable fatigue, with sets over a few hundred items, and subjects clearly must make considerable efforts to maintain their vigilance. In the case of 10,000 items, the cumulative effects of five days' viewing, as checked by the author, are extremely gruelling and unpleasant" (Standing 1973: 221)

[67] In the same paper, the authors review evidence which demonstrates the ability of two pigeons to acquire and recall, on a picture-response association task, 62,3% and 67,6% correctly out of a memory set of 3,037 and 1,978 images, respectively.

who themselves found that a horse could remember and apply categories learnt some 10 years earlier.

The conclusion to draw from the arguments in this section is that a very large mental repertoire is no prima facie practical or theoretical implausibility. I do acknowledge, though, that for people who are ontologically minimalist, the proliferation of concepts proposed in this chapter may cause some heartache. If an individual has at her disposal a large number of different concepts potentially activated by the word 'tired', with a potential to form many more, this entails that there is an equally large number of *tired* properties out there in the world for her to lock to. Similarly for any other context-sensitive lexical item, like 'upset', 'value', 'love', 'red', 'good' etc. This follows from the postulation that each concept gets its content from a corresponding property. But since properties, on the Fodorian account, are seen as mind-dependent, they are, as Fodor claims, "ontologically harmless" (Fodor 2000b: 352).

In sum, it may be concluded that the proliferation of mental symbols and properties is both cognitively and metaphysically inexpensive.

3.3.5. Type physicalism and functional correlation

A further worry someone evaluating the account I have proposed might have, is that the "potential activation" story relies too heavily on concepts being individuated by "geography", implicitly entailing a commitment to *type physicalism*. If a theory depends on properties like space to individuate concepts or the relation between concepts and words, and "space" is something physically realized in someone's brain structure, it seems that the concept activation story has implicitly assumed a view claiming that concept's physical-neuronal location partly determines what a concept can be.

The problem with this commitment is that very few theorists hold the type physicalist thesis to be empirically and theoretically viable, the main issue (at least for my purposes here) being that it rules out the possibility of mental states being multiply realizable within and across thinkers. So if a concept TIRED's Mode of Presentation is individuated by its location in the brain, it cannot be recognized as the same concept in another type of place. It will have to be the literally same neuron types that are activated from tokening to tokening of TIRED. But this is an empirical prediction that is likely to be too strong when brains are considered across times and thinkers.

Even if neuroscience has shown that one can make robust generalisations about cognitive systems with relative continuity (Bickle 2008), this does not necessarily hold to the

level of individual concept. Each tokening of a given concept TIRED will not be identifiable by exactly the same neural marker, given that brain structures are ever changing and not completely identical across people (see Putnam 1980 [orig. 1967] and Fodor: 1974 for the original multiple realizability arguments against type physicalism; see Kim 1993: chapter 16; Bickle 1998; Bickle 2008 and Smart 2008 for recent reviews).

On a theoretical level, type physicalism violates what Fodor calls the "generality condition", a version of the publicity constraint discussed in the previous chapter. He claims that "It ought to be an empirical issue which kinds of creatures have our kinds of minds; which is to say that it ought not be a priori that only creatures with our kinds of brains do" (2008: 90), but identifying the Mode of Presentation of concepts with specific brain structures rules out the possibility of e.g. thinking Martians with silicon brains a priori. Fodor strongly advises against taking for granted that "computationally homogeneous primitive Mentalese expressions *ipso facto* have neurologically homogeneous implementations; indeed, we had better take for granted that they often don't" (*ibid*). The generality condition he advocates effectively rules out typing mental symbols via locations in the brain, since there seems to be no way of specifying a symbol's location physically without appealing to neural structures. But appealing to neural structures will not do, since it seems to preclude the possibility of a cat or Martian sharing concepts like DOG or TIRED a priori.

But even though I have relied on such notions as "coordinates", "geography", "addresses", "domains" etc. in explaining the individuation of concepts, these needn't be specified by an appeal to actual neurological *locations*. The idea of symbols grounded in a "conceptual geography" may very well be cashed out functionally, meaning that it is the causal relations a symbol enters into which determine its Mode of Presentation (see Fodor 2008: 90-92).

This way of seeing identity relations can be illustrated by thinking about such mundane activities as letter recognition. How does one know when a letter is an *h* or a *b* when faced with a text written in sloppy handwriting, the symbol being indistinguishable from their orthography alone? One looks at the rest of the word and the sentence and sees what functional role the letter plays in the text in which it is a part. How does one know whether I think DOG on two occasions in which I employ different sets of neurons to do the thinking?

One looks at the rest of my thought and the thoughts that follow, observing what functional role the DOG token plays in the rest of my thinking and behaviour[68].

Fodor holds it to be "a great mystery, and not just in psychology, why so many different kinds of lower-level phenomenon converge to sustain the same high-level generalizations" (2008: 91), and confesses to not possessing the key to this "metaphysical puzzle" (2008: 92). But, he claims, the "puzzle isn't specific either to the relation between intentional psychology and its various computational implementations", which makes it all right to momentarily "pretend that it is not there" (*ibid*). Even though I ideally would have liked to have a more substantial story to tell about what exactly the nature of symbols is, I will therefore rest assured that it is fair to leave the issue outside the scope of my work here[69].

Nothing in my proposal about concept activation, then, entails a particular view on how mental symbols are physically realized, and I am open to the possibility of the "addresses" at which concepts are located being individuated e.g. temporally (by way of specific patterns and ranges of brain wave oscillations, see e.g. Buzsáki 2006), rather than spatially. What the nature of the content of the aforementioned lexical "coordinates" is will in turn depend on how exactly the MoPs are specified, and what story one tells here will have to rely on assumptions about information encoding from neuroscience. But I believe nothing in the account as I have presented it so far rests on the exact specifications here.

That said, I am fairly confident that however the issue of symbol individuation turns out in the end, the geographical metaphor I have suggested is not completely out of place. That at least aspects of the orthography of symbols may be specified by an appeal to neural domains[70] is a hypothesis shared by many contemporary realist approaches to the mind. Sperber and Wilson (1998), for instance, talk of concepts as addresses in long-term memory,

---

[68] If it is OK to individuate the MoP of concepts functionally to rescue generality, is it not then fine to do the same for their content in order to preserve publicity? No, because Fodor takes functionalism to be incompatible with CTM and constraints on concept possession. Functionalism runs into the circularity objections discussed in the previous chapter, and the fact that the computations in CTM are explained "by reference to such semantic notions as content and representation; a computation is some kind of content-respecting causal relation among symbols" (Fodor 1998a: 11) makes symbols necessarily prior to the computations they enter into. If one, however, wants a functional explanation of content identity, "some other way of saying what it is for a causal relation among mental representations to be a computation; *some way that does not presuppose such notions as symbol and content*" (*ibid*) is needed.

[69] Susan Schneider argues, in a series of papers (2009a, 2009b) and a forthcoming book (Schneider in press), for an individuation of symbol types in the Language of Thought in terms of "total computational role". She holds that once the properties of symbols are thus specified this calls for a revision of the whole Language of Thought program, entailing an appeal to concept pragmatism. But since symbols are individuated non-semantically (Schneider 2009b: 526), it's unclear to what extent this alleged pragmatism affects what the theory has to say about content.

[70] To continue the analogy with letter recognition: deciding whether something is a *b* or an *h* is not done solely by looking at its functional role. One also has to take *shape* into consideration. Similarly, the function of a LoT symbol has to be constrained by something other than the inferences it enters into.

as does Pietroski (2010) in his theory of how words get content. In cognitive science in general, one finds many advocates of the idea of the conceptual mind as organized according to (functional) domains (see e.g. Caramazza and Mahon 2006; Mahon and Caramazza 2009).

Fodor (2008: chapter 5), too, uses the analogy of a Newtonian attractor landscape to strengthen the idea of concept acquisition as a brute process of nomological locking. He suggests that the mind can be metaphorically thought of "like a sea" (2008: 159) where concepts are attractors ("whirlpools") distributed geographically across the ("naval") landscape. The concepts lock to stereotypic instances of properties ("boats") in a causal process according to relative closeness of the relation between the attractor and an instance. The better a particular exemplar is, the closer it will fall to the attractor and the "more likely it is that learning the stereotype is [empirically] sufficient for acquiring the corresponding concept; that is, for locking to the property that the corresponding concept expresses", Fodor (2008: 160) claims. The bedrock for the concept acquisition process is the innately specified "geometry of the attractor landscape"[71] (2008: 161). He suggests that "[w]hat's learned (not just acquired) are stereotypes (statistical representations of experience)" (2008: 162) but knowing the corresponding "stereotype of a concept, together with a specification of a creature's experience, does *not* suffice to determine whether that creature will acquire that concept. You also need to know the geometry of the creature's attractor space" (2008: 163).

One clearly sees, then, that the topography of the conceptual mind is vital on Fodor's story as well. The geometry of the conceptual space is what constrains what conceptual representations a thinker can form, and serves as the foundation on which the atoms of meaning are formed. But even though the topography has to be available prior to the process of acquisition, nothing is decided up front on what the geographical space looks like for a given thinker or its species. Fodor maintains that he has "no reason to deny that [the geometry of the conceptual space] can alter under the pressure of experience, learning, maturation, or any other mind-world or mind-body interactions" (2008: 163) and I wish to follow him on this. Like Fodor, I am not committing myself to any particular view on how the topography is organized, since "in principle, all that is required to be set innately is the *initial* layout of the attractor landscape. From there on, everything is negotiable" (*ibid*).

---

[71] Fodor launches this idea in order to show how informational semantics gets out of what he sees as the problem of unlearnable concepts, forming the basis of his notorious arguments in favour of "nativism" (Fodor 1981a; see also my chapter 2, section 2.3.3). Even though he admits that explaining concept acquisition in terms of an "attractor landscape" analogy says nothing about "how the locking is achieved; it's just to say that how it's achieved isn't to be explained at the intentional level" (2008: 163). I return to the nature of the locking mechanisms in part II of this thesis.

3.4. Doing without encoding – the broader picture

3.4.1. Theoretical ecumenicity and other advantages

Over the previous sections I have outlined the potential objections an account that dissociates concepts from words might face. I hope to have shown that even though the consequences of giving up on encoding in favour of "potential activation" are serious, the alternative can answer some of the principled objections against it. The reason for proposing the dissociation was the wish to abide by the publicity constraint on conceptual content, as well as following the logical consequence of the contention that the "semantics of Mentalese is prior to the semantics of English" (Fodor 2008: 218).

Besides respecting publicity and the theoretical primacy of thought, I think there are some further upsides to viewing words and concepts as located on two separate levels. Firstly, it allows for some ecumenicity in the study of language and meaning. Fodor, famously, has argued that

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

In *Concepts* he holds that

> "English inherits its semantics from the contents of the beliefs, desires, intentions, and so forth that it's used to express, as per Grice and his followers. Or, if you prefer (as I think, on balance, I do), English *has no semantics*. Learning English isn't learning a theory about what its sentences mean, it's learning how to associate its sentences with the corresponding thoughts" (Fodor 1998a: 9)

But if one buys the ideas I have been pushing, with the lexical entry containing normative constraints plus a fair bit of syntactic and lexical information, not everything semantic is plausibly captured by Fodor's conceptual atoms.

Even though the conceptual level still has ontological priority over the lexical information (it's not possible to learn a content word without there being a corresponding concept to activate) and most of what has traditionally been called the *meaning* of a word will be captured by concepts, there will be interesting aspects of the lexical level to be studied on my alternative conception of semantics as two-levelled. I mentioned the possibility of thematic structure being a part of the lexical-semantic level, and others of what linguists have taken to be structural features of meaning can also be seen as plausible candidates for being part of lexical content. Johnson (2004: 354) argues that research suggests "many reasons for

supposing that agentivity is somehow part of the linguistic structure of a verb like *sink*"[72], but holds that if there's no necessary isomorphism between words and concepts this need not entail "that the concept of sinking contains a similar bit of structure" (*ibid*).

Fodor has, at various points over the last few years, been engaged in a debate over the viability of the so-called "impossible words argument", which purports to show that some lexical items have complex structure (Fodor and Lepore 2002: chapter 6; Fodor and Lepore 2005; Hale and Keyser 1993; 2002; Johnson 2004; Mateu 2005). The reason why he cares about this issue is, as Johnson (2004) points out, that he accepts the Isomorphism Assumption, which holds that lexical structure is mirrored in the corresponding concepts. So if there is an argument to the effect that words are complex then it follows that concepts are complex, viz. non-atomic, too.

But if one gets rid of the isomorphism between words and concepts, as I have done, the assumptions linguists make about the content of lexical entries may very well be independent of the ones a philosopher of mind makes about concepts. Any arguments indicating the complexity of lexical entries is unlikely to have a bearing on conceptual content, and neither theories of concepts nor theories of words need be restrained by the requirement that their suppositions hold for both domains.

Another positive consequence of severing the tie between the lexical and the conceptual levels is that it can explain various dissociations between language and thought, thereby accounting for some robust findings in the psychological literature. Martínez-Manrique (2010) summarises several arguments from Vigliocco and Vinson (2007) which "seem to demand a theoretical distinction between conceptual and semantic levels of representation" (Vigliocco and Vinson 2007: 198). These are results from neuro-psychological studies which show that some patients with lesions display domain-specific semantic deficits only in linguistic tasks and not in non-verbal tasks (Hart and Gordon 1992; Cappa et al. 1998) and studies with healthy individuals where "language-specific effects of grammatical gender disappeared when the task did not require verbalization"[73] (Vigliocco and Vinson 2007: 198; Vigliocco et al 2006).

---

[72] Johnson draws on syntactic data from the passivisation of verbs (Baker et al. 1989) which is taken to show that there in verb phrases are subject positions which, even though the subject is unpronounced, still supports so-called *by* phrases and purpose clauses (Johnson 2004: 351). Johnson concedes that the data does not argue conclusively in favour of complex lexical structures, but holds that the lexicon is the prime candidate for being the origin of such structures, given what he sees as a lack of other options. See Fodor and Lepore (2005) for another interpretation of Johnson's data.

[73] The effects reported were from studies of elicited similarity judgments, where Italian nouns with the same grammatical gender were rated as more similar in verbal than in non-verbal tasks (see Vigliocco et al 2005).

Vigliocco and Vinson (2007) also claim that widely held assumptions within psychology of language (some of which have been discussed above), such as that people are taken to have more concepts than words, that there are some universalities of conceptual structures across space and time, and that polysemous words are not plausibly accounted for via ambiguous lexical entries, all speak in favour of there being a dissociation between words and concepts. Though I grant that there are a number of ways in which these facts can be incorporated and accounted for within other theoretical framework than the one I have proposed in this chapter (whether or not they appeal to isomorphism and/or encoding), I think the ease with which the "potential activation" theory I have discussed handles this data constitutes an argument in favour of developing it further.

3.4.2. Other non-isomorphic approaches to words and concepts

Though the story I have told so far is original to this thesis and therefore somewhat idiosyncratic, there are other suggestions parallel to mine in the current psychological, linguistic and philosophical literature. It may even seem as if the traditional stronghold of the Isomorphism Assumption is declining somewhat, with new and exciting empirical and theoretical investigations into the word-concept relation emerging from several theoretical disciplines. The considerations Vigliocco and Vinson raise, as well as other issues discussed by scholars in e.g. lexical semantics and pragmatics, have led a number of other theorists to propose accounts of word meaning in which the layer between lexical and conceptual content is dissociated to various degrees. Consequently, some of these theorists have proposed alternative conceptions of word meaning that don't rely on an isomorphic mapping.

For instance Vigliocco et al (2004; Vinson and Vigliocco 2008) have developed a model of the word-concept relation where concepts "comprise distributed featural representations" which are bound together by a lexical-semantic process (Vigliocco and Vinson 2007: 198; their account draws on some central insights from Damasio et al 2004). According to this so-called "Featural and Unitary Semantic Space" (FUSS) hypothesis, "[c]onceptual features (…) are bound into a separate level of lexico-semantic representations which serves to mediate between concepts and other linguistic information (syntax and word form)" (Vigliocco and Vinson 2007: 209). Vigliocco and Vinson suggest that the "organisation at this level arises through an unsupervised learning process (…) which is

sensitive to properties of the featural input" (*ibid*). The features are held by Vigliocco and Vinson[74] to be properties which taken together express the meaning of a word.

Evans (2009) has also argued for the separation of a lexical level from the conceptual domain, and claims that "semantic structure and conceptual structure form two distinct levels of representation, and do so because they inhere in two distinct representational systems" (2009: 43). Word meanings arise "by virtue of a dynamic exchange taking place between the linguistic and conceptual systems" (2009: 43-44) where the conceptual systems consist of cognitive models involving *frames* and *simulations* – notions he borrows from the work of Barsalou (1999; 2003).

Vicente (2010) has developed a decompositionalist account of word meaning which holds, contra classical approaches to meaning decomposition, that a "simple term in our language does not correspond to a complex concept with a definitional structure: to begin with, the relation between words and concepts is one-to-many" (2010: 80). Instead, he claims, "the complexes a token of a word corresponds to on a given occasion are typically built out of a determinate set of basic concepts" (2010: 81) as an end-product of a process of pragmatic search.

Recanati (2004) has suggested (but not endorsed) a view of word meaning he calls *Meaning Eliminativism*, which "gets rid of abstract meanings for [word] types, in favour of particular uses. The contextualized sense carried by the word on a particular use depends upon similarity relations between that use of the word (…) and past uses of the same word" (2004: 151)". Words contribute "semantic potential", seen as "the collection of past uses on the basis of which similarities can be established between the source situation (…) and the target situation", according to such an account (2004: 152). Another idea in the same (Wittgenstenian) spirit comes from Pritchard (2009: chapter 6), who claims that "our use of words is guided by a memory-form, which arises from interaction with objects/events, and which provides a constraint which we apply with judgement" about whether something can be correctly described by a given word. The idea is that perceptual features of objects in the world gets discerned and stored in memory, and that these "imagistic" memories form the basis of word use.

---

[74] The reliance on semantic features as what individuates meaning is something Vigliocco and Vinson share with many contemporary semantic accounts. They distance themselves from the accounts of their peers, however, in that Vigliocco and Vinson model their semantics on the explicit meta-semantic judgments of "multiple naïve speakers" (2007: 209). Though they take this to be an improvement on accounts that rely on investigators "a priori" choice of features, it presupposes the controversial assumption that conscious introspection provides one with pure data of actual semantic content (see my chapter 6, section 6.3.2 for discussion).

Finally, Paul Pietroski has, in a couple of recent papers (2009; 2010), suggested that word meanings be thought of as "instructions to fetch concepts", where the instructions implement a procedure which pairs signals with *monadic* concepts. The relationship between words and concepts may be one to one, many to one, or one to many (as with polysemy), according to Pietroski (2010: 251-252). He holds that "Human I-languages are *naturally acquirable procedures* that pair distinctively human linguistic signals—like the sounds of spoken English or signs of ASL—with the corresponding interpretations, whatever they are" (2010: 12).

What I have proposed as words potentially activating concepts shares some affinities with aspects of these accounts. I agree, for instance, with Vicente (2010) when he claims that pragmatics is responsible for selecting which part "of the cluster of concepts associated with a given lexical entry has to be active in the recovery of the thought expressed" (2010: 98). I also think that the contents of my lexical entries could be seen as implementable procedures that fetch and combine concepts, in the manner envisaged by Pietroski (2009; 2010). I hesitate to fully adopt any of these accounts as they stand, though, as they all stop short of filling in the gaps left by giving up on isomorphism/encoding.

For instance, the reliance on features (be they perceptual as on Pritchard's and presumably Recanati' accounts, or conceptual as in Vigliocco et al.'s theory) leads to the problem of specifying just what these features are and which of them are constituitive of a given word's meaning. This problem is particularly forceful when it comes to perceptual features, since objects falling into the same category are seldom identical to one another. One therefore has to rely on similarity of features; but the features themselves cannot be similar to each other, because this again would presuppose identity (see my chapters 1 and 2 for a fuller argument and discussion).What is needed is some sort of story about how these features are individuated, what physical properties characterise them (if they are perceptual) or how they otherwise get their content via e.g. denotations (if they are conceptual), but neither Vigliocco et al. nor Pritchard hint at how this can be done.

This worry also threatens Vicente, since he relies on complex clusters of concepts as being what individuate word meanings. But if the meaning of a word like 'tired' is a cluster of complex concepts, something needs to be said about which and how many of these concepts have to be possessed by an agent before she can be said to know the meaning of a word. Also lacking in Vicente's account is a story about how the individual concepts in the cluster get their content and what distinguishes them from each other (for some general arguments

against concept decomposition, see Fodor 1998a: chapter 3; Fodor 2008: 158ff, especially note 28; Fodor and Lepore 2005: chapter 9).

I see Evans' and Pietroski's theories as the most developed and interesting of the ones mentioned above, but they too are at loss in explaining in what way the mental items underlying word meaning get their content in the first place. Pietroski is agnostic about this question (suggesting tentatively that his account "leaves room for the externalist idea that interpretations are individuated by features of the environment" 2009: 19, n18), while Evans ultimately relies on Barsalou's notions of cognitive content, where "simulators" play the role concepts play in the Computational Theory of Mind (see Adams and Campbell 1999). What determines the content of these, the account says nothing about.

Normativity is, as far as I can see, a problem for several of these accounts. As I argued above, getting rid of isomorphism/encoding has as a consequence that one loses the explanation of how *the right mental item* can be picked out from a range of potential candidates on a specific occasion. If a polysemous word is to fetch a given one of a number of associated concepts (as on Pietroski's story), there has to be some (semantic, pragmatic or whatever) mechanism to indicate which is the correct or best one in that context. Normativity is also the biggest obstacle to pursuing the line suggested by Recanati. If words contribute semantic potential, and thereby potentially pick out lots of memory traces (think of how many encounters with e.g. chairs or dogs you have had) some way of picking out the most relevantly similar traces for a given context is needed – in addition to the required specification of what it takes for two traces to be 'relevantly similar'.

I do not necessarily raise these points as objections to Pietroski, Evans and the other accounts, as I acknowledge that they are works in various degrees of progress and that the scope of their research is different from the perspective I take in this thesis. I merely intend the issues I discuss in this section to highlight how getting rid of isomorphism/encoding between words and concepts is no small matter. The popular degree of support it receives from contemporary theorists, and the implicitness with which it figures in their accounts, help mask the heavy work-load it carries in explaining how the gap between language and thought is bridged in communication – no matter which theory of communication one works with.

I hope, then, to have adequately presented some of the important consequences of opting for an alternative explanation of the word-concept relation as well as providing enough justifications for my endeavours in pursuing such just an alternative. With the new picture of concept activation and the mapping between language and thought in place, I will proceed to consider what I in chapter 2 argued was the other main problem for the reliance of lexical

pragmatics on informational semantics/conceptual atomism. In the next two chapters, I will discuss the *argument from ontology*, looking at what resources the theorist relying on informational semantics has at her disposal in explaining concept acquisition and mechanisms of semantic access. In so doing, I will draw on both the explanatory potential of Relevance Theory and the theoretical benefits I hold the alternative view of the word-concept relation presented in this chapter to have, in order to explain how words denoting non-perceivable entities get their content.

3.5. Conclusion

In this chapter I have argued for and tried to develop a view of word meaning which, though it preserves many of the ideas from Relevance Theory, does not rely on semantic encoding or any other default mechanism in specifying the mapping between words and concepts.

I have advocated a theory of word meaning on which lexical items potentially activate a range of corresponding concepts that are formally distinct from each other. What determines which concept gets picked out by the tokening of a given word in a specific context is a pragmatic process governed by the search for relevance, according to the suggested account. I have claimed that each lexical entry contains, in addition to syntactic and phonological/orthographic information, *retrieval constraints* limiting the domains in which the pragmatic mechanisms are allowed to search. These constraints therefore confine the potential space in which one can look for the particular concept that will be the most relevant on a given occasion, explaining how meaning can be normative even when the idea of semantic encoding is lost.

The motivation for developing this alternative conception of word meaning was the desire to respect Fodor's publicity constraint on concept possession, which specified that all claims about differences in individuals' conceptual repertoires should be empirical. I have suggested that by not relying on semantic encoding, or any other type of word-concept isomorphism, the theory is not forced to take a stand on any questions about mental inventories or the relationship between language and thought a priori. This, I held, was in line with the ontological primacy of concepts assumed by the Language of Thought hypothesis, and gives the potential activation account a methodological advantage over theories that conceptualise the word-concept relationship as more or less direct.

The theory I have ended up with marks a significant break with traditional conceptions of word meaning, as well as clashing with ordinary intuitions about how words work. Treating words as having no inherent meaning, but as picking out instead something that has

to be determined in the specific context in which they are tokened, is radical, even more so than the position taken by Sperber and Wilson (1995).

One may imagine, given the existing objections to the relevance theoretic view of communication as taking impoverished semantic representations as input to heavy-duty pragmatic processing (see chapter 1), that what I have been advocating will go against many theorists' conceptions of how words contribute meaning to natural language utterances. At the very least, my claim that there is no way to predict the "literal" meaning of word types will be at odds with ordinary folk linguistics. Just looking at a random sample of letters to the editor of any newspaper will suffice to show that among the topics that engage people the most are what the "real meaning" of a given term is.

Few things get people so worked up as discussions about the correct and incorrect application of words, with generations of men and women constantly annoyed by others' "improper" uses. How can the account make sense of this, if the claim is that words do not have meaning at all? In this chapter, I have suggested that there is nothing in the intuitions people have about word meaning which indicates that they are necessarily semantic in nature. Following Fodor, I think it is "always up for grabs what an intuition is an intuition of" (1998a: 87), and I believe there are a range of good candidates for what underlies the intuitions people express in angry readers' letters, though none of them are good enough to be considered the default.

Furthermore, I should point out that even though the account I have proposed eschews the notion that words have meaning outside of a context, it does not necessarily follow that it is possible to token a word without any kind of conceptual content being activated all. My claim that no word activates a given concept by default does not entail that there is no automaticity at all to the process by which mental items are tokened. On the contrary, I think it is highly likely that the workings of the concept retrieval mechanisms are out of reach of cognitive control and therefore impossible to suspend. Once this clarification is made, it may be possible for the account I have proposed to explain intuitions about word meaning as arising from whatever concept is activated by a given word at a given time, or which concepts fall within the domain specified by the lexical retrieval constraints. Even while engaging in reflective, metalinguistic contemplation, one still finds oneself in a context, with factors about when a given word was last used, which of its meanings is the most frequent, etc. are likely to influence one's judgment about the particular meaning of the word.

Adding the views presented in this chapter to the central tenets of informational semantics and Relevance Theory outlined in chapter 1 and 2 provides at least some answers to

the questions with which I started this thesis. I wanted to know how people understand words, how they are used to convey meaning and how this meaning is acquired. So far, I have dealt with the nature of communicative processes and the metaphysics of meaning, but have had nothing to say about how meanings are acquired. The reason for this has been the separation Fodor makes between semantics and epistemology, or in other words, between conceptual content and whatever mechanisms sustain the content-constitutive link between mind and world.

Even though Fodor claims that an explanation of what these mechanisms of *semantic access* are and how they work is irrelevant to semantics, this does not mean that they are uninteresting for the purpose of understanding concepts. Indeed, Fodor (1998a: 75) claims that informational semantics is "untenable" without some account of the causal processes behind the nomic connections which individuate meaning. In the second part of this thesis, I will therefore investigate the means by which a particular kind of concept might be acquired within the framework of Fodorian semantic externalism.

I will focus on a class of concepts representing so-called *abstract entities*, things that people talk and think about without being able to see, touch or interact with. Assuming that it is possible (as Horsey 2006: chapter 5 does) to give an informational semantic account of how concepts representing properties instantiated by tangible entities are acquired via perceptual mechanisms, I will focus on cases where non-perceptual processes seem to provide the mechanisms that sustain semantic access. The aim will be to build on the groundwork I have laid so far, contributing to a more complete view of word meaning from the perspective of radical pragmatics and semantic externalism.

Part II:
Concept acquisition and the representation of abstract entities

4. Semantic externalism without metaphysical commitment: The argument from ontology and the representation of abstract entities

I have long believed that part of the reason we have inflation is that it makes it easier for us to con ourselves into thinking we are moving forward in our careers even when we are not
– Reader comment on the blog Pop Economics[75]

4.1. Introduction

So far in this thesis I have argued for an externalist/informational semantic view of mental content. I have claimed that concepts get their meaning from standing in a lawful relation to mind-external properties, which may or may not be instantiated by something in the world.

In chapter 2 I discussed a number of objections the informational semantic approach to concepts faces, and in this section of the thesis I take up one that I see as particularly challenging for any theory relying on a mind-world link to explain what primarily constitutes content. I will discuss what I call *the argument from ontology*, which holds that objects in the world, being too unruly and/or inaccessible to people's perceptual systems, are not the right kinds of things to supply semantic content in a naturalistic theory of the mind.

Versions of this argument have been put forward by a number of different theorists who use it to support an assorted set of claims. The argument can therefore be interpreted in numerous ways, and the above gloss is but rough simplification of a very complex set of contentions. Nevertheless, in what follows, I will take a closer look at what the argument from ontology purports to show and how it can be seen as raising problems for informational semantics. I will focus on two ways of presenting the argument, one which I attribute to Chomsky (2000) and one whose chief advocate I take to be Jackendoff (2002).

The notions I will develop and build on were introduced in chapter 2, where I argued that Fodor's version of externalism was neutral with respect to what is out there in the world and how people come to be acquainted with it. What I aim to do in this chapter is to consider exactly how, other than by perceptual acquaintance, one can lock to properties with instances that are not encountered, not realized or not realizable. My chief claim will be that people rely on each other to a considerable extent in the way they lock to a range of properties, so that

[75] http://www.popeconomics.com/2010/10/02/should-we-be-worried-about-inflation/

their concepts often get their content from a semantic link mediated *via talk*. The majority of this chapter will be concerned with saying something constructive about the nature of this process.

In considering the argument from ontology, I venture from the semantic/pragmatic domain in which I have been operating in the first part of this thesis. The work in this and the following two chapters will deal explicitly with the question of *concept acquisition* and the mechanisms people use to actually sustain the content of a range of different types of concepts. In chapter 2 (section 2.3.4), I showed how some theorists see the lack of a substantial story to tell about how semantic content is sustained as a significant shortcoming in Fodor's theory of concepts. I will therefore try to flesh out some ideas suggested by Fodor (1998a) about the manner in which semantic access can be sustained, thereby addressing these critics' worries.

The structure of this chapter is as follows: In section 4.2 I introduce the argument from ontology as it is presented by Chomsky (section 4.2.1) and Jackendoff (section 4.2.2). I discuss the relevance of the argument for informational semantics, and argue in section 4.2.3 that even though what is out in the world is (largely) irrelevant to semantic concerns, the theory needs to have some way of specifying how concepts representing so-called "abstract entities" get their content in order to explain concept acquisition[76].

In section 4.3, I outline a theory of how concepts with semantic content that is not plausibly seen as sustained via perceptual mechanisms are acquired. Here I develop a view of semantic access via "deference" based on the cognitive pragmatic approach of Relevance Theory. I start out, in section 4.3.1, by discussing some general motivations for complementing Fodor's nomic informational semantics with a theory of concept acquisition. In 4.3.2, I give an outline of the relevance theoretic view of communication as "mindreading", and present Sperber and Wilson's (1995) view of how linguistic interaction can result in the formation of new beliefs (section 4.3.3). Drawing on work by Sperber (1996; 1997), I show how some beliefs formed as a result of communicating with and trusting in others may be epistemically incomplete "reflective beliefs" (section 4.3.4).

---

[76] In the psychological literature on the representation of non-perceptual entities (which I'll return to discuss in chapter 6), it is normal to talk about "abstract concepts". I tend to avoid this terminology in order not to associate myself with a "Fregean" view of concepts which treats concepts themselves are abstract (see e.g. Glock 2009). This can be contrasted with the "subjectivist" view I adopt in this thesis (which is also the view most psychologists work with), where concepts are seen as mental particulars and constituents of propositional attitudes such as beliefs (see Margolis and Laurence 2007).

In section 4.3.5 I argue that acquisition of concepts representing entities that an individual has never encountered can be seen as derived from such reflective beliefs. I suggest that the postulation of a cognitive mechanism for the formation of reflective beliefs can explain how concepts for abstract entities may be disseminated across unlimited numbers of people in lengthy causal chains, without anyone ever having encountered instances of the corresponding property. In section 4.3.6 I introduce the notion of *theoretical inference* to explain how semantic access to non-perceived/non-perceivable entities is sustained at the end of these causal chains. I conclude the chapter by looking at how informational semantics can explain the acquisition of concepts expressing natural kinds (section 4.3.7) and raise some questions and problematic issues which follow from the account I have presented.

4.2. The argument from ontology

4.2.1. Version 1: Chomsky and the use of 'reference'

In informational semantics, mental content is seen as constituted by there being an appropriate, meaning-making lawful connection between the mind and the world. For Fodor (1998a), this is a relation of nomological locking that holds between a concept and a property. This concept-world relation is taken to hold in virtue of there being some sort of causal mechanism which somehow sustains the mechanisms of semantic access. The idea is that people, in going about the world, encounter objects, events and states of affairs which instantiate different properties. As a consequence of these encounters, they come to form concepts resonating to these very properties.

The view that the basis of meaning is some kind of link between what's out there in the world and something in our minds has an intuitive appeal, and Fodor's way of cashing out the sustaining mechanism in terms of causal connections is theoretically simple and psychologically plausible. But some theorists hold it to be *too* simple, since many of the objects of people's thoughts and utterances are perceptually inaccessible or otherwise outside their epistemic reach. The objection is that once we start to investigate what actual things people speak and think about, we see that the story cannot be as straightforward as externalists claim. There are many properties which are never instantiated in objects encountered or perceived, or which are not uniformly instantiated in any one particular object, and this makes it a mystery how people lock to whatever it is that their concepts express.

This is what I call the argument from ontology, different versions of which can be attributed to theorists like Chomsky (1975; 1993; 2000), Hinzen (2007), Jackendoff (2002; 2006), McGilvray (1998; 2002) and Stainton (2006). The core of the claim is that there is no

way of naturalizing a semantic relation of reference if the semantic theory is to take into account the diverse and unsystematic way in which the objects of people's talk and/or thought are realized in the world.

A particularly forceful version of this objection is due to Chomsky (2000; Stainton 2006 and Hinzen 2007 also use some of the same examples), who takes natural language data to show that there can be no simple reference relation between what people talk about and what is out there in the world. A closer examination of what pre-theoretically seem like perfectly ordinary and tangible objects reveals that people ascribe "curious properties" (Chomsky 2000: 37) to these entities in their linguistic practice. Using 'London' as an example, Chomsky claims that "such terms as 'London' are used to talk about the actual world, but there neither are nor are believed to be things-in-the-world with such properties of the intricate modes of reference that city name encapsulates" (*ibid*).

According to Chomsky, "Referring to London, we can be talking about a location or area, people who sometimes live there, the air above it (but not too high), buildings, institutions, etc., in various combinations" (*ibid*). He provides, among others, the following examples to show that the actual, physical location of London plays no role in determining the reference of the term:

74. London is so unhappy, ugly and polluted that it should be destroyed and rebuilt 100 miles away
75. London has remained the same, although it is now located elsewhere

Even with simple terms such as 'book', "we find that words are interpreted in terms of such factors as material, constitution, design intended and characteristic use, institutional role, and so on" (Chomsky 2000: 15) and "the use of language can attend in various ways to these semantic features" (2000: 16). "Suppose the library has two copies of Tolstoy's *War and Peace*, Peter takes out one, and John the other. Did Peter and John take out the same book, or different books?" Chomsky asks (*ibid*), and points out that the answer to this question depends on such non-semantic factors as what is being attended to at a given time, what the goals of the speaker and hearer are, and so on.

Chomsky takes these examples to be so numerous as to show that "even in the simplest cases, there is no word-object relation, where objects are mind-independent entities. There is no reference relation, in the technical sense familiar from Frege and Pierce to contemporary externalists" (Chomsky 2009b: 199). The upshot is that there can be no naturalistic study of the relations between words and the world. This claim is also endorsed by McGilvray (1998; 2002), who suggests that reference-relations are artefacts, "variable,

flexible and context-sensitive forms of human action [and therefore not] apt subjects for science" (2002: 74).

But it is not clear how much of a bearing these arguments from natural language have on the issue of content in informational semantics. Recall that for Fodor, concepts are in principle theoretically and ontologically prior to and independent from lexical items (see sections 2.3.1 and 3.4.1 of this thesis). The content of a conceptual item BOOK or LONDON is seen as constituted by a nomological relation between this item and the (mind-dependent) properties *being a book* or *being London*. The relation can be sustained by a variety of mechanisms, without there being anything in the theory which makes the concept rely on the lexical items 'book' or 'London' to in any way get their content. And whereas different tokens of the word 'London' may pick out different objects or aspects of an object in the world, this does not show anything about what the properties of the conceptual item(s) associated with the term are.

According to Edwards (2010a: 111), Fodor's view of content as constituted by a relation between a mental item, not a word, and the world is put forward in an explicit effort to develop a theory of meaning on naturalistically respectable grounds; Edwards adds that this strategy is also part of Dretske's (1981) and Millikan's (1993) externalist programs. Depending on what theory of natural language semantics and pragmatics one adopts, the context-sensitivity of lexical items and the flexibility of their use need not intrude on the domain of mental content at all. In any case, the notion of reference Fodor has in mind does not seem to be the one that Chomsky is targeting in his critique – even though it is often represented as such[77]. While Chomsky talks about 'reference' in a natural language sense (Chomsky 2000: 36), the notion of reference employed by Fodor involves a brute-causal process detached from the social practices and norms of a community (Fodor 2008: 216)[78].

---

[77] See e.g. McGilvray's introduction to Chomsky (2009a) where he presents the debate about the referentiality of words as being between Putnam, Kripke, Burge and *Fodor* and "rationalist-romantic" (RR) theorists in the Cartesian tradition (Mc Gilvray 2009: 8). In an endnote, though, McGilvray admits that the "RR theorist has no qualms about naturalistically determinate causal relations in the world-head direction. Relations of this sort figure in an account of acquisition, not use" (McGilvray 2009: 111, n6). No doubt, Fodor may be partly to blame for the confusion, in that he implicitly relies on an isomorphism between words and concepts. But as argued in the previous chapter, the Isomorphism Assumption is not something he is committed to, and informational semantics does not need to rely on it in any way.

[78] In a discussion of Chomsky's objection to the idea of a referential semantics, Ludlow (2003) proposes to distinguish between several meanings of the term 'reference'. Reference$_0$ is an internal relation between mental items (like inferential roles). On his account, reference$_1$ would "simply be some direct (perhaps causal) relation" between an item and the world, while reference$_2$ "takes the relation to be rather more complex, involving at a minimum a four-place relation that involves the speaker, the expression used, context, and aspects of the world" (2003: 142). The notion Chomsky is concerned with would then seem to be R_2, which Ludlow and other externalists rely on, but crucially Fodor, Dretske and Millikan do not.

This equivocation in the use of 'reference' is unfortunate, since it masks important points of agreement between the externalists of Fodor's type and Chomsky and his followers, who both hold that there is no word-world relation (primarily) responsible for giving words their meaning. For Fodor, and others working in the Gricean tradition where utterances express speakers' meanings/thoughts (see Fodor 2001), words get their content from some sort of internalist (be it isomorphic or not, see chapter 3, section 3.2.2) relation to mental items. The point Chomsky presses is that an explanation of the complex use patterns of ordinary lexical items like 'book' cannot appeal to a direct link between the word and something in the world. This is a perspective a Fodorian may very well share.

It is of course extremely interesting how what seems like one and the same lexical item 'book' can express so many different meanings, but informational semantics is in principle completely neutral on how that happens. Plausibly, it can be explained by appeal to a lexical pragmatic account, such as the one proposed by Wilson and Carston (2007), or, if one believes the story told about the word-concept relation in chapter 3 above, one might argue that 'book' potentially activates a range of different concepts locking to different properties. If analysed this way, the polysemy of 'book' is accounted for by claiming that it activates two distinct concepts representing *bookhood* in an utterance like 76:

> 76. The book he is planning to write will weigh at least five pounds if he ever writes it (Chomsky 2000: 16)

The concepts are distinct in that they will have different Modes of Presentation, even though they are located in the same region and activated simultaneously (see chapter 3, section 3.3.1 and 3.3.4, also chapter 5, section 5.3.2 of this thesis). What type of analysis one opts for here, though, and whether a pragmatic account of communicated meaning has its place in a naturalistic science (see the introduction to Carston 2002 for discussion) is immaterial to the issue of whether externalism can account for how people form concepts corresponding to such things as *books*[79].

4.2.2. Version 2: Jackendoff and the problem of abstract entities

In the previous section I tried to show that Chomsky's version of the ontological argument against semantic externalism, though interesting and substantial, does not raise a problem for a theory which treats concepts as theoretically independent of lexical items. However, there is

[79] I will return to some of Chomsky's examples and the way a single word can activate different concepts and concept types in the next chapter (section 5.3.2). For a fuller treatment of polysemy from a relevance-theoretic perspective, see Falkum (2010).

a second construal of the argument which seems to me to be more relevant for the present concerns. Ray Jackendoff (2002) has questioned the way the "predominant traditions in Anglo-American semantics and philosophy of language" take for granted a "common-sense" view that linguistic/mental expressions "say things about the world and have truth values based on their relation to the world" (2002: 294). Jackendoff argues that even though the presupposed notion of "the world" and the objects, events and states-of-affairs which populate it is entirely intuitive, "we refer routinely to all sorts of 'objects' that are not so simple to put our hands on" (2002: 300). Externalists, Jackendoff argues,

> "assert that we refer to 'objects in the world' as if this is completely self-evident. It is self-evident, if we think only of reference to middle-sized perceivable physical objects like tables and refrigerators. But as soon as we explore the full range of entities to which we actually refer, "the world" suddenly begins to be populated with all sorts of curious beasts whose ontological status is far less clear" (Jackendoff 2002: 303).

He spends several pages listing alleged problem cases: for instance, social entities (value, reputation, a PhD Degree, General Motors, the first dollar I ever earned), fictional and mythical characters (Sherlock Holmes, unicorns), geographical entities (Wyoming, the Mississippi river), auditorily perceived entities (the word 'banana', Mahler's Second Symphony) and virtual objects, whose ontological status he holds to be debatable at best[80]. Certainly, one could develop elaborate accounts of the semantic representation of each of these classes of entities, but according to Jackendoff's (2002: 303) version of the ontological argument, the result would be to radically distance "the notions of reference and 'the world' from direct intuition".

As far as I understand Jackendoff, he is calling for some kind of uniform, intuitive account of how concepts representing the types of ontologically troublesome objects above are acquired. Social and economic entities, fictional characters and the like are all *abstract* in the sense that they do not have a definite spatio-temporal realisation. But how can a proponent of e.g. informational semantics explain how we lock to the associated properties if we can never see or touch their instantiations? "What sense are we to make of the notion of

---

[80] Other theorists also take some of these cases to be troublesome for the externalist, and consequently suggest that mechanisms other than causal connections between the mind and the world underlie concepts linked to things like fictional entities. Rey (2005a; 2009a; 2011) suggests that internal role may have to account for the content of concepts of e.g. Sherlock Holmes, normative entities and what he calls Standard Linguistic Entities, such as phonemes and words. He stops well short, though, of endorsing Jackendoff's slippery slope argument against externalism (see Rey 2006), where the argument from ontology leads to the view that "it is necessary to thoroughly psychologize not just language, but also 'the world'" (Jackendoff 2002: 294). I will return to some of Rey's objections to externalism in the next chapter.

"grasping" an abstract object?" asks Jackendoff (2002: 298-299), following up his rhetorical question thus:

> "We know in principle how the mind "grasps" concrete objects; by constructing cognitive structures in response to inputs from the senses. This process has a physical instantiation: the sense organs respond to impinging light, vibration, pressure, and so forth by emitting nerve impulses that enter the brain. But an abstract object by definition has no physical manifestations that can impinge on the nervous system. So how does the nervous system "grasp" them? Without a careful exegesis of the term – which no one provides – we are ineluctably led toward a quasi-mystical interpretation of "grasping", a scientific dead end (Jackendoff 2002: 299)

Here, Jackendoff brings up an age-old philosophical problem, echoing critiques of the representation of abstract objects that have been around since Plato.

The problem of abstract objects comes in many different forms, but always takes as a starting point the idea of an abstract object as an entity which lacks spatio-temporal realisation (Liggins 2010: 67). Although there is no philosophical consensus on what exactly the class of abstract objects comprises (or whether it even makes sense to talk of abstract objects as a separate class from the class of concrete entities, see Rosen 2008; I will return to this issue in chapter 5, section 5.3.1) there are some objects that uncontroversially fit the definition: these include numbers and fictional characters.

There are two main problems for any account which assumes that there are entities which lack spatio-temporal realisation and that people can represent them. One problem is metaphysical, in that if it is claimed that there are entities which exist outside of space-time, there do not seem to be any obvious options about how they can exist. Where are abstracta located, if not in space-time? The second problem is epistemological and follows on from the metaphysical one. Since humans are inevitably located in space-time and abstract objects are seen as existing outside this realm, how can people come to have knowledge or beliefs about them (Benacerraf 1973)?

Even though "the contemporary philosophical debate over the puzzle has focused on the case of mathematical entities […] it is relevant to any philosophical theory which ascribes knowledge of abstract entities" (Liggins 2010: 67). The semantic theorist who holds that people can think and/or speak of properties with no spatio-temporally realized, or otherwise perceivable, instances therefore owes us an explanation of how concepts representing these properties are acquired. Swoyer (2008:27) points out that there are a few philosophers who have postulated "a cognitive faculty of intuition that provides some sort of non-causal access to numbers or other abstracta"; however, the nature of this intuition has never been explained,

"scientists have no inkling where it is located in the brain, and it has yet to turn up in any empirical studies" (*ibid*).

In fact, it is not even clear how a scientist interested in such a cognitive ability would proceed to investigate it: its operational logic would go against all our current knowledge about human biology. This follows from the fact that abstract entities "cannot affect our senses, our brains, or our instruments for measuring and detecting" (*ibid*), if they are seen as atemporal and non-spatial. Clearly, how big a problem for informational semantics this proves to be will depend on what one takes to be the metaphysics of the entities cited by Jackendoff as "ontologically curious". It may very well be, for instance, that quite a lot of the things he lists as abstract are encountered through some physical manifestation or another[81], and that these manifestations somehow provide (indirect) access to properties that people lock to.

But no matter what one thinks about the metaphysics of these entities, the theorist who relies on a purely causal connection between a mental item and the world to explain how semantic content is sustained has to tell some story about how acquaintance with things that cannot be seen underlies our concepts representing them. The aim of the rest of this chapter, then, is to take a closer look at Jackendoff's challenge and ask to what extent this version of the argument from ontology poses a problem for informational semantics.

4.2.3. Ontological and epistemological neutrality

As shown in chapter 2 (section 2.3.4), the informational semantic approach to mental content takes concepts to get their content from standing in a nomic relation to properties in the world. Concept possession is "constituted by there being the appropriate, meaning-making lawful relations between instantiated *doghood* and one's neural-cum-mental states" (Fodor 1998a: 76). Fodor emphasises, though, that "It's *that* your mental structures contrive to resonate to *doghood*, not *how* your mental structures contrive to resonate to *doghood*, that is constitutive of concept possession" (*ibid*). Fodor thereby holds questions about whatever mediates between the mind and the world to be independent from the semantic questions of which he is concerned. In one sense, then, Jackendoff's construal of the argument from ontology and his calls for an explanation of how concepts are formed do not apply to Fodor's particular kind of externalist semantics.

But, as pointed out in my brief review of this issue in chapter 2, the fact that questions about concept acquisition fall outside the scope of semantics if Fodor is right does not in any

[81] Cf. Prinz (2002: 148), who argues that "the failure to see how certain properties can be perceptually represented is almost always a failure of the imagination". See Dove (2009) for critical discussion.

way entail that an explanation of how concepts are formed would be uninteresting to conceptual atomism. Many scholars (such as Cowie 1999; Horsey 2006; Jylkä 2009; Landau 2000; Laurence and Margolis 1999b, Levine and Bickhard 1999) see the lack of a positive account of concept acquisition as a deficit to Fodor's proposed theory. Fodor, too, clearly thinks that an explanation of how semantic access is sustained should eventually be part of the theory.

Even though he prefers a "nomic-informational story about the metaphysics of broad content[82] to a causal-information one" (Fodor 1994: 54), it should hold true that "if there is a nomic connection between *doghood* and *cause-of-*DOG*-tokeninghood*, then there must be a causal process whose operation mediates and sustains this connection" (Fodor 1998a: 75). Somewhat more strongly, he says that "informational semantics is "untenable" unless there's an answer to questions like: 'how does (or would) the instantiation of *doghood* cause tokenings of DOG" (*ibid*).

I think Jackendoff is justified, then, in calling for an account of how the causal connection between concepts and entities of which it is not immediately obvious how they are instantiated. While it may be intuitively clear how we acquire concepts like DOG by way of using our perceptual mechanisms, the theory can then rely on the psychology of perception to come up with a story about how such concepts are formed (see Horsey 2006: chapter 5), it is presumably not as straightforward how the concepts VALUE and UNICORN acquire their content through a causal process.

But although Jackendoff's seemingly assumes the contrary, Fodor's version of semantic externalism does not necessarily need to rely on perception as being the only way to sustain semantic access. Even though, for Fodor, "perceptual mechanisms head the list of the ones that mediate our semantic access to *doghood*" there is an open-ended list of ways "*other than* perceiving dogs, that do, or might, sustain the meaning-making causal connection between dogs and their mental representation". This, Fodor highlights, is a point "that's important both for semantics and for epistemology" (1998a: 77).

It is not necessary to have directly encountered objects which instantiate the property of *being a dog* in order to acquire the concept DOG. One can acquire any such concept (or activate an existing one) without direct perception, relying instead on theoretical inference,

[82] I've taken the liberty of ignoring the distinction between *broad* and *narrow* content, which, even though it played an important role in much of Fodor's earlier writings about semantics, it's not an integral part of his newer work with which I'm concerned in this thesis. See Fodor (1994: chapter 2) and Cain (2002: chapter 6) for discussion.

technological extensions (radar, heat-detectors), non-technological proxies ("dog bells") or listening in on *mere talk*. Here, Fodor evokes the idea of *deference* introduced by Putnam (1975) and developed by Burge (1979), claiming that simply hearing about dogs from a peer may lead to the formation of a DOG concept locked to the property of *being a dog*.

According to the original proposal by Putnam, a person may rely on an expert to sustain the meaning-making connection between the concepts ELM or BEECH and actual elms and beeches. This reliance on the expert is what makes it possible to distinguish the two concepts – even though the person in question is not able to discriminate between the two types of trees upon encountering them. Burge suggests that the same, deferential mechanism may be at work with terms like 'arthritis', where a speaker depends on his linguistic community to fix the word's exact meaning. In line with this, Fodor holds that "relying on experts to mediate semantic access is a lot like relying on perception to mediate semantic access, except that the perceptions you are using belong to someone else" (Fodor 1998a: 78).

Like Burge, Fodor claims that there is no need for the mediator between the concept and the corresponding property to be an expert on the object in question, and (obeying the publicity requirement, see chapter 2, section 2.4.2) he maintains that there is no a priori constraint on who really counts as having a given concept. This preserves epistemic neutrality, since it leaves the door open for people to lock to properties with instances they have never encountered via talking to friends and neighbours, not just experts in the appropriate domain. This position has desirable consequences, since it ensures that the theory has no particular bias towards a specific way of locking to properties.

Fodor's liberal attitude towards the nature of semantic access ensures that he is not committed to the view that there have to be physically-instantiated counterparts to everything a person, call him A, can think of. If all A knows about a given entity is via the talk of B, this is enough for semantic access to be sustained. And the same holds for the next person in the causal chain. There is nothing in the account that stops person B, who A relies on as a mediator to the property a certain concept expresses, from in turn relying on person C to mediate semantic access. Moreover, person C is free to rely on person D, who in turn can rely on E, and so on, all the way until one reaches the end of the causal chain.

But what is at the end of this causal chain? It obviously cannot go on ad infinitum, since it is in the nature of causal chains that they necessarily have to end up somewhere. This somewhere will then be where the real epistemological and metaphysical problems of abstract objects arise. How did the person at the end of the causal chain encounter the object of his thought, if this is seen as existing outside space-time?

Fodor does not say, but it is worth noting that because of his view of properties, he is not committed to having any particular theory about the nature of abstract entities. Informational semantics, as well as being epistemically neutral, can be completely neutral as to the ontology of the objects of our thoughts. According to Fodor, our concepts DOG, WATER, A CONVINCING ARGUMENT and DOORKNOB are in some way locked to the properties *being a dog, being water, being a convincing argument* and *being a doorknob*. But these properties are constituted (inter alia) by the effect they have on minds like ours (Fodor 1998: 149).

This mind-dependence makes the many properties of "dubious ontological status" (Hampton 2000) *ontologically harmless* (Fodor 2000b: 352), so that someone who wants to accept Fodor's account on this point is not committed to any particular metaphysics of abstract objects. Whether something is *morally right*, for example, can be a fact of nature, a consequence of a religious credo or the result of an innate disposition, for all conceptual atomism cares.

The epistemological and metaphysical neutrality entailed by Fodor's version of externalism gives him a reply to Jackendoff's argument from ontology. Informational semantics is impartial, and has nothing to contribute to debates about the metaphysics of the objects of thoughts and utterances. The theory can account for the representation of a given entity, regardless of whether that entity has or could have actually existed in the world, and independently of whether someone acquiring such a representation has encountered instances of this entity or not.

Maintaining this neutrality makes it possible to kick the whole problem of abstract objects out of the domain of semantics. Although this makes perfect methodological sense (it surely should not be up to a theory of concepts to decide what is out there in the world) it fails to satisfy anyone interested in the actual process of how such ontologically curious entities as those mentioned by Jackendoff are represented, and what exactly underpins the acquisition of concepts corresponding to e.g. fictional characters or social entities. What I propose to do in the rest of this chapter, then, is to supplement the Fodorian account of concepts with an account of how people rely on communication with, and trust in, other people to acquire concepts corresponding to properties whose instances they have not encountered.

4.3. Communication, the acquisition of beliefs and concepts

4.3.1. Why care about the ontology?

Perhaps the first question to ask in developing the Fodorian account in order to better understand the representation of abstract objects is *why bother*? If the theory of semantic

representation provides the foundation for understanding how people acquire concepts expressing different types of entities regardless of their metaphysical status, why should one care about the ontology? If deciding what is out there in the world and how a person locks to a property does not bear on the semantics of the corresponding concept, it does not seem to be a task for the informational semantic theorist to give a theory of the epistemology and metaphysics of abstract objects.

But even though it may very well be that the issues raised by the so-called abstract entities is of no concern to a Fodorian semanticist, I think there are several good reasons for a theorist interested in cognition more generally to take up the argument here. Firstly, as it was argued above, Fodor's account of semantics would be more substantial and more convincing if it could appeal to "a detailed and plausible account of the psychology underlying concept acquisition" (Horsey 2006: 31). Also, as Horsey (following Laurence and Margolis 1999b) claims, without such a story of concept acquisition in place, it is not clear that the idea of concepts getting their content through locking to mind-dependent properties is any better than the radical nativist view attributed to the earlier Fodor (anno 1981a) according to which all concepts are innate (see section 2.3.3 of chapter 2 of this thesis).

Secondly, many of the things that can be thought of as abstract seem to play an incredibly important role in people's lives. Political, economic, religious and social entities exercise a profound influence over human life, in spite of (or perhaps because of) the fact that entities like democracy, inflation, piety and justice cannot be seen or touched. The question of exactly how many entities of this kind there are is of course a matter for empirical and metaphysical investigation. But, with Jackendoff (2002), I take it as fairly uncontroversial that the class of concepts without physically-instantiated counterparts in the world is not inconsiderable. It would therefore be desirable to know more about how people think with (and consequently act on) beliefs containing concepts which relate to political, economic, religious and normative entities, and whether these differ in their behaviour from perceptual concepts.

Thirdly, I think the sheer number and intellectual force of the people who cite the argument from ontology as a reason for excluding semantic externalism about (at least some types of) content from a naturalistic study of the human mind, makes it worth investigating how one might give a positive account of the acquisition of concepts that don't have a directly perceptual basis. The way I read both Chomsky's and Jackendoff's versions of the argument from ontology, they amount to a call for action, a demand that the externalist give at least a

sketch of how one can cash out intuitions about the mind-world relation in naturalistically respectable terms.

As far as I am aware, there are no complete treatments of the representation of abstract objects from an informational semantics perspective, and no systematic outline of how to address the problems they raise. There are some very interesting attempts in the literature to deal with a sub-set of the problem cases (especially Horsey 2006: chapter 4 and Edwards 2010a; I will return to a discussion of these in the next chapter), but a full-on defence of the theory and a thorough externalist explanation of how to deal with mental items corresponding to abstract entities is still lacking.

Part of this gap in the literature may stem from Fodor's reluctance to admit that critics of externalism have identified a real problem. As recently as in his (2008) book, he has claimed that "the crux of the problem of naturalizing reference is to provide a theory of perceptual representations" (Fodor 2008: 199). He thinks it is likely that

> "the paradigms of such might well be present-tense demonstrative thoughts (thinks: *That's a cat*). Having got a substantial number of cases of perceptual reference under control, the rest of the story might appeal to some or other sort of definite description to fix the reference of Mentalese terms that don't occur in present-tense perceptual thoughts (*the cause of that grinding sound*; *the guy I saw in the kitchen yesterday*; and so forth) (Fodor 2008: 199-200)

So in Fodor's view, an account of how people think about non-perceivable entities will follow close after the perceptual sciences have done their job.

But it is not clear that either an appeal to descriptions of the type Fodor cites or a more complete perceptual science will help address the problem of abstract objects in a satisfactory way. Even if definite descriptions may help with reference fixing in some cases, there are others which seem to be less good candidates for this treatment. What, for instance, is the content of the candidate descriptions fixing the reference of concepts such as DEMOCRACY, VALUE or LOVE? And if one could provide adequate descriptions locking people to the properties these concepts express, what are these descriptions cognitive status? Is one to assume that they are consciously entertained, Fodor's suggestion amounting to the claim that all concepts which do not have their content sustained via perception are learned via explicit instruction?

This would of course be a viable empirical claim, but I am not sure how well it holds up when faced with actual experimental data. When asking people what words in their vocabulary mean, one is likely to get wildly disparate or elliptical answers. Ordinary speakers seem particularly at a loss to explicate the meaning of their abstract object terms, as Wiemer-

Hastings and Xu (2005: 731) find in their study, where they note that "a few participants had extreme difficulty generating properties for abstract items". They think this "suggests that considerable parts of our conceptual knowledge of abstract entities may be difficult to express verbally" (*ibid*). But if the descriptions fixing semantic content are not consciously represented, this raises the question of how they arise and thereby aid the processes by which concepts lock to properties in the first place.

Fodor holds that in order to explain cases where semantic access is sustained via descriptions, all he needs "is something in the perceptual vocabulary that I can use to establish the locking. It might, for example, be a definite description of the referent of [an expression] E that is (as one says) rigid" (2008: 200). But without any constraints on what is required of such a description to get concepts successfully locked to properties, it is not clear how the theorist can make the step from a perceptually-based vocabulary to a story about concepts expressing non-perceivable entities. It seems, then, that questions about how semantic access to abstract objects is sustained remain unanswered, and a theory of how people acquire concepts like DEMOCRACY is still lacking. I will therefore go on to consider what I think are Fodor's more positive suggestions about alternative ways of sustaining semantic access, with the aim of filling in the gap left open by an eventual theory of how concepts for perceivable entities are acquired[83].

I will do so by trying to flesh out the idea of *deference* presented in Fodor (1998a), taking into account his claim that "phenomena like 'deference' to 'experts' […] belong neither to semantics nor to (cognitive) psychology, but to the pragmatics of linguistic communication" (Fodor 2008: 88, n57). The story I will tell therefore draws on some central insights from Relevance Theory and the Gricean foundation upon which it is built. In what follows, I aim to show how seeing communication as underpinned by "mindreading" provides a good basis for theorizing about content mediation via talk – thereby supplementing the Fodorian view of semantics with a more substantial theory of concept acquisition.

4.3.2. The role of mind-reading in communication

At the heart of Relevance Theory lies the Cognitive Principle of Relevance, which states that "Human cognition tends to be geared to the maximisation of relevance" (Sperber and Wilson

[83] To be fair to Fodor's suggestion: the ideas about locking to properties via theories which I will present in the last couple of sections in this chapter can be seen as variants on the theme of using definite descriptions to acquire concepts. What I am claiming, then, is not that Fodor's suggestion is untenable, only that more work needs to be done (in addition to completing the perceptual sciences) in order to understand the basic processes of abstract object representation.

1995: 260-66). All sorts of things, any input to the cognitive system, can potentially be relevant, as long as it connects with background information a given person has available to "yield conclusions that matter to him" (Wilson and Sperber 2004: 608). People can derive these worthwhile conclusions on the basis of sights and sounds in the environment, from looking at living things and inanimate objects, or by retrieving information from memory, and the search for relevance might be seen as one with we share with other creatures with minds, as well as our evolutionary ancestors.

But one potential source of information in our environment occupies a privileged position compared to other types of input, and that is *ostensive stimuli* produced by other humans. Sperber and Wilson suggest that instances of these (e.g. utterances) convey a presumption of their own optimal relevance (this is the Communicative Principle of Relevance). Someone coming across an utterance ostensively addressed to her will expect it, when interpreted in the intended way, to be worth the processing effort of understanding it and deriving (some of) its implications.

Thus, Sperber and Wilson (following Grice) argue that the hearer has to see the communicator as someone with *an intention* to communicate something. He has to exercise a capacity to infer "the speaker's intended meaning from evidence she has provided for this purpose" (Sperber and Wilson 2002: 3). In their framework, a speaker's meaning is a complex mental state involving both an informative and a communicative intention. Recognizing the speaker's informative intention, is a prerequisite for communication of this type to take place (see chapter 1, section 1.3.2). It follows that having a functioning metapsychological – or mindreading – capacity is a necessary condition for being a successful ostensive communicator.

Sperber and Wilson suggest that this capacity is part of what in the psychological literature is known as "Theory of Mind" (the term was introduced by Premack and Woodruff 1978): the ability to attribute mental states to others in order to explain and predict their behaviour. According to the relevance-theoretic view of cognition and communication, this capacity to represent the representations of others (or oneself), and to recognise their attitudes towards the representations they entertain, is an essential trait of human cognition and an evolutionary precursor to the emergence of human languages (Sperber 2000).

The m*etarepresentational capacity*, i.e. the ability to represent other representations (such as beliefs, desires, hopes, goals etc.), is one shared by all humans (setting aside pathologies) and employed routinely and automatically, without conscious effort. When the person sitting next to you reaches for a water bottle, you attribute to her the intention to pick it

up and the desire to drink from it, as well as the mental state of being thirsty (or bored, or nervous, depending on the context). When you see someone running past you as a bus approaches a nearby bus stop, you attribute to that person the intention of trying to catch the bus, as well as the future intention of hopping on the bus should it stop.

When you walk straight towards a stranger on a narrow sidewalk, you attribute to her the intention of wanting to pass you, rather than stopping right in front of you, or suddenly grabbing hold of you and giving you a hug. Based on this mind-reading capacity, you may predict what course of action she will take to best achieve her goal (take a step to the right and continue walking in the road, turn slightly sideways and let you pass etc.) and adjust your behaviour accordingly. Normally, none of this happens as a result of your going through a complex, explicit reasoning process. It just happens, in a fast, automatic and reliable way, and the claim in the huge literature on Theory of Mind is that the ability to make these intuitive inferences is what underlies a great deal of our interactions with others.

Though much more needs to be said about what exactly Theory of Mind is, who has it, what it consists of and so on, linguistic understanding is taken by many to involve the mindreading capacity in some form or other (see Sperber 2000; Sperber and Wilson 2002; for a recent review from a pragmatic and developmental perspective, see Chevallier 2009: 20ff). Sperber and Wilson trace the origin of this insight to Grice's work on inferential comprehension, and Wilson claims that it is relatively standard for researchers on pragmatics to adopt Grice's view that the comprehension process starts "from a metarepresentation of an *attributed utterance* and ends with a metarepresentation of an *attributed thought*" (Wilson 2000: 412, for a discussion of Grice's view of communication as based on the reading of intentions and the way it has been developed by Relevance Theory, see Allott 2008: chapter 1, and Wharton 2009: chapter 1).

Suppose that, in the course of a conversation, a hearer H starts from the following representation of an attributed utterance:

77. S is saying "I am home."

This, together with contextual information, may provide evidence for further thoughts of the following kind:

78. S believes that [S is home]
79. S intends H to believe that [S is home]
80. S intends H to believe that S intends H to believe that [S is home]

The essential link between utterance interpretation and mindreading can be seen in the fact that any utterance communicates not only a proposition, but also the speaker's attitude to that proposition. Thus, a declarative utterance such as 83 may be interpreted as asserting (i.e. intended to inform the hearer of the content of) the proposition expressed, as in 84 (Sperber and Wilson 1995: 247):

81. S: I'm home
82. H: S is asserting that [S is home]

But at other times, when faced with an interrogative or an imperative, for instance, the interpretations may involve other propositional attitudes, as in 85 and 87 :

83. S: Has the 9.05 bus passed yet?
84. H: S is wondering whether [the bus due to leave at 9.05AM has passed yet]
85. S: Please leave the room
86. H: S is requesting [H to leave the room]

Though there may be a canonical syntactic form used to communicate a corresponding attitude, an interrogative does not automatically communicate a request for information, or, more generally, a proposition embedded under an inquiring attitude (as evidenced by the existence of rhetorical questions). Nor is there any guarantee that declaratives are always used to assert, or inform the hearer of, the content of the proposition expressed:

87. S: You are home!
88. H: S is surprised that [H is home]

Here, the proposition that H is home would in most contexts be manifestly irrelevant to the hearer, and it is implausible that a speaker aiming to make her utterance relevant enough would have intended to inform H merely of this. What is important to note is that, in this and all other cases, the hearer will presume that U is intended to be relevant enough when interpreted in the intended way, and this will trigger a search for an interpretation that yields enough implications to make U worth his processing effort.

In this case, the search for relevance leads him beyond the interpretation on which S is merely asserting that H is at home, since this will lead to absolutely no cognitive effects (in normal circumstances[84]). According to Relevance Theory, H is therefore bound to go beyond

[84] This is of course not to say that there could not be a situation in which the *informing* attitude would be the most relevant. Imagine, for example, someone who has recently lost his eyesight and is taken back to his house from the hospital. Or one could think of cases where someone has been kidnapped and is dumped on the front porch of her apartment building after a ransom has been paid; she would then find it relevant to be told that she was home.

this minimal interpretation, and may instead construct one in which S intends to inform him that *S is aware,* or *surprised,* that [H is home]. This prediction follows from the relevance-theoretic comprehension procedure, which was outlined in chapter 1, section 1.3.1, states that a hearer should "a) Follow a path of least effort in computing cognitive effects: Test interpretive hypotheses in order of accessibility and b) Stop when your expectations of relevance are satisfied" (2004: 613).

4.3.3. The role of communication in acquiring beliefs

Now, depending on how much the hearer trusts the speaker (more on this topic later), she may use an utterance as a basis not only for attributing beliefs and other mental states to the speaker, but also for forming beliefs of her own. To illustrate this point, an utterance such as the following

89. S: I was stuck in traffic all morning

may lead the hearer to form the belief that

90. S is asserting that [S was stuck in traffic all morning]

If the hearer believes that S is speaking sincerely, this may in turn lead her to form the belief that

91. S believes [S was stuck in traffic all morning]

In most cases, this will be strong and reliable evidence for H to form the belief that

92. S was stuck in traffic all morning

Although some of our behaviour is based on beliefs formed as a consequence of direct perceptual access to the objects of the belief, we also rely on each other for information to a very significant degree.

Earlier today I was told by my girlfriend that

93. There is a great little café around the corner

This caused me to form the belief that the café around the corner from me was worth trying out. Consequently, when I fancied a coffee some hours later, I went around the corner and bought me a cup.

On the morning of November 5th 2008, I woke up and read on the internet that

94. Barack Obama wins the US presidential election

This caused me to form the belief that Barack Obama had won the US presidential election, which led me to form the further beliefs that Barack Obama would be the president of the US from January 2009, the US would have a Democratic president, John McCain had lost the 2008 presidential election, and so on.

Though the examples above may be mundane, the fact that I formed these beliefs is by no means trivial. In the first case, I had no direct perceptual evidence for the fact that there was good coffee to be found around the corner, and in the second, I had not overseen the counting of the votes. In both cases I used linguistic input produced by others to form beliefs about states of affairs I could not otherwise be sure were true. I had to use contextual information to recognise that someone was putting these utterances forward as statements of fact, and I had to trust the sources enough to accept these statements as true (or probably true).

Though "many – possibly most – human beliefs are grounded not in the perception of the things the beliefs are about, but in communication about these things" (Sperber 1996: 87), there is nothing automatic in the way beliefs about the beliefs of others lead to the formation of beliefs of one's own. Suspicions about the level of competence of one's fellow humans, or the compatibility of their preferences with one's own, or their motivations, or sincerity, or benevolence, may lead one to keep the beliefs they express embedded under the original propositional attitude rather than adopting them as beliefs of one's own.

If I thought my girlfriend had a terrible taste in coffee and only liked cafes where loud euro dance was played over the stereo, I would keep the belief expressed in 93 above embedded in the original attitude and represent it as:

95. S believes that [there's a great little café around the corner]

This embedded belief of hers, when combined with the contextual information about her taste, would lead me to form the belief that

96. There's a little café which probably serves bland coffee and plays loud music around the corner

Similarly, I may hear from someone who I know to be completely ignorant of sports that

97. Louis Armstrong won today's stage of the Tour de France

without automatically forming the belief that a dead jazz musician won a cycle race earlier today.

Sperber (1996; 1997) argues that it would make very little evolutionary sense for humans to have developed a capacity for mindreading if we then went on to accept as beliefs everything our conspecifics told us. He claims that as much as humans benefitted from developing the ability to attribute mental states, this would be worth nothing without some way of detecting incompetence and malevolence in others. Hence, the Theory of Mind needs to be supplemented by another capacity, which enables us to treat others as *fallible* sources of information.

In much of the evolutionary psychology literature, this ability is analysed as a "cheater-detection mechanism" (see e.g. Cosmides and Tooby 2005; 2010). However, in two recent articles, Sperber and his colleagues (Mascaro and Sperber 2009, Sperber et al 2010) have developed the idea that this capacity is linked to "a suite of cognitive mechanisms for epistemic vigilance" (Sperber et al 2010: 359) which develop gradually (ontogenetically), alongside the capacities for communication and Theory of Mind.

Though I will not go any further into the issue of trust and epistemic reservations, Sperber et al's suggestion that there are dedicated mechanisms for judging the trustworthiness of both the source of communicated information and its content provides a useful backdrop to the discussion of concept acquisition that follows. If they are right, this very ability may be what allows people to entertain beliefs about the mental states of others while at the same time keeping a critical distance from them.

For instance, if I keep hearing from people around me that

98. The Labour party is best fitted to run the British Government,

because of this I may form the belief that

99. Most people I know think that [the Labour Party should be in Government]

What Sperber et al suggest is that I may entertain this thought as an enduring part of my database of assumptions, and be fully aware that I do, yet still have the following belief:

100. The Liberal Democrats should govern the United Kingdom

Though the embedded belief in 99 is incompatible with the belief in 100, there is no contradiction in claiming that a certain individual may entertain both of them at the same time.

The reason for this is that they are beliefs of qualitatively different natures, stored in (functionally) different compartments of my mind. One forms part of a base of beliefs upon

which I am prepared to act, while the other is one I attribute to a group of others, but do not myself endorse (to use Sperber's vocabulary, it is not embedded in a "credal attitude"). Everybody has such metarepresentational beliefs, which they attribute to individuals they know, strangers, groups of friends, communities or even whole cultures. These may include folk sayings, "common sense" knowledge, generic statements and so on:

101. Scientists say [a glass of wine a day is good for the heart]
102. Political theorists claim that [communism doesn't work] (adapted from Sperber 1997)
103. Time Out New York says that [only taxi drivers and people from New Jersey drive in New York]
104. The French think that [the English are bad cooks]
105. Right-wing Republicans believe that [Barack Obama is enforcing communist policies in the US]

Any of these embedded beliefs may be accepted as true by a person who holds them metarepresentationally and trusts the source of the belief. If this is the case, Sperber claims, they will function in much the same way as the person's other unembedded beliefs, and contribute to her representation of the world. She will be inclined to act upon them and derive new beliefs on their basis, and they will play an important role in her cognitive economy as what Sperber calls *reflective beliefs*.

4.3.4. Reflective beliefs

Reflective beliefs are metarepresented beliefs where a "validating context" provides the thinker with justification for treating the embedded assumption as a true description of an actual state of affairs (Sperber 1997: 71). According to Sperber, "[t]here is an indefinite variety of possible validating contexts: reference to authority, to divine revelation, explicit argument or proof, etc" (1997: 72) and "they cause belief behaviours because, one way or another, the belief in which they are embedded validates them" (1996: 89). He contrasts these with *intuitive beliefs*, which are stored in a "data-base" and "freely used as premises in practical and epistemic inferences" (Sperber 1997: 68):

> Data-base beliefs are "intuitive" in the sense that, in order to hold them as beliefs, we need not reflect - or even be capable of reflecting - on the way we arrived at them or the specific justification we may have for holding them. The presence of a representation in our data-base causes us to treat it as data.

Though intuitive and reflective beliefs are both treated as true descriptions of the actual world, according to Sperber, they are of fundamentally different kinds, and therefore held apart in the cognitive architecture. He suggests that some of the constraints widely

thought to apply to an individual's inventory of intuitive beliefs - that they display mutual consistency, for instance - do not apply to the stock of reflective beliefs. Many reflective beliefs are only half-understood, or not even fully understandable, and "their content, because of its indeterminacy, cannot be sufficiently evidenced or argued for to warrant their rational acceptance" (Sperber 1996: 91). It follows that intuitive and reflective beliefs "achieve rationality in different ways" (*ibid*). While intuitive beliefs "owe their rationality to essentially innate, hence universal, perceptual and inferential mechanisms", reflective beliefs "are rationally held, not in virtue of their content, but in virtue of their source" (1996: 91-92).

Liberating the category of reflective beliefs from the many constraints traditionally placed on "genuine" beliefs in the epistemological literature lets them play a bigger role in explaining how people can think with and act on beliefs containing concepts which they do not fully grasp[85]. It also allows for an explanation of how people may act or speak inconsistently about the same thing on two different occasions. Take, for instance, the example I introduced above, of people around me claiming that

106. The Labour party is best fitted to run the British Government

Suppose now that I am completely ignorant of British politics, and furthermore that I am not even sure what "running a Government" means. Still, I trust the people around me to have better judgments about these matters than I do, and assume that they are uttering 106 with no intention of misleading me (I take them to be merely expressing their honest opinion). I then come to form the *reflective belief* that the Labour party is best fitted to run the British Government, embedded in the validating context *Most people I know think that*, as in 107

107. Most people I know think that [the Labour party is best fitted to run the British Government]

Trusting my friends' judgments, I express the embedded belief in 107 in conversation, and even act on it (I vote Labour in the next Parliamentary election, say), but it remains embedded in the validating context since I don't fully grasp the meaning of the words 'government' or even 'Labour Party'.

To take another example, consider a scenario in which I know absolutely nothing about matters of economy, and wonder why my bank account is so utterly devoid of cash after a recent trip abroad. A friend, Alex, explains to me that:

[85] It also has some important epistemological implications, some of which will be discussed in chapter 6. I will also return to the distinction between intuitive and reflective beliefs in chapter 5, section 5.3.3.

108. The value of the Norwegian currency is a lot lower now because of inflation

As a result of this interaction, I form the following metarepresentation:

109. Alex claims that [inflation has caused the Norwegian currency to lose its value]

Since I trust my friend and believe him to be well versed in finance, I (subconsciously) end up incorporating the content embedded in 109 into my repertoire of reflective beliefs. The fact that this is a belief of mine manifests itself in the way that I express it, and beliefs deriving from it, in conversation, without any particular attribution to my friend, as in the following utterance:

110. Inflation is making it really expensive for Norwegians to go abroad these days

Still, given that I do not fully understand the content of the original utterance, and I am not quite sure what 'inflation' actually means, my reliance on Alex will serve as the validating context for my belief – making it reflective rather than intuitive.

4.3.5. Reflective concepts

What I want to argue now is that the postulation of a category of reflective beliefs may help explain how people are able to acquire not only beliefs but also concepts via communication. Specifically, what I will claim is that it is precisely the capacity to entertain reflective beliefs, acting on them as if they were one's own, which underlies the way people acquire concepts denoting some items that are not directly perceivable. To see this, let's assume that in talking to my friend Alex, I came across the first utterance of the word 'inflation' I had ever heard. Still, even though the word was novel to me, I could make enough sense of it to use it as in utterance 110 above. How did I manage this?

The hypothesis put forward by Sperber is that upon encountering an utterance containing a novel word, I form a concept by opening up a new mental file, which will gradually become associated with encyclopaedic information, and index the newly formed concept with "the Mentalese equivalent of quotation marks"[86] (Horsey 2006: 148-149). Mental items thus indexed function as what Sperber calls *reflective concepts*, concepts whose content relies on an attribution to other individuals. In deferring to my friend in forming the

[86] The reference to quotation marks is not meant to be literal, since the Language of Thought/Mentalese version of quotation is seen as involving only an attribution of content, and not an attribution of linguistic form, as in natural language uses.

concept INFLATION, I rely on him to mediate to a corresponding property, Alex thereby sustaining my semantic access to *inflationhood*.

The fact that I can apply the word corresponding to the concept in slightly different contexts than the one in which I acquired it is linked to the fact that even though my knowledge of what *inflation* is remains minute, it is not completely non-existent. When opening up a new conceptual file, some bits of encyclopaedic information about the denotation of the concept invariably become associated with it, since new lexical items are never learned in a vacuum.

The hypothesis is, then, that in my first encounter with 'inflation', I open up a mental file marked «INFLATION» (indexing it with mental quotation marks), with something like the corresponding encyclopaedic information attached:

- «INFLATION» can cause a currency to be worth less
- The Norwegian currency has recently fallen due to «INFLATION»
- If it hadn't been for «INFLATION», I would have had more money in my bank account after my recent trip abroad

In using the lexical item 'inflation' (expressing my reflective concept «INFLATION») in a novel context, I rely on my friend to mediate to the content of the concept itself, but draw on the associated encyclopaedic information in order to communicate with it nonetheless[87]. As I hear more and more instances of the word in use, I will add new assumptions to my encyclopaedic entry, but the claim is that as long as I rely on other people to sustain semantic access, the concept itself will always remain reflective.

Once I have acquired the concept, and hence the ability to use a corresponding lexical item in conversation, I may spread the concept to other people who trust me as a reliable source of knowledge. Someone who hears me utter 110 ('Inflation is making it really expensive for Norwegians to go abroad these days') and thereby encounters the word 'inflation' for the first time, is likely to go through the same process as I did, and open up a conceptual file for the reflective concept INFLATION, with very limited encyclopaedic information attached to it, relying on me to mediate to the property the concept expresses.

It is the metapsychological, metarepresentational device that allows the meaning-making mechanism that mediates between the concept INFLATION and the property *inflation* to

[87] Note that this does not make the concept in any way dependent on, or constituted by, the encyclopaedic information attached to it. All that matters for the purposes of individuating content is the link to the individual through whom I encountered the corresponding word. The fact that the corresponding encyclopaedic entry allows me to communicate better is completely collateral to the Fodorian story about concept individuation, though it should be emphasised that it is still the beliefs, desires, hopes etc. that people have which are what a theorist turns to in order to explain or predict their behaviour. This is true even for the conceptual atomist (see Fodor 2008: 87; Margolis and Laurence 2003: 205; Sperber and Wilson 1995: chapter 2).

be sustained by proxy, and acquired and spread through conversation. The explanation I have given for INFLATION may be extended to other words where direct perceptual access to the property expressed by the corresponding concept is lacking, and I will return to this issue towards the end of this chapter.

Arguably, however, several of the most interesting questions about abstract entities and the way they are represented are left unanswered by this account. Most crucially, one would like to know what goes on at the end of the causal chain. Where deference stops, there has to be some kind of ability to lock directly to a property, and here the metaphysical and epistemological problem of abstract objects will arise again.

It may be noted that even if there were no available answer to this question, this would not be a problem for informational semantics, since the metaphysical puzzle of abstract objects (and the closely associated epistemological puzzle) is separable from the ontologically and epistemologically neutral theory of conceptual content. In the case of DOG, this physical input is (in most cases) provided by dogs in the immediate environment of the thinker, while in cases like INFLATION the physical input originates from some person (or newspaper or item on the TV or…) in the immediate environment *talking about* inflation. Whatever goes on beyond this step, what sustains the speaker's semantic access, is *inaccessible* to the hearer. For all she knows, the guy who just went on and on about the role of inflation in affecting the US export trade might have made the word up himself. So the theory is on safe ground in maintaining its ontological neutrality.

Nevertheless, without a solution to the original problem of abstract entities, we seem to be missing an important component in an account of how things which do not impinge on the human nervous system can lead to the formation of concepts. In what follows, I will therefore outline one possible view of what happens at the far end of the causal chain, what ultimately sustains our semantic access to such properties as democracy or inflation, and what this entails for the metaphysically neutral semantics I have advocated so far in this thesis.

4.3.6. At the end of the causal chain

I have so far argued that it is possible, using Fodor's informational semantics, to give an externalist account of the content of concepts denoting abstract entities. I have suggested a pragmatic framework that can contribute to understanding how these concepts are acquired in conversation and how the mechanism of semantic access is sustained across several individuals in a causal chain.

So far, nothing in the story I have told about concept acquisition relies on explicit learning of any sort. The metarepresentational story about reflective concepts thus explains nicely how people "just pick up" vocabulary from each other – without being told that "this word means such and such"[88]. I take this to be a positive feature of the theory, in that much of the vocabulary learned outside school or academic training is not traceable to any particular instance of explicit learning, and can only be accounted for by saying something about how repeated encounters with a word lead one to subconsciously "fill in" its meaning (Atran and Sperber 1991).

The account I have sketched above suggests that a person, on hearing a word for which she has no existing corresponding concept, and finding nothing in the immediate environment that might plausibly provide it with content, will form a reflective concept and attach to it encyclopaedic information taken from the context. The more contexts in which one encounters tokenings of the new word, the more information is added to the encyclopaedia, and the more information one has available to draw on in specifying the denotation of the concept. All this may happen without the person who acquires the concept ever having had direct perceptual contact with anything instantiating the property it expresses.

But this is not, of course, to say that new words with corresponding reflective concepts cannot be learned through good-old fashioned instruction. In fact, Sperber emphasises that quite often, reflective concepts are acquired through being "introduced by explicit theories which specify their meaning and the inferences that can be drawn on their basis" (1997: 79). This would be the case with many concepts encountered in scholarly writing, where the meaning of a term is specified by way of a description. A physics student may be introduced to the concept PROTON through hearing that a proton is a subatomic particle with an electric charge of +1 elementary charge, just as a student of economics may be explicitly informed about the meaning of the term 'inflation' by reading in a textbook that inflation is a rise in the general level of prices of goods and services in an economy over a period of time.

Though the account of how reflective concepts are acquired does not rely on explicit and conscious learning processes, nor does it rule out the possibility that such a process might underlie the acquisition of at least some concepts for some people. In fact, I now want to suggest that the use of explicit theories may be precisely what is responsible for generating concepts at the end of the causal chain, eventually mediating semantic access to e.g. *inflation*.

---

[88] It also avoids succumbing to the concept learning paradox which Fodor (1981a) claimed to show that all concepts have to be innate.

There is, as Fodor (1998a: 78) maintains, an open-ended list of ways in which semantic access can be sustained. One of these, which is particularly relevant to what goes on at the end of the causal chain, is *theoretical inference*. In my view, theory-construction (where the standard for something being a theory is not necessarily scientific) may very well be what ultimately mediates between concepts and properties in a range of cases where the properties have no directly perceivable instantiations.

Using INFLATION as a case study, the story about how it first came into being might be something like the following[89]: Once upon a time in the 19th Century, someone sat in a dark office in a bank in the Midwest of the United States and thought about the economic system of the nation. Paper currency, which had just been introduced by private banks, and so-called "Greenbacks", a type of government bond issued during the Civil War, could be exchanged for hard currency, thus serving as a proxy for real money (i.e. metal coins).

Our brilliant economist saw that the amount of money this paper currency would exchange for was not fixed. In a moment of extraordinary clear-headedness, he realised that dips and rises in value of the paper currency in circulation varied relative to the supply of gold and silver held by the banks and the state. Following the market over time, he noticed that this phenomenon was robustly systematic, and thought that there might be something in the economic system actually underlying the co-variation. For this thing he had now started to think about (which goes to show that he had formed a concept corresponding to the phenomenon), he coined the term "inflation" (*metaphorically extending* the word for the physical process of distending an object with air or gas). Once the economist started to use the word to describe certain states of affairs, his colleagues and friends were able to pick it up in the manner discussed earlier, and apply the concept underlying it in thoughts of their own.

But what was this entity in the economic system, and how did the economist get *locked* to the property the concept INFLATION resonates to? Well, whatever it was that the economist thought was responsible for the systematic variation; it was not *directly* perceivable and arose as an amalgam of economic, historical and social factors. Though each of these factors could perhaps be perceived as individual events or states-of-affairs, the complex whole made up of them was not accessible via other means than theoretical inference.

I take it that many of the things which fit the common-sense understanding of an abstract entity are like this, in that they stem from a combination of many different mental and physical events (*democracy*, *marriage*, *culture*, *French Culture*, the *Recession*, the *medal tally*

[89] I am basing the discussion fairly loosely on a review in Bryan (1997). The historical details of the story are not significant, as I am using it only by way of illustration.

of the Norwegian Winter Olympians are all plausible candidates). Sperber (2011) analyses entities of this type by appeal to a notion of *social/cultural cognitive causal chain*, and claims that a wide range of abstract phenomena can be characterised in terms of such causal chains.

Using the example of a cultural institution such as a folk tale, he argues that this can be characterised, on the one hand, in terms of the content of the tale, and on the other, in terms of its distribution in (relevantly similar) forms across a social cognitive chain of individuals. As Sperber puts it, "Add to it an extended distribution of a higher level representation with a normative content that prescribes, say, that this folktale is to be told on Christmas Eve", which indeed causes the telling of the tale on Christmas Eve, and one has "an elementary institution: a Christmas tale" (2011).

In Sperber's (2011) view, "More complex institutions, universities, churches, armies, markets for instance, involve the articulation of many more social cognitive causal chains with a much greater variety of changes in the environment, but the principle is the same". I believe that an account of this type can be extended to deal with the ontology of *inflation*. What an economic theorist studies when looking at markets is not a tangible phenomenon that can be accessed directly through one event or object that instantiates the property *inflationhood*. Rather, the concepts that express this property may be acquired as a result of the theorist generalising over a wide a wide variety of public representations, using his previous conceptual vocabulary to lock him to what he assumes is an actual, though non-perceivable property *inflationhood*.

4.3.7. Concepts for natural kinds

If my views on the ontology of inflation are correct, and if I am right in claiming that a lot of other so-called abstract entities have similar, complex ontologies, the story I have given of the genesis of the concept INFLATION could be extended to how other concepts people have formed for entities they have not seen or interacted with. Cases in point could be theoretical concepts, like FEMINISM or HEGEMONY, political ones, like DEMOCRACY or JUDICIARY, and cultural ones, like IDENTITY and PATRIOTISM. These may all be seen as concepts that were at one point introduced by scholars or politicians to capture complex political or social phenomena, before they were passed on and taken up in popular discourse[90].

The act of introducing and using concepts like these presupposes a relevant conceptual background, in that recognizing something as, e.g., *democracy* requires that the thinker

[90] Some theoretical concepts are, like Sperber (2011) claims, "borrowed and adapted from folk sociology", but still aim to capture entities with complex, intangible ontologies.

possesses a number of other concepts, such as GOVERNMENT, CONSENSUS, ELECTION etc. to build the theory from (Dove 2009: 418). The same goes for other theoretical concepts like FEMINISM, where recognizing something as instantiating *being feminism* (or *being feminist*[91]) presupposes a vocabulary consisting of e.g. POLITICAL RIGHTS, LEGAL RIGHTS, INDIVIDUALITY etc. that is previously individuated. If the relevant conceptual background is not in place (or the appropriate inferences between them not drawn) someone wishing to acquire FEMINISM will have to rely on others in so doing.

Note that the same goes for a range of entities that are not plausibly thought of as abstract, in that identifying something as instantiating e.g. *being proton, being an atom, being a cell, being a bacterium, being a cyst* etc. presupposes both prior theoretical competence, appropriate technological aids and the practical skills in order for the correlating concepts to be formed. Given that most people do not meet these requirements, the consequence is that a large majority of those who have the concept ATOM rely on deference to others to mediate to its content.

This way of seeing "the division of *conceptual* labour" has important implications for how other natural kind concepts, with instantiations people *do* perceive, such as WATER, STAR and even TIGER, as well. Fodor (1998a: chapter 7) distinguishes between locking to a natural kind via its superficial, phenomenological properties and locking to it *as a natural kind*. Locking to a natural kind "as such" depends on the link between one's concept and "the essence of the kind *not* to depend on its inessential property" (Fodor 1998a: 159). A given person thereby has two ways of acquiring a concept like WATER; either by locking to *water* via its appearance properties, or by getting "some expert to teach [her] a theory that expresses the essence of this kind" (Fodor 1998: 159).

Alternatively, if this particular person possesses the prerequisite mental vocabulary to construct a theory that will lock him to the essence of *water* (or *stars* or *tigers*) he may do the science himself (Fodor 1998a: 160). The two ways of locking are mechanistically and qualitatively different, and "support quite different counterfactuals [which] shows up (*inter alia*) in the notorious thought experiments about Twin-Earth" (Fodor 1998a: 157). Whereas a person locking to *being water* via its essence (by way of a theory or an expert) wouldn't think

[91] Surely, *being feminist* is concretely instantiated in lots of people (just like *being democratic* is instantiated in a number of countries). Why not claim, then, that the concept FEMINISM is derived on the basis of already possessing FEMINIST? I take it that recognizing someone as instantiating *being feminist* presupposes the prior individuation of assumptions in which the concept FEMINISM should form a part, just like knowing that two objects instantiate *being two* presupposes a prior conception of THE NUMBER TWO. If this is right, concepts like FEMINIST and DEMOCRATIC and so on are all necessarily secondary in the order of acquisition.

that XYZ is water "Homer [who locked to via its appearance properties] wouldn't have understood the question" (*ibid*).

Note too that on this way of construing the mind-world relation, anyone might be an expert on entities like *water*, or *star*, or even *inflation* or *feminism*, locking to these properties via theories. What one needs to do to acquire "a natural kind concept as a natural kind concept *ab initio* is (i) construct a true theory of the hidden essence of the kind; and (ii) convince yourself of the truth of the theory" (Fodor 1998a: 160). Who in the end possesses the cognitive and conceptual prerequisite for successfully locking to properties via theories will of course depend on the particular property in question. Acquiring PROTON by locking to *protonhood* via a theory will presumably be a tough task, in that it presupposes an already individuated inventory of concepts like ATOM, SUB-ATOMIC, ATOMIC NUCLEUS, ELECTRIC CHARGE etc. Other concepts acquired via theories, e.g. DEMOCRACY, will perhaps require less scientific sophistication.

It follows from this that mediating between people's concepts and properties like *being atom*, *being water* and *being inflation* may be many diverging causal chains, with different theorists serving as distinct end-points of each chain.

4.4. Conclusion

Fodor claims that the successful locking to essences of substances via theories depends on these theories *being true*. If you are a scientist concerned about finding the essence of a natural kinds, the eventuality of you being locked to the kind's essence via a theory amounts to your believing that "the theory locks you to such-and-suches via property that they have in every metaphysically possible world". The upshot of this theory construction, and your believing your own theory, is that "if the moon is blue, and everything goes as planned, you will end up with a full-blown natural kind concept; the concept of *such-and-suches as such*" (1998a: 160).

The state of the present natural sciences being what it is, the theories which lock scientists' concepts of WATER and ATOM (and thereby the concepts of those of us who rely on them in sustaining semantic access) are bound to be true and accurate, successfully getting them (and thereby us) locked to what I take to indeed be a natural kind property *being water*. But science has of course not always been what it is now. Many versions of the theory of what an *atom* is have been around in the scientific (and philosophical) community, but surely not all of them managed to lock people to the property of *being an atom*.

The question naturally arising out of this will be; what, if one agrees that there are true and false scientific theories, and while the true ones get people locked to actual properties, what is achieved by false or inaccurate theories? What does a misconstrued theory of *being an atom* mediate to? What content does ATOM have in the case where a false theory serves to sustain semantic access?

In the extension of these questions, one may further like to know how theoretical concepts where there plausibly is not any property for the theory to mediate to get their content. A scientist can, of course, not only be mistaken in devising a theory, but also in thinking about what is, has been or will be realized in the world. The history of science and philosophy have, of course, been full of examples of mistaken theoretical concepts, and many of the entities people think and talk about even today don't plausibly have a corresponding property for the concepts underlying their talking and thinking to express.

The case of mistaken theoretical concepts, such as PHLOGISTON (taken by the 17th Century scholar Johann Joachim Becher to be a fire-like element released by bodies during combustion), may be treated on par with other of what are called "mistaken concepts". These are entities like El Dorado and unicorn, which don't exist despite there at various points throughout history have been people who had theories and thoughts about these things. In the philosophical literature, the mistaken concepts are often grouped together with fictional entities, like Sherlock Holmes and Anna Karenina, or metaphysically impossible entities, like ghosts or elves, in that they all pose problems for a theory which relies first and foremost on externalism to explain semantic content.

Though I hold it to be plausible that many people acquire and have acquired concepts for these entities through talking to and relying on each other, what goes on at the end of these causal chains cannot be theory-construction in the sense of the word used to explain the genesis of INFLATION. There should be something that accounts for the difference between a concept that is hooked up to a natural kind like water and one that has no (simple or complex) counter-part in the world, like unicorn. Furthermore, something should explain the difference between concepts for entities that are believed by at least some to exist and those that everyone holds to be fiction.

In the next chapter I will use these observations to say something about the ontologically problematic category of fictional and mythical entities, and how the entities representing these are formed. I will use some of the conclusions drawn from the representation of inflation and build on the ideas of representation via deference and theories introduced in this chapter, with some crucial added clarifications and amendments. Even

though I will concede that there in many cases are differences between the representations of entities which exist and those that do not, I will aim to preserve the semantic account's ontological neutrality and thereby the methodological virtues of not committing semantics to any particular metaphysical thesis.

In this chapter, I have claimed that this methodological assumption serves to shield Fodor's version of informational semantics from arguments from ontology. I discussed two construals of these, both of which aimed to show that explaining the mediation of semantic content by the existence of a causal connection between the mind and the world is unfeasible for any theorist working within a naturalistic program. I held that both construals failed in providing decisive arguments against the ontologically neutral semantic externalist theory under scrutiny (Fodor 1998a; 2008), but suggested that the argument from ontology should serve as a prompt for the externalist to give a positive account of how non-perceptual entities are actually represented, acquired and employed in cognition.

I showed how informational semantics was compatible with a multiplicity of means of locking to properties, and used the case of inflation to highlight how semantic access could be fruitfully seen as sustained via deference. Employing the insights of cognitive pragmatics (Sperber and Wilson 1995), I showed how an understanding of the metarepresentational capacity underpinning human inferential communication could be applied in an analysis of how people acquire and spread concepts expressing entities they have not been in perceptual contact with.

Towards the end of the chapter, I argued that informational semantics could be used to explain what goes on at the end of a causal chain of deference, fleshing out the idea of locking to properties via theories. I held that, taken together, the elaborated idea of deference and the notion of locking via theory could explain quite a few of the cases where what is expressed by a given concept is seen as ontologically problematic.

What I will do in the next chapter, then, is to try to extend the theory proposed and defended in this chapter to representations that have nothing that is or could be realized in the world as counterparts. Accounting for the representation of these entities is taken, even by some externalists, so challenging as to suggest that non-externalist strategies are the best ways to deal with them. The purpose of the next chapter is to take issue with this claim, as well as coming up with a constructive account of the representation of other, ontological problematic entities.

5. Abstractness, "medium-abstractness" and fictional objects: Further investigations into the representation of non-perceptual entities

"You can't see it with your eyes, hold it in your hands
But like the wind that covers our land
Strong enough to rule the heart of every man, this thing called love"
- Jerry Reed

5.1. Introduction

In the previous chapter, I discussed the argument from ontology and its implications for the type of semantic externalism advocated by Fodor (1998a; 2008). Even though I claimed that informational semantics is neutral with respect to what is in the world and how people become acquainted with it, I suggested that the argument from ontology should encourage semantic externalists to provide a positive account of how various kinds of non-perceptual entities are represented and concepts corresponding to these entities formed.

Using what I saw as a paradigm example of an abstract object, *inflation*, I argued that pragmatic theory can help explain how communication mediates semantic in cases where there is no direct contact between an individual and what she thinks about. I also suggested that at the end of a causal chain of deference, the ability to lock to a property via theory can explain how semantic access is sustained.

What I now want to argue is that with this foundation in place, the representation of a range of other ontologically problematic entities may also be better understood. I will suggest that content mediation via deference or theory may provide some insights into the representation of so-called fictional entities, such as unicorns, ghosts and Sherlock Holmes, and of a class of entities I will refer to as *medium-abstract*.

This chapter has two main parts. In the first (section 5.2), I will discuss the case of fictional entities, taking as my starting point the discussions of these in Edwards (2010a), Horsey (2006) and Rey (2005a). I show how these authors, though sympathetic to Fodor's overall picture of concepts, regard the case of (at least some) fictional entities as so problematic for nomic-informational semantics that they conclude that the concepts supposed to pick them out are either *empty* or individuated *internalistically*. I critically evaluate their arguments before proposing my own informational semantic treatment of the problem cases. I argue that my account, which is intended to preserve ontological neutrality across the board,

is methodologically preferable to accounts which give distinct treatments to concepts for metaphysically possible entities and to those that are not.

In the second part of the chapter (section 5.3), I discuss another potential problem case for the theory of acquisition outlined in chapter 4. Returning to Chomsky's construal of the argument from ontology, I consider his claim that typically "words offer conflicting perspectives" (2000: 126), and use this to discuss a range of concepts whose contents are not plausibly seen as either wholly abstract or concrete. Here I will claim that concepts with contents such as HAPPINESS are unlikely to be acquired purely on the basis of perception (as concepts such as DOG presumably are), explicit instruction or theory construction. Medium abstract entities, which I take to include moral, emotion and so-called 'naive psychological' concepts, require a somewhat more nuanced treatment than what has been proposed so far.

The idea with this chapter is to show how the treatment of concepts for abstract entities can be extended to some cases that are not so obviously treated by the theoretical apparatus I put forward in the previous chapter. I will stick to the narrow-minded approach with which I have addressed and dealt with most issues so far, leaving potentially relevant literature on e.g. the metaphysics of fictional entities and the problems they pose for externalist theories (see Thomasson 2009 for a recent review of the philosophical literature on fictional entities) outside the scope of my discussion. I will instead keep focusing on the task of developing a positive account of the representation of non-perceivable entities, thus complementing informational semantics with a story of concept acquisition and the mechanisms that sustain semantic access to abstract objects.

5.2. Fictional entities

5.2.1. The problem of fictional entities

In attempting to extend the theory proposed in the previous chapter to a further range of ontologically problematic cases, it is immediately obvious that there are some obstacles to overcome. Even if it is granted that inflation is an abstract entity which can be represented by locking to the property *inflationhood* via a theory or by deference, and granted that perceptual contact underwrites the formation of concepts resonating to properties instantiated by a wide range of ordinary objects, there are still lots of cases left unaccounted for.

It is not safe to assume, for example, that what may be the case for inflation also applies to entities which no one regards as real. While people may agree that there is a property of *inflationhood* which is instantiated (albeit in a complex way) in the world, there are very few people who would hold such characters and places as El Dorado, or Sherlock

Holmes, or Zeus to actually exist. Also, all sorts of metaphysically or nomologically impossible entities, such as ghosts or unicorns, are deemed to remain creatures of the imagination, as there are no actual properties that they can be taken to instantiate. How, then, do the concepts associated with these terms get their content[92]?

To put the concerns of the last paragraph in a slightly different way: any attempt to extend the idea of a reflective concept to deal with the representation of fictional entities faces two main problems. Firstly, unlike democracy or inflation, unicorns are not real, and not generally regarded as real. Surely, this means that different types of concept are involved in the representation of these entities. Secondly, although it may be possible to claim that people acquire the concept INFLATION by locking to a property *inflationhood,* there are no instantiated properties such as *unicornhood*, and therefore no way to account for the content of the associated concept UNICORN in the first place. *Unicornhood, elfhood*, *ghosthood* and the like are metaphysically "defective" or nomologically impossible, and hence not the kind of thing that someone can nomologically lock to.

These are valid points, and I acknowledge that there are some important differences in the way that unicorns or Sherlock Holmes, on the one hand, and *inflation* or *democracy,* on the other, are represented. I want to address these concerns in turn, starting with the problem of impossible objects, which are often seen as wholly excluding an externalist semantics for fictional entities. Horsey (2006) and Edwards (2010a), who endorse informational semantics/externalism for concepts in general, claim that concepts of fictional entities are better analysed as having no externalist content at all. Similarly, Rey (2005a), argues that there is no way to account for the meaning of fictional concepts without appealing to the internal role they play in individual minds.

Against this, I want to argue that there is no need to postulate an internalist or empty semantics for fictional entities, since there is a plausible externalist solution suggested by Fodor's (1998a) ontologically and epistemologically neutral approach[93]. First, though, I will

---

[92] In my treatment of these entities, I ignore a couple of distinctions which some theorists may see as important, such as the one between fictional and mythical characters, and those entities which are picked out by proper nouns and those that are described by common nouns. Though I ideally would have liked to go deeper into the differences between these classes of entities, the account I propose does not discriminate between them and analyse concepts representing both Sherlock Holmes and unicorns in the same manner.

[93] I am not claiming that Fodor himself would endorse the position I take on the semantics of fictional entities in particular and abstract entities in general. As discussed in section 4.3.1, he seems to be fairly uninterested in the whole argument from ontology, simply expecting there to be "metaphysics enough for all our primitive thoughts, an instantiable property for every primitive predicate" (Rey 2005a: 79). What I suggest, though, is that Fodor's approach to semantics allows for the possibility of an externalist treatment, regardless of his actual views on the treatment of entities like ghost.

explain briefly why I do not endorse the solutions proposed by Edwards (2010a), Horsey (2006) and Rey (2005a).

5.2.2. Internalism about concepts for fictional entities

A nomologically or metaphysically impossible is one which cannot possibly be instantiated, and therefore cannot be causally connected (by law or otherwise) to minds like ours, at least not if one is aiming at a naturalistic account.

Horsey (2006) argues that if there are concepts of such nomologically impossible entities (he suggests that the concept GHOST is one), "informational semantics has no resources to account for the content of such concepts, other than by stipulating that they are non-atomic (that is, phrasal)" (2006: 169). Even though this solution may be the way to analyse impossible concepts like ROUND SQUARE, postulating a complex semantics for GHOST goes against the basic assumptions of conceptual atomism, according to Horsey. It will, moreover, force the atomist to "take the same approach for any concept of a nomologically impossible entity, of which there are plenty" (2006: 133). Treating all these concepts as exhibiting phrasal structure would thereby lead to the theory "losing its generality and appeal" (*ibid*). Thus, externalism comes up short in dealing with the representation of unrealisable entities.

Rey (2005a; 2005b; 2006; 2009a; 2011) too endorses the claim that externalism lacks the resources to account for concepts of fictional objects. He does not see "how we can hope to do the psychology of such empty thoughts, particularly the necessarily empty ones, without saying something about the role of these thoughts in a person's mind (2005a: 82). "[I]f you want to know about the nature of *ghosts*", Rey claims, "you look not to what people have been 'getting at' in the world, or even in any genuinely possible world, but to merely what they think" (*ibid*). He asks whether, "independent of our semantic desperation, there is any reason whatsoever to believe in such things [as *elfhood*, or *being divine*] and to laws and counterfactuals supporting them" (2005a: 79). Even though science assures us that there may be may be many instances of properties "that never happen to be instantiated in the actual history of the world, but which nonetheless need to be included as values of physical variables", "ghosts, elves and unicorns are surely not among them" (*ibid*).

In Rey's view, the case of fictional entities[94] shows that, *pace* externalists like Fodor, "some conceptual contents are individuated by some aspect of their role in our thought" (2005a: 87). As a result, Rey endorses a mixed view of content, where the concepts for some (e.g. fictional) entities are internalist, while others (e.g. those for more ontologically straightforward objects like cars and houses) are externalist.

Rey (2005a) is of course aware of the pitfalls of internalism and the arguments (many due to Fodor, see section 2.3.2 of my chapter 2 for a review) against appealing to conceptual roles in doing semantics. He admits to being "impressed by the sceptical challenges" internalist semantics face (2005a: 74), most notably the problem of deciding which roles individuate the content of a given concept. This task "may, indeed, be as difficult as isolating the principles of Chomskyan grammar" (2005a: 87), but it is motivated not only by the case of empty concepts, but also by logical concepts (NOT, ALL) and response-dependent concepts like FUNNY, according to Rey (2005a: 82).

Though I find Fodor's arguments against internalism convincing (even those against a conceptual role account of the logical vocabulary, see Fodor 2004), I will not rehearse the debate on the pros and cons of opting for internalism or externalism here. I will merely note a consideration someone starting out from a (Fodorian) externalist perspective may want to keep in mind in deciding whether or not to opt for a mixed account of content, when faced with the problem of empty concepts.

The point I have in mind is this: for someone like Rey, who wants to endorse a mixed view of semantics, concepts denoting fictional and impossible entities and those denoting real entities display a difference in *content types*. While the former have internalist content, i.e. meaning individuated in terms of internal roles, the latter has content in virtue of standing in a meaning-making relation to something in the world. But in what way does this difference manifest itself cognitively? Do internalist and externalist concepts behave differently in individual psychologies? I suggest that the answer to this is negative, and that such concession may lead the theorist down a slippery slope.

Intuitively, it does not seem to make much of a difference to people's concepts representing historical people, such as e.g. James I, Attila the Hun or Moses whether these characters actually existed or not, as long as those who think about them are unaware of the relevant facts. To get a clearer grip of this intuition, one could consider James I. In the actual

[94] Rey also holds that such things as geometrical and linguistic entities (e.g. morphemes) are problem cases for externalism, but I will not have space to discuss those in any detail, except for some brief remarks I make in footnote 108.

world, there is a lot of good evidence for the fact that James I was a real historical figure, but it is easy to imagine a nearby world where there never was such a person, despite him turning up in all the same historical records as we have in the actual world. In this nearby world the records allegedly showing him to be a British monarch in the 16th and 17th Century were falsified by scholars who, in an attempt to deceive their peers and readers, invented James I as a fictional character. The scholars' deceit has resulted in all the inhabitants in this nearby world falsely believing that James I was an actual, historical person. In this world people's concepts of James I have their content internally, if Rey's theory is assumed. In the actual world, the concepts people have of James I should be externalist, according to Rey's account. But would these concepts in any way differ in their cognitive behaviour from our world to the nearby one?

It seems to me that they would not[95]. The JAMES I thoughts of everyone in the nearby world will be the same as those entertained by people in the actual world, given that all other historical facts here are assumed to be identical between the two worlds. But the way internalist and externalist concepts can be expected to behave in the exact same manner raises the question of what justifies such a distinction. And if Rey takes internalism to be a live and viable option for the representation of fictional entities such JAMES I (in the world where the historical records about his reign are falsified), why cannot it to be taken to work for people's JAMES I concepts in the actual world? And if internalism can be made to work for actual JAMES I concepts, why cannot the treatment be extended to other concepts for historical people, and from there on to a range of other concepts?

In short, if there is no difference between the psychological roles concepts with different content types plays, is there any other way one can justify drawing the line between concepts that have internal from those having external content? Indeed, Rey (2005a; 2005b) appeals to independent metaphysical criteria in order to decide whether a concept has its

---

[95] Though these intuitions may waver, and depend heavily on "invented technical terms" such as *content* and psychology (Chomsky 2000: 153). Nevertheless, I think the James I type of cases make a substantial point about about the way an entity's existence/non-existence is sometimes irrelevant to the thoughts we have about that entity. This view is shared by Edwards (2010a: 103), who discusses a thought experiment in which we are to imagine a world in which George W. Bush does not exist. In this world, unbeknownst to everyone but a small group of insiders, a robot or a holographic image is responsible for all "Bush's" public appearances. Edwards asks whether the fact that people's conceptual representations of George W. Bush fail to refer in this scenario "make any difference to the role they play in our psychology?" Although he claims that in one sense the answer may be 'yes', "since there is a difference in the referential content of counterpart concepts across these two worlds […] in another sense, one that is perhaps more intuitive, the answer is 'no'" (*ibid*). He claims that "In the world in which the relevant concepts are empty, this fact isn't realized by any but a few in the inner circle. For everyone else, the relevant concepts play exactly the same role in thought that they do for their counterparts in world1" (*ibid*).

content determined externally or internally. He suggests that for anyone positing an externalist concept C, its content "better be specifiable independently of the state whose content it purports to provide" (2005b: 399). Rey believes there are good reasons to take objects like chairs and tables to exist, the concepts CHAIR and TABLE thereby being externalist, "since an optimal theory of the use of 'table' and 'chair' involves them applying to things that support certain sorts of objects, and those objects can (nearly enough) be identified with complex phenomena independently posited by physics" (2005b: 397).

But as Rey (2005b) also notes, there are lots of entities whose metaphysical status are not nearly as easy to determine. There are all sorts of objects which do not interact directly with our perceptual systems, and whose existence philosophers and empirical scientists are in dispute about. Think for instance about mathematical entities, such as numbers and sets, or folk philosophical notions such as consciousness and free will. And what about hypothetical entities in e.g. theoretical physics, such as the Higgs Boson, which existence is yet to be confirmed/disproved by data[96]? Should one hold the concepts that you and I have representing these entities have internal or external content? No one, it seems, is in a position to give a good answer to this question at the current stage of scientific development.

As I showed above, Rey claims that it is impossible to do the psychology of thoughts about fictional or otherwise non-existing entities without saying something about the role of these thoughts in a person's mind (2005a: 82). But this surely applies to all of the above concepts too. If someone wanted to know about the psychology of people's thoughts about James I, she would have to look at the inferences the JAMES I concept enters into, and what beliefs, hopes desires etc. in which it features. Similarly, investigating the Higgs Boson (if it were possible) would get the theorist nowhere near saying anything about people's psychologies and what role the concept THE HIGGS BOSON plays in cognition. I therefore think the metaphysical criterion is somewhat unsatisfactory in making the distinction between internal and external content, leaving the externalist with no case to be made for all sorts of concepts expressing theoretical entities to be given an informational semantic treatment.

[96] There are, of course, numerous concepts like these to be found not only in theoretical physics, mathematics and philosophy, but in all other scientific disciplines. Does my concept HEGEMONY have a referent? How about SIMULACRUM? And what about THE SINGULARITY, specified as a sequence which starts with the invention of a computer that is more intelligent than humans, which in turn will be able to create a machine that is more intelligent than itself etc., the sequence culminating in an ever greater intelligence (Chalmers 2010)? There are many theorists who believe that these concepts pick out actual or possible entities, and perhaps even more that think they do not, but the metaphysical criterion proposed by Rey is not very helpful in determining what contents such concepts have.

My fear of opting for a mixed view is, in sum, that by conceding the concepts for fictional entities, one has no available mechanism to stop the surge of conceptual contents which the internalist claims could just as well be specified by an appeal to conceptual role[97]. It would be methodologically preferable, then, to hang onto an analysis according to which all primitive concepts have externalist content. Doing so, I suggest, would also be in better tune with intuitions about there being no qualitative difference between the psychology of concepts for fictional from historical entities.

5.2.3. Concepts for fictional entities as 'empty'

Another option an externalist could pursue in dealing with fictional entities is simply to bite the bullet and claim that concepts like SHERLOCK HOLMES, which fail to pick out any actual individual, are devoid of content.

But as e.g. Scott (2003: 22) points out, this suggestion, on its own, "does not really help, since it makes all non-referring expressions refer to the same thing – namely, nothing". If the concept expressed by the name 'Sherlock Holmes' is EMPTY, and the same is true of names like 'Hannah Montana', or 'Anna Karenina' or 'Mighty Mouse', neither of which refers to anything, then these concepts should all fail to pick out anything, and therefore be truth-conditionally equivalent. If so, there is no thought SHERLOCK HOLMES SOLVES MYSTERIES IN 19TH CENTURY LONDON, only the thought […] SOLVES MYSTERIES IN 19TH CENTURY LONDON. The thought that is intuitively glossed as HANNAH MONTANA IS A SCHOOLGIRL WITH A SECRET IDENTITY AS A POP SINGER is merely the thought that […] IS A SCHOOLGIRL WITH A SECRET IDENTITY AS A POP SINGER, and so on.

But whether I think this last thought about Sherlock Holmes or Hannah Montana makes a difference to its truth value. The thought conveyed by an utterance of 'Sherlock Holmes is a schoolgirl' should come out false, while the thought conveyed by 'Hannah Montana is a schoolgirl' should come out true. Explaining the intuitive difference in truth value between these thoughts forces the externalist who regards fictional concepts as empty to choose between two options. She could either hold that, contrary to appearances, these thoughts, since they fail to pick out anything in the world, are not well-formed thoughts at all

[97] This, effectively, is the argument Segal (2000: chapter 2) makes for internalism about semantic content across the board. Horsey (2006), who concedes concepts for nomologically impossible entities to internalism, claims that he is in a better position to avoid the slippery slope, since he can make a separation between intuitive and reflective concepts, insisting that only reflective concepts can be "metaphysically defective". But depending on the eventual size of the possible class of reflective concepts, this potentially makes internalism a pretty large part of Horsey's semantic theory.

(this is the view Millikan 2000 opts for, according to Rey 2005a), or she could claim that the intuitive difference results from something other than their conceptual content.

According to Rey (2005a; see also Segal 2000: 33-39), however, the first solution makes an utter mystery of how these concepts come to play any role at all in the cognitive economy:

> "If apparent 'thoughts' with empty terms are not really thoughts at all, then how on earth are we to explain the often rational, *content sensitive* behavior of people (and their subsystems) that seem to have them? People pray, make sacrifices, and engage in often elaborate reasonings about gods, devils, elves, angels, ghosts, [and] empty thoughts interact in myriad *inferential* ways with non-empty ones (e.g. about churches, misdeeds, light, the sun), ways that surely require the empty ones to possess some *kind* of intentional content" (2005a: 81)

The conclusion seems to be that the theorist will be well advised to pursue other options for explaining the effects of empty concepts on thought. The solution developed by Edwards (2010a) involves appealing to the role a concept plays in the belief system to individuate it, while simultaneously trying to avoid the conceptual role semantics which Rey (2005a) invokes. His dialectical move here is subtle and elegant, and relies on a "distinction between what *constitutes* the reference relation [between a concept and a property] and the nuts-and-bolts of the causal processes that contingently *implement* (or *realize*, or *instantiate*) that relation" (Edwards 2010a: 101).

The key difference between Edwards' position and Rey's internalism is that while Rey takes the conceptual role of a fictional concept to constitute its content, Edwards separates the metaphysical content of a concept like SHERLOCK HOLMES (which he analyses as EMPTY) from the role the concept plays in thought. According to Edwards, fictional concepts only 'as-if-refer' to objects, where "the 'as-if' refers' locution is an indirect way to gloss relevant features of a concept's role; it is *not* a specification of content, or of conceptual structure" (2010a: 112). "In the case of concepts like UNICORN, where the concept holder's beliefs about unicorns "are sufficiently simple-minded", "it will be possible to characterize the relevant concepts' roles with the claim that the concepts have roles as-if they refer to horses with horns. For an agent with more complicated beliefs, we might need to add with lions' tails and/or with cloven hoofs, and so on" (Edwards 2010a: 104)

For Edwards, the conceptual atomist is allowed to make the distinction between actual and 'as-if' reference, while at the same time seeing conceptual role as fundamental to the behaviour of a concept, since "even a card-carrying referentialist about conceptual content, can and to my mind should, accept the importance of the relationship between a concept's

content and its role in mental processes" (Edwards 2010a: 100). Edwards claims that "even though the Concept Referentialist draws an important metaphysical distinction between a concept's (referential) content and its role, he/she has powerful reasons to maintain that a concept's role will, in some intuitive sense, reflect its content" (2010a: 101).

In Edwards' view, "The importance of [the relationship between a concept's content and its role] is something that Fodor himself tends to ignore, much to the detriment of his polemical position" (2010a: 100, n26). But although this may be true to some extent, the important referentialist point Fodor (1998a; 2004; 2008) repeatedly makes is that explaining the content of a concept is *prior* to saying something about its role in cognition[98]. Fodor holds that it is not possible to (non-circularly) specify the role a concept plays without appealing to an already individuated content (I take Edwards' use of 'reflect' in the quote above to suggest that he implicitly accepts this).

If Edwards wants to avoid succumbing to the circularity objection, he needs some way of explaining why an empty concept which 'as-if' refers to Sherlock Holmes differs in its cognitive role from another empty concept which 'as-if' refers to Hannah Montana. But what explains why a concept comes to have the role it does if not its content? No obvious candidates suggests itself, and without one it is not clear how Edwards solution gets the externalist any closer to solving the problem of fictional entities.

Edwards (2010a) tries to draw on the virtues of both referentialism and conceptual role semantics, avoiding the problems of each, thus having his cake and eating it. But it seems to me that no matter how one construes the relationship between the content of a concept and its role, the necessary primacy of content will force him to opt for either Rey (2005a)'s appeal to conceptual role semantics or Millikan's (2000) dismissal of thoughts with empty concepts as "not genuine thoughts". For the reasons already cited, I think that neither consequence is desirable. In what follows I will therefore propose my own externalist and conceptual atomist solution to the problem of fictional entities.

5.2.4. Semantic externalism and the representation of fictional entities

The claim I will make in this section is that what gives content to the concepts UNICORN, GHOST or SHERLOCK HOLMES are the respective properties *unicornhood*, *ghosthood* and *being*

[98] And Fodor does not, of course, claim that a concept's role is irrelevant to psychology. On the contrary, he points out that "on any reasonable story, you don't predict behavior (just) from the content of concepts (i.e. from their semantics…). You predict behavior from the galaxy of beliefs, desires, hopes despairs, whatever, in which the concepts are engaged. EVERYBODY agrees with that, referentialists very much included." (2008: 87, caps in original)

Sherlock Holmes. This claim immediately raises two (possibly related) objections, which I will address in order to flesh out my positive account.

The first objection is due to Rey (2005a; 2005b) and Horsey (2006) and was touched on above. The objection is that the properties mentioned are all *nomologically or metaphysically impossible,* and therefore not capable of providing the content of concepts, at least not in a naturalistic approach to cognition. Against accounts which rely on a profligacy of uninstantiated properties like *ghosthood*, Rey (2005a) enters "a certain methodological protest" and encourages a theorist to avoid introducing entities for which there is no independent evidence of their existence:

> Before we glibly resort to *elfhood* – or to *being phlogiston*, or *being divine* – and to laws and counterfactuals regarding them, we need to ask whether, independent of our semantic desperation, there is any reason whatsoever to believe in such things? I know of none. Our best theories of the world – biology, chemistry, physics, and cosmology – do not seriously make any room for such things or properties, or for laws or counterfactuals relating them (Rey 2005a: 79)[99].

But how terrible would it really be, from a naturalistic perspective, to postulate properties such as *elfhood*? Clearly, this depends on the theorist's metaphysical views. And as I have already emphasised several times in this thesis, Fodor (1998a) sees the properties that informational semantics relies on as *mind-dependent*. In other words, a property is (at least partly) constituted by the effect it has on minds like ours, something which is supposed to make it ontologically harmless (Fodor 2000b).

Fodor (1998a) claims that every primitive concept gets its content from standing in a lawful relation to a corresponding property. And if informational semantics is right, the implication is that "there must be laws about everything that we have [primitive] concepts for" (Fodor 1998a: 122). But if the properties corresponding to concepts like ELF are mind-dependent, the laws about *elfhood* are really laws about minds like ours. No particular view about the actual existence of elves is therefore entailed, and ontological neutrality is preserved.

I concede that to fully counter this objection I would ideally need to appeal to a substantial story about the metaphysics of mind-dependent properties. Having precious little to contribute with on that topic, I will merely settle for the contention that it is at least not a priori impossible that there are *psychological* laws about how people think about non-

[99] Edwards (2010a: 111-112) agrees with to this objection and quotes the same passage from Rey as an argument against the view that the problem of fictional entities could be solved by simply embracing the existence of the relevant objects.

perceivable entities. Even though I do not propose the existence of laws about supernatural beings as a solution to the problem of fictional entities, I think a case could be made for there being interesting, firm generalisations psychology could make about how creatures with minds like ours come up with complex thoughts and theories about such things as unicorns, ghosts and elves.

Arguing that mind-dependent such as *elfhood* provide the contents of concepts such as ELF does quickly lead to a second problem emerging, though. Would there not have to be something in the world that interacts with our minds if the story about a concept's locking to a property is even to get off the ground? Certainly, *doorknobhood* or *tablehood* can be said to be mind-dependent in the sense that whether something is a doorknob or a table depends on how our minds respond to certain objects. But in the case of *elfhood,* there is no actual object for our minds to interact with. Hence, there can be no law about how minds like ours resonate to *elfhood*, and no mind-dependent property of *elfhood*.

I believe this is much too strong a conclusion. As argued in the previous chapter, there are many entities people think and talk about, and many properties people take to be actual, although they are not instantiated in any unique physical form. I have argued that entities like inflation and democracy are like this, in that no single object, event or state of affairs instantiates *being democracy*. Rather, entities like democracy and inflation are constituted by an amalgam of many different psychologies and events, causal chains of private and public representations which cannot be accessed via direct perception.

The table in front of me instantiates *tablehood* and my neighbour's cat instantiates *cathood*. But there are no perceivable objects which are candidates for instantiating *inflationhood*. This is despite the fact that many of the causes and effects of *inflation* are perceivable and tangible. For instance, I take it that it is possible to experience the event of having a large amount of money in the bank reduce in value over a couple of days, or see that a plastic bag full of 50 billion dollar bills is needed to buy a loaf of bread[100]. These events are in a sense tangible, but they do not instantiate *inflation* in and of themselves.

Rather, in order to acquire the concept INFLATION one has to have some thoughts about the underlying causes and consequences of perceivable events, unlike what is required to possess a CAT concept, or a concept representing an event such as a tennis match. The encounter with a dollar bill with an exceptionally high face value does not provide anything

[100] As was the case during the period of hyperinflation in Zimbabwe in 2009. Similar consequences have been felt in a number of countries throughout the last 100 years, most famously in the German Weimar Republic and in Hungary after the Second World War (Salemi 2008).

sufficient for acquiring the concept INFLATION, any more than seeing people queuing at a polling station will suffice for acquiring DEMOCRACY.

I argued in the previous chapter that content mediation via talk (deference) or via theoretical inference were two plausible mechanisms by which people may acquire concepts of such things as *democracy* or *inflation*. I now want to argue, similarly, that people acquire concepts such as ELF and GHOST and SHERLOCK HOLMES by constructing theories about the properties *elfhood*, *ghosthood* and *being Sherlock Holmes,* or by relying on a causal chain of deference that ends with someone locked to these properties via a theory.

What I suggest is that the concept SHERLOCK HOLMES was originally formed when Sir Arthur Conan Doyle sat in a dark, damp room in London and thought up detective stories involving a character with such and such attributes (he had an extraordinary skill at deduction, he lived on Baker Street in London, he smoked the pipe regularly etc.). In so doing, Doyle constructed a complex theory of a person instantiating the property of *being Sherlock Holmes*, thereby locking to the property by way of theoretical inference. Conan Doyle's readers get locked to the same property by reading the detective stories in which he figures (or by watching the TV series, films, stage plays etc. based on them), thus relying on the author to sustain semantic access to that property via deference[101].

Similarly, there was at one point a story teller in Ancient Greece who sat and thought about animals and their attributes. He took individual traits from different animals, combining thoughts about the physical appearance of a horse, the strength of a bull, a horn with magical powers etc. into a theory of a mythical creature with specific abilities, thus locking to the property *unicornhood*. Having formed the UNICORN concept, the story-teller started thinking about the potential powers a creature instantiating the property of *unicornhood* would have, and the many events it could partake in and cause. He began telling stories about this mythical creature, as a result of which people formed reflective UNICORN concepts of their own, relying on the story teller for semantic access to the property their concepts expressed.

As time went on, more people started forming theories about unicorns, relying on different attributes that they took the creature to have. They relied on different experiences with different types of animals, and drew on these to form complex conceptual representations of unicorns. In turn, the stories they told were passed on to people who relied

[101] In discussing fictional proper names, I am sidestepping all the difficult philosophical questions about the semantic content of proper names for actual people, even though this may be an issue that is theoretically prior to that of characters in fiction. For a review of the literature on proper names, the problems they raise and a solution to these from a cognitive, relevance-theoretic perspective, see Powell (2010: chapter 3).

on the storytellers as mediators to the property, with the result that eventually lots of different concepts of *unicornhood* were formed in people of different cultures and creeds.

The ability to lock to properties via theory or talk can thus account for the genesis and spreading of both UNICORN and INFLATION type concepts. However, there is an obvious danger here: I may have been too liberal about both the notion of "theory" and what it takes to be a "property", and this may be seen as undermining the account. In what follows, then, I will try to account for the differences between the two types of concepts and the way they are entertained.

5.2.5. Descriptive vs. attributive concepts

Someone reading the story of the representation of fictional entities I have suggested in the previous section may object that I have ignored an important qualitative difference between concepts such as SHERLOCK HOLMES and INFLATION: a crucial aspect of the original idea of *deference* is lost when dealing with things that most people believe to be products of the imagination.

I may trust my economist friend to mediate accurately to the property *inflation*, and thereby endorse his claim that inflation actually exists, but I give no such endorsement in relying on Sir Arthur Conan Doyle to mediate to the property *being Sherlock Holmes*[102]. Similarly, though there may have been people in the past who believe in unicorns[103], contemporary possessors of a UNICORN concept are not disposed to think that at the end of the causal chain of deference there will be an actual expert on *unicorns*. It would be fair to claim, then, that there is a qualitative difference between the two types of concept which an adequate account should capture in some way.

Although I regard both UNICORN and INFLATION as reflective concepts, I want to analyse the difference between them by appeal to differences in the thinker's attitude towards the concepts. The reflective concepts expressing properties that a thinker believes have, have had or could have had instantiations in the actual world can figure in thoughts embedded

---

[102] Though people tend to attribute a sense of privileged "intellectual access" to the authors of some fictional characters, which, I guess, count as a form of expertise. Nicholas Allott has drawn my attention to Miguel de Cervantes's introduction to the second part of Don Quixote, where the author scolds another writer, who published a badly written sequel to the first Don Quixote under the pseudonym 'Alonso Fernández de Avellaneda'. Here, Cervantes tells his readers that "this Second Part of Don Quixote, which I now present you, is cut by the same hand, and of the same piece with the first. Here you have the Knight once more fitted out, and at last brought to his death, and fairly laid in his grave; that nobody may presume to raise any more stories of him"

[103] A bit of trivia: there are seven clear references to unicorns in the King James translation of the Old Testament, according to Shepard (1993: 41). In the majority of the subsequent translations, 'unicorn' has been replaced by 'wild ox'.

under a *credal* attitude (Sperber 1997). Those concepts expressing properties that are regarded as not possibly or plausibly instantiated can figure only in *attributive* thoughts: i.e. thoughts entertained as interpretations of beliefs attributed to someone else (Sperber and Wilson 1983: chapter 8, 1995: chapter 4).

Horsey (2006: 170) uses this distinction to analyse the word 'clairvoyant' (expressing the concept CLAIRVOYANT), which, when used by people who do not believe in supernatural psychic powers, can be glossed as something like "someone who claims that there are or is claimed to be clairvoyant". In Horsey's view, this use of the term does not entail any "commitment to the existence of actual clairvoyants, merely to the existence of individuals who are attributed with such abilities by themselves or others" (ibid). The difference between a truly deferential concept like INFLATION and a reflective concepts not capable of being embedded under a credal attitude is then that "deference to experts involves endorsing whatever the content of the expert's concept is. A reflective concept, however, may or may not involve an endorsement of the content", (Horsey 2006: 170) (as was originally pointed out by Sperber and Wilson 1983).

As well as explaining the difference between representations of actual and fictional entities, the dissociation between the content of a concept and the type of attitude in which it can be embedded may shed light on how people can think, talk and argue using a range of theoretical or religious concepts without endorsing their content. On the account I have proposed, it is entirely consistent for a philosopher to claim that "there are no propositions" or for a creationist to hold that "there was never a Big Bang", attributing the content of the concepts PROPOSITION and BIG BANG to people or groups of people who, unlike him, do believe in the existence of these entities. Similarly, an agnostic (or an atheist) may talk about God, or the divine, or the Holy Trinity, using the concepts GOD, THE DIVINE or THE HOLY TRINITY attributively, without being committed to any particular Christian metaphysics.

It is also important to note that, for the psychology of the holder of a given reflective concept, it will be irrelevant whether or not she is right about the actual existence of the entities in question. This holds both ways, in cases where entities are believed to exist despite the facts of the matter, and those where fictional characters turn out to be real. If it transpired that there really is no property *inflation*, and that the economists have been wrong all along in thinking that there the systematic variations in currency prices have a unique cause (as opposed, say, to several related causes), nothing about my concept or its role in my mental life would change as long I was unaware of the fact.

If it so happened that there really was a person with the exact attributes Sir Arthur Conan Doyle ascribed to his protagonist (imagine that the man supposed to be the inspiration behind Sherlock Holmes, Joseph Bell, turned out to be the exact model of the literary character), then the property *being Sherlock Holmes* would have been instantiated by an actual historical person[104]. The many readers of Sherlock Holmes detective stories who were unaware of this fact would still have the same reflective concept SHERLOCK HOLMES embedded under an attributive attitude, which would change to a credal attitude if the true facts came to light. Using religious concepts as an example, a consequence of this account would be that whether the theist or the atheist is right about the existence of God changes nothing about the way their conceptual representations of God behave in their mental life, if they were not aware of (or not convinced by) the metaphysical facts.

The way actual or potential existence of something corresponding to a concept makes no difference to its content nicely accommodates intuitions in the 'James I' cases above, where I held that nothing in the psychology of an individual changes if she is mistaken about the way the world is and ignorant of her mistake. There are also methodological advantages to this account, since it allows the semantic theorist to remain non-committal on all matters of actual existence, leaving it up to empirical sciences whether *ghosthood* is nomologically possible, or a person instantiating the property *being Sherlock Holmes* actually existed – as well as deciding on whether really there is something that instantiates the property *being God*.

5.2.6. Squaring mind-dependence with realism

Treating content-individuating properties like *elfhood*, *ghosthood*, *being Sherlock Holmes*, *inflationhood, being democracy, being feminism* etc. as mind-dependent means that, regardless of whether they are instantiated or not, these properties are constituted by the way we think about them. As Fodor admits, the price of holding a whole class of properties to be dependent on thinkers is "a touch of Wotan's problem. It turns out that much of what we find in the world is indeed 'only ourselves'. It turns out, in lots of cases, that we *make things be of a kind* by being disposed to take them to be of a kind" (1998a: 162).

If this is indeed the consequence of opting for an informational semantic approach to meaning[105], Fodor's externalism may go counter to the motivations which led many

[104] Ignoring caveats about personal identity, and conveniently disregarding the fact that being called 'Sherlock Holmes' may very well be taken as a necessary precondition for someone actually instantiating the property *being Sherlock Holmes*.

[105] It should not be taken for granted that the proliferation of mind-dependent properties necessarily follows from the central tenets of Fodor's informational semantics, and there may be other metaphysical theories that square

philosophers to favour traditional externalism. These include intuitions about the classic Twin-Earth arguments put forward by Putnam (1975), and a robust intuition shared by many philosophers that thoughts (and/or words) somehow *mirror* reality (Scott 2003: 31; Rey 2005a refers to this type of ontologically committed semantics as "strong externalism"). While I do not share this intuition, and think, with Chomsky (2000: 153ff), that Putnam's Twin Earth arguments are only persuasive relative to a theory-specific idea of content, it is worth pointing out that the view I have been advocating does not entail or assume metaphysical idealism.

Even if people, in their ordinary perceptually-mediated interactions with the world, can only represent entities in the way they appear to minds like theirs, there are other ways of locking to properties that are expressed by natural kind concepts. As Fodor puts it, we can describe "things that are alike in respect of the hidden sources of their causal powers, regardless of their likeness in respect of their effect on us", viz. we can describe "the world the way that God takes it to be" (1998a: 162). Theoretical inference is plausibly one such way of locking to properties that *are not* constituted by their effects on people's minds.

"We do science", Fodor says, "when we want to reveal the ways that things would be similar *even if we weren't there*. Idealists to the contrary notwithstanding, there's no paradox in this" (Fodor 1998a: 160). The more good science is done, the more people learn about what natural kinds like *water* are and the more they learn about the world. "Not philosophy, but science is the way to get Wotan out of his fly bottle", Fodor (1998a: 162) concludes[106].

Famously, the Ancient Greeks thought that the stars were holes in the heavenly canopy. If one accepts that there are two ways of locking to a natural kind property, either via a theory about the essence of a kind or its phenomenological properties (see chapter 4, section 4.3.6), this means that the Greeks had the concept STAR by standing in a relation to the appearance properties of stars. Most educated people today will have the same concept STAR[107], albeit they are locked to the essence of *starhood* via deference to astrophysicists who themselves have accurate theories about what a star *essentially is*[108].

---

just as well with what I've said so far in this thesis. Fodor offers such a view of ontology in a mere effort to look at "the geography that reveals itself if conceptual atomism is taken seriously" (1998a: 161).

[106] Wotan is a character from Wagner's *Ring Cycle* while "the fly-bottle" is a reference to a passage in *Philosophical Investigations* (309), where Wittgenstein claims that the aim of philosophy is "to show the fly the way out of the fly-bottle".

[107] Though a case could be made for the Greeks and the modern astronomers having STAR concepts with different contents. I think this is a line I am inclined to take, but do unfortunately not have the space or resources to argue for that here. Fodor (1998a: 157), responding to the question of whether or not Homer had the same concept of water that we do, claims that he does not "much care which you say as long as you like the general picture".

[108] The separation of externalist semantics from metaphysics also reveals how a Fodorian may deal with other necessarily uninstantiable properties, such as those expressed by "the familiar primitive concepts of Euclidean geometry that most high school students grasp, e.g. POINT, LINE, PLANE, CIRCLE" (Rey 2005a: 80). Even though

It is in principle possible for anyone with access to the right conceptual, theoretical and technological resources to lock to a property as it would be individuated by an omniscient being, thus finding out to far our actual use of concepts and thoughts about their objects match the way the world is. Given that science has achieved quite a lot over the last few thousand years, there should be a good few of people's thoughts which succeed in locking to properties via the essences of the entities that instantiate them. It is just that this is, as Fodor says, "a late and sophisticated achievement, historically, ontogenetically, and phylogenetically, and there is no reason to take it as a paradigm for concept possession at large" (1998a: 159).

5.3. The problem of *medium abstract* entities

5.3.1. The abstract-concrete distinction

In this chapter and the previous one, I have been concerned with entities that are sometimes taken to be problematic for the externalist, since they cannot be perceived and their representations therefore not acquirable on the basis of purely sensory mechanisms. In the philosophical and psychological literature, these entities are often labelled "abstract", on the ground that they have no spatio-temporal realisation. These entities can be contrasted compared with *concrete,* spatio-temporally realised entities, of which cars, refrigerators, dogs, buildings etc. are prototypical examples.

Concrete entities themselves are not as frequently brought up as counter-examples to externalist semantics, since it is assumed that these can be individuated via perception. It follows that externalism, given the right background view of how perception works, can tell some story about how concepts of concrete objects are formed via a causal process that systematically involves the sensory system.

Many writers have therefore implicitly relied on a presupposed distinction between abstract and concrete, holding that only entities which fall into the 'abstract' category are problematic for theories of language and/or knowledge which invoke a causal connection

---

we can conceive of a perfect triangle, "we could never be sensorily *presented* with any such thing – all perceptible points and lines have some thickness, and so no representation in our head could enter into causal relation with any such thing or property", according to Rey (*ibid*). If Rey is right, there can be no direct epistemic link between people's geometrical concepts and something in the world, but it does not follow that informational semantics is unable to account for such concepts as TRIANGLE and CIRCLE. It only means that people, when learning these concepts, do so either by deferring to books and teachers, or learn the actual geometry behind them, thus locking to e.g. *being a perfect circle* via theory.

between the mind and the world in explaining thoughts and beliefs[109]. So Rosen (2008) claims that

> "The abstract/concrete distinction matters because abstract objects as a class appear to present certain general problems in epistemology and the philosophy of language. It is supposed to be unclear how we come by our knowledge of abstract objects in a sense in which it is not unclear how we come by our knowledge of concrete objects (Benacerraf 1973). It is supposed to be unclear how we manage to refer determinately to abstract entities in a sense in which it is not unclear how we manage to refer determinately to other things (Benacerraf 1973; Hodes 1984)."

Rosen goes on to note that even though there is "a great deal of agreement about how to classify certain paradigm cases", the "challenge remains, however, to say what underlies this alleged dichotomy". He adds that "We may know how to classify things as abstract or concrete by appeal to 'intuition'. But unless we know what makes for abstractness and concreteness, we cannot know what (if anything) hangs on the classification".

Drawing on Lewis (1986), Rosen discusses several ways in which a theorist can try to explicate the distinction, but concludes that they are all unsatisfactory in one way or another. Using the traditional criterion of spatio-temporal realisation to determine whether an entity is abstract leads to a clash with intuitions when it comes to such entities as *the game of chess*. According to Rosen (2008), chess would come out as concrete according to this criterion, since it is generally assumed that it was "invented at a certain place and time (though it may be hard to say exactly where or when); that before it was invented it did not exist at all; that it was imported from India into Persia in the 7th century; that it has changed in various respects over the years, and so on".

Similarly, the psychologists Wiemer-Hastings and Xu agree that using "physicality as the distinguishing factor" between abstract and concrete is unsatisfactory. In particular, this criterion cannot

> "account for graded differences in concreteness. For example, most people perceive *scientist* to be more abstract than *milk bottle*, but both are perceivable physical entities. Likewise, most people perceive *notion* as more abstract than *ambiance*, but neither is a perceivable physical entity" (Wiemer-Hastings and Xu 2005: 720)

[109] In the philosophical literature, the problem of abstract entities is often construed as a metaphysical puzzle about existence (see Balaguer 2009; Burgess and Rosen 1997; Dorr 2008; Swoyer 2008; Yablo 2002) and its consequences for mathematical theory and mathematical theorizing (see Balaguer 1998; the collections of essays in Benacerraf and Putnam 1983 for the classical readings, and Irvine 2009 and Bueno and Linnebo 2009 for some recent views).

It seems that intuitions about physical realisation are hard to capture and convert into anything substantiating a clear-cut division between concrete and abstract. As Rosen (2008) points out, it then becomes a challenge for the externalist to decide which concepts are plausibly seen as formed on the basis of perceptual mechanisms and which are not. For the particular semantic externalist account I have advocated over the course of the last two chapters, the wavering intuitions about abstractness/concreteness make it hard to see to whether the solutions I have proposed for such things as *inflationhood* and *elfhood* generalise to other properties.

If one takes one's intuition as a starting point, there are many entities which may be best characterised as falling within the "abstract" group, in that they are not spatially or temporally realised, yet which somehow do not "feel" as abstract as inflation or ghost. Take for instance such things as love and happiness, or normative concepts like JUSTICE and MORALLY RIGHT. Certainly, it seems reasonable to claim that many people rely on others to mediate to *inflationhood*, but who mediates between you and *being love*? Whose theory do you rely on for your semantic access to moral entities? How do people lock to properties such as *being just*, *being a norm* or *being reciprocity*?

It is quite hard to see how a causal-historical account of the genesis and spreading of concepts like LOVE and JUSTICE would go. It seems to me wildly implausible that once upon a time there was an expert on matters of the heart who sat in a candle-lit room and hypothesised about a non-tangible property of love, locking to it via a theory that his contemporaries and their descendants rely on to mediate to the content of their LOVE concepts. It would be equally implausible to claim that the concept activated in the kindergarten kid who is denied the same amount of candy as his peers gets its content via a long causal chain ending up with some renaissance judicial scholar and his theory of justice.

Unlike democracy and inflation, entities like love, happiness and right/wrong are universal and universally important in all cultures. While democracy is a fairly recent invention, and inflation only arises in monetary economies, some system of norms is present in every culture, and all humans display a capacity for emotional response, barring severe pathologies.

However, these concepts are not really concrete either, given that love cannot be seen, touched or kicked around the room. It is impossible to decide whether something is morally right using only one's perceptual capacities, at least not in the same way as the perceptual capacity can give rise the thought HERE IS A DOG. I know what it is like to pat or play with a dog, and can recognize the look, scent and taste of entities like coffee without much effort.

But how does one recognize and point out justice or happiness? Surely not in the same manner as one can recognize an English setter.

In this section, I will address the problem of what I will refer to as *medium-abstract* entities: entities which, even though they are not spatially or temporally realised, are not abstract "in the same way" as the paradigm cases already discussed. I address this problem because I think that it raises important issues for semantic externalism, and that an account of semantic representation should do justice to the intuition that such things as *love* do not belong squarely in the class of either concrete or abstract objects.

In what follows, then, I will propose a solution to this dilemma, exploring the middle ground between representations of things that are paradigmatically concrete and those that are paradigmatically abstract. What I will suggest is that underlying the words people use to talk about such things as *love*, there are several concepts, some of which are plausibly seen as acquired via theory or deference, while others have an innate or perceptual basis. To explain the relationship between the single lexical items and the plurality of concepts that underlie them on my account, I will appeal to the mapping principles between words and concepts outlined in chapter 3 of this thesis.

5.3.2. Revisiting the argument from ontology

It may be somewhat surprising that it should be so hard to partition the world into abstract and concrete entities, especially if "abstract" is defined negatively as including everything that lacks a spatio-temporal realisation. From this definition it should follow logically that something is either abstract or concrete, leaving no room for a middle ground.

Though there may be many different reasons why it is so hard to make such a division, one particularly interesting suggestion is due to Chomsky (2000: 126). He claims that "Quite typically, words offer conflicting perspectives". For instance, an entity like a city can be both concrete and abstract: "London could be destroyed and rebuilt, perhaps after millennia, still being London". This shows that not only the physical attributes of the city, but also "The abstract character of London is crucial to its individuation", according to Chomsky (*ibid*).

People, too, can be thought of as both abstract and concrete in a similar way. For instance, "Tom Jones, though perfectly concrete, could be reincarnated as an insect or turned by a witch into a frog, awaiting the princess's kiss, but [still be] Tom Jones all along" (*ibid*). And even though there may be objects that people think and speak of as unambiguously concrete, "Proceeding beyond the simplest examples, intricacies mount" (Chomsky 2000: 127). Chomsky takes this point to apply even to perfectly ordinary objects like houses and

doors. “I can paint the door to the kitchen brown, so it’s plainly concrete”, but at the same time “I can walk through the door to the kitchen, switching figure and ground” (*ibid*).

As Chomsky puts it, houses “are concrete but, from another point of view, are considered quite abstractly, though abstractly in very different ways; similarly books, decks of cards, cities etc.” (2000: 36). “Books are concrete objects. We can refer to them as such”, he continues (2000. 20), but at the same time, it is perfectly possible to see them as abstract, as in utterances like “he wrote the book in his head, but then forgot about it”. In an utterance like “that deck of cards, which is missing a Queen, is too worn to use”, the deck of cards is “simultaneously taken to be a defective set and a strange sort of scattered ‘concrete object’, surely not a mereological sum” (Chomsky 2000: 21).

In the previous chapter, in discussing Chomsky’s view of ontology and its consequences for semantic externalism, I suggested that many of the cases he raises become less puzzling if one makes a theoretical separation between the concepts encoded by words and the concepts those words are used to communicate, as Wilson and Carston (2007) do. Better still, if one completely drops the isomorphism between words and concepts, a view I argued for in chapter 3, one is in a good position to account for the conflicting perspectives Chomsky takes words to offer.

In chapters 3 and 4, sections 3.2.2 and 4.2.1, I argued that informational semantics is neutral in principle on the issue of how words and concepts map onto each other. The relationship between lexical items and mental concepts may be one to one, one to a few or one to many, a matter to be determined empirically on a word by word, speaker by speaker basis. Extending the argument, I would now like to claim that there is nothing in the account I have outlined which entails that a given word invariably provides access to the same type of concepts, representing the same kind of entities in the same way. Underlying each lexical item there may be several concepts, each of which is locked to a different property or to the same property via distinct mechanisms of semantic access.

Taking Chomsky’s example of London, one may imagine that someone who has visited London and also seen it on a map has acquired a concept of London via direct perception[110]. Furthermore, one could imagine that this person, call him Boris, has read a lot about the politics and governance of London and therefore has beliefs about the relative financial and political autonomy of the city’s administration and its mayor. These beliefs contain a concept which is potentially activated by the lexical item ‘London’ (I’ll gloss this

[110] Lots of unanswered questions on the nature of *direct perception* are lurking here, but since the metaphysics perceptual processes is outside the scope of this thesis, I will not pursue the issue.

concept as Γ [Gamma]), but is formally distinct from and independent of the concept which he has acquired perceptually (glossed Δ [Delta])[111].

Using the vocabulary introduced above, one could say that Boris has a *reflective* concept, potentially activated by the lexical item 'London', which represents a particular property of the policies of running the city, and which he has acquired by having a theory of the political governance of London or relying on textbooks in deference to other people who have such a theory. Similarly, the perceptually acquired concept expressing the property *being London* can also be activated by hearing the lexical item 'London'. Which of the two concepts gets selected on a given occasion is determined by the relevance-theoretic comprehension heuristic and their relative degrees of activation: the reflective concept is most likely to be selected if the context involves thoughts about political systems, elections or the like, while the perceptually given one is most likely to be selected in the context of a conversation about architecture and nice places to travel.

Sometimes, as suggested in section 4.2.1, two concepts may be activated by the same word token, the interpretation of the following utterance being a case in point:

111. London is beautiful but terribly mismanaged

Here, both the "political-abstract" and the "physical" aspect of the city are relevant, with the result that two different sentences are formed in the Language of Thought, containing two different concepts activated by the same word. If Boris hears a friend, let's call him Ken, uttering 111, he is likely to end up with the following representations as a consequence:

112. KEN BELIEVES THAT [Γ IS BEAUTIFUL]
113. KEN BELIEVES THAT [Δ IS POLITICALLY MISMANAGED]

Of course, Chomsky's examples of *polysemy* are complicated, and it is not given that they can all be subject to the same treatment. It may be, for instance, that while some concepts underlying the same lexical item have distinct contents, other concepts, though formally discrete, are locked to the same property. Falkum (2010) has suggested that this may be the case for a range of cases of what she calls *systematic* polysemy, such as 'book' in 114 and 'window' in 115:

114. Roth's new book has a clever plot and an eye-catching cover.
115. Mary opened the window and crawled through it.

[111] I have glossed it this way to highlight the fact that these concepts are formally distinct, even though they are both potentially activated by the same lexical item (see my chapter 3, section 3.3.4 for discussion).

What accounts for the way two different aspects of *book* (its content and its appearance) and *window* (its physical object and aperture senses) are highlighted by one single word tokening of each sentence, are not differences in conceptual content but in perspective-taking, according to Falkum (2010: 220). She holds that

> On this highly suggestive approach, the meaning variation observed for window (and many of the other book/window cases) is not treated as an instance of ad hoc concept construction. Instead, window is treated as having a constant referential meaning across these different contexts, with some of its encyclopaedic information receiving extra activation in each case, resulting in the perception of the object being viewed from a specific 'perspective'.

One may imagine, then, that even though the concepts I have glossed above as Γ and Δ are stored as separate mental items, with their own files of encyclopaedic information attached, they may have the same stable content. What accounts for the difference is that something encoded in the concepts' *Modes of Presentations* provides cues or triggers towards what part of the corresponding property's instances are highlighted in a given context.

I will return to discuss what may lead to co-referential concepts having distinct MoPs in chapter 6 (section 6.2), but merely note for now that the account of word-concept mapping I have suggested leaves these issues open. The important point for present purposes is just that seeing the relationship between words and concepts as non-isomorphic allows for a treatment of Chomsky's interesting examples which does not rely on a word's having to denote only a single type of entity. Even if a concept activated by a word denotes a concrete object in one utterance, nothing stands in the way of the same word's activating a concept denoting something more "abstract" if used in another utterance.

The fact that a word can activate a number of different concepts of distinct types is what I think explains how words like 'love' and 'justice' may strike one as semi-abstract. Some of the concepts these lexical items activate are plausibly acquired on the basis of explicit reflection or instruction, possibly derived from highly-culture specific beliefs. They will then be reflective concepts. Other conceptual correlates of 'love' will have another basis, and I will follow Sperber (1996; 1997) in calling these *intuitive concepts*.

The challenge remains, though, to explain what exactly this "basis" for the formation of intuitive concepts is. As I argued above, it seems implausible to claim that the entity love can be perceived by the sensory system in the same way as an exemplar of dog can be. One cannot pat and groom love as one can a beagle, and nor does love smell, sound or look like anything. There has to be something else underlying the formation of intuitive concepts for love, justice, moral rightness.

My goal in the next sections is therefore to take a closer look at the notion of *intuitive concepts* proposed by Sperber, and to use this to give an account of how representations for moral, emotional and social concepts are formed.

5.3.3. Intuitive concepts

In a series of articles (Sperber 1985: chapter 2; 1996: chapter 3 and 4; 1997), Dan Sperber has argued that humans have two types of belief, which are "similar in some behavioural and epistemological respects, but different in cognitive organization and role" (1997: 67). On the one hand there are reflective beliefs, derived through testimony and "believed in virtue of second-order beliefs about them" (1996: 89). As outlined in the previous chapter, reflective beliefs are seen as embedded in a validating context; their content need not always be fully understood or understandable, and they are warranted by trust in the source of the belief.

On the other hand, there are *intuitive beliefs*, which are "typically the product of spontaneous and unconscious perceptual and inferential processes" (*ibid*). While reflective beliefs may contain vocabulary (*reflective concepts*) corresponding to non-perceivable objects and are locked to properties via deference or explicit, theoretical reasoning, "the mental vocabulary of intuitive beliefs is probably limited to that of *basic concepts*: that is, concepts referring to perceptually identifiable phenomena and innately pre-formed, unanalysed abstract concepts (of, say, norm, cause, substance, species, function, number, or truth)" (Sperber 1996: 89).

Intuitive concepts can be tentatively characterised as those acquired via perceptual access to objects directly instantiating the corresponding properties. According to Sperber's proposal, then, "our perceptual mechanisms assign basic-level concepts to sensory stimuli. The perceptual concepts can thus be broadly identified with the basic-level concepts studied extensively by Rosch [1978; Rosch et al 1976] and others", as Horsey (2006: 155-156) points out.

Sperber suggests, furthermore, that "spontaneous inferential processes derive intuitive beliefs from perceptual beliefs and from other inferentially derived intuitive beliefs" (1997: 78). It is not entirely clear in what way one can cash out the idea of "spontaneous inferential processes", but one possibility is following Horsey (2006) who, using the relevance theoretic idea of a concept's *logical entry* (see my chapter 2, footnote 34), suggests that meaning postulates contained in these entries for perceptual concepts give rise to intuitive "inferential concepts". Horsey (2006: section 4.3). Horsey's (2006: 157) proposal is that there are "A number of different conceptual domains [which] licence spontaneous inferences about the

entities relevant to the domain". These domains, plausibly seen "as encapsulated modules and sub-modules" (*ibid*), give rise to "spontaneous inferences that are governed by the meaning postulates and procedures attached to concepts" (2006: 159).

In somewhat the same vein, Mercier and Sperber suggest that *intuitive inferences* are "the direct output of all inferential modules [that] take place without attention to reasons for accepting these inferences" (Mercier and Sperber 2009: 165). Which concepts will be possible outputs of these spontaneous inferential processes, and thereby qualify as intuitive, will clearly depend on claims about whether and what modules the human mind contains[112], and which of these are intuitive. I will not engage any further with the details of this issue here, but note that the category of intuitive concepts is not limited to not only the purely perceptual ones, but may expanded to include representations derived *on the basis* of the perceptual vocabulary.

In addition to these types of processes, there may be other formats which give rise to intuitive concepts, according to Sperber. Going back to his (1996: 89) quote above, one sees that Sperber includes in his specification of the *intuitive* category "innately pre-formed, unanalysed abstract concepts (of, say, norm, cause, substance, species, function, number, or truth)". This is interesting, since it opens up the possibility that the human mind has access to a stock of mental concepts not derived via perception or inference from perceptual concepts.

It is, of course, a wholly empirical question what should be included in this category, and what precisely the requirements are for something to be *innately pre-formed* (see this chapter, section 5.4). But if any good candidates for universally present human concepts that are not acquirable via perception should be found, it appears that Sperber would group these in the category of intuitive concepts, along with those formed on the basis of perception.

5.3.4. Understanding 'love' and 'happiness'

With the introduction of two different classes of concepts, being intuitive and reflective respectively, which get their content via different epistemic capacities, I am now in a position to make some more concrete claims about the representation of what I have called "medium abstract" entities.

---

[112] The issue of whether the cognitive processes are "modular" is a highly debated one in the recent cognitive science literature. See Sperber (1996: chapter 6; 2001; 2005), Mercier and Sperber (2009), Barrett and Kurzban, (2006), Carruthers (2006) and Tooby & Cosmides (1992) for arguments in favour of massive modularity. For critical views, see See Fodor (2000a) and Weiskopf (2010a). The original arguments for modularity in Fodor (1983) were meant to apply to only the input systems, Fodor ruling out any appeal to modules to explain "the central thought system" because he sees thought to be much too flexible and context-sensitive to be accounted for by domain-specific, informationally encapsulated cognitive units.

To see how the introduction of a category of intuitive concepts may help with the analysis of such entities, however, it might be helpful to consider some data. The ordinary use of a term like 'love' seems to pick out a variety of types of love, depending on the situation in which the word is used. This is of course not unique to the notion of love: I argued in chapter 2 and 3, for instance, that an ordinary lexical item like 'tired' could activate a range of distinct mental items expressing different properties of *tiredness*. Moreover, each of these could lead to different contextual implications on different occasions, even if they were used in the same sentence. A similar one-to-many relation between words and concepts might also exist for other terms such as 'book', 'house' and proper names like 'London', as suggested above.

A corpus search for the noun 'love' reveals a similar type of variation as to what particular type of entity is being talked about. The following are meant to illustrate this heterogeneity (data extracted and adapted from the British National Corpus (examples 116-119), and the Oslo Multilingual Corpus (examples 120-121, references to source texts are in parenthesis)[113]:

116. Together in the cottage, she felt like exploding with love for him (K8R 1395)
117. It was just that she loved Eve as much as any mother could love a daughter (CCM 1071)
118. God in no way needs the world and He did not need to make it: creation is the expression of His pure outgoing love (A5P 34)
119. Second-hand shop discoveries can be turned into prized possessions with a little love and attention (C8A 1627)
120. Indeed, the Liberal Party was, after Marie-Louise and one other, the great love of his life (RDA1)
121. The odious feel of rough khaki on the backs of his knees and of his neck also inspired in Hartmann at this time his love of luxury (AB1)

The emotional state attributed to the protagonist in 116 is clearly different from the one in the other examples. There is no suggestion of explosiveness in the type of *love* involved in fidgeting with second-hand clothing, as in 119. And even though someone's love for an idea or a cause might have a similar intensity to the love for one's spouse (as indicated by 120), there is a certain romantic aspect missing from the *love* one may have for a political party.

This data, together with all sorts of other real-world examples, suggest that there are different aspects of *love* at play in the conversations and thoughts people have about love[114]. There is romantic love, a feeling which may come in different types depending on age, the

[113] All results were downloaded on April 12, 2010. The BNC "is a 100 million word collection of samples of written and spoken language from a wide range of sources" and is freely available for search on http://www.natcorp.ox.ac.uk/. See footnote 47 for details on the OMC.
[114] I acknowledge that it is an open question to what extent this should be mirrored in the conceptual content one ascribes to speakers/thinkers, cf. the discussion in section 5.3.2 of this chapter.

time it has persisted, culture, whether or not it is acknowledged by the other, and all sorts of other circumstances. There is parental love, which may be of different types depending on whether it involves the father or the mother, and which should be qualitatively distinct from the love of children for their parents[115]. Related to this may be a category of emotions best described as affectionate *love*, of the type one may feel for friends or non-parental family relations (or even pets).

Similarly, there is love for inanimate material things, such as artworks or souvenirs, and for immaterial things, like ideas, ideals and causes, where the emotional state being described may be different from case to case. Interestingly, there are also very culture-specific ideas of *love*, as be expressed in examples like 118. The Christian Bible's idea of God's love for mankind shares many aspects with other kinds of love, but also differs from them in important respects from them. Other notions of *love*, such as those involved in man's love for God and humans' love for each other, are also expressed in the Christian Bible.

What I am claiming is that some or all of these may involve distinct concepts (be they stable, or formed *ad hoc*). Some of them will be reflective, such as the one expressing God's love for mankind, which is supported by a theory based on beliefs specific to Christianity, or via deference to the Bible and the church (or both), while others will be intuitive.

The intuitive concepts – I take the concept of *mother's love* to be a good candidate – may be acquired in one of three ways: 1) via direct perceptual access to the object, 2) via an intuitive inference mechanism using the perceptual concepts as input, or 3) developed independently of perception on the basis of an innate format. All three manners of locking raise interesting possibilities for the analysis of concepts for medium abstract entities, and it is plausible that intuitive concepts representing *love* may be acquired either via perception or via an innate format.

There may be a plausible story about how a conceptual representation of *love* is formed by perceiving physiological and behavioural traits in oneself and others. It should in principle be possible to lock to a property via one's own emotional response (externalism about semantic content, I take it, has no problem with the external entity being located inside one's own body): for instance, having one's heart racing, or feeling satisfaction and comfort in the presence of a partner, or being sweaty and nervous, are all perceivable events or states.

---

[115] Bartels and Zeki (2004: 1162) present an interesting study showing that brain activations associated with romantic vs. maternal love "involve a unique and overlapping set of areas, as well as areas that are specific to each". For what it is worth, this might be seen as supporting the lexical-pragmatic claim that there are different concepts at play in the representation of the different forms of *love*.

Alternatively, an account on which some concepts of LOVE are derived on the basis of an intuitive inferential capacity or another innate format should be equally empirically tractable. Frank (1988), among others, has argued that the capacity for romantic love may be an evolutionary adaptation that enhances fitness for couples and their offspring, and could therefore be plausibly selected as part of human biology.

The same should be true of parental love, and it is not hard to come up with a story on which the offspring of creatures that instinctively develop parental love towards them are more likely to survive until reproductive age than those who do not receive such feelings from their parents. Some kind of parental affection would indeed seem to be highly advantageous evolutionarily, since it would give even the most egoistic caregivers a powerful incentive to ensure the health of their children. For human offspring, who have a prolonged period of helplessness compared to other mammals, this incentive would have to be comparatively bigger and therefore merit more powerful innate constraints.

Similarly, concepts linked to other emotion words may be the output of any one, two or all three of the potential intuitive concept construction mechanisms. As originally proposed by William James (1884) and Carl Lange (2010 [orig. 1885]) certain perceivable, physiological responses reliably co-occur with many human (and non-human animal) emotions: for instance, increased breathing frequency and tensed muscles (Prinz 2004). As pointed out by Damasio (1994; 2000), the same neurological mechanisms linked to emotions are also involved in the perception and regulation of the body, and this may lend support to a view of (many) emotions as cognitively represented by a set of perceptually derived emotion concepts (see Prinz 2004 for a possible account of what aspects of the perceptual input give these concepts their content; see Zinck and Newen 2008: 5 for a critique of Prinz's account).

Alternatively, one could claim that emotion concepts are directly or inferentially derived via an innate format. According to a dominant view in psychology, all humans have an innately determined set of so-called *basic emotions*, which in the original proposal were seen as including happiness, sadness, fear, surprise, anger, and disgust (Ekman et al 1969). Which emotions are seen as belonging to this set will depend on one's theory, and Ekman (1999) includes amusement, contempt, contentment, embarrassment, excitement, guilt, pride in achievement, relief, satisfaction, sensory pleasure and shame as part of the list. Regardless of what the set of basic emotions contains, a plausible case could be made for a set of corresponding concepts being somehow derived from these.

The grounds for Ekman's innateness claims (and those of other supporters of a basic emotional repertoire) are empirical findings providing evidence for a universal set of common

emotions. There is psychological and anthropological evidence that in all cultures and human groups, people display and are able to recognize anger, happiness, fear, and so on. However, it might be premature to regard emotion concepts as a unitary, universal output of either perceptual mechanisms or innate formats, since the same type of conceptual heterogeneity illustrated above for *love* is likely to be found with emotion concepts as well.

Even though there may indeed be a set of basic emotions lexicalized in all languages, they are invariably lexicalized and categorized differently. So an emotion such as *anger* may be expressible by one word in Language 1, two in Language 2, and three in Language 3, whereas in Language 4 the closest synonym may pick out a broader range of emotions than any of the ones lexicalized in languages 1-3 (see Wierzbicka 1999: chapter 1.6 for a systematic discussion of empirical data).

Also, just as with different types of *love*, the basic emotions come in a variety of forms and facets, since "not all instances of an emotion referred to by the same word (e.g., ''anger'') look alike, feel alike, or have the same neurophysiological signature" (Barrett et al 2009: 430)[116]. Barrett et al rightly point out that a range of emotions described by the same word may lead to radically different behaviour and physiological responses: for instance, the anger you feel when someone cuts you off in traffic will be qualitatively different from the anger you experience when a disobedient child breaks a rule, or when you hear the voice of a disliked politician. It would indeed be surprising if the vocabulary of English, or of all languages, perfectly captured the basic emotions as natural kinds, when there are such great variations in 1) the emotions described by these words and 2) the way in which they are categorized across languages.

Claiming that there is a heterogeneous set of distinct concepts correlating with emotion words exempts the theorist from having to make a forced choice among mutually exclusive accounts of how the emotions picked out by a single lexical item are mentally represented, and on what basis these representations are derived. Each emotion term will be capable of expressing a number of different concepts, some of which get their content from different, yet direct, perceptual processes (and will therefore be intuitive concepts), while others will acquire their content indirectly (and will therefore be reflective concepts). Just because people use a single term to describe something ('love'), it does not follow that there is a single mental concept which gets its content from a unique underlying entity.

---

[116] One should also expect to find a similar pattern with other words which have multiple conceptual correlates, like 'tired', 'book' and even 'dog'.

The externalist who is not bound by a one-to-one isomorphism between words and concepts is therefore in a position to explain all sorts of empirical findings without succumbing to what Barrett (2006b: 29, using the vocabulary of Gould 1977 and Lewontin 2000) refers to as "an error of arbitrary aggregation", i.e. the grouping of "emotional processing into categories that do not necessarily reveal the causal structure of the emotional processing".

This is particularly important when thinking about cases such as 'happiness', one of the six emotions in Ekman's original, basic repertoire. All humans (barring pathologies) experience *happiness* from time to time: it emerges early in development (by around the 3-month mark according to most studies; see Lewis 2008) and it is instantly recognizable in oneself and others across cultures (people are faster to recognize instances of *happiness* than of most other "basic" emotions, such as anger; Calvo and Marrero 2008; Calvo et al 2008; Damjanovic et al 2010). HAPPINESS, then, is a likely candidate for being an innately determined intuitive concept.

At the same time, however, there is a dissociation between *being happy* and *happiness*, in that the implications of the noun 'happiness' "far exceed those of the adjective" (Wierzbicka 1999: 53). Among the implications of *happiness*, some are highly culture-specific, and whether "happiness is a highly aroused state like joy or elation" or "contentment, tranquillity or peace of mind", or is best identified with engrossing engagement in an activity or rather with an "equanimity of spirit that even misfortune cannot disturb" (Averill and More 2000: 663) is likely to vary not only across cultures and times, but also from person to person.

Culture- and individual-specific notions of *happiness*, such as those which link it inextricably to romantic love, or to marriage, or to having children, or achieving material success (having lots of money, or owning expensive cars), or immaterial success (celebrity status, or a fancy job title) or to religious worship, or to specific experiences (travelling the world), may be represented by concepts introduced by explicit theories or through deference to others. Thus, some concepts correlating with the word 'happiness' will be plausibly analysed as *reflective* rather than intuitive.

If I am right, what accounts for intuition that 'happiness' is *medium abstract* is, on the one hand, the heterogeneity of concepts representing different types of *being happiness* properties, and on the other, the different possible ways of locking to these properties.

5.3.5. Moral and normative concepts

Another type of mental items which I took to be representative of the class of medium abstract concepts are those expressible by so-called *moral* or *normative* words. These are concepts that figure in moral judgments and evaluations, and correspond to words like 'justice', 'fairness', 'reciprocity' and so on. Though moral entities are highly abstract, in the sense that they cannot be perceived, the claim I will make in this section is that there are good reasons to hold that not all moral concepts are acquired on the basis of deference, learning or theory-construction.

There has been much discussion of whether moral judgments constitute a natural kind and if so what characterises them. Though I do not propose to venture into this debate, in what follows I will hint at some theoretical strategies an externalist might use to show how some moral concepts get their content, and at the same time explain what gives them the feel of being "medium abstract".

In traditional accounts, moral evaluation and judgment always emerge as outcomes of a process of explicit reasoning based on learned principles. In early work in developmental psychology, this reasoning capacity was assumed (e.g. by Piaget 1932; Kohlberg 1984) to develop in stages and gradually increase in sophistication during child and adolescent development. If this is right, on the model I have proposed, all the concepts used in a person's moral reasoning would be classified as reflective. If a child somehow had to be explicitly taught a set of underlying principles in order to think about moral dilemmas, moral concepts would have to be acquired in much the same way as INFLATION or DEMOCRACY.

Though I do not doubt that many of the concepts underlying moral judgments and evaluations can be plausibly seen as reflective (in the sense that they are culture-specific, acquired via teaching and/or sustained by proxy), there seems to be increasing evidence that not all the capacities underlying moral reasoning result from explicit learning. There are two main types of data which argue against the traditional picture: 1) evidence of widespread dissociation between moral judgment and moral reasoning, and 2) striking cross-cultural similarities in certain moral domains and in patterns of response to moral dilemmas[117].

[117] There is also a third line of experimental evidence emerging from pre- and post-linguistic developmental studies (Hamlin et al. in preparation; Bloom 2010 discusses some preliminary findings from Hamlin et al.'s work. See also Baumard et al. under revision). Some of the research in this area is indeed highly promising, but may also be problematic in that it is not really clear how to identify and classify moral actions non-linguistically. The data emerging from studies of speaking children is probably more informative, but here too there are uncertainties about what is really being investigating in experiments on moral decisions in children, since not all cognitive capacities related to moral decisions may develop at the same time. The underdevelopment of one or more components relevant to moral decisions, if not properly taken into account, may contaminate the data.

Empirical evidence of dissociation between judgment and reasoning in the moral domain was first systematically explored by Jonathan Haidt and colleagues (Haidt 2001; Haidt et al. 1993; Greene and Haidt 2002), who asked participants in a range of studies whether certain actions described in a vignette were permissible, and subsequently elicited justifications for the judgments given. What Haidt et al found was that actions triggering disgust responses in participants were likely to be judged highly immoral, but that people generally failed, or contradicted themselves, in their efforts to provide a justification for their judgments. For instance, a group of Westerners found it immoral for someone to eat the family dog that had been accidentally run over by a car, without being able to explain why. Another study which found strong, but not properly justified, moral condemnation of conscientious, non-harmful sexual relations between siblings has been replicated cross-culturally.

Cross-culturally, too, an equally robust effect of dissociation between judgment and justification has been found using varieties of the so-called *trolley problem* (Foot 1967; Kamm 1992; 1998; Thomson 1976). In experiments based on this problem, participants are presented with a range of artificial moral dilemmas involving an uncontrolled train and asked whether it is morally permissible to interfere with the course of action described. In a typical scenario, people have to judge whether it is permissible to press a switch which will redirect the train from a track where five people are in its path towards one where only a single individual is in its path. Participants overwhelmingly judge that this is morally permissible. In another version, people are asked whether it is permissible to throw a large individual onto the track to stop the train, killing him in order to save people further down the line. Participants overwhelmingly judge that this is morally impermissible.

This is consistent with what in moral philosophy has been called the principle of double effect, which states that "it may be permissible to harm an individual for the greater good if the harm is not the necessary means to the greater good but, rather, merely a foreseen side effect" (Hauser et al 2007: 3) In Hauser et al's (2007) survey, however, "a large majority of subjects failed to sufficiently justify their moral judgments, including a majority of those subjects who had been exposed to readings in moral philosophy" (2007: 16). This suggests that people may have robust moral intuitions which do not appear to depend on acquired and articulated moral principles.

Hauser et al elicited responses via internet surveys, so their pool of data contains responses collected from a wide variety of geographic, ethnographic and social backgrounds. Tests of people's intuitions in different trolley cases have also been carried out in

anthropological studies, which have tended to show that patterns of responses in the two cases described above are very similar across geographical locations and cultural backgrounds. This provides a second line of argument against the traditional "moral learning" model. If people across the globe, even in civilizations which have had little or no contact with outside human groups, give similar response patterns to moral dilemmas, something other than explicit learning from others will have to explain the similarities[118].

Evidence from other disciplines, such as behavioural economics and comparative law, supports this conclusion and reveals that people across the globe have similar expectations about reciprocity and fairness in the distribution of goods, and an unselfishly motivated inclination to punish those who violate these expectations (for a review, see Sripada and Stich 2006). Mikhail (2007, 2010a, 2010b) has also argued that the fact that "prohibitions of murder, rape and other types of aggression appear to be universal or nearly so, as do legal distinctions that are based on causation, intention, and voluntary behaviour" (2007: 1364) shows some form of unlearned, contentful moral capacity to be universally present in humans.

Sripada (2008) summarizes the experimental findings of recent years as showing that "moral norms exhibit a striking pattern of commonalities across human groups. Moral norms are not indefinitely variable or randomly distributed across human groups. Rather, there are certain kinds of norms that one sees again and again in almost all human societies" (2008: 322).

Even though most people take this to show that learned principles cannot be the whole story about what underlies human moral judgments and reasoning, what exactly is responsible for this universality, as well as the dissociation between moral justification and judgment is still a matter for debate. In particular, it is hard to find any agreement on 1) what degree of specificity should be ascribed to the faculty underlying moral reasoning, and 2) how far this faculty is specific to the moral domain. Patterns of similarity in responses across human groups do perhaps count in favour of a dedicated moral faculty such as the one proposed by Mikhail (2007; 2010a), Dwyer (2007; 2009) and Hauser (2006; Hauser et al 2008); but huge cultural variations in the content of moral norms and systematic differences revealed by

---

[118] Prinz (2008) rightly observes that the fact that a given trait is found cross-culturally does not prove that it is innate. It may be given an environmental explanation, or turn out to just be the result of the way societies which do not have the trait in question have been unable to survive in the long run. In the case of such things as organizing social life according to a normative system, or obedience to authority, though, it is hard to see how an environmental explanation could go. And maintaining that the ubiquity of normative traits is just the result of a search for the most effective social arrangement leads to the prediction that there have been many human cultures which have succumbed because of bad social organization. I know of no historical evidence to back this up.

anthropological and experimental psychological results might be seen as pointing in the opposite direction. The question is how these similarities and differences are best explained.

Mikhail, Dwyer, Hauser and others are currently developing a model of human moral cognition based on an analogy between morality and generative linguistics in the Chomskian tradition,first suggested by Rawls (1972). Their hypothesis is that humans may be innately endowed with moral principles, and that variations in moral behaviour and reasoning may be restricted to a fairly narrow "moral space" and accounted for by something like "parametric variation".

The idea, in a nutshell, is that an innately determined faculty of morality gradually develops in humans, processing moral input from the environment and constraining the class of potential outputs. What accounts for the cross-cultural similarities between moral behaviour is the fact that people share a moral faculty which can only produce a limited range of potential outputs, just as the human language faculty can only produce a fixed set of linguistic variations according to Chomskian linguistics (see Smith 2004: chapter 2). What accounts for the cross-cultural differences is the fact that people across the world are exposed to different input depending on the environment they grow up in, and this sets their moral parameters differently.

But depending on how the linguistic analogy is cashed out, it is not clear that there are sufficient grounds to postulate a faculty of morality. As Sripada argues,

> "There are certain *high-level themes* that one sees in the contents of moral norms in virtually all human groups – themes such as harms, incest, helping and sharing, social justice, and group defense. However, the *specific rules* that fall under these high-level themes exhibit enormous variability" (2008: 330).

He concludes that "The Principles and Parameters Model relies on a small set of discrete and relatively rigid parameters to explain moral norm variation and is thus ill equipped to explain this pattern of thematic clustering" (*ibid*).

Prinz (2008) has also argued that it is not possible to specify the content of universally held moral principles, pointing out a number of exceptions to alleged universals (e.g. prohibition of intentional harm against in-group members), which he claims will force a weakening of moral principles into mere *moral domains*. In his view, this reduces claims about the innateness of morality to near vacuity: "If the only moral universal is the existence of morality itself, then an adequate account of human moral psychology will have to focus on culturally learned rules to gain any purchase on how we actually conduct our lives" (2008: 383).

Instead of appealing to a moral faculty to explain cross-cultural similarities and differences in moral judgements, Prinz (2007) suggests that they may be the outcome of a set of non-moral psychological capacities, among them emotions and theory of mind. Various other possible explanations of similarities and differences in moral norms across time and space have been proposed in the literature, among them the "Affective Resonance Account" of Nichols (2004) and an account in terms of innate moral biases developed by Sripada and colleagues (Sripada 2008; Sripada and Stich 2006, Sripada, Stich, Kelly and Doris in preparation).

Haidt and colleagues (Haidt 2001; Haidt and Joseph 2004; Haidt and Bjorklund 2008) have proposed a *social intuitionist* account of morality, according to which one layer of moral reasoning, the output of a number of "moral modules", supplies the building-blocks on which moral reasoning takes place. Exactly how and how well these different theories explain the emerging data from moral psychology remains to be seen, and how best to account for moral reasoning and its similarities and differences across cultures will probably remain unanswered questions for some time to come.

What can this debate tell us about human moral cognition, and what are its implications for the distinction between intuitive and reflective concepts I have suggested in this chapter? Whether or not one favours an account in terms of innate moral norms, or a dedicated moral capacity, or argues instead that moral reasoning is developed on the basis of a palette of emotions, it does not seem possible to treat the whole of morality as involving transmission through learning, as on the Piagetian story. This provides some evidence for the conclusion that not all moral concepts are *reflective*. Some other capacity, then, has to be responsible for generating a stock of intuitive concepts. As with the emotion concepts discussed above, these might be the output of either intuitive inference mechanisms, or innate formats, or perception.

Whether any one of these processes will be sufficient to explain the acquisition of all moral concepts (whatever they are) will depend on what theory of moral cognition turns out to be the most scientifically convincing. But I claim that the category of intuitive concepts, as I have described it, provides the theorist with a means to explain the acquisition of moral concepts which is compatible with the major theoretical frameworks in moral psychology.

If Prinz' theory of the emotional construction of morality is right, we can explain how some moral concepts are intuitive since they are derived on the basis of perceptual concepts (in this case emotion concepts, which are derived via direct perception). If the nativist theories of either Sripada, Stich, Nichols and others, or the Moral Grammar proponents (Hauser,

Mikhail, Dwyer) are right, intuitive moral concepts are derived on the basis of some innate format, for instance a module or an intuitive inference mechanism, or whatever turns out to be the most empirically viable way of concretizing this somewhat vague idea of innate determination.

My claim is that what explains the intuitive difference between concepts denoting abstract objects such as *justice* and those denoting entities such as *inflation* or *feminism* is that words such as 'justice' give access to both intuitive and reflective concepts, while 'inflation' can only activate reflective concepts. While there is very little reason to think that there are ways of acquiring a concept such as INFLATION without relying on a theory, or on others to mediate semantic access, the recent findings in the moral psychology literature make it prima facie plausible that at least some moral concepts are either innately determined or perceptually derived.

5.3.6. Folk concepts and Theory of Mind

So far in this chapter, I have discussed a variety of entities generally described as "abstract" and shown how the associated concepts may be seen as getting their content. For a variety of theoretical and fictional entities, I have proposed a deferential treatment, on which concepts like INFLATION and SHERLOCK HOLMES lock to the corresponding properties via explicit theories, or through the oral or written testimony of others.

For moral and emotion terms, I have suggested that there are many underlying concepts, some of which may be acquired by proxy, while others are more likely to be derived on the basis of either perceptual capacities or innate formats, since there seems to be good empirical evidence that not all concepts in these domains involve reliance on others.

Quite plausibly, the analyses of words like 'love' and 'morally right' as potentially activating not only deferential/theoretical concepts, but also intuitive concepts derived on the basis of an innate format, can be extended to other types of abstract entities. Cases in point might include mental state words ('belief', 'idea', 'hope') and words expressing physical and material processes ('agent', 'motion', 'cause'). Even though people learn about physics and psychology later in life, some basic expectations about causal processes, motional paths and persistence of physical objects, and about the behaviour and motivations of other agents, appear underlie all interactions with the world from birth onwards.

Many psychologists interpret a wide range of recent experimental results as showing that infants have certain "naive" expectations about the physical behaviour of material objects from around 2 months of age. Looking-time studies reveal that infants are surprised when the

movements of objects fail to comply with principles of cohesion, continuity and contact (Spelke and Kinzler 2007; see also Spelke 2000; Baillargeon 2001 and Carey 2009 for reviews), and will look reliably longer at objects or scenarios that appear to violate e.g. spatio-temporal continuity.

Much research on animal cognition also shows that similar expectations are found in some non-human species, most notably the rhesus monkeys studied by Hauser and colleagues (Hauser and Carey 1998; 2003; see also Santos 2004). Carey (2009) takes the experimental results to indicate the presence of an evolutionarily old, innately endowed system of object cognition with a rich conceptual content, which is responsible for the ability to interact with physical bodies and objects in infants, human adults and non-human animals alike.

Though the evidence that this ability is grounded in an innate conceptual vocabulary is somewhat tentative (Carey 2009: chapter 2 and 3 gives some good arguments for the hypothesis, but concedes that the matter is far from settled), it strongly suggests that pre-linguistic humans and some non-linguistic animals have expectations about the movement of physical bodies. For these expectations to guide behaviour, they have to be mentally represented in some form, and within the framework of a Computational Theory of Mind, concepts are obvious candidates for carrying at least some of this information. Taken together, these assumptions suggest that infants and animals may have concepts denoting abstract things such as *motion* and *cause* which are not acquired through linguistic interaction.

Among the authors cited above, there is wide agreement that a format for representing objects and their movements may be part of what some theorists refer to as "human core cognition": that is, an innately-endowed domain-specific cognitive system which remains unchanged throughout development (see Carey and Spelke 1996; Carey 2009; Spelke and Kinzler 2007). Spelke and Carey also argue that there is an innately endowed domain-specific system for representing and analysing the behaviour of conspecifics, and a separate system dealing with magnitudes. It is plausible that these systems are linked to the formation of other unlearned concepts denoting abstract entities such as *agent, belief* and *idea*, on the one hand and amounts or magnitudes, on the other.

Carey (2009: chapter 5) reviews a wide range of studies testing infants' understanding of actions and events, which suggest that the representation of actions "in terms of attentional states, referential states and goal-directedness is part of core cognition" (2009: 186). She presents evidence that infants have a preference for assigning goals to perceived actions even of inanimate objects, and takes this to show that the conceptual representation of, among other things, *agents* and *actions* form part of this core domain.

Very soon after birth, infants seem to have a preference for human faces and voices as compared to other stimuli, and they follow the gaze of human adults from as early as 2 months (Hood et al 1998, Farroni et al 2004). This suggests that humans may have a capacity to recognize fellow humans before they have had time to internalize much (or any) information from the environment.

There is also experimental evidence showing that infants, from some time after their first birthday, develop an ability to entertain more sophisticated beliefs about the mental life of other agents. Onishi and Baillargeon (2005) report that 15-month olds keep track of an adult's belief about the location of an object, expressing surprise when the adult acts inconsistently with the belief by looking for the object in a place where they (i.e. the adult) had not seen it (see also Song and Baillargeon 2008). This surprise did not change as the location of the object changed, showing that that the baby was not simply tracking the actual location of the object, but was also tracking the mental state of the adult. Similar results were obtained by Surian, Caldi and Sperber (2007) with babies as young as 13 months. Other studies have found what they describe as a mentalizing ability in older pre-linguistic children, e.g. Southgate, Senju and Csibra (2007) and Southgate, Chevallier and Csibra (2010).

This line of evidence suggests that what is known in the literature as Theory of Mind (ToM, as discussed in the previous chapter) is present much earlier than is sometimes thought. For inferential theories of communication (like Relevance Theory, see Sperber and Wilson 1995; 2002), this is what underlies the human capacity for linguistic interaction, and ToM is seen as play a role in a wide variety of other cognitive capacities, such as empathising, moral theorising and so on. If this ability is part of a core cognitive (or otherwise innately determined) system in humans, Theory of Mind may provide the format in which concepts denoting abstract entities like *belief* and *idea* are derived.

It should be said that Surian, Caldi and Sperber's (2007) claim that Theory of Mind is already present at 13 months is controversial, and goes against what has been the received view in developmental psychology for the last 20 years. Most psychologists have assumed that "true" Theory of Mind emerges later than linguistic communication, and many think that it depends to some extent on successful language use (Astington 2006, Pyers 2006). This assumption is based on an impressive array of tests which show that children are not able to pass verbal false belief tasks before the age of four.

There is therefore some scepticism about whether the new studies really reveal more than a mere ability to attribute goals and *attention* (as opposed to beliefs) to other individuals. Moreover, the claim that pre-linguistic children are able to attribute both true and false beliefs

to others "leaves a great mystery to be solved – namely, understanding 2- and 3-year-olds' failures on the battery of theory-of-mind tasks that reflect a representational theory of mind" (Carey 2009: 211-212)[119].

Because the ToM problem is linked to a wide range of other theoretical issues, especially lexical acquisition, I will leave these questions unaddressed. Instead I will conclude this section with the claim that there is some empirical evidence suggesting that concepts denoting abstract entities like mental states may be derived from an intuitive format. If so, it is not, therefore, necessary to provide a story about how concepts like BELIEF and IDEA can be acquired via deference or by explicit theory construction. I see this as an encouraging result for the account of concept acquisition I have proposed, since providing such a story strikes me as extremely hard. It would also fail to do proper justice to the intuition that there is something that distinguishes people's concepts of mental states from those that express properties like *being feminism* and *being inflation*.

5.4. Conclusion

In this chapter, I have a given an externalist account of the representation and acquisition of concepts denoting entities that are seen as ontologically problematic. I have argued against an internalist account of concepts for fictional entities on methodological grounds, suggesting instead that these concepts get their content from explicit theories of mind-dependent properties, or via deference to those who hold such theories. In the case of words for "medium-abstract" entities, I have argued that these potentially activate a number of different concepts, some acquired by deference or theory, while others get their content via perceptual mechanisms or innate formats.

I have tried to show how such an analysis, as well as doing justice to intuitions about ontology, is in tune with empirical findings from current developmental, social and moral psychology. I hasten to point out, though, that my aim was not to give a Grand Unified Theory of the relationship between mind and world, or of mental representation in general. In

[119] A possible solution to this mystery might be found in the claim that traditional false belief tasks are too cognitively complex to be a reliable test of ToM (Bloom and German 2000). An alternative solution suggests that it may be the lack of *other* cognitive capacities which explains children's failure on standard false belief tasks. Mascaro and Sperber (2008) argue that treating communicated information as false presupposes *epistemic vigilance* (see also Sperber et al 2010). Epistemic vigilance is explained as the ability to judge communicators as unreliable and/or malevolent, an ability that does not start to develop before the age of 3/4. Mascaro and Sperber argue that "since younger children are more dependent on caregivers and have less choice in the selection of partners to interact with, they may be less willing and able to categorize people as malevolent" (2008: 377). This, together with the tendency of pre-schoolers "to make more stable attributions of internal traits to others on the basis of positive events than on the basis of negative ones" (*ibid*) may go some way towards accounting for their almost universal failure to pass traditional false belief tasks.

these last two chapters, I have merely discussed some theoretical possibilities available to the externalist who is worried by the argument from ontology. The goal was simply to make some suggestions about how concepts for different types of entities might be acquired.

Still, the representation of many types of entity is left unaccounted for, and I have avoided discussion of perhaps the most notorious cases of abstract entities; those from the domain of mathematics. This was a deliberate choice, since doing justice to the incredibly intricate debate on the nature of e.g. numbers would go well beyond my resources[120]. By focusing on a range of cases which do not figure as prominently in discussions of mental representation (e.g. theoretical entities), I also wanted to show how the argument from ontology seems to me to identify a serious issue that deserves much more detailed discussion by any externalist semantic theory interested in the mechanisms responsible for the acquisition of concepts and mediation of content.

In dividing representations into two sub-categories – intuitive and reflective – and defending a non-isomorphic treatment of the word-concept relationship, I hope to have done just enough to show how different types of entities, with distinct ontological statuses, can be seen as entertained. In the next chapter, I will look more closely at some of the theoretical consequences of making such a division, sketching some empirical predictions that follow from holding that different types of acquisition processes lead to the formation of different concept types.

No doubt a lot of conceptual ground-clearing remains to be done in order to arrive at a fuller understanding of exactly what the data currently being collected by psychologists and neuroscientists imply for the acquisition and structure of conceptual representations of different types of entities. Perhaps the most urgent need is for a clearer understanding of what it means to say that an intuitive concept gets its content from an "innate format". What does it take for a cognitive capacity to be considered innate and how direct does the relation between the format and a supposed intuitive concept have to be?

Prinz (2008) identifies three different types of innateness properties exhibited by different cognitive and physiological functions in different species. "Some innate traits are very rigid. They manifest themselves in a fixed way, and they are relatively impervious to change", he argues (2008: 370). The fear-reaction behaviour caused by the perception of

---

[120] For a collection of recent philosophical discussions on the metaphysics of mathematical entities and the epistemology of mathematics, see Irvine (2009); Bueno and Linnebo (2009) and the other references cited in footnote 109. For recent developmental and neuropsychological approaches to the representation of mathematical entities, see Dehaene (1997); Carey (2009)

sudden, loud noises or objects looming towards one may be examples of this type of innateness property. He argues that a type of innateness associated with the setting of pre-determined "parameters" is qualitatively different from this (2008: 371).

If a mental mechanism is parametrically innate, the appropriate kind of environmental input will determine which of a closed class of options is selected, so that while two or more traits are innate, only one will actually influence physiology and cognition. Moreover, some innate traits are like the starling's ability to imitate songs: "The actual songs are not innate" in that the variation of the output is open-ended "but they are the result of an innate song acquisition device", according to Prinz (2008: 371).

If one claims that a particular cognitive capacity is innate, the empirical consequences will depend significantly on which type of innateness property is involved. It would be possible to take a more neutral stance and claim that any cognitive mechanism which develops reliably independent of environmental input, or dependent on the right type of environmental conditioning, is innate (see e.g. Barrett 2006a). But determining which mental capacities satisfy this requirement is no easy task, since, as Prinz (2008) argues, the universal emergence of a certain cognitive/behavioural trait can often be explained away by appealing to similarities in type of input or interaction with another, unrelated mechanism.

Nevertheless, I believe the categories of intuitive and reflective concepts can serve as useful analytical and heuristic tools in both empirical and theoretical investigations of the human mind. In what way and to what extent the assumptions I have made and defended can contribute to an understanding of cognition will therefore be the focus of the next and final chapter.

6. Conclusion: Concepts, cognitive science and empirical inscrutability

6.1. Introduction

Throughout the preceding chapters, the topic of word meaning has been approached from a purely theoretical perspective. In this concluding chapter, I look at some implications of the framework I have defended for a more empirically oriented study of words and concepts. I consider how far it is possible to carry out a cognitive scientific investigation of meaning based on central relevance theoretic/informational semantic assumptions, and ask whether the account I have proposed can draw on data from theorists using other approaches to concepts.

I start out, in section 6.2, by summarizing some of the most important claims I have made in this thesis, outlining some implications of my view that different types of acquisition processes lead to the formation of distinct concept types. In section 6.2.1 I explain what I take to be the explanatory advantages of this view. In section 6.2.2 I suggest that treating differences in the epistemic capacities that sustain semantic access as leading to differences in concept types may help contribute to understanding why some notorious puzzles in the philosophy of language arise.

In section 6.3, I look at what a distinction among concept types may entail for empirical investigation, arguing in section 6.3.1 that the hypotheses I have made about differences between intuitive and reflective concepts are testable through predictions based on their characteristic features. Somewhat pessimistically, I suggest that the nature of concepts as described in this thesis is not directly accessible either to internal or external modes of investigation. And since the notion of 'concept' is highly theory-dependent, I claim, in section 6.3.2, that empirical findings in cognitive science do not easily generalise to outside the experimental context.

In section 6.4, I compare the informational semantic/conceptual atomist account of concepts with other approaches to the topic in cognitive science, asking whether Fodor's concepts could be subsumed and/or replaced by another theoretical entity. I conclude that whatever one's interests are in studying language and mind, there needs to be something, somewhere, that plays the crucial role assigned to concepts in Fodor's theory.

6.2. Differentiating between concept types

6.2.1. Implications of the intuitive/reflective distinction

In this thesis I have been concerned with the treatment of word meaning from the perspective of radical pragmatics/ informational semantics. Following Relevance Theory, I have suggested that words get their content from standing in a correspondence relation to mentally represented *concepts*.

Taking as my starting point Fodor's (1998a; 2008) view of concepts as mental particulars which, via their Modes of Presentation, "satisfy whatever ontological conditions have to be met by things that function as mental causes and effects" (1998a: 23), I examined the informational semantic hypothesis that concepts have content in virtue of standing in a nomic relation to something in the world. I was particularly interested in the implications of the informational semantic approach for the ontological commitments of its proponents, and considered to what extent the theory has the resources to explain the representation of entities that are not directly observable.

Throughout my discussion, I followed Fodor in distinguishing between the semantic content of a concept and the mechanisms by which that content is acquired and sustained. I showed how Fodor's conception of semantics relies on nothing more than a concept's locking to a property in actual or counter-factual circumstances in order to explain how content is individuated. Even though this "thin" construal of semantics might prevent the theory from explaining many facts of interest to the cognitive scientist investigating the relations between language and thought, I argued that it has the methodologically desirable consequence of preserving ontological and epistemological neutrality.

I suggested that despite Fodor's insistence on a clear division between theoretical domains, proponents of informational semantics are not exempted from saying something about the actual mechanisms by which content is acquired and concepts learned. Fodor (1998a: 75) himself regards informational semantics as "untenable" in the absence of answers to questions about how a concept is formed or tokened as a result of a causal process sustaining what he calls "semantic access". The second part of this thesis has therefore been devoted to exploring the possibilities opened up by Fodor's (1998a: 75-80) proposal that semantic access can be sustained in a multitude of ways. This "open-ended" list includes *theory construction* and *deference* as potential ways in which a concept can be locked to a property.

I have been particularly interested in the implications of this approach for the representation of so-called abstract entities, i.e. things people talk and think about even though they are not perceivable (in the intuitive sense that tables and dogs are perceivable, and love and inflation are not). I have suggested a framework based on Relevance Theory (Sperber and Wilson 1995; 2002; Wilson and Sperber 2004; Sperber 1996; 1997) which might explain the acquisition of a wide range of concepts whose content is not plausibly seen as sustained by perceptual mechanisms. Here, my aim was to complement interesting work on perceptually-based concept acquisition, such as the investigations in Horsey (2006: chapter 5).

What followed from the insights of Sperber and Wilson, and in particular the proposals in Sperber (1996; 1997), was the idea that there are two distinct types of concepts: intuitive and reflective. The intuitive concepts are acquired through encounters with "perceptually identifiable phenomena" (1996: 89) or derived on the basis of spontaneous inferential processes or innate formats. Reflective concepts, by contrast, are acquired via deference and/or introduced by explicit theories.

In chapter 4, I suggested that concepts acquired via communication or theory construction are *epistemically incomplete*, since the thinker has had no direct contact with instances of the property his reflective concept expresses, and therefore has no way of verifying what her concept applies to in the world. Sperber (1996) takes this to show that there is always a risk that thoughts featuring reflective concepts will be *mutually inconsistent*. Hence relying on others to mediate to properties when one does not have first-hand knowledge of their instances leads to the formation of sub-sets of beliefs with content that "cannot be sufficiently evidenced or argued for to warrant their rational acceptance" (Sperber 1996: 91). According to Sperber, the same situation arises when an individual relies on explicit theory to sustain semantic access. In his view, "even for physicists, the theory of relativity is a reflective belief; it is a theory, a representation kept under scrutiny and open to revision and challenge" (*ibid*).

By contrast, beliefs with constituents drawn only from the set of intuitive concepts are "concrete and reliable in ordinary circumstances" (Sperber 1996: 89), and are therefore not expected to give rise to contradictions or inconsistencies. Since these *intuitive beliefs* acquire their content via perceptual processes or spontaneous inference based on perceptual or

innately determined input[121], they "owe their rationality to essentially innate, hence universal, perceptual and inferential mechanisms" (Sperber 1996: 91-92).

Given these assumptions, it is easy to see that there will be important qualitative differences in the behaviour of reflective and intuitive concepts in the mental lives of individuals. The fact that reflective concepts may be epistemically incomplete makes it much more likely that someone acting on a repertoire of intuitive beliefs will behave more consistently than someone acting on beliefs with the same content held reflectively. The account can therefore distinguish different levels of competence among holders of distinct concept types with the same content. This, I think, is a desirable consequence, which goes some way towards explaining intuitions about differing levels of expertise among speakers with different types of experiences.

Another consequence of the assumption that reflective concepts play a role in cognition is that it allows people to have inconsistencies in their belief sets without this necessarily leading to a dismissal of their behaviour, or the behaviour of whole cultures, as irrational. As Sperber puts it, reflective beliefs are "rationally held, not in virtue of their content, but in virtue of their source" (1996: 92). While some sources will be regarded as credible, knowledgeable and benevolent and therefore serve as reliable content-mediators, others presumably will not. Relying on a well-educated friend to sustain semantic access to *inflation* will (in most cases) be good enough to warrant rational acceptance, while relying on a 5-year old or a pet frog as mediator will (in most cases) not.

The fact that reflective beliefs are not subject to a rational consistency requirement can, according to Sperber, provide a response to "those who see in the great diversity and frequent apparent inconsistency of human beliefs, an argument in favour of cultural relativism" (1996: 91). Sperber claims that "Those beliefs which vary across cultures to the extent of seeming irrational from another culture's point of view are typically reflective beliefs with a content that is partly mysterious to the believers themselves" (1996: 92), and the fact that "different people should trust different sources of beliefs – I, my educators, you, yours – is exactly what you would expect if they are all rational in the same way and in the same world, and merely located in different parts of the world" (*ibid*).

---

[121] Here I am simply assuming that intuitive beliefs cannot contain anything other than intuitive concepts and are acquired on the basis of direct perceptual contact. It is likely that Sperber sees the issue as more complicated, given that some of his (1997) examples of intuitive beliefs suggest that he is more lenient on the requirements for something to be a data-base belief. See Sperber (1996: 93ff) for discussion.

While it is widely believed that internal consistency is a crucial property of a cognitively useful belief set, a consequence of Sperber's theory is that it will not always be possible to verify the compatibility of a whole range of beliefs because some of their constituent concepts will be epistemically incomplete. Although this might go counter to classical theories of rationality[122], the postulation of a class of reflective concepts can help explain how people can have and act on beliefs they do not fully understand, and thus shed light on the tendency of individuals or groups to contradict themselves.

6.2.2. The Frege problem revisited

Despite the potential explanatory advantages of distinguishing between intuitive and reflective concepts, a critic might object that a distinction based on qualitative attributes of a concept goes against the spirit of Fodor's theory. In chapter 2, section 2.3.4, I showed how Fodor treats the nature of content mediation as irrelevant to semantic issues, since "it's *that* your mental structures contrive to resonate to [properties, such as] *doghood*, not *how* your mental structures contrive to resonate to *doghood*, that is constitutive of concept possession" (Fodor 1998a: 76). Moreover, Fodor rejects the idea that one manner of locking has priority over another, so that someone like Helen Keller, for whom it was not visual perception that sustained the meaning making connection between the concept DOG and the property *doghood*, can satisfy the conditions for DOG possession just as straightforwardly as a sighted person.

An important reason for making this claim is to preserve the publicity constraint on semantic content, as discussed in chapter 2, section 2.4.2. According to the publicity constraint, a semantic theory, "[b]arring very pressing considerations to the contrary" (Fodor 1998a: 29), should treat concepts as capable of being literally shared even by thinkers situated in different cultures, at different times, and/or equipped with different epistemic abilities. The possibility that different manners of locking may lead to different types of concepts thus

[122] Cf. Makinson's (1965: 205) *preface paradox*, which claims that the common practice of acknowledging in the preface to a book that the book contains false statements "appears to present a living and everyday example of a situation which philosophers have commonly dismissed as absurd: that it is sometimes rational to hold logically incompatible beliefs." (quoted in Allott 2008: 137, n110). The reason why inconsistencies have not been regarded as a plausible property of belief sets is that any arbitrary proposition follows logically from any two inconsistent propositions. "In psychologically realistic terms, then, a danger posed by inconsistency is that a system for generating valid inferences, fed a contradiction as input, may reach any conclusion whatever" (Allott 2008: 137-138). But as Allott (2008, following Cherniak 1986; O'Brien 2004 and others) points out, there is no way to practically achieve such a goal without it leading to "computational explosion", which means that inconsistencies will always be a property of large sets of beliefs as a whole. In practice, Allott argues, "the need for global consistency testing is avoided by (a) segregation of beliefs, (b) the way that cognition is set up so that we are more likely to form and store true than false beliefs, and (c) the distinction between long-term and short-term memory". (2008: 139).

seems to go against the epistemological neutrality of a semantic theory constrained by publicity considerations.

But if one adopts the nomic-informational semantics advocated by Fodor (1994; 1998a), in which questions about conceptual content and acquisition come apart, it does not automatically follow that the difference between reflective and intuitive concepts should lead to differences *in content*. After going through the many ways in which semantic access can be sustained, Fodor underlines the point that whether a concept gets its content via a deferential or a perceptual or other relation does not matter for purposes of semantics:

> "Just as I did not say that having perceptual mechanisms that connect dog sightings with DOG-tokens-in-the-belief-box constitutes your *having* the concept DOG, so I also did not say that the character of these mechanisms determines the *content* of your concept. How a concept achieves semantic access is one thing, what content the concept has is quite another. It is a chief virtue of informational semantics to distinguish between these two" (Fodor 1998a: 76)

A concept C_1 entertained intuitively by person S1 can therefore have a content that is type identical to the content of C_2, even when C_2 is entertained reflectively by another person S2. Instead, what accounts for the qualitative difference between C_1 as entertained by S1 and C_2 in the mind of S2 will be something about the way the two concepts are *syntactically realized*.

Returning to the notions introduced in chapter 1, section 1.4.4, it is easy to see how the distinction between reflective and intuitive concepts can be cashed out in terms of their having (functionally) different Modes of Presentations (MoP). There, I argued that, according to the individualist psychological notion of thought that Relevance Theory (following Fodor) operates with, a thought always comes in a physically-realized form in addition to having content. In section 2.3.3, I showed how Fodor appeals to these MoPs for help in dealing with the so-called Frege problems, where concepts with co-refererential contents can be informative in locutions of the type a=a ('The Morning Star is the Evening Star').

According to Fodor (2008: chapter 3), two concepts expressing exactly the same property can still have distinct roles in an individual's mental life, since "it is the syntax, rather than the content, of a mental state that determines its causal powers" (2008: 70). Two concepts, even if they have the same referent, can be acquired on different occasions and therefore result in the formation of two different mental files. As long as these files are not merged, the person with such co-referential representations will have distinct sets of inferences and attitudes connected with each concept. But the question remains: what explains why these cases arise in the first place? How come different encounters with an object will

lead to the formation of two co-referential concepts in some situations, while it in other contexts different encounters cause the generation of a unitary concept?

A full explanation of how and why Frege cases arise, if Fodor is right, will plausibly have to appeal to a range of different (psychological, epistemic, perceptual etc.) factors, and only further research will show if a systematic theory of this can be provided. But even though I have nothing close to a complete story to tell of how this happens, I suggest that the idea that distinct ways of sustaining semantic access may lead to concepts with different "indexing" can contribute to addressing the issue of what leads to co-referential concepts emerging in cognition.

Let me briefly explain what I mean by this. Fodor suggests that "the world, and all other worlds that are nomologically near-by, arranges things so that the syntactic structure of a mode of presentation reliably carries information about its causal history" (1994: 54). It follows that the syntactic structure of a concept will carry some information about how its content is sustained. Since the cognitive system processes symbolic strings on the basis of syntactic information, the system will be sensitive to *protocols* (or "tags", to borrow vocabulary and a metaphor from Rey under review), which somehow indicate how it is to be handled, in much the same way metadata attached to computer documents (.doc, .rtf., .pdf) tell the system which programs are appropriate for accessing them. The hypothesis that different acquisition mechanisms may lead to concepts of distinct (functional-)syntactic types is not only available to the proponent of informational semantics, then, but may even contribute to an understanding of how exactly Frege cases and other semantic puzzles arise in the first place[123].

As argued above, the fact that reflective concepts are sometimes epistemically incomplete means that the individual who entertains them is not always able to check them for mutual consistency. It follows that two reflective concepts, or one reflective concept and one intuitive concept, even if they express the same property and therefore have the same content, may be cognitively isolated from each other. If someone acquires a concept (e.g. by talking to others) with a content that is beyond her epistemic reach, she will always run the risk of later acquiring a concept that expresses the same content, without recognising that the two concepts are co-referential.

Just to take a concrete example, one could imagine a junior high classmate (call her Sharon) of the young man called Robert Allen Zimmerman. After hearing Robert's

[123] For a defence of computational explanations of the Frege cases against arguments purporting to show that they have to be intentional, see Schneider (in press: chapter 8),

performances in the school theatre, Sharon remembers him as a terrible writer and a musician devoid of talent. Years later, she hears a record playing at a friend's house by one 'Bob Dylan'. On listening to it, she finds the lyrics most compelling and the music mesmerising. She comes to think that BOB DYLAN IS A TALENTED SINGER AND WONDERFUL LYRICIST. Of course, Sharon's classmate Robert and Bob Dylan are one and the same person, and her concepts ROBERT ALLEN ZIMMERMAN and BOB DYLAN have the same referent, although she has not yet realized this.

Sharon therefore has beliefs about her classmate/the folk singer which, from an omniscient, third-party perspective. can be seen as mutually inconsistent. She thinks that ROBERT ZIMMERMAN IS A TERRIBLE WRITER but at the same time believes that BOB DYLAN IS A WONDERFUL WRITER. According to the theory as I have presented it, this may happen because Sharon first acquired a concept ROBERT ZIMMERMAN locked to the property that her classmate instantiated, with semantic access sustained via perceptual mechanisms. By contrast, Sharon's co-extensive concept BOB DYLAN was reflective, since she only formed the mental item as a result of listening to one of his records. She therefore had no direct access to the property that gave content to her BOB DYLAN concept, and no way of telling that Dylan and Zimmerman were in fact one and the same.

If someone were to tell Sharon about Dylan's Minnesota origins, his real name or some other fact that allowed her to make a connection between the artist and her former class mate, the reflective concept BOB DYLAN would merge with her intuitive concept ROBERT ZIMMERMAN, ending up as a unitary mental file. Though the thought leading to this merger would be BOB DYLAN IS ROBERT ALLEN ZIMMERMAN, and this would have the logical form A=A if only content is taken into account, the mechanisms by which she sustains access to the same property are responsible for generating a syntactically formal distinction.

Granted, the distinction between reflective and intuitive concepts is not enough to explain how the full range of Frege cases arise. However, I think it is likely that a better understanding of syntactic kinds of concept based on the different epistemic capacities that sustain their content may take informational semantics at least part of the way to explaining the emergence of co-referential concepts[124]. As Sperber claims, "Reflective beliefs are a loose

[124] The remaining part of the explanatory burden, once the epistemic capacities have been individuated as finely as is appropriate, will probably have to be carried by the metaphysics, and a theory of what attributes is necessary for someone to instantiate a property, as well as what encounters with which attributes should be taken to suffice for someone to count as having a concept resonating to this property. It may be, for instance, that in cases where two co-referential concepts expressing *being Robert Allen Zimmerman/Bob Dylan* are both acquired

family of derived attitudes that are continuous with other reflective attitudes of a non-credal kind (1997: 82)", and it should be possible to observe finer distinctions on the side of beliefs containing perceptually derived concepts as well.

I also think the idea that distinct mechanisms of semantic access lead to the formation of syntactically distinct concepts may help address some other notorious semantic puzzles. Kripke (1979) discusses the case of Pierre the Monolingual Frenchman, who has heard a great many good things about the city of London, which he, like other Frenchmen calls 'Londres'. Through hearing all these fine things, he comes to believe that LONDON IS PRETTY (which he expresses in French as 'Londres est jolie'). Years later, Pierre moves to London, and ends up living in a particularly unattractive part of the city, as a result of which he forms the belief that LONDON IS NOT PRETTY. However, as Pierre the Monolingual Frenchman turns into the Pierre the Bilingual French-English speaker, he begins to mix with the local crowd, picking up his new language from them. He never translates between English and French, and therefore does not make the connection between his ghastly new home-town London and the much admired place he knows only as Londres.

"Does Pierre, or does he not, believe that London is pretty" asks Kripke (1979), who takes this puzzle, in much the same way as the Frege cases, to raise problems for a directly referential view of the content of proper names. But assuming the story of syntactic individuation of co-referential concepts, and granted that the notion of a reflective concept may help us understand why some mechanisms of semantic access can lead to potential inconsistencies, the informational semanticist has the resources to solve the puzzle.

We might assume, first, that Pierre, on hearing about this wonderful place Londres, formed a corresponding concept reflectively without having direct access to anything instantiating it. Then, when he arrived in England and took up residence in (say) Stratford, he formed a different concept intuitively, which, although it happened to express the same property as a concept he already possessed, never merged with the one he had acquired by deference. What allows this to happen is that the syntactic structures of the MoPs of the two concepts expressing *being London* carry information about the qualitatively distinct causal processes by which they were acquired[125].

---

perceptually, something could be said about the holder having acquired these by way of coming across different attributes of Zimmerman/Dylan on different occasions.

[125] The particular solution I have presented to this puzzle shares some affinities with other solutions which compartmentalize beliefs/belief constituents into different domains. For instance, it is claimed by Thomason (2010) that "it is better to treat belief as a loosely related family of related modalities". In his view, "a modular account of belief seems to be a necessary condition for resolving [Kripke's Pierre] problem".

It turns out on this picture that knowing the content of someone's stock of concepts is not enough by itself to predict someone's "behavior, predilections, inferences etc." (Fodor 2008: 86). In order to say something about how a given person is likely to act, one needs to look instead at "the galaxy of beliefs, desires, hopes, despairs, whatever, in which the concepts are engaged" (2008: 87). However, this is not a deficiency of the theory, in that the thin conception of semantics that Fodor operates with was never meant to explain cognition in general. As I pointed out in chapter 2, Fodor acknowledges that informational semantics would have to be supplemented by a variety of theories from different domains in order to explain much of what someone interested in the mind would like to know about.

An account of what Fodor calls "the central thought system" will ultimately have to take into consideration a wide range of facts, quite a few of which will be idiosyncratic to individual thinkers, while others may be expressed as generalisations across groups of thinkers. Similarly, the precise details of the mechanisms by which semantic access is sustained are left to theories of the psychology of perception, epistemology or pragmatics to flesh out. How thoughts are expressed and communicated in public languages is also a separate issue from the individuation of content, as I argued in chapter 3 above. A detailed theory of the procedures by which speakers/hearers translate back and forth from natural language to the language of thought will have to take into account facts from not only philosophy of mind, but also linguistic semantics, pragmatics, philosophy of linguistics and cognitive science.

In this thesis, I have tried to make a modest contribution to this endeavour by supplementing informational semantics with insights from pragmatics and philosophy of linguistics. My aim has been to arrive at a better understanding of two issues I take to be of interest to the wider scientific study of language and thought. The first is the issue of how the relationship between concepts and lexical items, between private and public languages, should be understood. The second relates to how concepts representing non-perceivable entities are acquired and semantic access sustained.

If the merger of radical pragmatics with informational semantics is successful, and the ideas I have presented are coherent, they may take us one small step closer to a theory of word meaning. It may be hard, though, to see the practical applications of this unified semantic/pragmatic theory, or to decide how best to explore these ideas further. I will therefore conclude with some brief comments on the prospects for an empirical investigation of concepts and word meaning.

6.3. An empirical science of concepts?

6.3.1. Deriving predictions about concept types

Although my starting point in this thesis was language use, since my main field of interest is the pragmatic processes underlying communication, much of my work here has revolved around the nature of concepts and how they get their content. This is a consequence of the fact that, according to Relevance Theory and other pragmatic frameworks in the Gricean tradition, the conceptual level is seen as ontologically prior to the use of words in communication. Words function first and foremost to express the thoughts a speaker intends to communicate, and thoughts have content which is independent in principle of the words that can be used to express them[126].

Throughout the preceding chapters, I have argued that adopting Fodor's version of the Computational Theory of Mind as a hypothesis about the individuation of thought also entails a view of concepts as *theoretically* prior to words, from which it follows that data from natural language and language use cannot be directly brought to bear as evidence about conceptual content. This has both positive and negative consequences. It makes it possible, for instance, to account for dissociations between mental and linguistic abilities, and leaves open to empirical investigation the question of whether creatures without language can have systematic and productive thought (chapter 2, section 2.3.1). I have also argued that the theoretical primacy of concepts makes informational semantics immune to counter-arguments based on the variability and context-sensitivity of natural language communication (chapter 4, section 4.2.1).

At the same time, however, I claimed that considerations based on the primacy of concepts should encourage us to look for ways to cash out the relationship between words and concepts that do not rely on semantic encoding or any other type of default mechanism specified a priori (chapter 3, section 3.2.2). As an alternative, I proposed that the word-concept mapping scheme should be seen as one of *potential activation*, which entails that for any given word there is range of potential candidate concepts which could serve as its meaning on a given occasion. Moreover, I suggested that none of these candidates was plausibly seen as the default meaning of a word independent of any context. Though this

[126] Of course, this will not be the case for all concepts *in practice*, if it is true, as I have claimed, that some concepts are acquired on the basis of deference. Concepts like INFLATION, then, will in some sense depend on the word 'inflation' and its use in the community. Alhough this may lead to linguistic relativism for some concepts, in that having thoughts involving e.g. INFLATION or FEMINISM will depend on the language one speaks, it strikes me as completely uncontroversial and close to what Reines and Prinz (2009) call "trivial Whorfianism".

move preserves Fodor's publicity principle, it has the negative consequence that concepts become somewhat inscrutable for the empirical scientist.

Johnson (2004: 354) claims that "Relatively little is currently known about the structure and individuation of concepts. Worse yet, in the present case [without assuming word-concept isomorphism] we cannot address this issue by looking to the structure and individuation of linguistic expressions". If I am right in claiming that there are several candidate concepts that can be expressed (or formed ad hoc) by the use of a single word depending on the situation of utterance, and that there is no way to specify which concepts are the literal meanings of words independently of any context, it becomes an open question what ordinary intuitions and introspective data about "word meaning" show. As far as I can see, nothing in the details of the account I have given can tell the theorist in what way linguistic intuitions can help to decide issues of representation, unless some further assumptions are added[127].

And even though the syntactic structure of a concept is supposed to carry reliable information about its causal history, there is no reason to suppose that people have direct access to this information. On the account Fodor proposes, the actual acquisition of concepts is supposed to be a brute-causal process which happens "below the radar" of intentional mechanisms (Fodor 2008: 152). The formal properties of thought are therefore as inaccessible to introspection as are the rules or principles governing the syntax of our languages. Just as one cannot close one's eyes and introspect syntactic trees, complex clusters of grammatical properties or whatever constrains such syntactic operations as wh-movement, so one cannot figure out, just by asking oneself or others, what kinds of concepts one uses to do one's thinking.

An empirical scientist may imagine that some day, if science ever reaches the required level of technological sophistication, we may be able to identify and individuate different concepts and concept types by looking at physiological data from the brain (measuring e.g. blood flow via functional Magnetic Resonance Imaging, or electrical activity, as in electroencephalography). But even given the right type of technology, there is nothing in the account I have sketched so far which gives any hints about what one should be looking at in

[127] To be sure, the version of the pragmatic account of word-concept activation I have told in this thesis is not much different from the alternative, relevance-theoretic account of encoding in this respect. So Sperber and Wilson (1995: 193) claim that "Semantic representations of sentences are mental objects that never surface to consciousness. If they did, they would seem entirely uninteresting (except, of course, to semanticists). Semantic representations become mentally represented as a result of an automatic and unconscious process of linguistic decoding".

identifying a concept. Recall that for Fodor, a concept's Mode of Presentation has a functional specification (Fodor 2008: 89ff; see also my chapter 1, section 1.5 and chapter 3 section 3.3.5), and there will therefore be no "neural marker" or anything comparable which might help a cognitive scientist to individuate any particular concept.

All these considerations may seem to suggest that, if Fodor is right, the science of concepts is predestined to remain a theoretical exercise. But the fact that a phenomenon is inaccessible to direct investigation does not mean that it cannot be studied. And even though concept types are individuated by their functional role, there is no reason why these roles should not be specified in such a way that some empirical predictions can be extracted from the account. In fact, the division I have made between intuitive and reflective concepts, and the different properties which Sperber (1996) takes to characterise them, may take us some way towards making some concrete predictions about how the concepts underlying a given word are individuated syntactically.

Though I will not attempt to propose anything close to an experimental paradigm, one might consider the property of epistemic incompleteness which is supposed to hold for concepts acquired via deference. If Sperber is right to claim that beliefs containing reflective concepts can give rise to inconsistencies in particular sub-sets of mental representations, it follows that people acting on these beliefs should sometimes behave inconsistently. They will also be more likely to express conflicting beliefs, without being able to identity these conflicts, or even being able to resolve them when they are pointed out.

I also think it is possible that the intentional manipulation of an individual's beliefs and actions will be easier when these beliefs contain reflective rather than intuitive concepts (as has been suggested by Allott 2006). It should be comparatively less demanding to get research subjects to agree with implausible, contradictory or even self-contradictory claims if these activate concepts that the subjects have not acquired through direct perception. If reflective beliefs may be epistemically incomplete, it should also be easier to get subjects to go against their previously stated beliefs when these contain reflective concepts rather than purely intuitive one.

It also follows from my account that beliefs containing reflective concepts should display much greater cross-cultural variability than intuitive beliefs. Assuming that people's perceptual systems and capacities are relatively invariable across populations, beliefs featuring concepts acquired on the basis of encounters with observable objects or events that instantiate a given property should be comparatively more stable than reflective beliefs across individuals and cultures. Another consequence may be that word meanings which correlate

with reflective concepts change faster than those which potentially activate only intuitive concepts. This derives from the fact that the phenomena in the world which presumably instantiate the properties expressed are not perceptually accessible and therefore cannot "tie down" the words which activate reflective concepts in the same way as perceptual input would.

Furthermore, one might hypothesize that words which potentially activate reflective concepts will typically be acquired later in development than those which correlate with concepts acquired via direct perception or on the basis of innate formats. Since mastering language to at least some extent is a prerequisite for concept acquisition via either deference or theory construction, reflective concepts will necessarily emerge at later stages of development than intuitive ones.

So here are some testable consequences of distinguishing between types of concept linked to different properties. Along with other predictions derivable from my account, these may help to bring the study of concepts out of the purely theoretical domain and thereby contribute to a better understanding of how meaning is constituted and content acquired. Of course, these predictions need sharpening, and there will be practical problems about how to carry out actual investigations which are not so obviously dealt with based on just the theory I have presented in this chapter. But I take it that even a minor contribution to the issue of how word meaning can be studied empirically from an informational semantic viewpoint would be a welcome development in the notoriously theoretical and complex issue of concept individuation.

6.3.2. Concepts and ordinary intuitions

Though Sperber's work on the intuitive/reflective distinction has inspired some research in psychology and anthropology, there are to my knowledge no studies which systematically test the predictions derivable from this distinction. However, some of my suggestions above, for instance the idea that beliefs containing reflective concepts should display greater cross-cultural variability than intuitive beliefs, harmonize quite well with anecdotal, folk-anthropological data. Moreover, empirical studies of child-language acquisition can also be seen as confirming the prediction that concepts representing non-perceived entities are acquired comparatively late.

Gleitman (2009: 240) summarizes findings in developmental psychology as showing that "The first-learned 100 or so words are – animal noises and 'bye-bye's excluded – mainly terms that refer in the adult language to whole objects and other kinds, mainly at some

middling or "basic" level of conceptual categorization". In her view, "This is one of the most robust effects in the literature of language learning, and is seen again and again cross-culturally and cross-linguistically" (2009: 251). Unfortunately, despite a flurry of cognitive scientific interest in the topic of concepts and word meaning the last few years, there is little other data that have a direct bearing on the account I have proposed and the predictions I made above.

The reasons for this were touched on above, and are linked to the fact that concepts are not directly observable. Consequently, investigations into the domain of concepts and conceptual representations have to rely on assumptions about how these are individuated at the data collection stage, which means that the results are highly theory-dependent. Which of the many different views of concepts in the literature (classical theories, prototype theories, exemplar theory, theory theories, inferential role theories, perceptual symbol theories and so on) a cognitive scientist subscribes to will inevitably provide a filter through which she interprets the experimental data.

A couple of examples drawn from some of the literature on the representation of abstract objects may serve as illustrations. In a series of experiments, Barsalou, Wiemer-Hastings and colleagues (Barsalou and Wiemer-Hastings 2005; Wiemer-Hastings and Xu 2005; Wiemer-Hastings et al 2001) have elicited speakers' judgments about the content of a range of representations of objects at various degrees of abstractness. These range from concepts representing highly concrete items (TREE, CAR, SOFA), through concepts that Barsalou and Hastings (2005: 129) describe as seeming "somewhat concrete but more abstract than typical concrete concepts" (COOKING, FARMING, POSSESSION, DAY), to medium-to-highly-abstract ones (HAPPINESS, EXPECTATION, TRUTH, FREEDOM). In the different studies, speakers' intuitions were elicited about properties of these entities, contextual elements likely to co-occur with them, and suggestions about specific contexts in which these entities may figure.

What the researchers found was that although concepts for abstract entities show some similarities with concepts for concrete objects (for instance, Barsalou and Wiemer-Hastings 2005: 134-135 claim that grasping both types of concept presupposes an understanding of "situational content", i.e. the contextual elements which have to be present for the concept to apply correctly), "abstract concepts involve qualitatively different types of properties from concrete concepts" (Wiemer-Hastings and Xu 2005: 731). Wiemer-Hastings and Xu find that "abstract concepts have fewer intrinsic and proportionally more relational properties" (*ibid*) and that these properties are less specific than those linked to concrete concepts. Barsalou and Wiemer-Hastings (2005) also find support for their hypothesis that abstract concepts are more

cognitively complex than concrete concepts, and that their content is distributed across a wider range of situations.

In an earlier study, Hampton (1981) also elicits responses from participants on "conceptual features" and uses these to develop his account of concepts as *polymorphic*. He defines a polymorphous concept as "one in which an instance belongs to a certain category if and only if it possesses a sufficient number of a set of features, none of which need be common to all category members" (1981: 149). This work also incorporates elements from prototype theory and defends the view that the more of a designated set of features possessed by a given object aptly described by a certain category label, the more typical the instance will be of that category[128].

But the claim that these data provide insights into the nature of conceptual representation rests on two main assumptions, neither of which is compatible with the view of mental representation defended in this thesis. The first is that concepts are decompositional, and the second is that ordinary speakers have direct, conscious access to this conceptual content. The assumption that concepts can be broken down into smaller bits that wholly or partly constitute their content is common in the psychological literature (see Moss et al 2009: 218ff for review), and it has a long and rich history (variants of it are often labeled "classical" theories of concepts, and according to Laurence and Margolis 1999a: 10, "it would be difficult to overstate the historical predominance" of such theories). But as explained in my chapter 2, Fodor sees the prospects for such componential analyses as very bleak (see also Fodor 1998a: chapter 3 and chapter 5; Laurence and Margolis 1999a), and he ends up defending the view that conceptual content is exhausted by the existence of an appropriate lawful relation between a concept and a property.

And as mentioned in the previous section, the processes by which concepts lock to properties are sub-intentional[129] and therefore not directly accessible to conscious introspection. On Fodor's story, concept acquisition is "a kind of thing that our sort of brain tissue just does" (2008: 152), and the assumption that introspection will yield direct, uncontaminated information about semantic access is therefore untenable for a Fodorian. Empirical work on concepts designed to confirm/disconfirm Fodor's hypotheses about

---

[128] For some recent experimental work on the notion of polymorphous concepts, see Stukken et al (2009) and Dry and Storms (2010). Interestingly, a somewhat similar methodology to the one being applied in the contemporary psychological literature on abstract entities also forms part of the study of ordinary speaker intuitions about *truth* in Næss (1938).

[129] An exception to this are cases where concepts are locked to properties by way of e.g. explicit theories (see Fodor 1998a: 78-79 for discussion).

meaning will therefore have to find some other way to access the workings of thought, unless, of course, one wants to opt for another, alternative theory of what a 'concept' is.

6.4. Conclusion: 'concept' as a heterogeneous concept

As Edwards (2009: 290) points out, Fodor is in a minority in the philosophical and psychological community in treating concepts as mental items that have their content individuated solely by a mind-world mechanism. Indeed, informational semantics/conceptual atomism "has met with substantial resistance, to put it mildly. Most philosophers and cognitive psychologists are prepared to dismiss the approach as a nonstarter; evidently, many assume that the view does not warrant explicit dismissal".

Though there may be several reasons (both theory-internal and theory-external, philosophical and sociological) for this unpopularity, I think the fact that informational semantics makes concepts empirically inscrutable might be one explanation. The thinking may be (though this is just speculation) that if concepts are unstructured entities which are outcomes of a sub-intentional, brute-causal interaction between the mind and the world, they will also be intractable for both empirical science and introspection, and are therefore completely uninteresting as objects of study. Many theorists, especially in cognitive science, therefore adopt one or other of the competing theories of concepts, regarding them as constituted instead by inferential/conceptual roles, knowledge about categories or featural components, or a combination of these.

Interestingly, there may be some vocabulary issues to resolve here, since it seems that Fodor's view of concepts as the constituents of thoughts diverges in important ways from the metaphysics and job description concepts are given in cognitive science. To take just one example, here are Russell and Lemay (2004: 492), with a somewhat loose characterisation of what concepts are and do:

> "Concepts are mental processes that transform the raw data of experience into manageable units (grouping them into categories and ordering them along dimensions). Concepts thus serve cognitive economy and are involved in perception, memory, thinking, solving problems, and any other psychological process. Concepts are tools, primitive versions of which can be found in infants, and sophisticated versions of which are part of advanced science. Concepts can also be thought of as parts of a larger network of assumptions and other cognitive skills. Thus any concept is 'theory-laden'"

Machery, in a précis of his recent review of different approaches to concepts and their role in philosophy and psychology (Machery 2009), claims that in cognitive science "a concept of x is a body of information about x that is stored in long-term memory and that is used by default

in the processes underlying most, if not all, higher cognitive competences when they result in judgments about x" (2010: 195-196).

Machery contrasts this with the philosophical view of concepts, which typically invokes whatever cognitive mechanism "allows people to have propositional attitudes (beliefs, desires, etc.) about the objects of their attitudes" (2010: 199). According to Machery, these distinct views "are not meant to answer the same questions and are thus not competing", since philosophers and psychologists working on concepts have "in fact entirely different goals". (2009: 34). And even though,

> "it would certainly be nice to have a correct philosophical theory of concepts and a correct psychological theory of concepts […] a psychological theory of concepts would not be incomplete for failing to explain how one can have propositional attitudes about the objects of our attitudes; *mutatis mutandis* for a philosophical theory of concepts" (2009: 37).

I think Machery is both right and wrong about this. I agree with him that there are distinct notions of concepts at play in the philosophical, psychological and cognitive science literature, and that these might be to some extent independent of each other. But there is also a question of ontological priority which Machery does not address, and I would claim, with Rey (2009b, 2010), that the psychological notion of 'concept' is at least partly dependent on the philosophical one.

According to Rey (2009b),

> Generalizations and explanations of, e.g., cognitive development, fallacies in reasoning, vision and language understanding (to take some of the more successful areas of recent psychology) -- all these presuppose concepts as shared constituents of the propositional attitudes the explanations concern. It's not clear how even to describe the phenomenon of the Müller-Lyer illusion unless we can presume that people share a concept of *longer than*, or the gambler's fallacy, without them sharing *more likely*. Concepts seem to be natural kinds at least to the extent that they are the kinds of entity over which psychology generalizes (Rey 2009b)

Rey suggests, following up on a claim in an earlier paper of his (Rey 1985), that it is important to distinguish "the concept of something from merely the (epistemic) conceptions of it that have been too much the focus of the psychological research Machery reviews" (Rey 2010: 222). Concepts are what may be shared and what may remain stable between different thinkers, and thus what allows the psychologist to make generalisations about behaviour, or more generally, "intentional explanations in any viable cognitive psychology" (Fodor 1998a: 29).

In Fodor's view, what makes psychological explanations intentional "is that they appeal to covering generalizations about people who believe that such-and-such, or people who desire that so-and-so, or people who intend that this and that, and so on" (*ibid*). While it should not be taken for granted that these explanations necessarily involve *Fodorian* concepts, psychologists do, as Edwards (2010b: 210) points out, "often appear to presume something in the vicinity of representational content without making this explicit. Consider, for example, how natural it is to talk about exemplars or prototypes being *of or about* a category".

Without some way of understanding how these representations are individuated and acquired, there does seem to be something missing from the psychologists' view of concepts as "bodies of information". Note that the reverse is not true, since according to the Fodorian account of concepts, representations are ontologically prior to the bodies of knowledge in which they figure. If one assumes that "concepts are constituents of mental states" (Fodor 1998a: 6) and accepts the atomist view that what concepts you have is "conceptually and metaphysically independent of what epistemic capacities you have" (*ibid*) then it is possible to provide a theory of how concepts have their content without saying anything about what role they play in cognition.

If Machery is right to claim that there are different notions of 'concept' at play in the philosophical and psychological literature, Fodor's view of 'concepts' is independent of psychological research that purports to deal with the same entity. But if Rey and Edwards are right to see the psychological view of 'concepts' as implicitly relying on a more basic representational notion, then not only are philosophical notions of 'concept' independent of the psychological ones (but not vice versa), they are also *prior to* them.

In short, the view of concepts as whatever is general and stable enough for psychological generalisations to be possible seems to be indispensable to any theory of cognition. Whatever one thinks of particular approaches to concepts and the heterogeneous use of the term in the literature[130], there is a genuine need for a theoretical entity that captures and helps us understand the key cognitive functions that psychologists and philosophers alike

[130] Machery (2009; 2010) argues that the class of concepts psychologically construed "divides into at least three fundamental kinds of concepts –prototypes, exemplars, and theories" where "prototypes are bodies of statistical knowledge about a category, a substance, a type of event, and so on." He sees exemplars as "bodies of knowledge about individual members of a category (e.g., Fido, Rover), particular samples of a substance, and particular instances of a kind of event (e.g., my last visit to the dentist)" while theories are "bodies of causal, functional, generic, and nomological knowledge about categories, substances, types of events, and the like." (2010: 201). From this, he makes a bit of a leap to claiming that "Cognitive scientists might be better off renouncing the very notion of concept. Rather, they should use theoretical terms introduced to refer to the fundamental kinds of concepts – namely, prototype, exemplar, and theory". For critical discussion, see the *BBS* reviews following Machery's (2010) précis article, especially Margolis and Laurecnce (2010), Samuels and Ferreira (2010) and Weiskopf (2010b).

are interested in studying. For any serious psychology to take place at all, the scientist is in need of some notion which allows her to generalise over populations. It therefore seems to follow that there must be something, somewhere in the theory, that abides by the publicity constraint Fodor suggests for concept possession.

Moreover, thought does seem to display features which indicate systematicity and productivity. Someone interested in saying something about the cognitive mechanisms which make this possible must either appeal to mental items that are compositional or otherwise explain away the need for compositionality (Weiskopf 2010b). As Fodor (1998a; 2004; 2008) insists, there are very few good candidates for this job other than atomic concepts. Even though the philosophical and psychological views of concepts might not be in competition, the explanation of how minds can contain something which yields productive, systematic, shareable and stable meaning should at least not be irrelevant to psychological concerns.

I think all this shows that philosophical accounts of concepts (whatever one wants to call them) are theoretically indispensable. The fact that treatments such as Fodor's are often dismissed outright masks this important point, leading many people studying the workings of the mind to ignore the significant insights at the core of Fodor's philosophy of mind. I therefore agree with the sociological speculation of Edwards (2009: 304) that "Being able to retreat to the claim that Fodor's alternative theory is a non-starter is an all too convenient excuse for avoiding hard questions about the nature and centrality of [constraints on concept possession]".

Even though Fodor's view of concepts is somewhat inaccessible to empirical research, and even though informational semantics needs to be supplemented by insights from a range of different theoretical domains in order to make significant, constructive contributions to the study of the mind, this does not mean that conceptual atomism/informational semantics is uninteresting. It merely means that the Fodorian account of concepts starts on a particularly tricky piece of the enormous puzzle that is understanding the mind. But if Fodor's ideas are correct, this piece might just be essential in order to get the rest of the picture right.

References

Adams, F. and Campbell, K. 1999: Modality and abstract concepts. *Behavioral and Brain Sciences,* 22, 610.

Allott, N. 2006: The role of misused concepts in manufacturing consent: A cognitive account. In De Saussure, L. and Schulz, P. (eds) *Manipulation and ideologies in the Twentieth Century: Discourse, language, mind.* Amsterdam: John Benjamins.

Allott, N. 2008: *Pragmatics and rationality*. PhD Thesis, University College London.

Astington, J. W. 2006: The developmental interdependence of theory of mind and language. In Enfield, N. J. and Levinson, S. C. (eds) *Roots of human sociality: Culture, cognition and interaction.* Oxford: Berg.

Atran, S. and Sperber, D. 1991: Learning without teaching: Its place in culture. In Landsmann, L. T. (ed) *Culture,schooling and psychological development.* Norwood, N.J.: Ablex Pub. Corp.

Averill, J. R. and More, T. A. 2000: Happiness. In Lewis, M. and Haviland-Jones, J. M. (eds) *Handbook of emotions, 2nd edition.*

Bach, K. 1994: Conversational impliciture. *Mind & Language,* 9, 124-162.

Bach, K. 2000: Review of J. A. Fodor 'Concepts: Where cognitive science went wrong'. *Philosophical Review,* 109, 627-632.

Baillargeon, R. 2001: Infants' physical knowledge: Of acquired expectations and core principles. In Dupoux, E. (ed) *Language, brain, and cognitive development: Essays in honor of Jacques Mehler.* Cambridge, Ma.: MIT Press.

Baker, M., Johnson, K. and Roberts, I. 1989: Passive arguments raised. *Linguistic Inquiry,* 20, 219-251.

Balaguer, M. 1998: *Platonism and anti-Platonism in mathematics,* Oxford: Oxford University Press.

Balaguer, M. 2009: Platonism in metaphysics. In Zalta, E. N. (ed) *The Stanford encyclopedia of philosophy (Summer 2009 Edition).*

Barrett, H. C. 2006a: Modularity and design reincarnation. In Carruthers, P., Laurence, S. and Stich, S. P. (eds) *The innate mind: Culture and cognition.* Oxford: Oxford University Press.

Barrett, H. C. and Kurzban, R. 2006: Modularity in cognition: Framing the debate. *Psychological Review,* 113, 628-647.

Barrett, L. F. 2006b: Are emotions natural kinds? *Perspectives on Psychological Science,* 1, 28-58.

Barrett, L. F., Gendron, M. and Huang, Y.-M. 2009: Do discrete emotions exist? *Philosophical Psychology,* 22, 427-437.

Barsalou, L. W. 1982: Context-independent and context-dependent information in concepts. *Memory & Cognition,* 10, 82-93.

Barsalou, L. W. 1983: Ad hoc categories. *Memory & cognition,* 11, 211-227.

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

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

Barsalou, L. W. 2003: Situated simulation in the human conceptual system. *Language and Cognitive Processes,* 18, 1-1.

Barsalou, L. W. and Wiemer-Hastings, K. 2005: Situating abstract concepts. *Grounding cognition: The role of perception and action in memory, language, and thinking*, 129–163.

Bartels, A. and Zeki, S. 2004: The neural correlates of maternal and romantic love. *NeuroImage,* 21, 1155-1166.

Baumard, N., Chevallier, C. and Mascaro, O. In revision: “It’s not fair!”: The sense of justice in young children. University of Oxford.

Beck, J. 2010: Sense, mentalese, and ontology. Ms., University of St. Louis, Wa.

Begby, E. Under review: Semantic minimalism and the 'miracle of communication'. University of Oslo.

Benacerraf, P. 1973: Mathematical truth. *The Journal of Philosophy,* 70, 661-679.

Benacerraf, P., & Putnam, H. 1983. *Philosophy of mathematics: Selected readings*, Cambridge University Press.

Bickle, J. 1998: *Psychoneural reduction:The new wave,* Cambridge, Mass.: MIT Press.

Bickle, J. 2008: Multiple realizability. In Zalta, E. N. (ed) *The Stanford encyclopedia of philosophy (Fall 2008 Edition).*

Blakemore, D. 1994: Echo questions: A pragmatic account. *Lingua,* 4, 197-211.

Bloom, P. and German, T. P. 2000: Two reasons to abandon the false belief task as a test of theory of mind. *Cognition,* 77, B25-B31.

Bloom, P. 2010: The moral life of babies. *New York Times online*, http://www.nytimes.com/2010/05/09/magazine/09babies-t.html

Blutner, R. 1998: Lexical pragmatics. *Journal of Semantics,* 15, 115-162.

Blutner, R. 2004: Pragmatics and the lexicon. In Horn, L. R. and Ward, G. L. (eds) *Handbook of pragmatics.* Oxford: Blackwell.

Boghossian, P. 2005: Is meaning normative? In Beckermann, A. and Nimtz, C. (eds) *Philosophy – science – scientific philosophy.* Paderborn: Mentis.

Borg, E. 2004: *Minimal semantics,* Oxford: Clarendon.

Bowerman, M. and Choi, S. 2003: Space under construction: Language-specific spatial categorization in first language acquisition. In Gentner, D. and Goldin-Meadow, S. (eds) *Language in mind: Advances in the study of language and thought.* Cambridge, Mass.: MIT Press.

Bryan, M. F. 1997: On the origin and evolution of the word 'inflation'. *Economic Commentary.* Cleveland: Federal Reserve Bank of Cleveland, http://www.clevelandfed.org/research/commentary/1997/1015.pdf

Bueno, O. and Linnebo, Ø. (eds) 2009: *New waves in philosophy of mathematics,* Basingstoke: Palgrave Macmillan.

Buleandra, A. 2008: Normativity and correctness: A reply to Hattiangadi. *Acta Analytica,* 23, 177-186.

Burge, T. 1979: Individualism and the mental. In French, P. A., T. E. Uehling and H. K. Wettstein (ed) *Studies in metaphysics. Midwest studies in philosophy, vol. IV.* Minneapolis: University of Minnesota Press.

Burge, T. 1986: Individualism and psychology. *Philosophical Review,* 95, 3–45.

Burgess, J. P. and Rosen, G. A. 1997: *A subject with no object: Strategies for nominalistic interpretation of mathematics,* Oxford: Clarendon Press.

Burton-Roberts, N. 2007: Varieties of semantics and encoding: Negation, narrowing/loosening and numericals. In Burton-Roberts, N. (ed) *Pragmatics.* Basingstoke: Palgrave.

Buzsáki, G. 2006: *Rhythms of the brain,* Oxford: Oxford University Press.

Cain, M. J. 2002: *Fodor: Language, mind, and philosophy,* Cambridge, UK ; Malden, MA: Polity.

Calvo Garzón, F. 2000: State space semantics and conceptual similarity: Reply to Churchland. *Philosophical Psychology,* 13, 77-95.

Calvo, M. G. and Marrero, H. 2009: Visual search of emotional faces: The role of affective content and featural distinctiveness. *Cognition & Emotion,* 23, 782-806.

Calvo, M. G., Nummenmaa, L. and Avero, P. 2008: Visual search of emotional faces: Eye-movement assessment of component processes. *Experimental Psychology,* 55, 359-370.

Cappa, S. F., Frugoni, M., Pasquali, P., Perani, D. and Zorat, F. 1998: Category-specific naming impairment for artefacts: A new case. *Neurocase,* 4, 391-397.

Cappelen, H. and Hawthorne, J. 2007: Locations and binding. *Analysis,* 67, 95-105.

Cappelen, H. and Lepore, E. 2005: *Insensitive semantics:A defense of semantic minimalism and speech act pluralism,* Malden, Mass.: Blackwell Pub.

Cappelen, H. and Lepore, E. 2006: Shared content. In Lepore, E. and Smith, B. C. (eds) *The Oxford handbook of philosophy of language.* Oxford: Clarendon.

Cappelen, H. and Lepore, E. 2007: Relevance Theory and shared content. In Burton-Roberts, N. (ed) *Pragmatics.* Basingstoke: Palgrave Macmillan.

Caramazza, A. and Mahon, B. Z. 2006: The organisation of conceptual knowledge in the brain: The future's past and some future directions. *Cognitive Neuropsychology,* 23, 13-38.

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

Carey, S. and Spelke, E. S. 1996: Science and core knowledge. *Philosophy of Science,* 63, 515-533.

Carnap, R. 1950: *Logical foundations of probability,* Chicago: University of Chicago Press.

Carruthers, P. 2006: *The architecture of the mind: Massive modularity and the flexibility of thought,* Oxford: Clarendon.

Carston, R. 2002: *Thoughts and utterances: The pragmatics of explicit communication,* Oxford: Blackwell.

Carston, R. 2004: Explicature and semantics. In Davis, S. and Gillon, B. S. (eds) *Semantics: A reader.* Oxford: Oxford University Press.

Carston, R. 2006: Linguistic communication and the semantics/pragmatics distinction. *UCL Working Papers in Linguistics,* 18, 37-69.

Carston, R. 2008a: Review of E. Borg 'Minimal semantics'. *Mind & Language,* 23, 359-367.

Carston, R. 2008b: Optional pragmatic processes or optional covert linguistic structure? *UCL Working Papers in Linguistics,* 20, 143-156.

Carston, R. 2009: Relevance Theory: Contextualism or pragmaticism? *UCL Working Papers in Linguistics,* 21, 19-26.

Chalmers, D. 2010: The Singularity: A philosophical analysis. *Journal of Consciousness Studies,* 17, 7-65.

Cherniak, C. 1986: *Minimal rationality,* Cambridge, Mass.: MIT Press.

Chevallier, C. 2009: *La communication dans le syndrome d'asperger*. PhD Thesis, Université Lumiere - Lyon 2.

Chomsky, N. 1975: *Reflections on language,* New York: Pantheon Books.

Chomsky, N. 1993: *Language and thought,* London: Moyer Bell.

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

Chomsky, N. 2009a: *Cartesian linguistics: A chapter in the history of rationalist thought, 3rd edititon,* Cambridge: Cambridge University Press.

Chomsky, N. 2009b: The mysteries of nature: How deeply hidden? *Journal of Philosophy,* CVI, 167-200.

Churchland, P. M. 1986: Some reductive strategies in cognitive neurobiology. *Mind,* 95, 279-309.

Churchland, P. M. 1998: Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered. *Journal of Philosophy,* 95, 5-32.

Collins, J. 2007: Syntax, more or less. *Mind,* 116, 805-850.

Connolly, A. C., Fodor, J. A., Gleitman, L. R. and Gleitman, H. 2007: Why stereotypes don't even make good defaults. *Cognition,* 103, 1-22.

Cosmides, L., Barrett, H. C. and Tooby, J. 2010: Adaptive specializations, social exchange, and the evolution of human intelligence. *Proceedings of the National Academy of Sciences,* 107, 9007-9014.

Cosmides, L. and Tooby, J. 2005: Neurocognitive adaptations designed for social exchange. In Buss, D. M. (ed) *Evolutionary Psychology Handbook.* Hoboken, N.Y.: Wiley.

Cowie, F. 1998: *What's within?: Nativism reconsidered,* New York: Oxford University Press.

Damasio, A. R. 1994: *Descartes' error: Emotion, reason, and the human brain,* New York: G.P. Putnam.

Damasio, A. R. 2000: *The feeling of what happens: Body and emotion in the making of consciousness,* London: W. Heinemann.

Damasio, H., Tranel, D., Grabowski, T. J., Adolphs, R. and Damasio, A. 2004: Neural systems behind word and concept retrieval. *Cognition,* 92, 179-229.

Damjanovic, L., Roberson, D., Athanasopoulos, P., Kasai, C. and Dyson, M. 2010: Searching for happiness across cultures. *Journal of Cognition and Culture,* 10, 85-107.

Dehaene, S. 1997: *The number sense: How the mind creates mathematics,* Oxford: Oxford University Press.

Dorr, C. 2008: There are no abstract objects. In Sider, T., Hawthorne, J. and Zimmerman, D. W. (eds) *Contemporary debates in metaphysics.* Oxford: Blackwell.

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

Dretske, F. I. 1981: *Knowledge and the flow of information,* Oxford: Blackwell.

Dry, M. J. and Storms, G. 2010: Features of graded category structure: Generalizing the family resemblance and polymorphous concept models. *Acta Psychologica,* 133, 244-255.

Dudai, Y. 1997: How big is human memory, or on being just useful enough. *Learning & Memory,* 3, 341-365.

Dwyer, S. 2007: How good is the linguistic analogy? In Carruthers, P., Laurence, S. and Stich, S. P. (eds) *The innate mind: Culture and cognition.* Oxford: Oxford University Press.

Dwyer, S. 2009: Moral dumbfounding and the linguistic analogy: Methodological implications for the study of moral judgment. *Mind & Language,* 24, 274-296.

Dyvik, H. 2005: Translations as a semantic knowledge source. *Proceedings of The Second Baltic Conference on Human Language Technologies,* 27–38.

Edwards, K. 2009: What concepts do. *Synthese,* 170, 289-310.

Edwards, K. 2010a: Concept referentialism and the role of empty concepts. *Mind & Language,* 25, 89-118.

Edwards, K. 2010b: Unity amidst heterogeneity in theories of concepts. *Behavioral and Brain Sciences,* 33, 210-211.

Ekman, P. 1999: Basic emotions. In Dalgleish, T. and Power, M. J. (eds) *Handbook of cognition and emotion.* New York: John Wiley

Ekman, P., Sorenson, E. R. and Friesen, W. V. 1969: Pan-cultural elements in facial displays of emotions. *Science,* 164, 86-88.

Evans, V. 2009: *How words mean: Lexical concepts, cognitive models, and meaning construction,* Oxford: Oxford University Press.

Fagot, D. and Cook, R. G. 2006: Evidence for large long-term memory capacities in baboons and pigeons and its implications for learning and the evolution of cognition. *Proceedings of the National Academy of Sciences,* 103, 17564-17567.

Falkum, I. L. 2010: *The semantics and pragmatics of polysemy: A relevance-theoretic account.* PhD Thesis, University College London.

Farroni, T., Massaccesi, S., Pividori, D. and Johnson, M. H. 2004: Gaze following in newborns. *Infancy,* 5, 39-60.

Fauconnier, G. and Turner, M. 2002: *The way we think: Conceptual blending and the mind's hidden complexities,* New York: Basic Books.

Fodor, J. A. 1974: Special sciences (or: The disunity of science as a working hypothesis). *Synthese,* 28, 97-115.

Fodor, J. A. 1975: *Language of thought,* New York: T. Y. Crowell.

Fodor, J. A. 1981a: The present status of the innateness controversy. In *Fodor (1981b)*.

Fodor, J. A. 1981b: *Representations: Philosophical essays on the foundations of cognitive science,* Cambridge, Mass.: MIT Press.

Fodor, J. A. 1983: *The modularity of mind: An essay on faculty psychology,* Cambridge, Mass.: MIT Press.

Fodor, J. A. 1987: *Psychosemantics: The problem of meaning in the philosophy of mind,* Cambridge, Mass.: MIT Press.

Fodor, J. A. 1990: *A theory of content and other essays,* Cambridge, Mass.: MIT Press.

Fodor, J. A. 1994: *The elm and the expert: Mentalese and its semantics,* Cambridge, Mass.: MIT Press.

Fodor, J. A. 1997: Connectionism and the problem of systematicity (continued): Why Smolensky's solution still doesn't work. *Cognition,* 62, 109-119.

Fodor, J. A. 1998a: *Concepts: Where cognitive science went wrong,* Oxford: Clarendon Press.

Fodor, J. A. 1998b: *In critical condition: Polemical essays on cognitive science and the philosophy of mind,* Cambridge, Mass.: MIT Press.

Fodor, J. A. 2000a: *The mind doesn't work that way: The scope and limits of computational psychology,* Cambridge, Mass.: MIT Press.

Fodor, J. A. 2000b: Replies to Critics. *Mind & Language,* 15, 350-374.

Fodor, J. A. 2001: Language, thought and compositionality. *Mind & Language,* 16, 1-15.

Fodor, J. A. 2003: *Hume variations,* Oxford: Clarendon Press.

Fodor, J. A. 2004: Having concepts: A brief refutation of the twentieth century. *Mind & Language,* 19, 29-47.

Fodor, J. A. 2008: *LOT 2: The language of thought revisited,* Oxford: Oxford University Press.

Fodor, J. A. 2009: *What Frege got wrong (with some help from Quine).* Paper presented at the University of Oslo, September 24, http://www.csmn.uio.no/podcast/Fodor1.html

Fodor, J. A. and Lepore, E. 1992: *Holism:A shopper's guide,* Oxford: Blackwell.

Fodor, J. A. and Lepore, E. 2002: *The compositionality papers,* Oxford: Clarendon Press.

Fodor, J. A. and Lepore, E. 2005: Impossible words: A reply to Kent Johnson. *Mind & Language,* 20, 353-356.

Fodor, J. A. and McLaughlin, B. 1990: Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition,* 35, 183-204.

Fodor, J. A. and Pylyshyn, Z. W. 1988: Connectionism and cognitive architecture: A critical analysis. *Cognition,* 28, 3-71.

Foot, P. 1967: The problem of abortion and the doctrine of double effect. *Oxford Review,* 5, 5-15.

Frank, R. H. 1988: *Passions within reason: The strategic role of emotions,* New York: Norton.

Frege, G. 1918: Der gedanke: Eine logische Untersuchung. *Beiträge zur Philosophie des deutschen Idealismus,* 1, 58-77.

Frege, G. 2010 [orig. 1892]: On sense and nominatum [Über Sinn und Bedeutung]. In Martinich, A. P. (ed) *The philosophy of language (5th edition).* Oxford: Oxford University Press.

Fretheim, T. 2008: *Some critical remarks on the explicature–implicature distinction in relevance theory.* Paper presented at the 23rd Scandinavian Conference of Linguistics. Uppsala University.

Furman, O., Dorfman, N., Hasson, U., Davachi, L. and Dudai, Y. 2007: They saw a movie: Long-term memory for an extended audiovisual narrative. *Learning & Memory,* 14, 457-467.

Gauker, C. 2003: *Words without meaning,* London: MIT Press.

Gleitman, L. 2009: The learned component of language learning. In Piattelli-Palmarini, M., Uriagereka, J. and Salaburu Etxeberria, P. (eds) *Of minds and language: A dialogue with Noam Chomsky in the Basque country.* Oxford: Oxford University Press.

Glock, H.-J. 2009: Concepts: Where subjectivism goes wrong. *Philosophy,* 84, 5-29.

Glüer, K. and Wikforss, Å. 2009: Against content normativity. *Mind,* 118, 31-70.

Goodman, N. 1966: *The structure of appearance, 2nd edition,* Indianapolis: Bobbs-Merrill Co.

Gould, S. J. 1977: *Ontogeny and phylogeny,* Boston, Ma.: Harvard University Press.

Greene, J. D. and Haidt, J. 2002: How (and where) does moral judgment work? *Trends in Cognitive Sciences,* 6, 517-523.

Grice, H. P. 1989: *Studies in the way of words,* Cambridge, Mass.: Harvard University Press.

Groefsema, M. 2007: Concepts and word meaning in Relevance Theory. In Burton-Roberts, N. (ed) *Pragmatics.* Basingstoke: Palgrave.

Gross, S. 2001: Review of J. A. Fodor 'Concepts: Where cognitive science went wrong'. *Mind,* 110, 469-475.

Haidt, J. 2001: The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review,* 108, 814-834.

Haidt, J. and Bjorklund, F. 2008: Social intuitionists answer six questions about moral psychology. In Sinnott-Armstrong, W. (ed) *Moral psychology: The cognitive science of morality, intuition and diversity.* Cambridge, Ma.: MIT Press.

Haidt, J. and Joseph, C. 2004: Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus,* 133, 55-66.

Haidt, J., Koller, S. H. and Dias, M. G. 1993: Affect, culture, and morality, or: Is it wrong to eat your dog? *Journal of Personality and Social Psychology,* 65, 613-628.

Hale, K. and Keyser, S. J. 1993: On argument structure and the lexical expressions of syntactic relations. In Hale, K. and Keyser, S. J. (eds) *The view from building 20: Essays in linguistics in honor of Sylvain Bromberger.* Cambridge, MA.: MIT Press.

Hale, K. and Keyser, S. J. 2002: *Prolegomenon to a theory of argument structure,* Cambridge, Mass.: MIT.

Hall, A. 2008: Free enrichment or hidden indexicals? *Mind & Language,* 23, 426-456.

Hall, A. 2009: 'Free' enrichment and the nature of pragmatic constraints. *UCL Working Papers in Linguistics,* 21, 93-123.

Hamlin, J. K., Wynn, K., Bloom, P. and Mahajan, N. In preparation: Third-party reward and punishment in infants and toddlers. Yale University.

Hampton, J. A. 1981: An investigation of the nature of abstract concepts. *Memory & Cognition,* 9, 149-156.

Hampton, J. A. 2000: Concepts and prototypes. *Mind & Language,* 15, 299-307.

Hanggi, E. B. and Ingersoll, J. F. 2009: Long-term memory for categories and concepts in horses (Equus caballus). *Animal Cognition,* 12, 451-462.

Hart, J. and Gordon, B. 1992: Neural subsystems for object knowledge. *Nature,* 359, 60-64.

Hattiangadi, A. 2006: Is meaning normative? *Mind & Language,* 21, 220-240.

Hattiangadi, A. 2009: Some more thoughts on semantic oughts: A reply to Daniel Whiting. *Analysis,* 69, 54-63.

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R. and Mikhail, J. 2007: A dissociation between moral judgments and justifications. *Mind & Language,* 22, 1-21.

Hauser, M. D. 2006: *Moral minds: How nature designed our universal sense of right and wrong,* New York: Ecco.

Hauser, M. D. and Carey, S. 1998: Building a cognitive creature from a set of primitives: Evolutionary and developmental insights. In Cummins, D. D. and Allen, C. (eds) *The evolution of mind.* Oxford: Oxford University Press.

Hauser, M. D. and Carey, S. 2003: Spontaneous representations of small numbers of objects by rhesus macaques: Examinations of content and format. *Cognitive Psychology,* 47, 367-401.

Hauser, M. D., Young, L. and Cushman, F. 2008: Reviving Rawls' linguistic analogy: Operative principles and the causal structure of moral actions. In Sinnott-Armstrong, W. (ed) *Moral Psychology: The cognitive science of morality,* Cambridge, Ma.: MIT Press.

Hinzen, W. 2007: *An essay on names and truth,* Oxford: Oxford University Press.

Hjelmslev, L. 1966: *Omkring sprogteoriens grundlæggelse,* København: Akademisk Forlag.

Hodes, H. 1984: Logicism and the ontological commitments of arithmetic. *Journal of Philosophy,* 81, 123–149.

Hood, B. M., Willen, J. D. and Driver, J. 1998: Adult's eyes trigger shifts of visual attention in human infants. *Psychological Science,* 9, 131-134.

Horn, L. R. 1984: Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In Schiffrin, D. (ed) *Meaning, form and use in context: Linguistic applications.* Washington DC: Georgetown University Press.

Horsey, R. 2006: *The content and acquisition of lexical concepts*. PhD Thesis, University of London.

Irvine, A. D. (ed) 2009: *Philosophy of mathematics,* Amsterdam: Elsevier.

Jackendoff, R. 1992: *Languages of the mind: Essays on mental representation,* Cambridge, Mass.: MIT Press.

Jackendoff, R. 2002: *Foundations of language: Brain, meaning, grammar, evolution,* Oxford: Oxford University Press.

Jackendoff, R. 2006: Locating meaning in the mind (where it belongs). In Stainton, R. (ed) *Contemporary debates in cognitive science.* Oxford: Blackwell.

James, W. 1884: What is an emotion? *Mind,* 9, 188-205.

Johnson, K. 2004: From impossible words to conceptual structure: The role of structure and processes in the lexicon. *Mind & Language,* 19, 334-358.

Johnson, M. 1987: *The body in the mind: The bodily basis of meaning, imagination, and reason,* Chicago: University of Chicago Press.

Jylkkä, J. 2009: Why Fodor's theory of concepts fails. *Minds and Machines,* 19, 25-46.

Jönsson, M. L. and Hampton, J. A. 2008: On prototypes as defaults. *Cognition,* 106, 913-923.

Kamm, F. M. 1992: *Creation and abortion: A study in moral and legal philosophy,* Oxford: Oxford University Press.

Kamm, F. M. 1998: *Morality, mortality: Death and whom to save from it,* Oxford: Oxford University Press.

Kaplan, D. 1989: Demonstratives. In Almog, J., Perry, J., Wettstein, H. K. and Kaplan, D. (eds) *Themes from Kaplan.* Oxford: Oxford University Press.

Keil, F. C. and Wilson, R. A. 2000: The concept concept: The wayward path of cognitive science. *Mind and Language,* 15, 308-318.

Kim, J. 1993: *Supervenience and mind: Selected philosophical essays,* Cambridge: Cambridge University Press.

Kjoll, G. 2007: *The "content" of 'content': Applying the lexical pragmatics of Relevance Theory*. MA Dissertation, University of London.

Kjoll, G. 2009: Review of W. Hinzen 'An essay on names and truth'. *Journal of Linguistics,* 45, 227-231.

Kohlberg, L. 1984: *The psychology of moral development: Moral stages and the life cycle*. San Francisco: Harper & Row.

Kripke, S. A. 1979: A puzzle about belief. In Margalit, A. (ed) *Meaning and use.* Dordrecht: D. Reidel.

Kövecses, Z. 1986: *Metaphors of anger, pride and love: A lexical approach to the structure of concepts,* Amsterdam: John Benjamins.

Lakoff, G. 1971: On generative semantics. In Steinberg, D. D. and Jakobovits, L. A. (eds) *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology.* Cambridge: Cambridge University Press.

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

Lakoff, G. and Johnson, M. 1999: *Philosophy in the flesh: The embodied mind and its challenge to Western thought,* New York: Basic Books.

Landau, B. 2000: Concepts, the lexicon and acquisition: Fodor's new challenge. *Mind & Language,* 15, 319-326.

Landauer, T. K. 1986: How much do people remember?: Some estimates of the quantity of learned information in long-term memory. *Cognitive Science,* 10, 477-493.

Lange, C. G. 2010 [orig. 1885]: *The emotions [Om sindsbevægelser]*: PoD: Nabu Press.

Laurence, S. and Margolis, E. 1999a: Concepts and cognitive science. In Margolis, E. and Laurence, S. (eds) *Concepts: Core readings.* Cambridge, Mass.: MIT Press

Laurence, S. and Margolis, E. 1999b: Review of J. A. Fodor 'Concepts: Where cognitive science went wrong'. *British Journal for the Philosophy of Science,* 50, 487–491.

Levine, A. and Bickhard, M. H. 1999: Concepts: Where Fodor went wrong. *Philosophical Psychology,* 12, 5-23.

Levinson, S. C. 1983: *Pragmatics,* Cambridge: Cambridge University Press.

Lewis, D. K. 1986: *On the plurality of worlds,* Oxford: Blackwell.

Lewis, M. 2008: The emergence of human emotions. In Lewis, D., Haviland-Jones, J. M. and Barrett, L. F. (eds) *Handbook of emotions 3rd edition.* New York: Guilford.

Lewontin, R. C. 2000: *The triple helix: Gene, organism, and environment,* Cambridge, Mass.: Harvard University Press.

Liggins, D. 2010: Epistemological objections to Platonism. *Philosophy Compass,* 5, 67-77.

Ludlow, P. 2003: Referential semantics for I-languages? In Antony, L. M. and Hornstein, N. (eds) *Chomsky and his critics.* Oxford: Blackwell.

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

Machery, E. 2010: Précis of 'Doing without concepts'. *Behavioral and Brain Sciences,* 33, 195-206.

Mahon, B. Z. and Caramazza, A. 2009: Concepts and categories: A cognitive neuropsychological perspective. *Annual Review of Psychology,* 60, 27-51.

Majid, A., Boster, J. S. and Bowerman, M. 2008: The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition,* 109, 235-250.

Majid, A., Bowerman, M., Staden, M. V. and Boster, J. S. 2007a: The semantic categories of cutting and breaking events: A crosslinguistic perspective. *Cognitive Linguistics,* 18, 133-152.

Majid, A., Gullberg, M., Staden, M. V. and Bowerman, M. 2007b: How similar are semantic categories in closely related languages? A comparison of cutting and breaking in four Germanic languages. *Cognitive Linguistics,* 18, 179-194.

Makinson, D. C. 1965: The paradox of the preface. *Analysis,* 25, 205–207.

Margolis, E. and Laurence, S. 2003: Concepts. In Stich, S. P. and Warfield, T. A. (eds) *The Blackwell guide to philosophy of mind.* Oxford: Blackwell.

Margolis, E. and Laurence, S. 2007: The ontology of concepts: Abstract objects or mental representations? *Noûs,* 41, 561-593.

Margolis, E. and Laurence, S. 2010: Concepts and theoretical unification. *Behavioral and Brain Sciences,* 33, 219-220.

Martí, L. 2006: Unarticulated constituents revisited. *Linguistics and Philosophy,* 29, 135-166.

Martínez Manrique, F. 2010: On the distinction between semantic and conceptual representation. *Dialectica,* 64, 57-78.

Mascaro, O. and Sperber, D. 2009: The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition,* 112, 367-380.

Mateu, J. 2005: Impossible primitives. In Werning, M., Machery, E. and Schurz, G. (eds) *The compositionality of meaning and content: Foundational issues.* Frankfurt: Ontos.

McGilvray, J. 1998: Meanings are syntactically individuated and found in the head. *Mind and Language,* 13, 225-280.

McGilvray, J. 2002: MOPs: The science of concepts. In Hinzen, W. and Rott, H. (eds) *Belief and meaning: Essays at the interface.* Frankfurt am Main: Hansel-Hohenhausen.

McGilvray, J. 2009: *Introduction to N. Chomsky* 'Cartesian Linguistics, 2nd Edition'*,* Cambridge: Cambridge University Press.

McLaughlin, B. 2009: Systematicity redux. *Synthese,* 170, 251-274.

Mercier, H. and Sperber, D. 2009: Intuitive and reflective inferences. In Evans, J. and Frankish, K. (eds) *In two minds.* Oxford: Oxford University Press.

Mikhail, J. 2007: Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences,* 11, 143-152.

Mikhail, J. 2010a: *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral judgment,* Cambridge: Cambridge University Press.

Mikhail, J. 2010b: Is the prohibition of homicide universal? Evidence from comparative criminal law. *Brooklyn Law Review,* 75.

Miller, J. D. and Tanis, D. C. 1971: Recognition memory for common sounds. *Psychonomic Science,* 23, 307-308.

Millikan, R. G. 1993: *White Queen psychology and other essays for Alice,* London: MIT Press.

Millikan, R. G. 2000: *On clear and confused ideas: An essay about substance concepts,* Cambridge: Cambridge University Press.

Moss, H. E., Tyler, L. K. and Taylor, K. R. 2009: Conceptual structure. In Gaskell, M. G. (ed) *Oxford handbook of psycholinguistics.* Oxford: Oxford University Press.

Murphy, G. L. 1996: On metaphoric representation. *Cognition,* 60, 173-204.

Murphy, G. L. 1997: Reasons to doubt the present evidence for metaphoric representation. *Cognition,* 62, 99-108.

Murphy, G. L. 2002: *The big book of concepts,* Cambridge, Mass.: MIT Press.

Neale, S. 2007: Heavy hands, magic, and scene-reading traps. *European Journal of Analytic Philosophy,* 3, 77–132.

Newmeyer, F. J. 1996: *Generative linguistics: A historical perspective,* London: Routledge.

Nichols, S. 2004: *Sentimental rules: On the natural foundations of moral judgment,* Oxford: Oxford University Press.

Noh, E.-J. 2000: *Metarepresentation: A relevance-theory approach,* Amsterdam: John Benjamins.

Næss, A. 1938: *"Truth" as conceived by those who are not professional philosophers,* Oslo: Jacob Dybwad.

O'Brien, D. P. 2004: Mental-logic theory: What it proposes, and reasons to take this proposal seriously. In Leighton, J. P. and Sternberg, R. J. (eds) *The nature of reasoning.* Cambridge: Cambridge University Press.

Onishi, K. H. and Baillargeon, R. 2005: Do 15-month-old infants understand false beliefs? *Science,* 308, 255.

Ostertag, G. 2008: Review of J. Stanley 'Language in context: Selected essays'. *Notre Dame Philosohical Reviews,* 25/05, http://ndpr.nd.edu/review.cfm?id=13183

Pagin, P. 2006: When does communication suceed? Ms., University of Stockholm.

Pagin, P. 2008: What is communicative success? *Canadian Journal of Philosophy,* 38, 85-115.

Pagin, P. and Westerståhl, D. 2010a: Compositionality I: Definitions and variants. *Philosophy Compass,* 5, 250-264.

Pagin, P. and Westerståhl, D. 2010b: Compositionality II: Arguments and problems. *Philosophy Compass,* 5, 265-282.

Paseau, A. Forthcoming: Resemblance theories of properties. *Philosophical Studies*.

Peacocke, C. 1992: *A study of concepts,* Cambridge, Mass.: MIT Press.

Peacocke, C. 2000: Fodor on concepts: Philosophical aspects. *Mind & Language,* 15, 327-340.

Perry, J. 1997: Indexicals and demonstratives. In Hale, B. and Wright, C. (eds) *A companion to philosophy of language.* Oxford: Blackwell.

Piaget, J. 1932: *The moral judgement of the child,* London: Routledge.

Pietroski, P. 2000: Euthyphro and the semantic. *Mind & Language,* 15, 341-349.

Pietroski, P. 2009: Lexicalising and combining. Ms., University of Maryland.

Pietroski, P. 2010: Concepts, meanings and truth: First nature, second nature and hard work. *Mind & Language,* 25, 247-278.

Pinker, S. 1994: *The language instinct: The new science of language and mind,* London: Allen Lane.

Pinker, S. 1999: *Words and rules: The ingredients of language,* London: Weidenfeld & Nicolson.

Pinker, S. 2007: *The stuff of thought: Language as a window into human nature,* London: Allen Lane.

Powell, G. 2010: *Language, thought and reference,* Basingstoke: Palgrave Macmillan.

Premack, D. and Woodruff, G. 1978: Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences,* 1, 515-526.

Preyer, G. and Peter, G. (eds) 2005: *Contextualism in philosophy: Knowledge, meaning, and truth,* Oxford: Clarendon.

Preyer, G. and Peter, G. (eds) 2007: *Context-sensitivity and semantic minimalism: New essays on semantics and pragmatics,* Oxford: Oxford University Press.

Prinz, J. J. 2002: *Furnishing the mind: Concepts and their perceptual basis,* Cambridge, Mass.: MIT Press.

Prinz, J. J. 2004: Which emotions are basic? In Evans, D. and Cruse, P. (eds) *Emotion, evolution, and rationality.* Oxford: Oxford University Press.

Prinz, J. J. 2005: The return of concept empiricism. In Cohen, H. and Lefebvre, C. (eds) *Handbook of categorization in cognitive science,* Oxford: Elsevier.

Prinz, J. J. 2007: *The emotional construction of morals,* Oxford: Oxford University Press.

Prinz, J. J. 2008: Is morality innate? In Sinnott-Armstrong, W. (ed) *Moral psychology: The evolution of morality, adaptations and innateness.* Cambridge, Ma.: MIT Press.

Prinz, J. J. 2011: Regaining composure: A defense of prototype compositionality. In Werning, M., Hinzen, W. and Machery, E. (eds) *The Oxford handbook of compositionality.* Oxford University Press.

Pritchard, T. 2009: *The pattern of word meaning: How words enable us to express content*. PhD Thesis, King's College London.

Pustejovsky, J. 1995: *The generative lexicon,* Cambridge, Mass.: MIT Press.

Pustejovsky, J. 1998: Generativity and explanation in semantics: A reply to Fodor and Lepore. *Linguistic Inquiry* 29, 289-311.

Putnam, H. 1975: The meaning of 'meaning'. *Philosophical papers, volume 2.* Cambridge: Cambridge University Press.

Putnam, H. 1980 [orig. 1967]: The nature of mental states. In Block, N. (ed) *Readings in philosophy of psychology: Volume 1.* Cambridge, Mass.: Harvard University Press.

Pyers, J. E. 2006: Constructing the social mind: Language and false-belief understanding. In Enfield, N. J. and Levinson, S. C. (eds) *Roots of human sociality: Culture, cognition and interaction.* Oxford: Berg.

Quine, W. V. O. 1963: *From a logical point of view: Logico-philosophical essays,* New York: Harper-Row.

Rawls, J. 1972: *A theory of justice,* Oxford: Clarendon Press.

Reboul, A. 2008: Review of N. Burton-Roberts (ed) 'Pragmatics'. *Journal of Linguistics,* 44, 519-524.

Recanati, F. 1993: *Direct reference: From language to thought,* Oxford: Blackwell.

Recanati, F. 2001: What is said. *Synthese,* 128, 75-91.

Recanati, F. 2004: *Literal meaning,* Cambridge: Cambridge University Press.

Recanati, F. 2007: It is raining (somewhere). *Linguistics and Philosophy,* 30, 123-146.

Reines, M. F. and Prinz, J. J. 2009: Reviving Whorf: The return of linguistic relativity. *Philosophy Compass,* 4, 1022-1032.

Rey, G. 1985: Concepts and conceptions. *Cognition,* 19, 297-303.

Rey, G. 2005a: Philosophical analysis as cognitive psychology: The case of empty concepts. In Cohen, H. and Lefebvre, C. (eds) *Handbook of categorization in cognitive science.* Amsterdam: Elsevier.

Rey, G. 2005b: Mind, intentionality and inexistence: An overview of my work. *Croatian Journal of Philosophy*, 15, 389-415.

Rey, G. 2006: The intentional inexistence of language - but not cars. In Stainton, R. (ed) *Contemporary debates in cognitive science.* Oxford: Blackwell.

Rey, G. 2009a: Concepts, defaults, and internal asymmetric dependencies: Distillations of Fodor and Horwich. In Kompa, N., Nimtz, C. and Suhm, C. (eds) *The A Priori and its role in philosophy.* Paderborn: Mentis.

Rey, G. 2009b: Review of E. Machery 'Doing without concepts'. *Notre Dame Philosohical Reviews*, 15/7, http://ndpr.nd.edu/review.cfm?id=16608

Rey, G. 2010: Concepts vs. conceptions (again). *Behavioral and Brain Sciences,* 33, 221-222.

Rey, G. 2011: Externalism and inexistence in early content. In Schantz, R. (ed) *Prospects for meaning.* de Gruyter.

Rey, G. Under review: We aren't all "self-blind": A defense of a modest introspectionism. University of Maryland.

Rives, B. 2009: The empirical case against analyticity: Two options for concept pragmatists. *Minds and Machines,* 19, 199-227.

Rosch, E. 1978: Principles of categorization. In Rosch, E. and Lloyd, B. B. (eds) *Cognition and categorization.* Hillsdale, N.J.: Erlbaum.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M. and Boyes-Braem, P. 1976: Basic objects in natural categories. *Cognitive Psychology,* 8, 382-439.

Rosen, G. 2008: Abstract objects. In Zalta, E. N. (ed) *The Stanford encyclopedia of philosophy (Fall 2008 Edition).*

Russell, J. A. and Lemay, G. 2004: Concepts of emotion. In Lewis, M. and Haviland-Jones, J. M. (eds) *Handbook of emotions, 2nd edition.* New York: Guilford.

Salemi, M. K. 2008: Hyperinflation. *The concise encyclopedia of economics.* Library of Economics and Liberty, http://www.econlib.org/library/Enc/Hyperinflation.html

Samuels, R. and Ferreira, M. 2010: Why *don't* concepts constitute a natural kind? *Behavioral and Brain Sciences,* 33, 222-223.

Santos, L. R. 2004: 'Core knowledges': A dissociation between spatiotemporal knowledge and contact-mechanics in a non-human primate? *Developmental Science,* 7, 167-174.

Schneider, S. 2009a: LOT, CTM, and the elephant in the room. *Synthese,* 170, 235-250.

Schneider, S. 2009b: The nature of symbols in the language of thought. *Mind & Language,* 24, 523-553.

Schneider, S. In press: *The Language of Thought: A new philosophical direction,* MIT Press.

Scott, K. 2008: Reference, procedures and implicitly communicated meaning. *UCL Working Papers in Linguistics,* 20, 275-301.

Scott, S. 2003: *Non-referring concepts*. PhD Thesis, Carleton University.

Segal, G. 2000: *A slim book about narrow content,* Cambridge, Mass.: MIT Press.

Segal, G. 2009: Keep making sense. *Synthese,* 170, 275-287.

Shepard, O. 1993 [orig. 1930]: *The lore of the unicorn,* New York: Dover Publications.

Smart, J. J. C. 2008: The Identity Theory of mind. In Zalta, E. N. (ed) *The Stanford encyclopedia of philosophy (Fall 2008 Edition).*

Smith, N. V. (ed) 1982: *Mutual knowledge,* London: Academic Press.

Smith, N. V. 2004: *Chomsky: Ideas and ideals,* Cambridge: Cambridge University Press.

Smith, N. V. and Tsimpli, I.-M. 1995: *The mind of a savant:Language-learning and modularity,* Oxford: Blackwell Publishers.

Song, H.-J. and Baillargeon, R. 2008: Infants' reasoning about others' false perceptions. *Developmental Psychology,* 44, 1789-1795.

Southgate, V., Chevallier, C. and Csibra, G. 2010: Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science,* 16, 907-912.

Southgate, V., Senju, A. and Csibra, G. 2007: Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science,* 18, 587-592.

Speaks, J. 2009: The normativity of content and 'the Frege point'. *European Journal of Philosophy,* 17, 405-415.

Spelke, E. and Kinzler, K. D. 2007: Core knowledge. *Developmental Science,* 10, 89-96.

Spelke, E. S. 2000: Core knowledge. *American Psychologist,* 55, 1233-1243.

Sperber, D. 1985: *On anthropological knowledge: Three essays,* Cambridge: Cambridge University Press.

Sperber, D. 1996: *Explaining culture: A naturalistic approach,* Oxford: Blackwell.

Sperber, D. 1997: Intuitive and reflective beliefs. *Mind & Language,* 12, 67-83.

Sperber, D. 2000: Metarepresentations in an evolutionary perspective. In Sperber, D. (ed) *Metarepresentations:A multidisciplinary perspective.* New York: Oxford University Press.

Sperber, D. 2001: In defense of massive modularity. In Dupoux, E. and Mehler, J. (eds) *Language, brain, and cognitive development: Essays in honor of Jacques Mehler.* Cambridge, Mass.: MIT Press.

Sperber, D. 2005: Modularity and relevance: How can a massively modular mind be flexible and context-sensitive? In Stich, S. P., Carruthers, P. and Laurence, S. (eds) *The innate mind: Structure and content.*

Sperber, D. 2011: A naturalistic ontology for mechanistic explanations in the social sciences. In Demeulenaere, P. (ed) *Analytical Sociology and Social Mechanisms.* Cambridge University Press.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G. and Wilson, D. 2010: Epistemic vigilance. *Mind & Language,* 25, 359-393.

Sperber, D. and Wilson, D. 1983: Draft of *Relevance*. Ms., University College London.

Sperber, D. and Wilson, D. 1985/1986: Loose talk. *Proceedings of the Aristotelian Society,* LXXXVI, 153-71.

Sperber, D. and Wilson, D. 1995: *Relevance: Communication and cognition,* Oxford: Blackwell.

Sperber, D. and Wilson, D. 1998: The mapping between the mental and the public lexicon. In Carruthers, P. and Boucher, J. (eds) *Language and thought.* Cambridge: Cambridge University Press.

Sperber, D. and Wilson, D. 2002: Pragmatics, modularity and mind-reading. *Mind & Language,* 17, 3-23.

Sperber, D. and Wilson, D. 2008: A deflationary account of metaphor. In Gibbs, R. W. (ed) *The cambridge handbook of metaphor and thought.* Cambridge: Cambridge University Press.

Sripada, C. and Stich, S. 2006: A framework for the psychology of moral norms. In Carruthers, P., Laurence, S. and Stich, S. (eds) *The innate mind: Culture and cognition.* Oxford: Oxford University Press.

Sripada, C., Stich, S. P., Kelly, D. and Doris, J. In preparation: Norms: Psychology, evalution and philosophy. University of Michigan.

Sripada, C. S. 2008: Nativism and moral psychology: Three models of the innate structure that shapes the contents of moral norms. In Sinnott-Armstrong, W. (ed) *Moral psychology: The evolution of morality, adaptations and limitations.* Cambridge, Ma.: MIT Press.

Stainton, R. 2006: Meaning and reference: Some Chomskian themes. In Lepore, E. and Smith, B. C. (eds) *The Oxford handbook of philosophy of language.* Oxford: Oxford University Press.

Standing, L. G. 1973: Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology,* 25, 207-222.

Stanley, J. 2000: Context and logical form. *Linguistics and Philosophy,* 23, 391-434.

Stanley, J. 2002: Making it articulated. *Mind & Language,* 17, 149-168.

Stanley, J. 2007: *Language in context: Selected essays,* Oxford: Clarendon.

Stukken, L., Verheyen, S., Dry, M. J. and Storms, G. 2009: A new investigation of the nature of abstract categories. Presentation at the *Cogsci 2009: The annual meeting for the cognitive science society.* VU University Amsterdam.

Surian, L., Caldi, S. and Sperber, D. 2007: Attribution of beliefs by 13-month-old infants. *Psychological Science,* 18, 580.

Swoyer, C. 2008: Abstract entities. In Sider, T., Hawthorne, J. and Zimmerman, D. W. (eds) *Contemporary debates in metaphysics.* Oxford: Blackwell.

Szabó, Z. G. (ed) 2005: *Semantics vs. pragmatics,* Oxford: Clarendon.

Thomason, R. H. 2010: Belief, intention, and practicality: Loosening up agents and their propositional attitudes. Paper presented at the *Rutgers University Center for Cognitive Science Colloquia Series*, March 2.

Thomasson, A. L. 2009: Fictional entities. In Kim, J., Sosa, E. and Rosenkrantz, G. S. (eds) *A companion to metaphysics.* Oxford: Wiley-Blackwell.

Thomson, J. J. 1976: Killing, letting die, and the trolley problem. *The Monist,* 59, 204-217.

Tooby, J. and Cosmides, L. 1992: The psychological foundations of culture. In Barkow, J., Cosmides, L. and Tooby, J. (eds) *The adapted mind.* Oxford: Oxford University Press.

Vega Moreno, R. E. 2007: *Creativity and convention: The pragmatics of everyday figurative speech,* Amsterdam: John Benjamins Pub.

Vicente, A. 2010: Clusters: On the structure of lexical concepts. *Dialectica,* 64, 79-106.

Vicente, A. and Martínez-Manrique, F. 2010: On Relevance Theory's atomistic commitments. In *Explicit communication: Essays on Robyn Carston's pragmatics.* London: Palgrave.

Vicente, B. 2005: Meaning in relevance theory and the semantics/pragmatics distinction. In Coulson, S. and Lewandowska-Tomaszczyk, B. (eds) *The literal and non-literal in language and thought.* Frankfurt am Main: Peter Lang.

Vigliocco, G. and Vinson, D. P. 2007: Semantic representation. In Gaskell, M. G. and Altmann, G. (eds) *The Oxford handbook of psycholinguistics.* Oxford: Oxford University Press.

Vigliocco, G., Vinson, D. P., Lewis, W. and Garrett, M. F. 2004: Representing the meanings of object and action words: The featural and unitary semantic space (FUSS) hypothesis. *Cognitive Psychology,* 48, 422-488.

Vigliocco, G., Vinson, D. P., Paganelli, F. and Dworzynski, K. 2005: Grammatical gender effects on cognition: Implications for language learning and language use. *Journal of Experimental Psychology,* 134, 501.

Vigliocco, G., Warren, J., Siri, S., Arciuli, J., Scott, S. and Wise, R. 2006: The role of semantics and grammatical class in the neural representation of words. *Cerebral Cortex,* 16, 1790-1796.

Vinson, D. P. and Vigliocco, G. 2008: Semantic feature production norms for a large set of objects and events. *Behavior Research Methods,* 40, 183-190.

Von Fintel, K. and Matthewson, L. 2008: Universals in semantics. *Linguistic Review,* 25, 139-201.

Wedgwood, D. 2007: Shared assumptions: Semantic minimalism and Relevance Theory. *Journal of Linguistics,* 43, 647-681.

Weiskopf, D. A. 2009a: Atomism, pluralism, and conceptual content. *Philosophy and Phenomenological Research,* 79, 131-163.

Weiskopf, D. A. 2009b: The plurality of concepts. *Synthese,* 169, 145-173.

Weiskopf, D. A. 2010a: Concepts and the modularity of thought. *Dialectica,* 64, 107-130.

Weiskopf, D. A. 2010b: The theoretical indispensability of concepts. *Behavioral and Brain Sciences,* 33, 228-229.

Wharton, T. 2009: *Pragmatics and non-verbal communication,* Cambridge: Cambridge University Press.

Whiting, D. 2007: The normativity of meaning defended. *Analysis,* 67, 133-140.

Whiting, D. 2009: Is meaning fraught with ought? *Pacific Philosophical Quarterly,* 90, 535-555.

Wiemer-Hastings, K., Krug, J. and Xu, X. 2001: Imagery, context availability, contextual constraint, and abstractness. *Proceedings of the 23rd annual conference of the cognitive science society,* 1134–39

Wiemer-Hastings, K. and Xu, X. 2005: Content differences for abstract and concrete concepts. *Cognitive Science,* 29, 719-736.

Wierzbicka, A. 1999: *Emotions across languages and cultures: Diversity and universals,* Cambridge: Cambridge University Press.

Wilks, Y. 2001: The "Fodor"-FODOR Fallacy bites back. In Busa, F. and Bouillon, P. (eds) *The language of word meaning.* New York: Cambridge University Press.

Wilson, D. 1995: Is there a maxim of truthfulness? *UCL Working Papers in Linguistics,* 7, 197-212.

Wilson, D. 2000: Metarepresentation in linguistic communication. In Sperber, D. (ed) *Metarepresentations: A multidisciplinary perspective.* New York: Oxford University Press.

Wilson, D. 2003: Relevance Theory and lexical pragmatics. *Italian Journal of Linguistics,* 15, 273-291.

Wilson, D. 2009: The conceptual-procedural distinction: past, present and future. Presentation at the conference *Procedural meaning: Problems and perspectives.* University of Madrid, October 17.

Wilson, D. and Carston, R. 2006: Metaphor, relevance and the 'emergent property' issue. *Mind & Language,* 21, 404-433.

Wilson, D. and Carston, R. 2007: A unitary approach to lexical pragmatics: relevance, inference and ad hoc concepts. In Burton-Roberts, N. (ed) *Pragmatics.* Basingstoke: Palgrave Macmillan.

Wilson, D. and Sperber, D. 1993: Linguistic form and relevance. *Lingua,* 90, 1-25.

Wilson, D. and Sperber, D. 2002: Truthfulness and relevance. *Mind,* 111, 583-632.

Wilson, D. and Sperber, D. 2004: Relevance Theory. In Horn, L. R. and Ward, G. L. (eds) *The handbook of pragmatics.* Oxford: Blackwell.

Yablo, S. 2002: Abstract objects: A case study. *Philosophical Issues,* 12, 220-240.

Young, D. 2006: *Sound without semantics: A representational approach to the semantic underdeterminacy problem*. PhD Thesis, University of Newcastle upon Tyne.

Zinck, A. and Newen, A. 2008: Classifying emotion: A developmental account. *Synthese* 161, 1-25.