Sławomir Wacewicz

Concepts as Correlates of Lexical Labels. A Cognitivist Perspective.

CONTENTS

PART II

THEORETICAL FOUNDATIONS OF STUDY OF CONCEPTS

PART III
CONTEMPORARY APPROACHES TO CATEGORISATION AND CONCEPTUAL STRUCTURE

Introduction

Objectives and Methods

The primary objective of this work consists in providing a typology and a critical examination of the key contemporary approaches to the topic of concepts and conceptual structure in its relation to categorisation. In the course of the text, I advance and defend two specific main theses. Firstly, concepts – at least for the purposes relevant to cognitive scientific research – are most fruitfully understood as 'lexical categories', in the sense of *mental representations with lexical correlates*. Secondly, concepts, so conceived, have internal structures, contrary to the influential proposal put forward by conceptual atomists. By way of conclusion, I suggest that quantitative categorisation models from other content domains (e.g. perceptual categorisation), such as exemplar models, may be the best suited to revealing the internal structures of concepts.

The other major goal, which can be considered auxiliary, consists in a comprehensive and epistemologically informed discussion of the cognitive perspective on the study of language, its utility and validity. What is worth stressing is the broad construal of 'the cognitive perspective', which embraces but also largely transcends cognitive linguistics. In accordance with the spirit of Cognitive Science, it extends to cover all research that is both founded on strong mentalistic and representational assumptions and relevant to understanding human language processing – thus being open to insights from experimental psycholinguistics, cognitive psychology, neurolinguistics, philosophy of language and mind, as well as a number of related fields.

The character of the present book is theoretical. In view of the breadth of the thematic scope of this work, I pursue the two major goals presented above mostly by way of surveying and synthesising contemporary research in the cognitivist tradition. However, contemporary and historic research from other traditions is presented as well, not just to seek the due theoretical distance that is

necessary for this type of academic work, but also in order to provide a proper background. Despite the theoretical character, in the course of the text I devote substantial effort to grounding the theorising in available empirical findings, whenever such results come as relevant. This work is based (largely but not exclusively) on a review and analysis of literature in the English language that dominates contemporary international research on the topic of categorisation and concepts. Philosophically, it builds to a substantial degree on the theoretical achievements of the Anglo-Saxon analytic tradition (but again, not to the exclusion of other relevant approaches).

At various points in the course of this dissertation, I stop to discuss and clarify matters of terminology. The definitions of several key terms assumed in this work, such as category, concept, and mental representation, are concisely stated in the glossary at the end of this text.

Profile and Scope

This thesis can be classified as having a historical-systematising profile. I put the views of particular influential authors, as well as larger intellectual approaches, into perspective and broken down into components, with the exposition of underlying philosophical commitments. The survey and analysis of contemporary research into the issue of concepts and categorisation are set in the appropriate historical context. This context, however, is necessarily overall rather than exhaustive, for reasons related to the breadth of the issue under consideration: in practice, most large-scale theoretical problems in the history of Western thought can be claimed to have relevance to the question of concepts. I have chosen to give priority to those thinkers whose contributions can be seen as foundational for occidental epistemology or inspirational for later analytic philosophy of mind and language, including Plato and Aristotle, John Locke, Immanuel Kant, and Gottlob Frege.

As for the current intellectual background, the thesis follows closely Noam Avram Chomsky's general philosophical assumptions regarding the nature

of language, i.e. strong mentalism, as well as the crucial methodological postulate of psychological reality[1]. Indeed, spelling out the consequences of the conflict between the mentalistic (internalist) and the non-mentalistic (externalist) perspectives becomes a central motif of this dissertation, discussed in detail in a separate chapter but recurring throughout the text. The views of Ray S. Jackendoff, a linguist with both generative and cognitivist inclinations, are also often referred to in a similar context. Among the key philosophical issues considered in this work that have been developed by contemporary analytic philosophy are those related to the ontological status of conceptual contents – a question which leads to a polemical discussion with the argumentation advanced by Hilary W. Putnam. Finally, the scrutiny of the current empirical research regarding categorisation focuses on the experimental findings from cognitive psychology, most prominently those by Eleanor Rosch (formerly Heider) and her collaborators and continuators, as well as the group of researchers associated with Douglas L. Medin.

Perhaps the most central researcher in the context of this dissertation is the linguist and philosopher Jerry Alan Fodor, for the past two and a half decades affiliated with Rutgers University. This prominence results from both the personal importance of Fodor as a leading cognitive scientist and philosopher of Cognitive Science, and from the relevance of multiple threads of his research. Fodor's views are quoted and discussed regarding several main issues of this work, such as the internalistic perspective in the study of language, the methodological soundness of interdisciplinarity, the ontology of concepts, and the requirements on a theory of concepts. What is more, Fodor's atomistic theory of conceptual content, often seen as a major contender, is reported and then critically addressed in a separate chapter.

Structure

[1] While remaining noncommittal on several more detailed premises, such as the autonomy of syntax or the existence of Universal Grammar.

This thesis assumes a three-part organisation, with the parts devoted, respectively, to the perspective of study, the object of study, and the analysis of the relevant theoretical approaches to the issue of conceptual structure and categorisation.

Part I – Research Perspective

The first part of this work aims at the presentation of the research perspective as well as a distanced discussion by way of contrasting it with viewpoints external to it. The initial chapter has a preparatory character, having as its objective an introduction of Cognitive Science; most importantly, in the historical aspect of its development over the past several decades, as well as more contemporarily, in the aspect of its relation to the cognitive study of language. It also sets up and critically examines the *representational* and *interdisciplinary* context relevant to the remaining part of this work.

I trace back the history of Cognitive Science to its birth from the research on Artificial Intelligence (Alan M. Turing and others) and memory (George A. Miller), and most importantly, the linguistic as well as philosophical contributions of Noam A. Chomsky. Two ways of understanding Cognitive Science are presented, with the first one, concentrated on the study and simulation of symbolic, computationally explicit processes, being now complemented with – and to an extent replaced by – a different approach, stressing the importance of a bodily and environmental context of cognition, as well as the role of nonsymbolic representational format. There follows a diagnosis of the present status of Cognitive Science, and in particular of the question of its interdisciplinarity, leading to a suggestion that the canonical descriptions of Cognitive Science in terms of its member disciplines fail to do justice to its present nature.

The issue of interdisciplinarity is explored in more detail, with focus placed on the methodological reservations often raised against it. After acknowledging some of the risks associated with it, I defend the idea of

interdisciplinary cooperation, both in general and specifically in the context of the study of the mind. Crucially, I intend the section on the strengths of interdisciplinarity to highlight the mutual relevance of cognitive linguistics (narrowly construed) and Cognitive Science: especially, how data from widely different disciplines of Cognitive Science can enrich, complement and validate purely linguistic data. An important role in this context is played by the examples of actual research; in particular, the examples inspired by George P. Lakoff's study of conceptual metaphor are backed up by several layers of converging empirical nonlinguistic evidence from a range of disciplines.

Chapter Two of the present dissertation seeks to substantiate, on independent grounds, a crucial research decision, that is the assumption of the intrasystemic understanding of concepts and categories. The intrasystemic perspective is evaluated as an alternative to the more routinely taken externalistic perspective. The guiding motivation behind this thread is the avoidance of the *petitio principii* fallacy, i.e. the validation of the intrasystemic standpoint merely on the basis of its being a necessary consequence of the presupposed cognitivist commitments.

I formulate this theoretical problem referring mainly to the framework set up by Noam Chomsky. The rivalling, externalistic perspective is then introduced, leading to the discussion of the reasons for the understanding of concepts as nonmental, abstract beings existing independently of individual minds. Gottlob Frege's influential account is presented in order. I explain the motivations behind his antipsychologism but resist the construal, common in the literature on concepts, of concepts as entities ontologically corresponding to Fregean senses.

The next step in the discussion of the perspective of study consists in the exposition of the overarching debate between externalism and internalism of conceptual content. Particular attention is devoted to a meticulous treatment of terminological distinctions, with a view to avoiding frequent misunderstandings resulting from the terminological intricacies in this area. The presentation of the content of the externalistic doctrine is based on the central example of Hilary

Putnam's "Twin Earth" thought-experiment. The rest of the chapter serves to spell out the consequences of such a position: extant and novel arguments against it are combined, ultimately leading to the rejection of this view, and thus reinforcing the internalistic position.

Part II – Object of Study

The second part of the thesis is concerned with the object of study, that is the topic of concepts and categorisation. These are introduced and depicted in a possibly general and theory-neutral way before being approached specifically from the cognitivist and mentalistic point of view adopted in this work. Terminology, again, plays a central role, and terminological decisions are carefully justified.

Chapter Three deals with the key notions of the thesis: concept, categorisation, mental representation. A maximally broad construal of the notion of concept is offered as a starting point, with an extensive list of conditions of 'concepthood' imposed by different theoretical outlooks; it serves as a broad background for the subsequent delimitation of the scope of study in Chapter Four. An important interim conclusion of this part of the work is that at least some of the criteria of concepthood might be impossible to reconcile within a single research perspective.

The notion of categorisation is treated in a more historical way, but in this case, too, a broad and inclusive construal is established. The fundamental role of (so broadly understood) categorisation for all kinds of cognitive activity is highlighted. Important in this context is the acknowledgement of the continuity between higher-level, linguistics categorisation and low-level, perceptual categorisation. With respect to the notion of mental representation, the most significant task to be achieved is the juxtaposition of the traditional philosophical understanding of this term with a modified and more contemporary cognitivist one, proving more functional in the area of today's Cognitive Science.

Chapter Four is pivotal to the construction of the entire thesis. In this chapter, I make and substantiate in detail several decisions related to terminology. Furthermore, I delimit the exact scope of this work to concepts as understood by Cognitive Science, that is considered from the mentalistic perspective. Most importantly, I advance the central argument regarding the nature of the relation between concepts and word meanings.

The proper scope of this work is restricted to categorematic concepts, in particular such lexical concepts that are expressed by nominal lexemes containing a single lexical morpheme. The special status of categorematic concepts – their psychological reality as a separate category – is documented based on empirical data from psycholinguistic and neurolinguistic research. Subsequently, the central ontological assumption derived from the cognitivist perspective is formulated: concepts and categories are understood here in an internalistic and individualistic way, as mental entities having the nature of representations. After reviewing a set of possibilities present in the literature on this subject, the definitional relation between concepts and categories is established in the following way: concepts are those categories that possess a lexical correlate (which can be understood as an entry in the mental lexicon).

The numerous theoretical problems resulting from the decisions described above are addressed in order; among them the controversial consequence that the cognitive systems of nonlinguistic organisms are denied concept possession. The wealth of mental representations unequivocally ascribed to such cognitive agents can be reinterpreted in terms of nonconceptual content, so that these systems can be said to have mental representations, but not fully fledged concepts. I adduce a broad range of evidence from linguistics (analyses by Ray Jackendoff, Steven A. Pinker, and others), psycholinguistics (Elisabeth S. Spelke and collaborators, Susan Carey), and general Cognitive Science (Andy Clark) that is intended to support this distinction as a factual rather than purely nominal one – a distinction that reflects the actual ontogenetic influence of language acquisition on the development of the conceptual system. Another section is devoted to showing

that the proposed direct linking of the conceptual repertoire to the lexicon need not produce strongly Whorfian consequences. This linking, however, makes it possible to deal with one of the most refractory problems faced by mentalistic theories of concepts, namely that of the shareability of concepts.

Tying together concepts and lexical items in such a straightforward, but principled way is a novel proposal that shows promise for a more rigorous use of 'concept' as a theoretical term with a unified meaning across the Cognitive Sciences.

Part III – Analysis of Theoretical Approaches

The third part of the present dissertation considers the particular approaches to of conceptual content in the aspect of categorisation. Accordingly, it constitutes the bulk of this dissertation. Chapter Five comprises a review of the classical theory of categorisation. This review is accomplished mostly from a historical position; however, it leads to conclusions regarding the present utility – or, more precisely, the severe limitations – of this approach. The discussion of the relevant views of the major figures in the history of philosophy – from Plato and Aristotle to the British empiricists to the logical positivists – is aimed at illustrating the unrivalled historical dominance of the generalised 'classical approach' to concepts and categorisation; it is then complemented by summing up the major modern directions of criticism of this stance (notably, by the philosopher Ludwig Wittgenstein and the linguist William Labov). This necessary review of well-known historical positions is followed up by an extended critical commentary and re-evaluation of the classical view. In those sections, I reveal its certain hidden ontological assumptions (being a possible reason for its incompatibility with the cognitivist perspective), and secondly, argue against the attempts to restore its utility for the cognitivist conceptions – the issue of psychological essentialism being perhaps the only viable area for its revival. I also underscore one specific theoretical problem – the problem of the *format of features* into which a concept is decomposed, as opposed to the way of (de)composition.

Chapter Six brings the analysis – and later, a refutation – of the influential position of conceptual atomism, championed most prominently by Jerry Fodor. Both the discussion and the rebuttal of this particular standpoint are important because it constitutes a major contender theory with respect to conceptual content – one that remains incompatible with mainstream Cognitive Science. Concepts, it is argued in Chapter Six, are most fruitfully construed as entities possessing complex internal structures, contrary to the atomistic position.

The substance of the first part of this chapter is comprised of the presentation of Fodor's rich and interconnected doctrine regarding the nature of mind and concepts. Among the main topics reviewed are those of folk psychology, broad and narrow mental content, modularity of mind, nativism, and language of thought (mentalese). The establishing of such a context allows the atomistic view to be seen, not as an isolated theoretical position, but rather as a direct consequence following naturally from the above doctrine. In the polemical treatment of conceptual atomism, simplistic arguments sometimes levied against Fodor's view are discarded. The criticism is focussed on the relative fruitlessness of this theoretical outlook, rather than its falsehood in any more absolute sense.

Chapter Seven concludes the dissertation. The theoretical assumptions as well as advantages of the so-called similarity-based approaches to concepts/categorisation are discussed. The exemplar view is suggested as an underestimated approach that maximises the potential advantages of the similarity-based approaches. The other general similarity-based view, the prototype approach, is not considered in detail. Rather, the discussion is focussed on shedding light on the underlying tenets of this broad group of views, in particular, on elucidating the role of the notion of similarity. Secondly, the differences between the very popular prototype view and the relatively undervalued exemplar view are spelled out. Thirdly, the prospects of applying exemplar-based models specifically to the tasks of modelling lexical categorisation are considered in the convention of research postulates.

Acknowledgements

I thank Professor Piotr Stalmaszczyk. I would like to express not only my gratitude for the years of tutoring and patient support during my PhD studies in Linguistics, but also my deep admiration for him as an academic. I thank Professor Aleksander Szwedek for inspiration and sparking my interest in Cognitive Science. I extend my sincere gratitude to staff members and colleagues from the Department of Philosophy; in particular, I thank Assistant Professor Tomasz Komendziński for his kindness and generosity in sharing the resources of his legendary library, and Professor Urszula Żegleń for the intellectually formative years under her excellent tutoring. I thank Professor Zdzisław Wąsik, whose comments on the earlier draft of this thesis were penetrating, but also most enlightening. Words of gratitude are also extended to the authorities of the Department of English, Professor Mirosława Buchholtz and Professor Waldemar Skrzypczak, as well as to Professor Przemysław Żywiczyński, for their invaluable advice and encouragement. Finally, I thank my Parents for their unfailing support.

PART I

INTERNALISTIC PERSPECTIVE ON LANGUAGE IN COGNITIVE SCIENCE

In the first part of this work, I consider questions related to the perspective of study. Chapter 1 is devoted in full to the presentation of Cognitive Science – including but substantially exceeding cognitive linguistics as its specific subfield – and to discussing the vantage point that it provides for the study of language. In my exposition of Cognitive Science, I begin with identifying its guiding theoretical assumptions and tracking down its historical roots, before sketching out its contemporary picture. In order to supply an externalised theoretical perspective, I examine several lines of criticism against Cognitive Science; in particular, I discuss the hazards of interdisciplinarity, and the reasons for which they are outweighed by its benefits.

The function of the second chapter is to demonstrate the validity and robustness of the mentalistic/internalistic perspective that lies at the heart of Cognitive Science. The character of this part of my thesis is primarily philosophical, bringing into focus metatheoretical issues indigenous to the fields of philosophy of language and philosophy of mind. I diagnose the reasons for the historically dominant character of the alternative view on the ontological status of language and the methods of its study, while not failing to notice the complementary rather than rivalling character of those two approaches. Finally, I critically address the position known as externalism regarding conceptual content. The specific sub-goal of this part of my work is a rebuttal of Hilary Putnam's thought experiment, which provides the chief motivation for the anti-cognitivist consequences of the externalist position.

1. History and profile of Cognitive Science

1.1. Introduction

All academic effort derives its meaning and significance from being related in systematic ways to a larger body of research. Thus, it is incumbent on the author to define their undertaking against the background of a larger-scale tradition. Consciously locating one's inquiry in a broader scientific landscape gives the project its identity, necessary for a number of reasons. The researcher inherits a 'frame of mind': an intellectual legacy that, although partly implicit, always forms scaffolding for the progress of further research, and provides one with an indispensable toolkit of methods by which to arrive at the solution of outstanding problems. Another important factor is the awareness of long-term research goals. The presence of such a long-term objective, even one that may seem distant and illusionary, ensures that one's work does not become what is known as 'mere Baconian fact-gathering' or 'porcupine research'[2], but – if indirectly – helps achieve some eventual utility.

The framework of the present modest work can best be described in most general terms as *contemporary Cognitive Science*. The rationale for the choice of such a broad paradigm as the background has to do with the very nature of Cognitive Science as a superdiscipline, a theme that will be dealt with more extensively in the following sections.

---

[2] Sir Francis Bacon (1561–1626), English philosopher and politician, is commonly considered as an early precursor of the approach to science that heavily emphasised bottom-up, inductive way of gathering information, at the expense of neglecting the overall guiding theoretical perspective.

The term 'porcupine research' is an informal derogatory expression for a type of study consisting in repeated replication of a research procedure with only slightly changed experimental conditions, thus generating very little new insight.

The general discussion provided in this chapter is necessary for several reasons. Firstly, the cognitivist perspective on language is sometimes taken to be coextensive with cognitive linguistics. As is explained in the sections to follow, broadly construed cognitive linguistics – despite having played a pioneering role in laying the foundations for the development of Cognitive Science at large – constitutes only one of its several major components. Secondly, over the five decades of its development Cognitive Science has undergone a slow but systematic transformation, having evolved into a field of study quite different from the original Cognitive Science from which it should be distinguished. Thirdly, the foundational principle of interdisciplinary collaboration, although very firmly established in Cognitive Science, continues to be seen by some as methodologically suspect – a criticism that is raised particularly frequently during conference panel discussions.

Consequently, in this chapter I undertake to address the questions of:

- what contemporary CS is,
- where it comes from (both directly and in terms of general intellectual legacy),
- how it locates itself on the landscape of views regarding human cognition
- what goals it strives to achieve,
- how it is related to the study of concepts,
- why it remains a methodologically sound enterprise

1.2. Cognitive Science: definitions and basic assumptions

Cognitive Science (sometimes: *the Cognitive Sciences*; commonly abbreviated in literature to *CS* or *CogSci*) is notoriously resistant to definition except on a very high level of generality. Reasons for such elusiveness might include the high dynamics of its development over the past several decades and at present, and its cutting across the traditional boundaries of academic disciplines. It appears that Cognitive Science is in fact better explained through

description and instantiation, and not through stipulation. Symptomatically, the single most authoritative reference work in the field, *The MIT Encyclopedia of the Cognitive Sciences* (edited by Wilson and Keil, 1999), avoids formulating a definition, and the editors "prefer to let the volume speak largely for itself" (p. xiii). One may also remember that if, so to say, definitions dislike Cognitive Science, the relation is reciprocal: it is a tradition closely associated with Cognitive Science that has brought about the rejection of definitions as the exclusively adequate modus operandi for characterising the meaning and use of natural language concepts (a point developed in detail in Chapter 5). Thus, this field might be more accurately conceived of in terms of 'family resemblance', with its many constituent disciplines, methodologies and goals brought together through a network of relations, without a precise set of universal features shared by all 'members'. By no means does this make the notion vacuous; (contemporary) Cognitive Science retains the necessary level of integrity for this name to be meaningful, functional, and intuitively clear.

The adjective 'contemporary', meant in the sense of Lakoff and Johnson (1999), is important. As indicated before, a problem working to the same effect of making precise definitions unwieldy is the fact that Cognitive Science is a living organism in the process of development. During the half of century of its history, Cognitive Science has undergone certain changes – although without being transformed into something radically different – and some narrower uses of this name, making reference to a more specific tradition, might differ from its present, broader understanding. This point will receive a more adequate treatment in 1.4.5. Here, I will depict and, later, relate my research to Cognitive Science in its present shape.

The initial sentence of the entry 'Cognitive Science' in *The Stanford Encyclopedia of Philosophy* (the largest and most up-to-date specialised reference work in philosophy) captures most succinctly the overall nature of the article's subject (Thagard 2006):

> Cognitive Science is the interdisciplinary study of mind and intelligence, embracing philosophy, psychology, artificial intelligence, neuroscience, linguistics, and anthropology.

A Companion to Cognitive Science, although offering an introductory formulation (Bechtel et al. 1998: 3):

> Cognitive Science is the multidisciplinary scientific study of cognition and its role in intelligent agency. It examines what cognition is, what it does, and how it works[,]

immediately qualifies it as 'premature', and complements it with a longer list of provisos. In *The Blackwell Dictionary of Cognitive Psychology* Cognitive Science is described in a somewhat less concise, though by no means comprehensive, manner (Eysenck 1990):

> Cognitive Science refers to the interdisciplinary study of the acquisition and use of knowledge… [it] was a synthesis [of computer science, information processing psychology, and generative linguistics] concerned with the kinds of knowledge that underlie human cognition, the details of human cognitive processing, and the computational modeling of those processes. There are five major topic areas in Cognitive Science: knowledge representation, language, learning, thinking, and perception.

All the above (and many other) examples converge, at the most schematic level, on a common superdefinition of Cognitive Science as the *interdisciplinary study of mind*. Thus, two terms emerge as fundamental:

(1) *mind* as the object of scientific study, and
(2) *interdisciplinary cooperation* as a critical methodological postulate.

The immediate elaborations on, and instantiations of, the two notions include, for (1), intelligence, thought, knowledge, mental processes, information processing, perception, conception and memory, the acquisition, storage, and use of mental representation, etc.; and for (2), the list of subdisciplines with linguistics, psychology, philosophy, neurology, and artificial intelligence (conflated with computer science for terminological convenience) recurring as the core scientific branches.

1.3. Basic tenets of Cognitive Science

Extending on the above minimal definition, it appears to be possible to list a number of traits that jointly provide a framework of CS's fundamental tenets. Basing on the literature reviewed in the sections to follow, I suggest several assumptions whose status appears to be central. Cognitive Science:

a) deals with the subject matter of *cognition*, or *the mind* of an individual,
b) broadly defines the mind as a *representational* system that processes information,
c) is profoundly *naturalistic* in its approach to the mind,
d) claims that the mind, although necessarily *physically implemented*,
e) can be described on several *levels*,
f) in the relation between the cognitive agent and its environment it focuses on the *internal processes* of the former,
g) is founded on the idea of interdisciplinary collaboration

Those views, although themselves capacious and potentially heterogeneous, are not universally agreed upon, and sometimes even contested within the field; nevertheless, they can be given as the guiding mainstream beliefs upon which Cognitive Science rests. In the next sections, I add more substance to the above skeletal description by singling out and the scrutinising CS's basic assumptions. The discussion to follow is intended as a contrastive, picturing Cognitive Science

in its relation to the alternatives offered by other outlooks. A larger section is devoted to the historical development of Cognitive Science and the transition from earlier to more contemporary trends. The discussion of interdisciplinarity is postponed until the end of this chapter so that it follows naturally from the historical considerations, with which it is closely connected.

1.3.1. Cognition

Cognition is, again, a rather broad term with rich and varied connotations in different intellectual approaches. Most generally, *cognition* (from Latin *cognoscere*, to learn, to know, or to recognise) refers to the process of acquisition or transformation of knowledge by an intelligent subject, or to the results of this process. Antoni Podsiad (2000: 652–654) distinguishes cognition (Polish: *poznanie*) as a process (cognising) and cognition as a product (the contents of mind). Cognition as a process is further divided by Podsiad into direct and indirect, and into theoretical, ethical, and creative; prototypically, cognition is direct and theoretical. According to this author, human cognition is characterised by being conscious (and sometimes reflexive), assimilative, intentional, and aspective.

As *obiectum reale*, cognition appears to be extremely elusive. Historically, the study of cognition may be considered as largely identical to the study of knowledge or 'the processes of the mind'. In the occidental tradition, cognition has generally been construed as individualistic, representational and linguaform. For example, Descartes'[3] "*cogito ergo sum*" expresses a solipsistic intuition that the only thing know to the subject are their own (internal, individual) thought processes; what is more, although *cogito* is supposed to capture a primitive, prelinguistic intuition, it can be expressed only in language. To John Locke[4], knowledge "is the perception of the agreement or disagreement of two ideas. Knowledge then seems to me to be nothing but the perception of the connexion

[3] René Descartes (1596–1650), Renatus Cartesius, French philosopher and mathematician.

[4] John Locke (1632–1704), an English philosopher and anatomist.

of and agreement, or disagreement and repugnancy of any of our ideas. In this alone it consists"[5] (1999 [1960]: 515). This underscores the representational character of cognition, which always has to be mediated by 'ideas'. The great conceptual contribution from Immanuel Kant[6] was the appreciation of the active, constructive role of the subject in the process of cognition.

More controversial than the 'shape' of cognition has been the *source* of cognition. For instance, in Plato's nativistic doctrine, 'true' cognition, that is the kind of cognition leading to properly understood knowledge, was essentially re-cognition, or *anamnesis*, that is the rediscovery of the knowledge already (latently) present in one's memory[7]. In contrast, the empiricist position is summed up in a widely cited passage from Locke (1999: 87): "All ideas come from sensation or reflection. Let us then suppose the mind to be, as we say, white paper, void of all characters, without any ideas:—How comes it to be furnished?... To this I answer, in one word, from experience". The debate between the nativistic and empiricist factions still continues in today's Cognitive Science.

Cognition was traditionally thought of as rational rather than driven by emotion, passive rather than creative, and constrained to 'higher' mental processes such as reasoning or planning, as opposed to 'low-level' mental processes such as perception, or bodily processes such as sensation or motor control[8]. As a result, it was also considered as a human rather than animal trait.

---

[5] Discussed in Dębowski (2000: 35).

[6] Immanuel Kant (1724–1804), a German philosopher.

[7] Plato (427–347 BC), a Classical Greek thinker, one of the classics of the Western intellectual tradition. Author of the theory of anamnesis. See Tatarkiewicz (2003: Vol. 1. 59–74).

[8] Cf. the entry "cognition" in *Oxford Companion to Philosophy* (ed. Ted Honderich): "…[t]raditionally this has been regarded as the domain of thought and inference, marking the contrast with perceptual experiences and other mental phenomena such as pains and itches. Sensations, perceptions, and feelings are all distinguished from episodes of cognition since they provide input to the domain of thinking and reasoning but are not thoughts themselves. More recently, cognition has been conceived as the domain of representational states and processes

More currently, all of the above distinctions are beginning to be seen as arbitrary, and the study of emotion, of constructive character of mental processes, 'animal cognition' and 'motor cognition' form important subfields of contemporary Cognitive Science[9].

On the present analysis, contemporary CS assumes cognition to consist in information processing (as such being amenable to a quantitative approach and formal analysis), to be representational, to be a natural rather than supernatural phenomenon and an individual rather than collective phenomenon. These assumptions are critically examined below.

1.3.2. Representationism and presentationism

An absolutely central assumption of Cognitive Science is that cognition is representational in nature, with mainstream CS being organised around a more specific, symbolic representation format (see e.g. Steven Pinker 1995 [1994]: 55–82). Representationism, however, is only one view on the character of cognition, its most prominent alternatives being:

- behaviourism
- the cybernetic perspective
- presentationism

---

studied in cognitive psychology and *Cognitive Science. These are phenomena involved in thinking about the world, using a language, guiding and controlling behaviour. The new definition embraces some aspects of sensory perception where this involves representations of a spatial world and the intelligent processing of sensory input." (1995: 138)

This is also testified to by the apparent transition from contrastive to inclusive use of 'cognition' and 'perception' in cognitive-scientific literature.

[9] See, e.g. the relevant entries in *The MIT Encyclopedia of the Cognitive Sciences* (1999, ed. Wilson and Keil).

Behaviourism[10] is a doctrine that used to be popular in the first half of the twentieth century (more in the US than in continental Europe), according to which entities that by their nature are not objectively observable by the researcher are not a legitimate object of scientific study. George Botterill and Peter Carruthers (1999: 4–20) remark that it is necessary to distinguish between logical behaviourism that denied the *existence* of mental states[11], and methodological behaviourism that merely denied the possibility of scientific study thereof. Both versions of behaviourism can be thought of as strongly antirepresentational.

A different approach comes from cybernetics. For example, Aaron Sloman (1993: 70) proposes that the mind be regarded as a control system, "involving many interacting control loops of various kinds, most of them implemented in high level virtual machines, and many of them hierarchically organised". Such a rendering captures the crucial intuition that a mind is a mind only in so far as it belongs to some concrete cognitive agent, whose meaningful, goal-oriented behaviour it controls (see section 1.4.5.4.). Still, this perspective is not broadly popular in contemporary literature.

Józef Dębowski (2000) scrutinises a general perspective that most directly opposes representationism, that is *presentationism*. Presentationism opposes the mainstream view of modern philosophy that cognition is mediated by 'ideas' that stand for the object of cognition in the external world. Paradoxically, one strand of presentationism is extreme idealism: scepticism about the existence of external reality and considering *ideas themselves* as the only (both direct and ultimate) possible object of cognition (Dębowski 2000: 47). But a more standard version of

[10] The term 'behaviourism' was proposed in 1913 by the American behaviourist psychologist John B. Watson (1878–1958), who based on earlier insights from the Russian physiologist Ivan Pavlov (1849–1936) and American zoologist Herbert Spencer Jennings (1868–1947).

[11] Particularly important in this context was Gilbert Ryle's (1951 [1949]) *The Concept of Mind*, where he aimed at showing that predicating existence of mental states was committing a category-mistake (see section 3.2.2.).

presentationism relies on the elimination of the notion of idea as an intermediary, with cognition construed as direct and/or intuitive.

Thomas Reid[12] is considered to be the leading proponent of presentationism in modern philosophy. However, it is important to distinguish between the representational theory of *knowledge* and the related but distinct representational theory of *perception*. The representational theory of perception is a doctrine maintaining that our perception of objects, rather than being direct, is mediated by some third-party beings external to both the subject and the actual objects. Those intermediaries in the process of perception were assumed to represent the objects (hence the name), and were variously construed, e.g. as *simulacra* by Lucretius[13] (see also Dębowski 2000: 27–29) or *sense-data* by G. E. Moore[14].

Finally, one should add that the very notion of *representation* (hence, mental representation) itself turns out to be quite problematic. Traditionally, representation is a kind of sign that stands for ('represents') some element of the external world. The 'external world', however, is a dubious entity both in terms of its aprioristic partitioning into 'elements', and reliable cognitive access to it; this provokes reservations against mental representation so conceived, especially from cognitive linguists (see 3.3., 4.2.3.). Also, this notion is much too narrow for Cognitive Science. *Stanford Encyclopedia of Philosophy* agrees that a broader construal is necessary, embracing "…thoughts, concepts, percepts, ideas, impressions, notions, rules, schemas, images, phantasms, etc. — as well as the various sorts of <subpersonal> representations postulated by Cognitive Science." (Pitt 2006) I would argue that the only truly important criterion is in fact *stability*: for example, Edward Nęcka et al. (2006: 26–27) conclude that any

[12] Thomas Reid (1710–1796), a Scottish philosopher, proponent of "common sense" philosophy.

[13] Titus Lucretius Carus (99–55 BC), a Roman poet and materialistic philosopher.

[14] George Edward Moore's (1873–1958), a British philosopher.

cognitive structure whose identity through time remains stable enough for it to be consistently deployed and redeployed counts as a mental representation, and mental processes are defined in terms of operations on those structures.

Thus, somewhat confusingly, (broadly defined) mental representations need not actually represent any entities from the extrasystemic reality. What a mental representation is, essentially, is *any temporally stable entity or postulated entity, individuated on the representational level of description* (see below) of the cognitive system. Section 3.3. provides a more detailed description of mental representation as understood specifically in the context of this book.

1.3.3. Naturalism and physical character of mind

Cognitive Science's approach to mind is strictly naturalistic, and it fully subscribes to the naturalistic consensus in the philosophy of mind. At a minimum, this equals a belief that the mind is a purely natural phenomenon that can be exhaustively studied with the methods of natural sciences, and its full explanation is possible, at least in principle, without any resort to non-natural explanatory categories. A very important immediate consequence is the straightforward rejection of ontological mind-body dualism (i.e. of Cartesian dualism). Even though functionalism – historically and even currently the most influential approach within Cognitive Science – abstracts away from the details of the physical implementation of the mind, it is still a rather obvious axiom that the mental must be strictly dependent on the underlying physical substrate (the Mind-Body problem[15]).

In the philosophy of mind, the precise nature of this dependence is debated (see Żegleń 2003 for an exhaustive discussion, or Jackendoff 1992 [1987] – Chapter 1, for a shorter review of particular positions). Still, it is commonly

[15] Some, e.g. Chomsky (2000: 109–110), point out that the Mind-Body problem is misguided: for this question to be intelligible there would have to exist a satisfying theory of the 'Body' (the physical substrate), which we do not yet have. This, however, remains a minority opinion.

assumed that, at a minimum, it must amount to some form of *supervenience*[16]. The specific technical statement of supervenience can itself take several forms, but a common core idea it expresses is that no difference on the mental level – e.g. between two hypothetical minds – is possible without a difference on the level of the physical substrate; identity of the physical ensures the identity of the mental. Contemporary Cognitive Science, however, manifests a growing bias towards more straightforwardly reductionist or eliminative accounts, more in tune with (and very clearly influenced by) today's neuroscience.

Similarly, Cognitive Science shows considerably less interest in the refractory 'philosophical' questions that by their nature hold little promise of being successfully approached with scientific methods. For example, the problem of *qualia* (the ineffable, intrinsic, subjective, first person, 'what's-it-like' quality of individual conscious experience – e.g. the special qualitative 'feel' of the redness of a perceived red apple), while recognised as a fully legitimate problem within the philosophy of mind, is excluded from, or at best peripheral to, mainstream Cognitive Science[17].

1.3.4. Levels of description

Already implicit in the above passage is the claim that cognitive systems can be analysed on several levels, a claim that most authors (e.g. Stillings et al. 1995: 7–8) consider as central to Cognitive Science. The mind, understood in representational terms in concord with 1.3.2., is one of them (the intermediate),

[16] The term "supervenience", though dating back to Aristotle, was reintroduced into the modern academic context by the American analytic philosopher Donald Davidson (1917–2003). For details see Żegleń (2003: 75).

[17] The topic of qualia is a major issue in the philosophy of mind, whose best known and most accessible illustration is Frank Jackson's (1982) example of "Mary The Color Scientist". However, I will not develop it further because of its marginal importance for Cognitive Science at large.

the physical implementation discussed above is another, the third level being the system's behaviour.

Cognitive Science commonly utilises the schema of three levels of description that are arranged in a hierarchy and are at least in principle translatable into one another – despite the present explanatory gaps. With regard to humans, the behavioural level is implemented at the mental one, and this in turn, at the physical, that is in the biology of the brain (cf. e.g. Green et al. 2000 [1996]: 5–7). Thus, the brain 'realises' and causes the mind, and the mind causes behaviour. This mirrors a more general, well-known distinction, motivated by the computer metaphor and devised by David Marr: into computation, algorithm and physical rendering (discussed in Galton 1993: 122). What individuates and separates the levels is – rather than the differences in their ontologies – the presence of explanatory gaps between them, leading to the application of its own generalisations and methods of investigation for each of the three levels. The striking consequence of a partial autonomy of each of the levels is a theoretical possibility of abstracting the mental and reconstructing it without the necessity of exactly duplicating the intricacies of the underlying physical substrate. This insight is at the heart of functionalism and served as a chief inspiration for Artificial Intelligence.

It must again be emphasised that the assumption of distinct levels of description need not, and should not, lead to any ontological claims about the existence of distinct 'levels of reality'.

Table 1 Three-level research program

Level	*Questions*
Task	How are natural languages structured? What must people know and what must they know how to do in order to produce and understand human speech?
Algorithmic	How is knowledge of language represented in the mind? What computational processes are involved in producing and understanding speech?
Physical	How are these representations and computational processes implemented in the hardware of the brain?

Table 2 Sources of evidence for the three levels

Level	*Example sources of evidence*
Task	Judgments of native speakers
	Which strings of words are grammatical and which are not?
	What meanings can a sentence have and not have?
Algorithmic	Developmental psychology
	How do children acquire language?
	What are the common patterns of language development?
	Cognitive psychology
	How do adults react to linguistic stimuli under controlled conditions?
Physical	Clinical studies
	What kinds of brain injuries and diseases cause language deficits?
	What specific language deficits are caused by specific brain injuries and diseases?
	Anatomical and functional studies
	What parts of the brain are involved in language use?
	How are these parts interconnected?

Fig. 1. Levels of description in Cognitive Science with regard to language. Source: Sam Scott 2006: 557–558.

1.3.5. Internalism (Individualism)

A central property of cognition is that it concerns the mental processes contained within the mind of an (idealised) *individual*. Cognitive Science has inherited this individualistic assumption, whereby the environment, and especially the social plane is seen as secondary and resultant, and the explanations of the totality of mental phenomena, including the knowledge of others, must be given in terms of the operation of a self-contained cognitive system of a single agent. This principle is known as *methodological solipsism*[18].

[18] *Solipsisim* (from Latin *solus* – alone, and *ipse* – self) – a philosophical stance according to which the cognising subject is itself the only possible object of cognition (or even the only existing object).

Methodological solipsism is a principle stating that the only legitimate object of study in scientific psychology (and Cognitive Science) are the inner (mental) states of an individual, effectively proceeding as if the world external to the individual did not exist. The term was coined by Hilary Putnam, and the principle itself is developed at length by Jerry Fodor (1981).

In one sense, the traditional Cognitive Science actually grew from the rejection of the external world: that is, not so much of its existence as its relevance. Especially on the 'mind is computation' view all that mattered was taken to be located inside the cogniser's head or, more precisely, his computational mind (in case the cogniser should have a CPU rather than a head). This view is known by the name *internalism* (or *individualism*). Later, with increasing prominence of new disciplines such as situated robotics, there came the realisation that the idealised requirement of complete isolation from the environment was overstated. For example, the proponents of the *extended mind* theory, Andy Clark and David J. Chalmers (1998), point out that drawing the precise border between the cognitive system and its environment is normally far from obvious (see section 2.2.2.); and Bechtel et al. (1998: 86–97), in their introduction to the *Companion to the Cognitive Science*, conclude that a more promising perspective on exploring the mind is in the process of its interaction with the external world.

Nonetheless, the focus on the cognitive agent's inner states must remain a key dogma. At least from the point of view of Artificial Intelligence, computer modelling, and robotics, this remains the only feasible approach; otherwise the modelling of an intelligent system would be insurmountably difficult. As explained by Stevan Harnad (2002), unless the task of the modeller is just making the programme or robot, his task becomes effectively to make a model of the entire world.

The internalistic commitment forces one to reflect upon the relation of Cognitive Science to the external world, especially in the context of a scientific perspective. Such a stance might, at least *prima facie*, be difficult to square with Cognitive Science's strong aspirations to qualify as a genuine *science*. This is so because 'doing science' is sometimes thought to be based, if implicitly, on deeply realistic underpinnings, that is on the belief in existence of objective, cognitively penetrable external reality (and possibly on an even stronger epistemological position of naïve realism). The relation between the

independently cognising subject and the reality external has been worked out in *constructivist approaches*. Founded on the theorising of Immanuel Kant and experimental work by Jean Piaget, constructivist theories have been more currently developed by George Alexander Kelly[19], Peter Berger and Thomas Luckmann, Ernst von Glasersfeld[20], and John Searle. Zdzisław Wąsik (2006: 23) summarises the approach of social constructivism, pointing to the level of shared social reality collectively constructed by individual subjects:

> To the extent that one person employs a construction of experience which is similar to that employed by another, his/het processes may appear as psychologically similar to those of the other person. Such insights into the personal nature of meaning may result from the social view of language and communication: (a) Meaning is a human construct, and as such it is dependent of the person who makes it. (b) Meaning cannot be passed on as an entity in the same manner as meaning bearers. It does not reside in words, symbols or appeal signals with which individuals express their emotional and conceptual contents. Language therefore, must be seen as a behavioral system which triggers communicating activities within the cognitive domains of particular communicating individuals.

Ernst von Glasersfeld (2008 [1984] {1981}: 7–8) remarks that the socially constructed reality, although intersubjectively shared, is not something existing in an objective and independent way: "[t]hanks to professional burglars, we know only too well that there are many keys that are shaped quite differently from ours but nevertheless unlock our doors… From the radical constructivist

---

[19] George Alexander Kelly (1905–1967), an American psychologist, author of the Personal Construct Theory. The Reader is referred to: Kelly, George Alexander 1955. *The Psychology of Personal Constructs Volume 1. Theory and Personality*. New York: Norton.

[20] The Reader is referred to: Berger, Peter L., Thomas Luckmann 1966. *The Social Construction of Reality*. Garden City, NY: Doubleday.

Glasersfeld, Ernst von 1995. *Radical Constructivism: A Way of Knowing and Learning*. London – Washington: The Falmer Press.

point of view, all of us — scientists, philosophers, laymen, school children, animals, indeed any kind of living organism — face our environment as the burglar faces a lock that he has to unlock in order to get at the loot".

John Searle (1995) observes that methodological solipsism does not in and of itself preclude the possibility of having thoughts about multiple agency. He takes *collective intentionality* as a primitive and unreduced phenomenon: the collective "we intend" is primary, and is not reducible to the individualistic "I intend and I believe that you intend, and you intend and you believe that you intend", etc. According to Searle (1995), collective intentionality is the fundamental property that makes it possible to create of social facts as distinguished from brute (physical) facts by means of constitutive rules ("X counts as Y in C"). On his account, the world of institutional facts has a real existence (translatable into the individual mental experiences of each of its participants), even if it is logically secondary to the world of brute facts.

1.4. History

As with most any philosophical subject, the roots of the interest in the mind – especially the human mind – can be traced back to the philosophers of the ancient world. On a certain level of generality, Cognitive Science is continuous with those early theories, such as the atomistic account of perception or the conceptions of the soul, reasoning, and knowledge formulated by Plato and Aristotle; on all other levels, however, it is radically different, using an entirely redefined set of basic notions, and altogether different methods.

Most sources (e.g. Buss 2004: 50–52, Chuderski 2002, Thagard 2006, Green et al. 2000) identify the origins of Cognitive Science with the scientific and socio-scientific developments of the 1950s, championed by Noam Chomsky, George Miller, and AI theorists, and commonly called 'the first cognitive revolution'. Particularly interesting in this context are the departures from this convention, where the 1950s are referred to as 'the second cognitive revolution'.

The ‘first cognitive revolution’ then comes to signify some earlier period in a way indicative of what might count as the prehistory of Cognitive Science.

1.4.1. Prehistory

Chomsky himself (2005) is inclined to see ‘the first cognitive revolution’ in the intellectual developments of the seventeenth and eighteenth centuries. It would be interesting to follow his inspiration and develop this line of thought. On the one hand, this period was characterised by a general change of academic sentiment, an effective transition from the legacy of scholasticism to the origins of modern science (ushered in, most importantly, by Isaac Newton[21]), and materialistic and deterministic accounts of the world (e.g. Julien La Mettrie, Pierre-Simon Laplace[22]). On the other, those two centuries witnessed a series of philosophical breakthroughs: the first idea of the human mind as a machine (Thomas Hobbes[23]); strong nativism of René Descartes[24] and his establishing of the primacy of epistemology; the doctrine of epistemological empiricism developed by John Locke, David Hume, and George Berkeley[25], with principal

[21] Sir Isaac Newton (1643–1727), English physicist, philosopher, mathematician, and astronomer. The success of his classical mechanics was a driving force behind the rise in prominence of the scientific worldview. See Tatarkiewicz (2003 [1931]: 66–69).

[22] Julien Offray de La Mettrie (1709–1751), French philosopher and mathematician, a defender of materialism and atheism; Pierre Simon Laplace (1749–1827), French mathematician and astronomer, a leading proponent of determinism. See Tatarkiewicz (2003: Vol. 2. 101–103).

[23] Thomas Hobbes (1588–1679), English philosopher, best known for his naturalistic views in ethics and social and political theory, also a materialist and determinist. See Tatarkiewicz (2003: Vol. 2. 50–53).

[24] René Descartes (1596–1650), Renatus Cartesius, French philosopher and mathematician. Best known for his rationalistic epistemology, and also for his work in analytic geometry. See Tatarkiewicz (2003: Vol. 2. 32–44).

[25] John Locke (1632–1704), English philosopher and anatomist; David Hume (1711–1776), British philosopher and historian born in Scotland; George Berkeley (1685–1753), Irish bishop and philosopher. Locke, Hume and Berkeley are considered to be the three major figures of British empiricism. See Tatarkiewicz (2003: Vol. 2. 76–90).

focus on the matters of broadly understood cognition; the 'Copernican' revolution of Immanuel Kant[26].

In contrast, Andrzej Klawiter (2004: 116) uses the term 'first cognitive revolution' in relation to the rise of scientific psychology in the nineteenth century in Germany. Several researchers turned out to be of significance in the process of extricating psychology from philosophy and establishing the former as a separate and rightful scientific discipline, distinguished by its own methodology focussed on experiment and observation. Following Keith Holyoak's (1999: xli–xliii) short review, one can mention especially Hermann Helmholtz, a natural scientist whose achievements include a theory of colour vision, and Hermann Ebbinghaus, who pioneered experimental research on memory. The single most important figure, however, was Wilhelm Wundt, due to his contributions of both intellectual (rigorous methodology) and institutional nature (the founding of the first institute of psychology)[27]. His influence was additionally reinforced by his many students – such as Edward Titchener – who popularised and further developed this discipline, especially in the United States.

1.4.2. Germination

Although aspects related to the Second World War are usually neglected in the accounts of the later advent of Cognitive Science, quite possibly it was this military conflict that provided the catalysing impulse for the emergence of this tradition in 1950s. This was due to a confluence of a number of factors, most of them related to computer science.

Firstly, the war generated a powerful need for increasingly advanced military technology, leading to projects such as ENIAC, the first electronic

---

[26] Immanuel Kant (1724–1804), German philosopher renowned for his groundbreaking contributions into ethics, aesthetics, epistemology, and ontology. Often considered to be a precursor of Constructivism in theory of knowledge. See Tatarkiewicz (2003: Vol. 2. 120–142).

[27] Wundt's (and his disciples') multifaceted importance is acknowledged by many commentators, e.g. Holyoak (1999), and Thagard (2006).

computer (as described by e.g. Vernon Pratt – quoted in Chuderski 2002: 4). Insights gained from the work in cryptography, computing machinery and arms industry proved to be seminal for the later development of Artificial Intelligence; among the people engaged in efforts related to military activities were such AI pioneers as Alan Mathison Turing (the breaking of the Enigma code[28]), Claude Elwood Shannon (fire control systems), John von Neumann (the Manhattan Project)[29].

Margaret Boden (2006: 200) observes yet another factor: in 1942 a series of annual conferences was started under the auspices of Josiah Macy Foundation. The Macy conferences were originally established with a view to developing ways of effective prevention against outbreaks of world wars in the future. In practice, the seminars turned out to be a platform for exchange of ideas and interdisciplinary collaboration between the leading intellectual of that time, especially in the area of what later developed into cybernetics.

1.4.3. Beginnings

According to most commentators (e.g. Adam Chuderski 2002), the breakthrough year marking off the beginnings of the cognitive movement was 1956, when two conferences on information processing were organised in the United States. A summer conference at Dartmouth College, also dubbed "the Constitutional Convention", brought together (among others) John McCarthy, the later inventor of LISP programming language; Claude Shannon, the co-author of modern information theory; Marvin Lee Minsky, the later director of the first AI project; Allen Newell and Herbert Alexander Simon, the later proponents of the strongly functionalistic Physical Symbol System Hypothesis, by which the mind is a

[28] Part of the credit for the breaking of the code is due to Polish cryptographers for their previous inroads into this task.

[29] The pivotal role of Turing, Shannon, and von Neumann, among other early information and computer scientists, is discussed by Margaret Boden (2006: 200–206).

physically implemented system that performs manipulations on symbols[30]. This event started the field of Artificial Intelligence, which was named after the title of the conference.

Later that year, a conference was held at the Massachusetts Institute of Technology that featured the talks from Newell and Simon, the psychologist George Armitage Miller, and the linguist Noam Chomsky (see Miller 2003). Newell and Simon presented Logic Theorist that had its debut at Dartmouth; it was the first programme that could formally prove logical theorems and thus approximate human (logical) reasoning[31]. Miller presented his research on the limitations of short term memory that led to the publication of the immensely influential paper "The Magical Number Seven, Plus or Minus Two" (Miller 1956). Chomsky introduced the theory of generative grammar, based on his 1955 doctoral thesis, and later expanded in Chomsky 1957. The latter two talks referred to human cognitive capacities in a way that suggested a possibility of their computational implementation[32].

At that time, the intellectual landscape was dominated by the behaviouristic tradition. Behaviourism began and was developed in the first half of the twentieth century, possibly as a reaction to 'unscientific' introspectionist psychology of some of the disciples of Wundt[33], and to highly speculative theorising within the emerging Freudian tradition. Central to mainstream behaviourism was not so much the denying of the existence of mind (as is sometimes claimed), but rather the conviction that mental processes, as hidden inside the 'black box' and in principle inaccessible to objective observation,

[30] See Boden (2006, section 6.iii.) for extensive coverage.

[31] Cf. Boden 2006 (section 6.iii.).

[32] "I left the symposium with a conviction, more intuitive than rational, that experimental psychology, theoretical linguistics, and the computer simulation of cognitive processes were all pieces from a larger whole and that the future would see a progressive elaboration and coordination of their shared concerns." (Miller 2003: 143).

[33] The precise methodological status of introspection is still debated. See e.g. Eysenck and Keane (2000 [2002]: 3–4).

could not be a legitimate object of scientific inquiry – as opposed to directly observable input (stimulus, reinforcement) and output (behaviour).

Notwithstanding its popularity and successes – e.g. in therapy programmes and animal studies – by the mid 1950s the general behaviouristic paradigm had begun to receive increasing criticism. One source of such criticism were strictly philosophical arguments, notably from David Lewis, Peter Geach, and Roderick Chisholm (quoted in Żegleń 2003: 56), such as the lack of any straightforward link in humans from 'dispositions' to overt behaviour. Another was existing empirical research, mostly by European psychologists, that could not be reconciled with behaviourism. Holyoak (1999: xliii–xlv) mentions the names of the German psychologist Wolfgang Köhler (and the Gestalt movement in general), the English memory researcher Frederic Charles Bartlett, the Soviet psychologist and neuropsychologist Alexander Romanovich Luria (Александр Романович Лурия), the Swiss developmental psychologist Jean Piaget, and even the American psychologist Edward C. Tolman, himself a behaviourist researching the creation of cognitive maps in rats. A fair case can be made for the inclusion of the American psychologist Jerome Bruner, for his work on categorization (e.g. Bruner et al. 1999 [1956]). The research of the abovementioned scholars was difficult to interpret with a behaviouristic toolkit, but instead seemed to evidence the reality of underlying complex mental representations[34].

Still, the decisive blow against behaviourism came from linguistics. In the famous review of B. F. Skinner's *Verbal Behavior*, Chomsky (1959) showed that the reduction of human higher mental processes to the basic behaviouristic concepts of stimulus, response and reinforcement was inadequate, and that the

[34] A second vein of intellectual opposition to behaviourism can be traced back to early research in *ethology* (by such researchers as the Austrian zoologist Konrad Lorenz [1903–1989], and the Dutch zoologist Nikolaas Tinbergen [1907–1988]), who demonstrated the existence of *innate capacities* in animals; this contradicted the basic tenets of behaviourist learning theory, which assumed no innate component to the learning process (see e.g. Holyoak 1999).

Skinnerian terms for describing verbal behaviour (e.g. *controlling stimulus*) were simply a disguise for traditional semantic terminology (e.g. *reference*). Further, he argued convincingly that the only viable account of human language acquisition must be constructed in representational terms. The review is today popularly considered to have been a watershed and a transition point.

However, as Thagard (2006) observes, the proper institutional foundations for Cognitive Science had not been laid out until the 1970s. It was then that the journal *Cognitive Science*[35] first came out (1977), followed by the creation of the Society of Cognitive Science[36] (1979) and launching of a series of cognitive-scientific conferences (1979). In the 1980s, the first courses in Cognitive Sciences were started[37].

1.4.4. Early and classical Cognitive Science

From today's perspective one can speak of two traditions in Cognitive Science, one more aprioristic and relatively narrow, the other more recent, more empirical and very inclusive. Lakoff and Johnson (1999: 11–12) call attention to the fact that

> [t]he term <cognitive> has two very different senses (…). A confusion sometimes arises because the term <cognitive> is often used in a very different way in certain philosophical traditions. For philosophers in those traditions, <cognitive> means only conceptual or propositional structure.

---

[35] http://www.sciencedirect.com/science/journal/03640213

[36] http://www.cognitivesciencesociety.org/

[37] In Poland, the development of Cognitive Science *qua* 'Cognitive Science' has been markedly delayed. The journal „Kognitywistyka i Media w Edukacji" (*Cognitive science and the media in education*) first came out in 1996, the Polish Society of Cognitive Science (http://www.ped.uni.torun.pl/kognityw/kognitywistyka.htm) was formed in 2001, and the first MA study programme was organised in 2005 at Adam Mickiewicz University in Poznań.
At the same time, the parallel developments in individual relevant fields, such as AI or Cognitive Linguistics, were not characterised by such a noticeable delay.

Such an understanding of the adjective 'cognitive', with consequences for what counted as Cognitive Science, was characteristic of the former approach, with which it still is sometimes associated.

Commentators (e.g. Żegleń 2003: 40–41) generally agree that at its birth, Cognitive Science was inspired by formal sciences and animated by computer science and AI, building on Alan Turing's (1950) seminal statement of the idea of an intelligent machine. In fact, Turing's exact claim was quite moderate, i.e. that – given a certain redefinition of thinking[38] – no fundamental reasons existed for which *bona fide* thinking machines could not be constructed. Conversely, the development of digital computer architectures provided a convenient metaphor in terms of which human cognitive processes could be understood.

Thus, early visions in Cognitive Science revolved around a rather strict reading of the computer analogy, on which human minds worked like serial digital computers. The natural recapitulation of Artificial Intelligence in the philosophy of mind was the doctrine of *functionalism*, and especially *machine state functionalism*, first explicitly stated by Hilary Putnam (1975 [1960]) in his "Minds and Machines". According to functionalism, the mind can be regarded as a kind of computer software implementable in any kind of sufficiently powerful hardware, the physical constitution of the latter being irrelevant. Cognition was regarded literally to consist in formal symbol manipulation, being in its essence independent of the body.

Such a stance had another consequence – of making cognition independent of the body in a related, but distinct sense. It favoured a natural assumption that cognition was limited to higher mental processes, such as logical reasoning and language, which were largely independent and in principle separable from 'lower-level' – and thus more 'bodily' – processes, such as

[38] More precisely, the crucial element was the operational *criterion* for 'thinking', constructed along the lines of what later became known as the Turing Test: the machine should be able to convince human judges that it was a human being and not a machine (Turing 1950).

perception, proprioception, or motor control. The prototype of 'intelligent behaviour' and its default model was *verbal* behaviour, unconnected to its agent's successfully functioning in the real, physical world[39].

1.4.5. Contemporary Cognitive Science

Contemporary Cognitive Science is a continuation of early, or 'classical' Cognitive Science in most major respects: representationalism, naturalism, and an internalistic perspective. Without radically departing from the spirit of the earlier trends, it has nevertheless grown into something qualitatively different. In the sections to follow, I will outline – all too briefly – what I consider to be the main directions of departure of modern Cognitive Science from the earlier tradition. The reader should bear in mind the recapitulatory character of the sections to follow, which are focused on sketching out the directions and profile of changes to the Cognitive Sciences rather than detailed discussion of the phenomena in question.

1.4.5.1. Chinese Room argument

Arguably the most influential criticism of the idea of thinking machines (the 'strong' AI view) and against functionalism has been John Searle's argument, showing that thinking is *not* simply symbol manipulation, and illustrated with the thought experiment of the 'Chinese room' (Searle 1980: 417–419)[40]. The crux of Searle's well-known argument consists in a simple but powerful insight that even

---

[39] The functionalist equation of cognition with explicit, formal, abstract and disembodied symbol manipulation has been repeatedly pinpointed by George Lakoff (e.g. Lakoff and Johnson 1999) and formulated as a chief objection against this doctrine.

[40] The design of the experiment requires that a person knowing no Chinese is locked in a room together with detailed written instructions of how to perform formal operations on Chinese symbols meaningless to the person. This person could then transform certain strings of Chinese symbols ('questions') into other strings of Chinese symbols ('answers') without any genuine 'understanding', but in a way that would indicate understanding and intelligence in the audience that would receive the answers.

though computers and humans could be potentially functionally equivalent, the very notion of function is always observer-relative. Thus, the behaviour of the machine or programme, regardless of what it consists in, *by definition* cannot be intrinsically meaningful, but instead is only interpreted as such by human observers.

Most scholars accept the argument; however, there is a controversy regarding its precise power and scope. On at least one viable interpretation, the effect of this thought experiment is limited to clarifying our linguistic intuitions regarding the use of the word 'thinking', with no devastating consequences against the general soundness of strong AI. Minimally, the argument does not question the validity of 'weak' AI, that is the idea of machines' accurately *simulating* thought processes; as such, it does not pose any theoretical threat against broadly defined CS.

1.4.5.2. Connectionism

Although connectionist architectures have been proposed as early as the 1940s[41], it was not until the 1980s that the long-dominant classical approach to AI began to be effectively challenged by the rapid development of this alternative approach: *connectionism*. The classical approach (termed 'Good Old Fashioned Artificial Intelligence', or GOFAI[42]) relied on serial but very fast computations over symbols performed by a single central processor. In contrast, connectionism relied upon architectures of neural nets compiled from large numbers of richly interconnected simple units, operating by the adjusting of the weights of connections (whether automatically – by in-built feedback mechanisms – or by external intervention, e.g. backpropagation), without resorting to any explicitly

[41] McCulloch-Pitts neural nets, developed by Warren McCulloch and Walter Pitts. Discussed in detail in Boden (2006, Chapter 4.iii).

[42] According to *Wikipedia*, the popular term GOFAI was coined by the influential philosopher of Artificial Intelligence, John Haugeland. The Reader is referred to Haugeland, John 1985. *Artificial Intelligence: The Very Idea*. Cambridge, Ma: MIT Press.

present algorithm. Vital is the fact that the basic units in such nets are *subsemantic*: individual nodes in the net are not semantically evaluable (do not have their own meanings), and units of meaning such as concepts emerge as a result of activations of larger portions of the net[43].

Connectionism struggles mostly with arriving at a satisfactory account of rule-based cognitive operations characterised by systematic productivity. On the other hand, the strengths of connectionism over the rival approach include more realistic patterns of degradation in response to error/damage (*graceful degradation* versus the catastrophic failure of classical systems), *spontaneous generalisation* of patterns, and better *sensitivity to context*. Nonetheless, the most important advantage of this approach has to do with its greater *psychological/biological reality*[44]. The simple example of an electronic versus a mechanical clock, which behave identically despite being implemented differently, shows that equipotentiality does not necessarily imply the identity of the implementing mechanism. Connectionist networks, while being very far from constituting viable models of the (aspects of the) actual working brain, exhibit at least some analogies to it. As a consequence, they display more characteristics in common with the brain, such as spontaneous learning by association, or graceful degradation (the destruction of a part of the network does not lead to catastrophic failure, but rather to a gradual decline in performance).

It is tempting to conflate the distinction between classical and contemporary Cognitive Science with the distinction between GOFAI and

---

[43] "Information [is] stored in several interconnected units, rather than in a signle location… information is distributed rather than concentrated." (James L. McClelland, cited by Waldemar Skrzypczak (2006: 12).

[44] The idea of psychological reality was proposed and championed by Noam Chomsky (see http://www.chomsky.info/interviews/1983----.htm).

Interesting in this context is the opinion of Henryk Kardela (2006a: 197–198), who also quotes Anna Wierzbicka, that Chomsky himself has not remained faithful to this methodological principle, largely as a result of his focus on syntax to the neglect of the semantic aspects of language.

connectionism. Still, this would not be appropriate for two rather basic reasons. Firstly, the debate is still in progress. For instance, Ric Cooper (2000 [1996]: 45–48), taking stock of this debate voices a representative opinion that as yet, connectionism has not gained the upper hand. Secondly, such a conclusion would fail to do justice to other important differences between classical and contemporary CS, which are discussed in turn[45].

1.4.5.3. Neuroscience

The study of the brain has always been acknowledged as an essential element of the Cognitive Science jigsaw, and even early work in neuroscience was seen as important[46]. In practice, however, for a long time neuroscience failed to have any substantial impact on the other component fields (a fact that is iconically represented on the diagram b) in Fig 2.). Two main reasons are likely to have been responsible. Most importantly, as was already discussed, functionalism lent support to the idea of study of the mind at a level of description that abstracted away from the details of the cerebral implementation, thus making neuroscientific data merely supplementary. The other factor was related to the very limited possibilities of the study of the working brain: the situation of the

---

[45] Other sources considered important for the question of the relation between classical AI and connectionist architectures include the foundational book by Rumelhart, McClelland and collaborators (Rumelhart, David E., James. L McClelland, the PDP Research Group 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Cambridge, MA: MIT Press.), as well as the volume edited by Pinker and Mehler (Pinker, Steven, Jacques Mehler (eds.) 1988. *Connections and Symbols*. Cambridge, MA: MIT Press.), including essays from Jerry Fodor and Zenon Pylyshyn and Steven Pinker and Alan Prince.

[46] For example, the 1949 book by Donald Hebb – *The Organization of Behavior; a Neuropsychological Theory* – is ranked as the 4th most influential work in Cognitive Science in the twentieth century, according to *Millenium Project*, Center for Cognitive Sciences, University of Minnesota:

http://www.cogsci.umn.edu/OLD/calendar/past_events/millennium/final.html

brain scientist could be compared to that of "the oceanographer probing the sea from ashore"[47].

The 1980s and especially the 1990s have witnessed both the decline of functionalism and breakthroughs in the technological possibilities for brain imaging (see Wójcik 2006; Medin et al. 2000 [1992]: 26–28). In particular, the noninvasive techniques of Positron Emission Tomography (PET) and Functional Magnetic Resonance Imaging (fMRI) have made it possible to formulate a much more accurate functional geography of the working brain; and Transcranial Magnetic Stimulation (TMS) has allowed the neuroscientists to directly demonstrate causality from the activity of specific brain circuits to particular behaviours[48]. The increase of knowledge about the brain has lead to the growing importance of 'neurological reality'[49], that is the requirement that cognitive models be made compatible with the principles of the operation of the brain. Moreover, it has opened new fields of cognitive-scientific study ripe for interdisciplinary collaboration. One good example may be research summarised by Gary Marcus and Simon Fisher (2003) related to the degenerative cerebral effects of the point mutation of the FOXP2 gene that result in a range of deficits,

---

[47] A comparison used by Tadeusz Marek in a talk at the conference „Neuronauki – W kierunku wyjaśniania funkcjonowania człowieka" [Neurosciences – Towards explaining human functioning] in Kazimierz Dolny, Poland, 26.06.2005.

[48] PET measures the local cerebral blood flow (which mirrors the metabolic, and presumably cognitive, activity of the parts of the brain) by detecting radiation of the radioactive isotopes injected into the subject's bloodstream. fMRI measures the local metabolic activity by detecting the level of oxygen in the blood. TMS produces a magnetic field that induces local disruptions in neuron excitability, simulating the effect of 'turning off' and 'turning on' individual narrow areas of the cerebral cortex.

See a short overview by Jan Wójcik (2006): http://www.kognitywistyka.net/mozg/badania.html

A discussion of the two most popular techniques, PET and fMRI, focussing on the differences in methodology and functionality, is offered by Buckner and Logan (2001).

[49] See, e.g., Grodzinsky (2003) for an attempt at characterising the ways in which neurological data cohere with the categories devised in lignuistic description.

notably in the processing of syntax, building over earlier studies (e.g. that of Myrna Gopnik, 1997) by combining data from neuroscience, genetics, and linguistics.

1.4.5.4. Cognitive systems are *real* systems

The above statement that "cognitive systems are real systems" is intended to reflect the spirit of several distinct theoretical traditions that nevertheless – on some level of generality – share the appreciation of the fact that the only type of existing cognitive systems belongs to real, flesh-and-blood, biological organisms. The "goal" (design principle) behind those organisms was not "thinking" or "thinking clearly", but successful survival and reproduction. The goal of the cognitive system of an organism, then, is not producing comprehensive or truthful representation of their environment, but producing opportunistically successful behaviour under the constraints of limited resources, such as time and scope of attention. Applicable here, in a general context, is Eleanor Rosch's concept of 'cognitive economy' (Rosch 1988a [1978]). This bears very profound implications for constructing the models of cognitive systems, suggesting that, generally, the most fundamental function of cognition is guiding action. More specific implications of that outlook are, for example, that more emphasis should be put on subconscious but fast feedback loops (e.g. in perception) that short-circuit conscious but slower central processes (e.g. Clark 2006).

As stated above, the realisation that 'cognitive systems are real systems' also directs one's attention to the fact that they are *evolved* systems (i.e. having an evolutionary history of adaptation to specific environments – see Anderson 2005). The basic constructional design of the nervous systems of all natural cognitive agents – including humans – is a result of the long process of natural selection for efficiency in surviving and reproducing in a given environment. Such considerations point to the relevance of evolutionary studies, which will be mentioned in 1.4.6.1., but also studies in diverse disciplines once thought to be

tangential to narrowly understood CS, such as animal cognition or animal communication.

Finally, (natural) cognitive systems are *embodied* systems. Crucially, this is not merely the trivial sense of having *a* body as a necessary physical substrate for the mind; in this context, embodiment means having a *specific kind of body* which directly influences one's mental processes in specific ways[50]. *Embodiment* is itself a rapidly growing and already very diverse research perspective that merits a separate discussion; its various threads are developed in philosophy (e.g. the continuators of Maurice Merleau-Ponty[51], such as Shaun Gallagher 2005), linguistics (e.g. George Lakoff 1990 [1987]), or general CS (e.g. Francesco Varela et al. 1999 [1991]). As a particularly important recent trend, for many theorists in the broadly understood ECS (Embodied Cognitive Science), for instance Jordan Zlatev (2007) or Michael Anderson (2005), cognition is best described as *situated* and *enactive*. This means that there is no fundamental dualistic distinction into the organism on the one hand and on the other, the passive, pre-given, objective world that is represented in the organism's mind. Rather, as described by Francisco Varela et al. (1999: 202), "…organism and environment are mutually enfolded in multiple ways, and so what constitutes the world of a given organism is enacted by that organism's history of structural coupling". This follows the Gibsonian[52] insight that the organism's own world, rather than being merely indirectly represented, is directly enacted in accordance with its individual sensorimotor capabilities.

1.4.5.5. Interdisciplinarity

[50] Tim Rohrer enumerates ten distinct senses of the term 'embodiment' as used in the Cognitive Sciences (quoted in Kardela 2006: 227–228).

[51] Maurice Merleau-Ponty (1908–1961), a French philosopher and phenomenologist.

[52] James Jerome Gibson (1904–1979), an influential American psychologist. His name has come to stand for an antirepresentational and interactionist perspective on perception. See Barsalou (1992: 17–18) for a short discussion in relation to the context of Cognitive Science.

Almost all sources and reference works (e.g. Bechtel et al. 1998: 94, Rapaport 1996, Stillings et al. 1995: 1, Thagard 2006, Kalisz et al. 1996: 58) depict Cognitive Science principally in terms of its member disciplines. The canonical list consists of linguistics, psychology, philosophy, computer science with AI, and neurobiology, sometimes being complemented with (cognitive) anthropology and education, with a penta- or hexagon being a convenient means of illustrating the relations between them (Fig. 2).

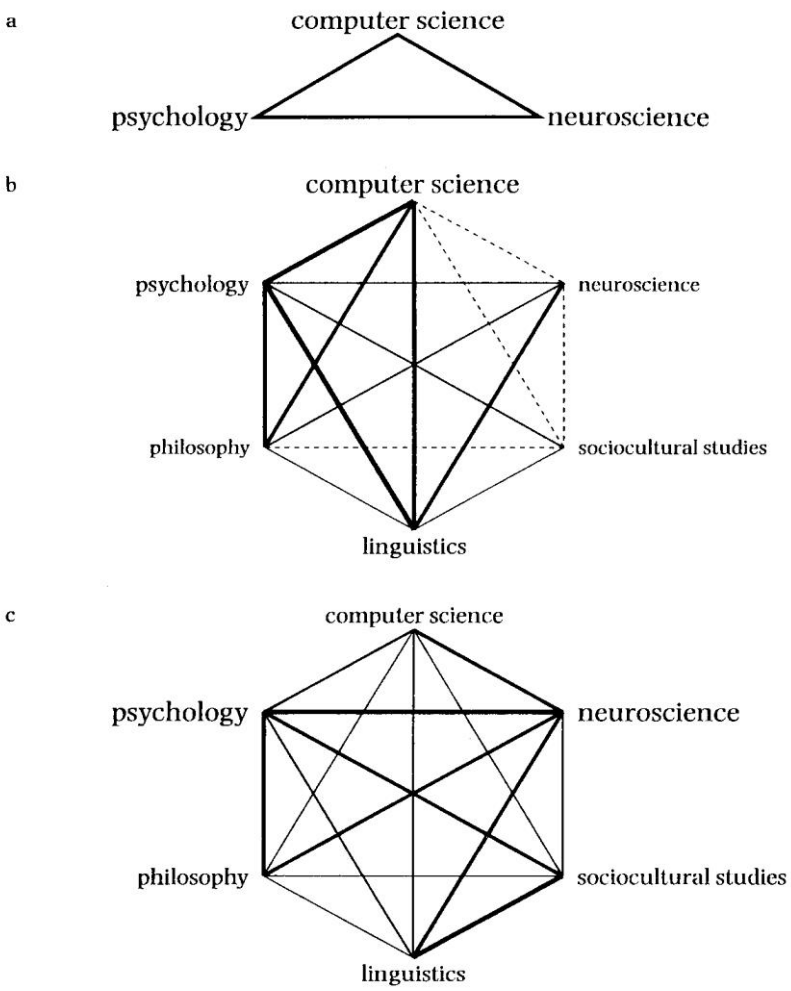

Fig. 2. Contributing disciplines and interdisciplinary connections during three different stages in the development of Cognitive Science. Source: Bechtel et al., 1998: 94.

This traditional designation no longer seems to be accurate. Contemporary Cognitive Science is these five or six disciplines – and much more. The reasons

for which the above descriptions are simplistic stem from *expanding*, which takes place whenever the original research problem during its analysis turns out to recursively involve sub-problems which require expertise from other fields; according to Klawiter (2004: 111–115), it commonly characterises Cognitive Science research. As a result, present-day Cognitive Science has greatly broadened its scope, effectively incorporating a large number of existing as well as newly established fields of study.

Again, reference to example will be useful: *The MIT Encyclopedia of the Cognitive Sciences* (1999, ed. Wilson and Keil) contains 471 short articles, ranging from "qualia" to "X-bar theory" to "primate amygdala" to "evolutionary psychology of sexual attraction", that are in a large part difficult to subsume under any particular discipline label[53]. At a closer inspection of today's Cognitive Science, one may find discipline boundaries not so much being transcended, as in fact dissolving. Consequently, an adequate description of contemporary Cognitive Science must be rather general, almost to the effect of the statement by Terry Winograd (cited in Rapaport 1996) – that "<Cognitive Science> is a broad rubric, intended to include anyone who is concerned with phenomena related to the mind".

This inclusiveness, however, is not absolute. Not everyone "concerned with phenomena related to the mind" should by default count as a 'cognitive scientist'. Firstly, the naturalistic commitment (1.3.3.) remains essential. In this context, characteristic is the omission from the *Encyclopedia* of the entries for phenomenology or Edmund Husserl, despite their unquestionable relevance in terms of the mind as a subject matter[54]. Secondly, cognitive scientists are

[53] Still, the layout of the introductory section of the tome adheres to the classical division: the six major Introductions are for Philosophy, Psychology, Neurosciences, Computational Intelligence, Linguistics, and Culture – Cognition – Evolution.

[54] It would be imprecise to state that phenomenology has been excluded from Cognitive Science: rather, it has been redefined and included into strictly naturalistic projects such as neurophenomenology (e.g. Gallagher 2005) or heterophenomenology.

committed to psychological reality. For example, work on designing programmes or robots that could achieve human-like levels of performance on certain cognitive tasks (e.g. text processing or detecting objects) but would perform such tasks in a different way that humans would (e.g. using 'mental' operations inaccessible to humans), while counting as Artificial Intelligence, should not be considered Cognitive Science because it would not offer insights into how they are performed by real cognitive agents.

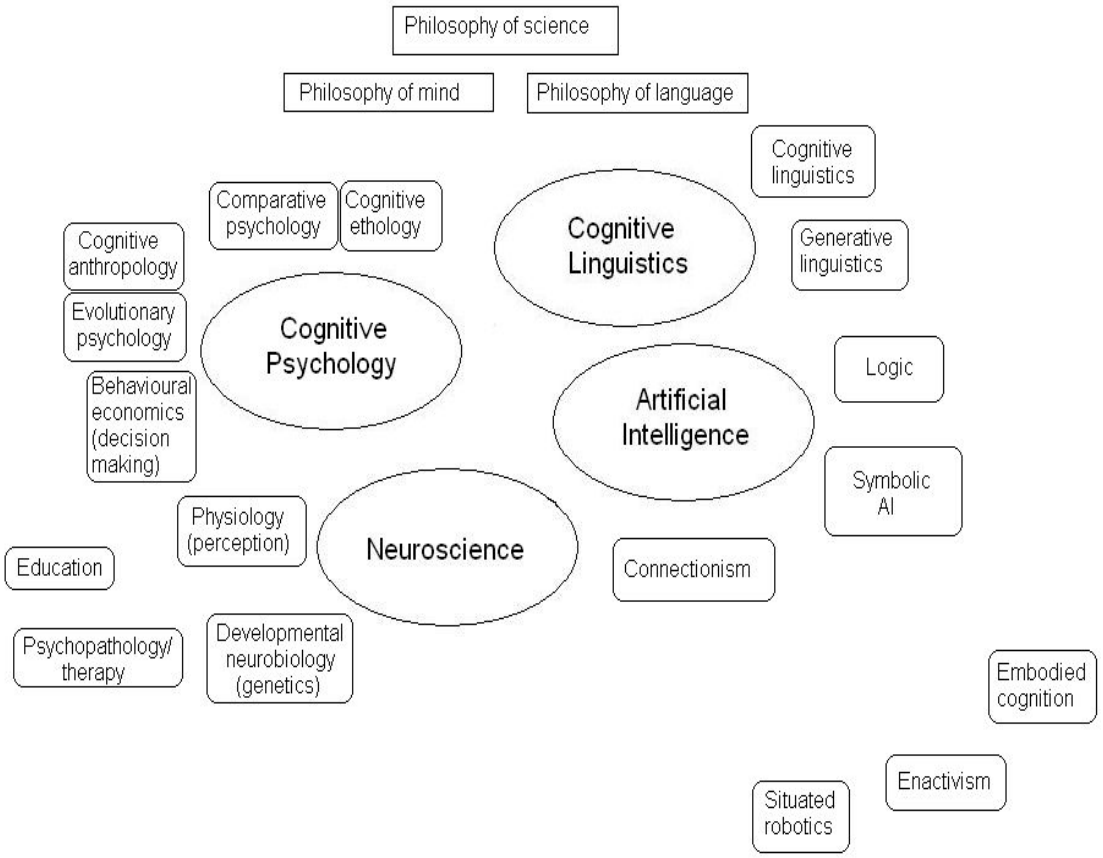

Fig. 3. The landscape of *contemporary* Cognitive Science, including disciplines, subdisciplines and paradigms. The list should not be treated as either systematic or exhaustive but rather as illustrating the inclusiveness of the thematic scope.

1.4.6. Interdisciplinarity: methodological notes

Kalisz, Kubiński and Buller (1996: 61, italics in the original) make straightforward a belief that is characteristic of both early and contemporary Cognitive Science:

> *Beware of false cognitivists! (…) Even though they declare their commitment to clarify the essence of human cognitive processes, they refrain from transcending the historically delimited bounds of their disciplines. (…) Cognitive Science must remain by its very nature interdisciplinary, or, better, superdisciplinary*.

Without any doubt, interdisciplinary cooperation has been the foundational, defining feature of Cognitive Science throughout its history. Johnson-Laird's (1980: 71) programmatic statement that "[w]e should reject the view that Cognitive Science is merely a clever ruse dreamed up to gain research funds – that it is nothing more than six disciplines in search of a grant-giving agency" is based on the assumption that it is interdisciplinarity that allows Cognitive Science become more than the sum of its component parts.

Interdisciplinarity, however, is a controversial notion that tends to raise suspicion, not least because in today's science it has grown to be somewhat of a 'buzzword'[55]. There are in fact strong reasons for this caution.

One of the few methodological postulates universally applicable to all kinds of scientific activity is the requirement for frequent reporting of one's results to a larger forum in order to bring them under unbiased revision. Interdisciplinarity may become an obstacle in putting this into practice, as the work of such character is more difficult to assess. Another problem with interdisciplinary research involves the technical difficulties in the dialogue between disciplines: while the object of study might be common, problems often

[55] Of course, the idea interdisciplinary research as such is not new in linguistics or the study of the mind. For example, Zdzisław Wąsik (e.g. 2003, Chapter 2), discussing the various senses of 'language' concludes that language as *obiectum reale* is an inherently heteronomous entity, which calls for a rich variety of perspectives, and not only linguistic ones.

arise from the incommensurability of specific discipline methodologies. For example, *consciousness*, *function*, *information*, *intelligence*, and *intentionality* are only a few examples from the long list of terms that have substantially different meanings and operational definitions in at least two – and usually more – member disciplines[56]. Thus, on the face of it, the adoption of an interdisciplinary perspective is threatened by the risks of incongruity and conceptual-terminological confusion. In this section, I will address those dangers.

To begin with, it is perhaps not sufficiently appreciated that many barriers blocking interdisciplinary communication have been effectively obliterated with the rapid progress in information technology over the past decade. Formerly, a scientist entering a new field was constrained by the scarcity of accessible resources that were usually both fragmentary and outdated (if simply because of the editorial lag). The popularisation of electronic data transmission, together with increasingly sophisticated techniques of presentation, have helped to overcome informational isolation among disciplines, and have made it possible for the researchers to become acquainted on the spot with the latest academic-level knowledge within, as well as outside, the researcher's particular bailiwick. Most certainly, the necessity for an appropriate background of scholarship has remained fundamental. But the elimination of the basic logistic problems that pestered interdisciplinary cooperation has been a crucial improvement.

What is more, frequently enough interdisciplinarity is not a matter of choice, but of necessity. The science of cognition seems to constitute a paradigmatic example; and Merlin Donald (2004: 248) argues explicitly that „the human mind is too complicated for any single field to master, and in this case, interdisciplinarity is not a luxury, but a necessity." Some domains of study are by

[56] Both the reservations against the idea of interdisciplinarity and the terminological difficulties of interdisciplinary dialogue have been repeatedly witnessed by me during my participation in cognitivist conferences, especially the conferences accompanying the annual conventions of The Polish Cognitivist Society organised in Toruń (2002–2006) and later in Poznań (2006–2007).

themselves inherently dependent on cooperation across traditional disciplines. Such is the nature of applied human sciences, e.g. glottodidactics or human resources management, which must combine psychology of the individual or of group relations with specialised knowledge of a particular subject. In some other fields, theorising must at a minimum be compatible with the key developments in the related areas, since an *experimentum crucis* in one discipline can ramify throughout another and even invalidate whole theoretical approaches therein. The above is particularly true of natural sciences operating on different levels of description. Still, even contemporary philosophy of mind – though itself non-empirical – because of its highly naturalistic profile cannot afford to ignore current empirical findings in psychology and neuroscience; one example may be the immediate relevance of Benjamin Libet's (e.g. 1999) studies to the problem of 'free will'.

As to the benefits resulting from interdisciplinary perspective, they appear to be quite obvious. Interdisciplinary research has more capacity for being profound, comprehensive and fruitful. Communication and cooperation between members of distinct but related specialties might yield fresh insights and, more importantly, lead from predictions in one area of inquiry directly to testable hypotheses in another. In the following sections, I will illustrate the above points by reviewing two examples of different generality illustrating successful interdisciplinary study: one from general Cognitive Science, the other from cognitive linguistics.

1.4.6.1. Interdisciplinarity and Modularity of Mind

The first example concerns the idea of the modularity of mind, which – although latent in the writings of earlier authors and in one sense dating back to Franz Joseph Gall's phrenology[57] – was effectively started with the 1983 book by Jerry

[57] Franz Joseph Gall (1758–1828) was a German anatomist and the founder of phrenology – a parascientific discipline popular in the early nineteenth century, founded on the assumption that people's inclinations or talents could be known by the study of their skulls.

Fodor. Himself a philosopher of mind, Fodor used his background to synthesise insights from several fields of Cognitive Science; in particular, he drew on Chomsky's observations on language acquisition (e.g. Chomsky 1975), as well as a broad range of experimental evidence from psychology, psycholinguistics, and neuroscience.

On Fodor's (1983, 2000) account, the mind consists of sensory transducers, central processor(s), and several modules that link the two. Modules are input/output systems that are domain specific (dedicated only to their specialised tasks), informationally encapsulated (do not draw on module-external information or resources during processing), automatic, rapid, relatively localised in the brain, and relatively developmentally rigid ('innately specified'); they relay information from the senses to central processor(s). Central cognition, responsible for belief-fixation, is itself *nonmodular* and therefore difficult to study and model computationally (Fodor 1983: 101–119; Fodor 2000).

In the following years, the results of the interdisciplinary synthesis completed by Fodor have fed back into the various areas of Cognitive Science, with the effect of:

a) inducing an ongoing theoretical debate,
b) catalysing productive empirical research,
c) the original idea being co-opted, and further developed in different directions.

In the overall debate, almost all disciplines of Cognitive Science provide relevant data, for instance language acquisition (e.g. a summary of double dissociations in general-domain cognitive deficits and language-specific deficits in Pinker 1995 [1994]), developmental psychology (e.g. Karmiloff-Smith's [1994] idea of

---

Fodor (1983: 12–23) discusses Gall's phrenology's influence on the doctrine of modularity; he rejects the assumption of the morphology of the brain influencing the shape of the skull, but retain his assumption about the narrow cerebral localisation of cognitive functions.

"representational redescription" in the ontogeny of humans), and neuroscience (e.g. Buller and Hardcastle's 2000).

In linguistics, the concept of modularity has recently been employed in models of language processing (usually with modifications: e.g. Jackendoff 2002), and in producing conclusions regarding the evolutionary emergence of the human capacity for language (Bickerton, e.g. 1998).

The idea of the modularity of mind – although substantially transformed – also became to a large extent constitutive of evolutionary psychology. Modules are viewed by evolutionary psychologists to be specialised mental functions, largely innate, evolved in response to the species-specific cognitive demands present in the evolutionary history of *Homo sapiens*. On this account, also called 'massive modularity', it is central cognition (higher level cognitive mechanisms, such as belief-fixation and decision-making) that has a modular structure. This contradicts Fodor's original claim and has sparked a debate (centred on Fodor's [2000] response to Pinker [1997]).

In turn, evolutionary psychology is itself inherently interdisciplinary. In addition to the core of cognitive psychology, it must rely on areas of biology as diverse as population biology and genetics. The key notion, Environment of Evolutionary Adaptedness (EEA) is a theoretical construct that requires a synthesis of data related to extinct hominids (paleoanthropology), their material culture (archaeology), contemporary hunter-gatherer societies (anthropology) and general sociology. In terms of the 'output', research within the framework of evolutionary psychology generates testable hypotheses regarding human sexual strategies and mate choice (Buss 2004, chapters 4–6 and 10–11), human deontic reasoning (e.g. Cosmides and Tooby 2006), and rapid cognition and decision making (e.g. heuristic reasoning, see Todd and Gigerenzer 2000).

Overall, in terms of generated research and theoretical interest, the idea of modularity has turned out to be immensely successful, and its success has been directly dependent on interdisciplinary cooperation on all stages of its development.

1.4.6.2. Interdisciplinarity and Conceptual Metaphor

The other case in point concerns the existence of (cognitively real) conceptual metaphor and the evidence for it supplied by Lakoff and Johnson[58]. On their view – foundational to the whole discipline of cognitive linguistics – metaphor is not primarily a mere way of speaking, but the main conceptual device by means of which we use the experiences and conceptual structures from one domain to make inferences about another. Lakoff and Johnson (1999: 81–88) begin by founding their claim on three kinds of supporting linguistic data, but supplement them by further, mostly extralinguistic, evidence from six major sources (some, admittedly, questionable, but others cogent). These are: psychological experiments[59], historical semantic change, spontaneous gesture studies, language acquisition studies, sign language metaphor studies, and discourse coherence studies. No individual source may seem conclusive on its own, but thanks to the interdisciplinary perspective, a range of converging evidence has been produced whose collective power is compelling.

A striking recent confirmation of the robustness of the conceptual metaphor MORAL PURITY IS PHYSICAL PURITY comes from a recent study by Zhong and Liljenquist (2006). Especially noteworthy are the study's strong interdisciplinary underpinnings, synthesising both theoretical and experimental evidence that cuts across all three levels of description described in 1.3.4. Zhong and Liljenquist cite:

---

[58] The theoretical perspective of Lakoff and Johnson was introduced in *Metaphors We Live By* (1980), but their *Philosophy in the Flesh* (1999) most comprehensively discusses its methodological and philosophical foundations as well as consequences.

[59] The authors mention studies with nine distinct methodologies; however, at least one study quoted by them admits an alternative interpretation. They report improved speed in the recognition of the test sentence in those subjects that had been primed with a metaphorical sentence, as compared to subjects primed with a non-metaphorical sentence. But this effect can be explained by reference to purely semantic content of both the priming sentences.

a) literary evidence (Lady Macbeth washing her hands after crime; this could be extended to Pontius Pilate),

b) cultural evidence (practices of symbolic purification common in most of the world's major religions),

c) the closely related case of moral and physical disgust (which recruit overlapping brain areas, involve common facial expressions, and may involve common behavioural reactions, such as puking)

d) the identity of words and expressions used to refer to aspects of physical and moral purity.

The four-part experiment by Zhong and Liljenquist (2006) complements this list with a fifth line of evidence:

e) thinking about unethical behaviours has a very strong tendency to prime words and behaviours related to physical cleansing; also, the act of physical cleansing reduces the need for other means of restoring moral self, e.g. by way of atonement.

In my opinion, the interdisciplinary perspective, simply indispensable in some cases, proves to be beneficial in many others. Even though it is burdened by potential drawbacks, these should not be viewed as inevitable. In particular, one must pay special attention to ensure that the definitions of the concepts used across disciplines are precise and do not allow ambiguity, and be careful to rely on such sources that minimise the risk of bias or obsolescence of knowledge brought from other areas. If the above requirements are met, and especially in the fields where sound foundations for interdisciplinary research have already been laid, advantages clearly prevail over possible dangers.

1.5. Summary

The goal of the first chapter of this work was to introduce and explain the character of the 'cognitivist perspective' of this work. I introduced Cognitive Science as the modern, naturalistic, interdisciplinary study of the mind. In the context of Cognitive Science, language research is seen not as an autonomous enterprise, but rather as an undertaking highly interrelated with and having profound implications for other disciplines interested in information processing in the mind.

I sketched out a historical picture of the development of Cognitive Science, with its thematic and methodological evolution. *Contemporary* Cognitive Science was given more prominence, as opposed to an earlier, more historical understanding of this label. In later sections, particular emphasis was laid on the idea of interdisciplinary research implied by Cognitive Science, its efficiency, and the minimisation its associated risks. I defended the reliance on interdisciplinarity; this was achieved partly by an in-principle argument concerning the recent increase in the accessibility of academic research, but mostly by analysing in considerable detail two relevant successful applications of this idea.

2. Intrasystemic and extrasystemic principles of concept individuation

In the second chapter of my work, I further develop questions related to the perspective of study. The specific goal to be achieved in this chapter is to discuss in detail and substantiate the crucial research decision, that is the choice of the psychological point of view on concepts, whereby they are understood as a kind of mental representation internal to individual cognitive systems. This decision is already implicit in the cognitivist approach presented in Chapter 1; however, as a fundamental issue, it will be supported by a more exhaustive argument.

2.1. Existential status of concepts

This section (2.1.) develops questions related to the ontological dimension of concepts. The structural framework for the contents presented in this section is informed by extant analyses by the philosophers of mind, notably those by Eric Margolis and Stephen Laurence (especially Laurence and Margolis 1999, 2007, and Margolis and Laurence 2006).

2.1.1. I-language and E-language[60]

It is commonly agreed that, when it does not contradict basic intuitions from natural language in any fundamental way, researchers are at liberty to define their notions of interest in a way that they find most theoretically productive. With language (analogically, concepts in language) this is not unproblematic. With respect to the question of how language exists, people seem to be endowed with a dual set of pre-theoretic intuitions, conflicting but both apparently valid: it

[60] The distinction into I-language and E-language was introduced by Noam Chomsky (e.g. 1986). This distinction is explained in detail below.

appears reasonable to accept language as existing *in abstracto*, independently of particular users, but also as a biologically grounded, ontogenetically developed ability possessed by individual speakers.

This split is reflected in theoretical approaches to language and its study, with Cognitive Science being firmly dedicated to the latter stance. Since the publication of Chomsky's 1986 book *Knowledge of Language: Its Nature, Origin, and Use*, the above distinction has usually been presented in terms of the opposition between E-language and I-language. Chomsky's (1986) terminology can be illustrated as follows:

The studied entity	Its theory	The relevant science
$S_0 \approx$ LAD	UG (Universal Grammar)	C-linguistics
S_s = I-language	(generative) grammar	
E-language	grammar	[A-linguistics]

Table 1. A simplified outline of metatheoretical distinctions in Chomsky 1986 (especially pages 1–52).

The highlighted area in Table 1. stands for *the faculty of language*, comprising both S_0 (the initial state of the language faculty) and S_S (a generalised stable state of the language faculty), as well as the transitional states in between. These are the states of the faculty at different stages in the ontogeny of an individual. S_0 is the hypothesised state of this faculty in a newborn child and for most purposes corresponds to the hypothesised innate, genetically transmissible and species-specific Language Acquisition Device (LAD); its theoretical description is Universal Grammar (UG). *I-language* (internalised language) is a stable state of this faculty, i.e. the state in a competent native speaker of a given language; its theory is simply the grammar of a given language. The name for the study of the

faculty of language committed to psychological adequacy and thus sensitive to psycholinguistic data is *C-linguistics* (for *cognitive linguistics*). *E-language* (externalised language) is typically understood as the totality of actual and potential utterances, in abstraction from language users. *A-linguistics* (for *abstract linguistics*) is the enterprise to characterise E-language with maximised simplicity and descriptive accuracy but with no commitment to psychological reality[61].

It is already evident that in the context of today, several points require clarification regarding the above set of terminology. Firstly, the present understanding of the term "cognitive linguistics"[62] is strikingly and paradoxically different, carrying a distinctive anti-Chomskyan connotation. Currently, this name has come to indicate a particular paradigm within contemporary linguistics, one that has grown out of the dissatisfaction with, and criticism of, successive versions of Chomsky's generative grammar (see also Jackendoff 2002). Today, cognitive linguistics and generative linguistics are popularly seen as rivalling and antagonistic, though they perhaps need not be[63]. It may be sobering to realise that both those approaches are 'cognitive' in the relevant sense, as they share the same central commitments of Cognitive Science: most importantly mentalism, and also representationalism and broadly construed computationalism (though the preferred character of computation may be different in each case). Indeed,

[61] See also Jackendoff (2002: 29, footnote 6) for a brief explication of the difference between the distinctions into *I-language*/*E-language*, and those into *competence* and *performance* (also introduced by Chomsky), and into *langue* and *parole* (introduced by Ferdinand de Saussure).

[62] The notion of "C-linguistics" is proposed and discussed in Chomsky (1986: 34–36), who bases his analysis on the "Leading Questions" put forward in an earlier paper by the linguistic philosopher Scott Soames.

[63] Such is, for example, the opinion of Ronald W. Langacker (1987: 28), when he speaks of what he calls the "exclusionary principle": a natural – but mistaken – tendency to assume that a given problem or phenomenon has only one 'correct' solution or approach, so that embracing it must imply excluding all the others.

from the outside perspective they are often rightly seen as variants of the same overarching paradigm (see, e.g., Andrzej Pawelec 2005: 41–42, 147–153).

A special issue that deserves mention as the reason of many controversies is Chomsky's understanding of language (more precisely, "the faculty of language") in a specific, syntax-centred way. Chomsky has remained adamant in his defence of his core philosophical conviction regarding the nature of language: when properly understood, the human species-specific capacity for language is in its essence equivalent to a generative computational core, responsible for the hierarchical productivity of syntax and largely dissociable from the rest of cognition. This has ramifying consequences. The latest large debate involving Chomsky concerned the topic of the evolutionary emergence of language. In this polemic, Chomsky, Marc Hauser and Tecumseh W. Fitch (Hauser, Chomsky, Fitch 2002; Fitch, Hauser, Chomsky 2005) were opposed by Steven Pinker and Ray Jackendoff (Pinker and Jackendoff 2005, Jackendoff and Pinker 2005) who attributed the bulk of the controversy not to evolutionary issues but precisely to the differences in the guiding assumptions on the nature of language (for critical discussion, see Wacewicz 2012, Wacewicz and Żywiczyński 2014).

2.1.2. I-concepts and E-concepts

Despite the abovementioned restrictedness of Chomsky's I-language, there exist no fundamental reasons against using his insight and terminology to other aspects of the study of language, the study of concepts being of primary importance. Precisely this is done by Ray Jackendoff (e.g. 1999 [1989]: 305–313, 1996: 539–542), who, on the analogy with *I-language* and *E-language*, speaks of *I-concepts* (as opposed to *E-concepts*) and of their study, *I-semantics* (as opposed to *E-semantics*).

In his discussion of this matter, Jackendoff (1999 [1989]: 309) illustrates it with an apt quotation from the American philosopher of language, David K. Lewis:

> I distinguish two topics: first, the description of possible languages or grammars as abstract semantic systems whereby symbols are associated with aspects of the world; and second, the description of the psychological and sociological facts whereby a particular one of these abstract systems is the one used by a person or population. Only confusion comes of mixing these two topics.

Accordingly, two major approaches to concepts in language can be distinguished following the criterion given above: psychological (in this sense, 'psychological' is treated as synonymous with 'cognitive' and 'mentalistic') and nonpsychological (in certain contexts variously referred to as 'philosophical', 'logical' or 'semantic')[64].

The semantic perspective maintains that concepts are not psychological entities but rather that they exist *in abstracto* – independently of them being entertained by any particular mind. It dates back to as early as Plato's strong realism about abstract ideas, being later defended notably by Gottlob Frege (especially 2001a [1952] {1892}, 2001b [1956] {1918}; see 2.1.3.). Still another influential thinker that could be named in this context is Karl Raimund Popper (e.g. 1978)[65], whose doctrine of three worlds or 'kingdoms' postulated an independent level of existence for the world of abstract ideas (as long as they are in a format suitable to be expressed in language). The semantic perspective and thus stands in clear contrast to the psychological perspective assumed in this

[64] See also the entry „Psychologizm w logice" [Psychologism in logic] in Marciszewski 1970: 232.

[65] Karl R. Popper's 1978 lecture is a follow-up developing specifically the issues related to the ontological status of the „three worlds" that were introduced in a previous, larger work (Popper, Karl Raimund 1972. *Objective Knowledge: An Evolutionary Approach*. Oxford: The Clarendon Press.).

work; their respective foci of study can thus be characterised as I-concepts and E-concepts[66].

The two most critical observations are that, firstly, the decision between the study of I-concepts or E-concepts is unavoidable (but see Laurence and Margolis 2007 for a partly differing view), and secondly, that it is nonempirical, being entirely goal-driven, that is, dependent on the class of problems that are to be solved. As the two approaches remain clearly distinct in terms of the ontology of the object of study, this translates into their divergent theoretical interests. These are: theory of mental representation, inference and reasoning, and (developmental) conceptual change for the cognitive approach; and truth conditions, hermeneutics, and (diachronic) conceptual change for the semantic approach[67]. It is important to stress once again that no general questions of legitimacy arise for either perspective, only a division of research labour.

As stated above, the semantic and the psychological perspectives view concepts as two different kinds of beings. This, however, does not necessarily imply that their methodologies are likewise entirely dissociable, and that the one does not bear any relevance to the other. Firstly, the goal of the psychological/cognitivist approach clearly does not consist in idiosyncratic descriptions of individual minds, but rather in achieving some sort of generalisations valid (that is, valid *ceteris paribus*: when one excludes the

---

[66] In a more historic context, of relevance at this point is the mediaeval scholastic discussion between the positions of realism and nominalism (with *conceptualism* sometimes proposed as a middle ground between them). The nominalist Wilhelm of Ockham (Occam), c.1288–c. 1348, is considered to be a central figure in this debate.

At the same time, one must not forget that the mediaeval discussions of the problem of universals were partly incommensurable with today's naturalistic ontological inquiries as a result of being set in a distinct theological context.

[67] This list is based on a conference talk and a book (2007) by Robert Piłat.

Jackendoff (1996: 540–541) proposes a much broader list of the "basic questions for a theory of I-semantics": 1) the nature of meaning, 2) correspondence to language, 3) correspondence to world, 4) brain instantiation, 5) developmental questions, 6) evolutionary questions.

variables identified as 'irrelevant') for all human minds, or at least for a large subset of human minds, e.g. the minds of people who participate in a given linguistic community. It is implicit that a certain isomorphism in all relevant respects must hold between individual minds. Despite concepts being mental entities uniquely possessed by individuals, their description must nevertheless be accomplished in an objective way that abstracts from concrete individuals – which is again analogous to the study of I-language – and exploits the assumed isomorphism. Secondly, concepts as vehicles of thought and as meanings of linguistic expressions have semantic properties arising from the relation to the external world: if not by reference to the elements of the external world, then at least by entering the causal chain in the system's interactions with the world.

Nevertheless, the semantic and the psychological approaches – while related and to a degree mutually penetrable – show very little promise for a successful synthesis. In this respect, I share the opinion of Robert Piłat (Piłat 2007: 46–52, 341)[68] and reject even the limited optimism of Laurence and Margolis (2007). This fact reinforces even further the necessity for an unequivocal decision between them in any theoretically interesting study of concepts.

2.1.3. Gottlob Frege: metaphysical views and their influence

Finally, the equal viability of both the psychological and the semantic approaches to concepts requires a comment. This is the case largely because of the status of the logician Gottlob Frege and the enormous influence of his works, in which the psychological view was explicitly questioned.

---

[68] In his recent comprehensive analysis of the subject matter of concepts, Piłat (2007) remains sceptical not only regarding the prospect of the unification of both perspectives, but also regarding one of the approaches ultimately superseding the other. In this author's opinion, the problem area of concepts is inherently and irredeemably two-dimensional, and an account from only one perspective can never be exhaustive, leaving out important problems that can only be formulated within the other perspective.

To Frege, anything that could aspire to the name of 'concepts' must be patently nonpsychological; to him, the tasks that people associate with concepts cannot be fulfilled by any kind of mental entities. The mental beings, which he terms *ideas*, are highly idiosyncratic and thus have little intersubjective stability; they are also of an imagistic (not linguistic) format, perhaps akin to percepts or even qualia. In short, ideas are entirely unsuited for the role of grounding propositional contents. This later function must be carried out by a different type of entity, namely *senses*[69].

Even in a brief discussion of Frege, it is critical to avoid the terminological pitfalls resulting from the intricacies of his nomenclature, magnified by the troubles with translation. Fregean *concepts* (*Begriff*) are functions (Frege 1960a [1952] {1891}: 30) in the logico-mathematical sense, that is they map arguments to one of the two truth values – True, if something "falls under" the concept, or False, in case it does not (Frege: 1960a: 30). In and of themselves, they are 'unsaturated', that is require a proper name or its equivalent to become 'saturated', thus forming a complete sentence in the logical sense[70] (Frege 1960a: 24). Functions are linguistically rendered as predicates, e.g. 'is a horse'.

It is sense (*Sinn*) that should be considered to be the Fregean category closest to concepts as understood in this work. This is so because senses are the kinds of being that have cognitive import, and they correspond to/underlie/constitute the meanings of names as distinguished from their references (*Bedeutung*), i.e. the denoted objects in the real world – broadly

[69] "A painter, a horseman, and a zoologist will probably connect different ideas with the name <Bucephalus.> This constitutes an essential distinction between the idea and the sign's sense, which may be the common property of many and therefore is not a part or a mode of the individual mind. For one can hardly deny that mankind has a common store of thoughts which is transmitted from one generation to another." (Frege 2001a: 8)

[70] I.e. a *proposition*. Proposition is a fundamental term in logic and linguistics. Marciszewski (1970: 261) explains that „a proposition [Polish: „sąd w sensie logicznym"] is the meaning of a declarative sentence… that which is common to a certain class of psychological experiences".

understood so as to include abstract objects. Senses make it possible to reflect the epistemic element of meaning, e.g. the fact that coreferential expressions with different senses are not substitutable in intensional contexts[71]. Finally, 'meaning' might itself be confusing, since the German term *Bedeutung* (normally translated as 'reference' – e.g. in Frege 2001a) in German literally means 'meaning'.

Note that on the original Fregean account, *senses* can be assigned either to names or to complete propositions, *not* to classes of things. This needs to be mentioned for reasons of clarity, since it may be felt that senses, given such a limitation on one hand and the link between concepts and categorisation on the other, have little relevance for the discussion of concepts. Despite this potential reservation, when looking into the matters of ontology, another motivation must be seen as deciding: the fact, already observed above, that it is the level of senses on which meaning is cognitively available to the cogniser. Likewise, in the philosophical-linguistic literature it is widely accepted that in this respect, concepts (as discussed in contemporary literature) should be compared to Fregean senses rather than Fregean concepts (e.g. Peacocke 1995).

It appears that Frege's concern was primarily motivated by the issues of stability and shareability of meaning. Undeniably, successful communication is possible between different people and across different points in time (e.g. understanding old texts). What Frege took great pains to underscore was that despite the presence of fine-grained individual differences in understanding, those differences were (to him) decidedly second-order, whereas on the level coarse-grained enough for logic to capture, there evidently appeared to exist "a common store of thought" that was invariant both through time and across particular natural languages (Frege 1960a, 1960b [1952] {1892}: 46 footnote 1).

---

[71] For example, in „Oedipus knows that he wants to marry *a*" or „Oedipus thinks he sees *a*", substituting *a* by *b* can change the sentence's truth value even if, referentially speaking, *a=b*, e.g. *a* = Jocasta, *b* = Oedipus' mother: „Oedipus thinks he sees Jocasta" may be true with „Oedipus thinks he sees Oedipus' mother" being false. Similar cases are sometimes referred to as 'Fregean puzzles'.

For Frege, the only way to explain this fact was through endowing senses with an independent ontological status.

Thus, on Frege's account any mentalistic perspective on concepts (as units of reasoning, linguistic meanings, etc.) must be deficient in that it fails to explain the crucial issue of the stability and shareability of meanings. However, on the present analysis, this appears to be the *only* such difficulty. In other words, my claim is that the challenge from the Fregean perspective can be *reduced* to the issue of the stability and shareability of meanings and that apart from it, there are no other compelling reasons for the rejection of concepts as psychological entities. It follows that the defence of the mentalistic perspective, which is the aim of this part of the dissertation, should consist in removing this particular objection. For the purposes of this work, it is sufficient to show that a mentalistic view can be reconciled with all the required characteristics of concepts, including shareability and stability of meaning. Specifically, this can be achieved by means of type/token relation, and the appropriate explanation shall be offered in section 4.2.6.

One could also remark that Frege's account might play down, but does not *eliminate* the alleged instability of the cognitive aspect of meaning. For concepts, or senses, to play any part in cognition or language, they must at some point be employed by individual cognitive agents, and at this point psychology retains its relevance. It could be argued that when a person 'grasps' a sense or a complete thought, the very act or relation of grasping now becomes psychological and might be accomplished in a way highly variable from person to person. Thus, far from removing idiosyncrasies, such a solution merely relegates them to the background.

What is worth emphasising is that the above remarks should by no means be taken as carrying disparaging undertones against Frege. A responsible evaluation of a system of ideas can only be conducted with regard to a larger context; most importantly, to the theoretical goals that this system is devised to accomplish. In the case of Frege, these were goals related primarily to logic and

philosophy of mathematics, with only secondary applicability to the description of natural languages. It is only later that his ideas have been transplanted to analytic philosophy in general, where they continue to exert enormous influence.

2.2. Internalist and externalist principles of content-individuation

The argument described in the previous section is supposed to show why it is legitimate (indeed, necessary) for Cognitive Science to treat concepts, not as *abstracta*, but rather as mental entities internal to particular cognitive agents. This question is an ontological one, regarding the kind of existence that could be predicated of concepts.

A different question is methodological, and it concerns the extent to which it is viable to study concepts (as well as other aspects of cognition) taking into consideration only these factors that are purely system-internal, with the consequent exclusion of any system-external factors. Note that this question is related to, but distinct from, the former, ontological one and arises once we have already agreed to view concepts in the mentalistic way illustrated above: concepts *are* mental representations over which computational operations can be defined.

One of the postulates of Chapter 1, where Cognitive Science was introduced as a theoretical framework for this text, was that at least in mainstream Cognitive Science, the *internalist* stance was the only feasible perspective. Nevertheless, an extremely influential argument by Hilary Putnam (1997 [1975]), later followed by Tyler Burge (1979), has established a case for the alternative view, *externalism*, giving rise to an ongoing theoretical debate. Not surprisingly, many (e.g. Burge 1986, also Żegleń 2003: 166) share the opinion that the externalistic stance may call into question the philosophical soundness of the premises of internalistically oriented study of language and mind. If externalism is true and so – as Putnam (1997: 227) famously stated – "meanings just ain't in the head", then the psychologism of Cognitive Science might find itself on shaky ground.

Of the two alternatives, internalism is the natural stance in the sense that it corresponds to our standard, pre-theoretical intuitions. Therefore, it is best defined negatively, i.e. in opposition to externalism, even if there is a co-dependence, since externalism is a weaker position. The exposition of externalism, in turn, crucially depends on the thought experiment introduced by Putnam in his classic text. The experiment serves as an indispensable *intuition pump*[72] without which it is impossible to formulate the externalistic standpoint, and as such, it is canonically referred to by commentators on the issue of semantic internalism/externalism (cf. e.g. Gabriel Segal 2000, Katalin Farkas 2003, Joe Lau 2006). For those reasons, I begin with presenting Putnam's argument in the section 2.2.1., but postpone my discussion and the consequent rebuttal until 2.2.3. This is due to the fact that misunderstanding regarding internalism/externalism is rife (cf. Farkas 2003). Numerous overlaps and intricacies in terminology together with the subtleties of distinctions provoke confusion, therefore it is necessary to clarify the terminology related to those issues, as well as dispel the most common misapprehensions. Precisely such will be the function of the intervening section 2.2.2.

2.2.1. Externalism: arguments by H. Putnam

In his 1975 classic, "The meaning of <meaning>", Hilary Putnam constructs a though experiment[73] in which the reader is invited to envisage an imaginary planet, Twin Earth (Putnam 1997 [1975]: 223). Twin Earth is Earth's particle-for-particle equivalent, and is fully identical to Earth in all respects, so that for all

[72] „Intuition pump" is a term devised by the American philosopher of mind Daniel Dennett, and applied by him in a negative context: basically, an intuition pump is simply a rhetorical tool (e.g. Dennett 1998: 24). See also the next footnote.

[73] Strictly speaking, a less dignifying name of 'speculative argument' would be more appropriate. In 'thought experiments', the 'experimental' element is missing, as the outcome of the 'experiment' is left for our intuitions to decide. Consequently, thought experiments have no scientific value and are treated with increasing reservations within philosophy. See 2.2.3.

objects, persons, events, etc. on Earth, exact Twin counterparts exist on Twin Earth. Consequently, the above is also true of a person Oscar, who by this hypothesis has an exact Twin duplicate, Twin Oscar, or "$Oscar_2$" (Putnam 1997: 224) – as well as of his internal constitution and mental states, which by this hypothesis have exact counterparts in the internal constitution and mental states of Twin Oscar (1997: 224). The only exception from this rule, and thus the one and only difference between the two planets, concerns the nature of *water*: the colourless, odourless substance known to us by the name of 'water', found in rivers, lakes, seas, and oceans, forming precipitation, turning into ice or steam under certain well defined conditions, quenching thirst, and having the chemical structure H_2O, on Twin Earth has exactly the same properties but a very different underlying chemical structure, being instead a highly complex compound, for convenience abbreviated to 'XYZ' (1997: 223–225).

Since Oscar and Twin Oscar are exact, particle-for-particle doppelgängers (but see 2.2.3.), according to Putnam (1997: 224), their bodies and thus mental states are identical[74], as are the intensions of the words they use. Still, when Oscar thinks of water and uses the word 'water', this word refers to the liquid with the underlying chemical structure H_2O, and when Twin Oscar thinks of water and uses the word 'water', this word refers to the liquid with the underlying chemical structure XYZ (Putnam 1997: 224). Even though 'what's in the head' remains the same, the extension of 'water' does not, and so the *meaning* of 'water' is different in each case. This, together with similar complementary examples[75], leads to the conclusion that the individuation of the meanings of words does not depend solely on the internal states of the cognitive

[74] An additional assumption might be necessary regarding the identity of Oscar's and Twin Oscar's mental states (narrowly understood – see 2.2.3.), but in contemporary Cognitive Science and philosophy of mind this is already guaranteed by the identity of their bodies (see 1.3.3.)

[75] Other examples adduced by Putnam include: aluminium/molybdenum (1997: 226–227), elm/beech (pages 226–227 and further in the text) and jadeite/nephrite (p. 241).

agent, but is also sensitive to factors external to that person's body; a conclusion summarised by Putnam's widely-cited dictum: "[C]ut the pie any way you like, <meanings> just ain't in the head" (1997: 227).

Extending the example further, if both planets are imagined in the year 1750, that is a few decades before the birth of modern chemistry, the difference between H_2O and XYZ is undetectable (Putnam 1997: 224–225). But, at least according to Putnam, the difference in the meaning of the word 'water' obtains nonetheless, and the inhabitants of Earth and Twin Earth refer, respectively, to H_2O and XYZ, even though they do not know that. This assumption is made in order to show the historical stability of meaning, which is dependent on a given word's extension[76] in the real world and independent of epistemic factors.

Note, too, that in its original formulation, the argument and its conclusion concern specifically the meanings, or 'contents', of words in natural language rather than mental states. Still, both the argument and its conclusion transfer naturally to the meanings ('contents') of concepts and, in consequence, mental states (e.g. Segal 2000: 24). Such a theoretical step has indeed been taken by Putnam himself (1995 [1981]: 18–19), with the 'H_2O/XYZ' distinction replaced with the 'elm/beech' version.

2.2.2. Common misunderstandings concerning internalism and externalism about content

Internalism as a theoretical term usually refers to one of the following: *internalism in ethics* (moral internalism), *internalism in epistemology* (internalism about justification), *internalism in the philosophy of language and philosophy of mind* (internalism about content). These three views, although loosely related, are distinct positions rather than varieties of one overarching view. The internalism relevant to the present work is internalism about content, or 'meaning'.

[76] The notion of *extension* (and related notions) is defined in section 3.1.3.

Internalism in ethics concerns the closeness of the relation between moral opinion and motivation for action (e.g. Björnsson 1998: 9). It maintains that the two are inseparable: having a moral belief necessarily implies being motivated for taking a certain course of action.

Internalism in epistemology concerns the conditions on which a belief can be properly said to be justified (e.g. Conee and Feldman 2001, Goldman 2001). According to internalism, all factors relevant to establishing whether a given belief is justified are internal to the subject or, on a stronger construal, all such factors are consciously accessible to the subject.

Within internalism about content, two strains can be identified, one of which – *semantic internalism* – regards the meanings of linguistic expressions, the other regarding the contents of concepts, and thus of the mental states composed of those concepts. Nonetheless, as explained by e.g. Joe Lau (2006), the name *semantic internalism* is sometimes extended to cover internalism about content *tout court*, thus encompassing its 'mental' subtype as well. According to Lau 2006 (who also quotes Colin McGinn), both those strains are best regarded as essentially one theoretical position coming in two versions, with internalism about mental content not constituting a separate view, but rather being a natural extension of internalism about expressions from the philosophy of language to the philosophy of mind. In line with the thematic focus of this work, I will concentrate on the question of concepts, and consequently, on internalism about mental content.

Internalism about content is sometimes also termed *individualism* (cf. Lau 2006; and Robert Anton Wilson [1999] discussed this term under such a heading in *The MIT Encyclopedia of The Cognitive Sciences*). Unfortunately, 'individualism' as a term is also burdened with a broad range of quite discrepant meanings in many disciplines including ethics and economy. The term 'internalism' will be used in the rest of this work.

Internalism about mental content – and, accordingly, externalism about mental content – can be stated in an alternative, but logically equivalent way, as

views regarding the question of *content individuation*, i.e. how the contents of mental states should be individuated. Whereas internalism holds that the contents of the mental states of a cognitive system can be individuated purely on the basis of factors that are internal to this system (what is 'in the head'), the externalist account allows the environment to play a role in the individuation of mental states. This follows directly from Putnam's contention that $Oscar_1$ and $Oscar_2$ have identical psychological states but that their states have different *meanings* (1997: 224 and henceforth in that text). In other words, internalism can be construed as the view claiming that being in a given mental state is an intrinsic property of the system (as e.g. *being spherical* is an intrinsic property of an object), and not a relational property (as e.g. *being someone's wife* is)[77]. In still other words – using the notion of supervenience introduced in section 1.3.3. – on this view, the content of a given mental state supervenes exclusively on the physical states/events of the system entertaining that state, and on no physical states/events external to that system. It follows that for any two physically identical individuals (most importantly, individuals whose *brains* are micro- and macrostructurally identical), they would necessarily entertain exactly the same mental states (i.e. states with the same contents) regardless of any relations to their external environments. That is, the contents of their mental states would continue to be the same even if they were embedded in different environments: physical identity of their cognitive systems would alone be enough to ensure the identity of the contents of the mental states. In contrast, the externalist position maintains that it is possible for two physically identical individuals (most importantly, their brains being exact, particle-for-particle physical duplicates) to

[77] The distinction of properties into relational and intrinsic might be somewhat hard to define rigorously. Given enough inventiveness, any property can be argued to be a relational property, e.g. *being black*, seemingly intrinsic, can still be said to depend on a number of external factors such as lighting conditions or perceptual capacities of observers. Still, I assume the distinction to be intuitively clear, and sound enough for a productive application.

have different (contents of) mental states, the difference resulting from their being embedded in different environments.

The difference between the internalist and the externalist positions regarding content can be further explained as follows. One may assume that a person's belief that, for example, 'the cat is a kind of animal' supervenes on a specific neuronal configuration *A* in her brain. On the internalist standpoint, if the configuration *A* is re-created in some other person, that other person will necessarily have the same belief, i.e. that 'the cat is a kind of animal', with the same content. On the externalist standpoint, this is *possible but not necessary*: it is at least possible that for that person to have that mental state, the re-creation of the configuration *A* may not be sufficient and that the appropriate relation to her environment may also be required.

As noted above, internalism and its rival view, externalism, often fail to be construed correctly, with some misconceptions being rather profound. Two characteristic misconceptions of this kind, almost self-evidently false but still relatively widespread[78], are: taking internalism to claim that the environment produces no effect on the cognitive system, and taking externalism to claim that only external factors play a part in determining content. I will consider these major fallacies and then turn to relatively minor problems.

Internalism does *not* stipulate that the external environment has no effect on the cognitive system and its mental states. Unless one proposes the cognitive agent to be a Leibnizian monad[79], it is trivially true that the environment does act

---

[78] A good example is the relevant entry, in the popular reference work *Wikipedia*, either incomplete or imprecise, or erroneous (permanent link): http://en.wikipedia.org/w/index.php?title=Internalism_and_externalism&oldid=79791497#Semantics

[79] Gottfried Wilhelm Leibniz (1646–1716), a German philosopher and mathematician, the author of the theory of monads. In Leibnizian philosophy, monads are the basic, atomic elements of the universe. The interactions between monads are only apparent and lack any real causal character since each monad is a totally self-contained whole, thus being perfectly independent of any other monads.

causally on the cognitive agent, thus affecting him in diverse ways; denying this would turn internalism into a clearly untenable position. Internalism remains neutral with regard to what kinds of external influences lead to what kinds of mental states. For instance, a person may acquire the belief that 'the cat is a kind of animal' through inferring it himself from previously learned knowledge, or, in a fortuitous, but still nomologically possible way, through electrical stimulation of the cerebral cortex. The influence of the environment is relatively indirect in the former case, very direct in the latter, but is clearly present in both. All such cases are perfectly compatible with internalism as long as at *any specific point in time* the person in question has this belief exclusively in virtue of what is inside his body, without any additional requirement of standing in a particular relation to his environment.

Externalism, in turn, should not be construed as claiming that the content of a mental state is determined *solely* by external, relational factors. While internalism requires that all determinants of the content of a mental state must be system-internal, externalism does not insist on the exact opposite; for externalism to hold, it suffices that *at least some* determinants are system-external. Everyone, including externalists, must agree with the trivial fact that it is system-internal states that play the decisive role in determining content, and any statement to the contrary is easily disproven. Two people (not to mention e.g. a person versus an animal or an object), despite being embedded in identical spatiotemporal environments, could nevertheless have drastically different mental states precisely in virtue of the differences in their internal constitution, in particular in their cerebral tissues (e.g. one of them being a stroke victim) or in their current electrical activity of their brain (e.g. one being asleep).

Furthermore, as observed by e.g. Lau (2006), there are problems concerning the scope of internalism and externalism. Neither view lays claims to being valid for the totality of concepts, and, consequently, for all mental states. Mental states whose contents are logically or analytically true, e.g. (believing, hoping, fearing, etc. that) 'a cat is a cat' or 'no lie is true' favour the internalist

approach, as there seems to be no environmental difference capable of rendering such content true for one person but false for another person – even independently of the physical identity of those persons. On the other hand, for contents involving various forms of overt indexicality[80], e.g. including deictic concepts such as HERE, THIS, YOUR, etc., internalism appears to be insufficient. The content of the mental state 'The keys are over there' undoubtedly depends partly on the spatial location of the person entertaining that state. What follows, the meanings of such concepts, and consequently of the propositional contents they figure in, are strictly dependent on the external contexts, and hence 'broad', or 'wide'[81].

Finally, externalism about mental content frequently becomes associated with varieties of the "extended mind" theory, developed by Andy Clark and David Chalmers (e.g. Clark and Chalmers 1998), whose consequences lead to so called 'active externalism'[82]. The association is not unfounded (in fact, it is explicitly encouraged by Clark and Chalmers), but here it is important to appreciate the differences between the standard (i.e. 'passive') and 'active' versions of externalism, especially in the context of their relation to 'standard' internalism.

Active externalists subscribe to Putnam's dictum that "the meanings ain't in the head", but they understand it in a markedly different sense. For example, when a person performs a cognitive operation (*mental*, thus system-internal operation), she may choose to exploit certain physical features of her body and/or

[80] Indexicality is the essence of externalism: an equivalent way of formulating externalism about the contents of 'standard' lexical concepts is by proposing that they contain a covert indexical component (cf. Putnam 1997: 233–235).

[81] Strictly speaking, theoretical strategies can be proposed that qualify or amend that last conclusion. This, however, need not concern us here, since deictic concepts do not constitute standard lexical concepts and therefore are peripheral to the interests of this work.

[82] Cf. a short discussion by Żegleń (2003: 166–167) who also classifies 'active externalism' as a kind of externalism in general, but stresses the qualitative difference between the 'active' and 'passive' kinds.

her environment in order to aid herself in the completion of those tasks. Clark and Chalmers (1998: 8–9) give the example of rotating objects in a game of Tetris, or adding numbers; this can easily be extended to other examples such as planning the next move in a game of chess, doing the shopping, etc. Physical instead of mental rotation or calculating the sum with the help of one's fingers or a pocket calculator instead of purely mentally – and analogically, moving the actual piece instead of deciding the next move without looking at the chessboard, and relying on one's notebook (or, in future, a memory chip) instead of one's 'hippocampal' memory – are almost invariably more effective, and may sometimes be indispensable for achieving the goal. Thus, Clark and Chalmers assert that human cognitive processes extend beyond the body, and that the mind, although conceived in a physicalist way, certainly extends beyond the brain. In other words, for some purposes it is more fruitful to think of the cognitive system as comprising not just the central nervous system, but the whole body, plus perhaps certain reliably present features otherwise classified as elements of the 'external environment'.

In my opinion, rather than lending support for standard externalism, active externalism is best seen as addressing a separate question, and one that is neutral regarding the status of internalism, namely, the question of the boundaries of the cognitive system. If the boundaries of the cognitive system become extended to embrace its body and some of its immediate environment, the meanings may not be in the head, but arguably remain inside the cognitive system. Such a construal does not invalidate internalism in any way because it allows the contents of the mental states to be individuated purely system-internally and independently of system-external relations[83].

[83] Two more ancillary considerations seem relevant about the compatibility of internalism with the „extended mind" approach. Firstly, it should be noted that its proponents are usually very modest in extending the boundaries of the cognitive system and do not claim that „the mind is everywhere" (see e.g. Anderson 2005). Secondly, the extended mind approach is generally considered to be more successful in dealing with low-level cognitive processes such as

2.2.3. Case against externalism

A relatively direct, but at the same time somewhat less interesting, way of rehabilitating the internalist focus of (broadly constructed) Cognitive Science is through making partial concessions to externalism: agreeing to its overall validity, but questioning its scope. Content, it may be claimed, can be conceived of in two ways, 'narrow' and 'broad'[84], each having certain theoretical usefulness. One variety (call it 'wide' or 'broad' content) indeed has truth-conditional criteria of individuation, which means that it is sensitive to external factors and differs for Oscar and Twin Oscar in Putnam's 'H_2O/XYZ' thought experiment described in 2.2.1. But this does not eliminate another notion of content (call it 'narrow' content), one which retains purely psychological criteria of individuation (i.e. remains locally supervenient; Segal 2000: 18) and stays the same for both Oscar and Twin Oscar (e.g. Botterill and Carruthers 1999: 132, 137). The notion of narrow content reflects the fact, fundamental to the construction of the externalist argument, that Oscar and Twin Oscar are psychologically identical in important ways, or in Putnam's (1997: 221) own words, are in the same psychological state[85].

The key consideration is the putative sameness of Oscar's and Twin Oscar's mental states in terms of their *causal powers*, in the sense of their roles in the mental life of the individual. It is very hard – and perhaps even impossible – to imagine any difference in behaviour between $Oscar_1$ and $Oscar_2$ that could be occasioned by the difference in the wide contents of their mental states. Their different 'broad' contents notwithstanding, it is uncontroversial that their

---

perception or motor control, but less successful for higher-level, „representation hungry" problems engaging linguistic or conceptual operations (see e.g. Clark 2006).

[84] For a discussion of wide versus narrow content see, e.g., Segal 2000.

[85] Although on a rigorously externalist interpretation transferred to mental contents, he would not be allowed to say this, since the mental states of the 'twins' have different potential truth values, and so are different ('wide') states – see Farkas 2003: 190.

behaviours connected to water-related mental states will be identical: if Oscar/Twin Oscar is thirsty and he believes there is water in the glass, this belief will prompt him to behave in the same way (drink water) regardless of whether his mental state refers to H_2O or XYZ. Psychological explanations, by their very nature, are defined over what is cognitively accessible to the subject, hence over contents that are narrow[86].

Accordingly, the straightforward but also less interesting possibility for defending the programmatic focus of Cognitive Science on the internal environment of the cognitive agent – rather than on the many possible aspects of its relation to its external environment – is by declaring that Cognitive Science simply does not need to be interested in wide contents. However, a more interesting alternative consists in the repudiation of externalist claims, which can be achieved through denying the validity of the central thought experiment that lays the foundations for the entire externalist position.

As already noted in the sections above, in terms of their theoretical status a certain disparity holds between internalism and externalism. The former position has the advantage of being the default view that one naturally accepts in the absence of contrary evidence – indeed, it was only after Putnam's (1997 [1975]) paper that an alternative to it could be formulated. In contrast, externalism requires that our original intuitions be reformed, which is achieved by the means of the thought experiment described above. Therefore I conclude that the case for externalism, to a very considerable extent, stands or falls with the validity of the quoted argument: if it can be targeted for convincing criticism, externalism loses its main rationale.

[86] Cf. Segal (2000: 121): "I think that psychology as it is practiced by the folk and by the scientists, is already, at root, internalist… The basic apparatus of psychology does not mandate externalism. So ascriptions of content that are made when practicing good, correct psychology are already internalist: the contents they attribute are already narrow".

It can be demonstrated that Putnam's 'H_2O/XYZ' thought experiment displays numerous shortcomings of various gravity. Some of these are of serious nature; they will be discussed in turn and proposed to form conclusive evidence against this type of thought experiments, thus invalidating at least one strain of externalistic theory. Other faults are related to the experiment's 'packaging', not its conceptual core. They are relatively minor and could be repaired or eliminated simply by restating essentially the same experiment in a different way. Still, the status of the 'H_2O/XYZ' experiment as externalism's flagship example requires that such limitations be addressed as well.

Firstly, one may consider a relatively trivial objection mentioned by many commentators (e.g. Lau 2006, Segal 2000): it is not technically possible for Oscar and Twin Oscar to be physically identical in the way proposed in Putnam's argument. This is due to the fact that the bodies of people contain water, that is H_2O on Earth and, presumably, XYZ on Twin Earth. The fact that body water accounts for more than half of a person's weight makes the difference between Oscar and Twin Oscar rather significant. This, of course, is a minor objection, but at the same a symptomatic one, since it hints at more fundamental underlying limitations.

In what follows, however, I formulate and develop several reservations of a much more serious nature. The first such objection, when put bluntly, is that if its conclusions are to be applied to the real world, the experiment makes little sense. When one is willing to follow the proposed situation to its immediate logical consequences, it turns out to generate a series of unacceptable side effects. A 'world' is a holistic system, wherein the substitution of one element for another is likely to result in a domino-like cascade of non-trivial corollaries. For example, replacing all H_2O with XYZ would in fact require the replacement of all compounds capable of producing water in chemical reactions, such as acids and bases, with some XYZ-compatible counterparts, and eventually supplanting entire chemistry with a new one (with ramifying consequences for physics, biology, etc.). The alleged conceptual possibility of the experiment exploits

human epistemic limitations: it is 'conceptually possible' only to the extent to which people are unable to immediately think of all the relevant consequences ensuing from the experiment's design.

As an illustrative example, I propose to consider a different thought experiment, involving a Twin Earth exactly like our own in every respect, including, for instance, the form and exact contents of physics textbooks – but lacking the phenomenon of gravity. The unacceptable consequences may be somewhat more immediately evident in this latter case, but both experiments are essentially the same – both are conceptually possible only in the sense that it is indeed possible to formulate them verbally without immediate discomfort. In short, Putnam's 'H_2O /XYZ' experiment is very clearly nomologically impossible[87].

In fact, acknowledging the legitimacy of arguments framed in terms of twin worlds has the unpleasant consequence of leaving no nonarbitrary ground for discrediting other counterfactual arguments, however absurd. Consider Putnam's contention that after the discovery of the chemical structure of water (H_2O) no change to the meaning of 'water' took place: 'water' had referred to H_2O even before the invention of modern chemistry because on Earth, the liquid in question had always had the structure 'H_2O' (Putnam 1998: 108–109, reasserted in Putnam 1995: 24). But actually, there is no way of knowing this with absolute certainty. One can at least imagine the possibility that before 1750, water on Earth might have been XYZ, only to turn miraculously into H_2O several years later. This possibility, however grotesque, is on the whole no less improbable than Putnam's original example. Giving credit to Twin Earth

[87] Similarly, it is not at all clear whether it is possible that a substance could have a total spectrum of phenomenological properties (not just some of them) identical to those of H_2O, but a different chemical structure. According to Thomas Samuel Kuhn (cited in Segal 2000: 25), it is at least very likely that H_2O is the only nomologically possible chemical structure capable of displaying precisely those properties. The same point is made and then further developed by Fodor (1994: 28–30).

eccentric experimental design automatically prevents one from discarding such bizarre considerations as nonsensical.

More importantly and more productively, Putnam's argument can be shown to rely on invalid intuitions derived from referential semantics. In my opinion, the problem with Putnam's argument originates from his fallacious assumptions regarding the privileged status of natural sciences in describing reality: to Putnam, properties as described by a natural science such as chemistry seem to have a different, and 'stronger', standing that those accessible phenomenally. On Putnam's view, properties of the latter type do take part in the formation of meaning by entering the words *stereotype* (Putnam 1997: 230, 247–252, 269), but they are 'weaker' in that do not decide about the word's *extension*. Extensions are drawn along the lines established by a natural science (here: chemistry; but see the reservation below). In other words, Putnam presupposes that the way in which all samples of H_2O belong to the same type is in some way more important or more basic than the way in which all samples of liquids-having-the-phenomenal-properties-of-water belong to the same type.

I consider such a presupposition to be fundamentally misguided. The chemical constitution of a sample of liquid is never decisive and may be marginal to its being appropriately classified as water, except in a very limited, technical sense. For instance, standard samples of the liquid naturally referred to as water always have H_2O content lower than 100%, sometimes significantly. One such example is the sea: consider for instance "[t]he sea is water" offered by De Beaugrande and Dressler (1981: 145) as an example of a trivially true, informationally void sentence. At the same time, substances higher in H_2O content that e.g. the sea might not be water. Bottled H_2O containing only trace amounts of artificial sweeteners and dyes is a *drink*, not *water*. A 0,0000003‰ solution of botulinum toxin would be *poison*, and most definitely *not water*,

despite being almost absolutely pure H_2O (although the causal history might be relevant here as well, cf. Chomsky's [2000: 127–128, 148–149] 'tea' example[88]).

Note, too, that pure H_2O is actually *not* water. It is *pure water* or *distilled water* (the latter may be considered lexemic), and at least some of its functional and thus inferential properties are dramatically different from those of water. As opposed to water, distilled water should not be drunk in excessive quantities (possible adverse health effects due to mineral depletion), as opposed to water, distilled water does not conduct electricity, etc. Yet another distinction concerns the state of aggregation: steam or ice is H_2O, but according to our most straightforward, everyday linguistic intuitions, it is not *water*. In fact, it has been shown experimentally (a study by Barbara Malt, quoted in Ahn et al. 2001: 64–65) that as a matter of everyday language use, there is only somewhat loose correlation between a substance being called 'water' and it's H_2O content, and by no means a strict correspondence.

Some (but not all) of the above reservations could be countered by the use of hedges, e.g. stating that "steam is *really*/*actually* water". Such indeed is Putnam's strategy when he notes that "water" is a natural kind term having a continuum of senses that are nevertheless grouped around a 'true', core sense (1997: 239–40). On closer inspection, however, this solution turns out to be the question, since it stipulates that a 'true' or 'core' is established by a natural science; it does not reflect prototypical language use but instead requires accepting precisely the presuppositions about the privileged status of scientifically-based explanations reviewed above. Chomsky (2000: 148) complains of the specialised, technical character of the terms "*extension, reference, true of, denote*" as applied by Putnam, but it appears that his

[88] "Suppose cup_1 is filled from the tap. It is a cup of water, but if a tea bag is dipped into it, that is no longer the case. It is now a cup of tea, something different. Suppose cup_2 is filled from a tap connected to a reservoir in which tea has been dumped (say, as a new kind of purifier). What is in cup_2 is water, not tea, even if a chemist could not distinguish it from the present contents of cup_1". (Chomsky 2000:128).

reservation can be extended to cover Putnam's treatment of the term 'water' itself. Putnam's analysis cannot be made to work unless one has already presupposed the validity of the scientistic vision of modern chemistry as 'carving nature at its joints', and agreed to restrict one's use of 'water' to a compatible technical sense (however it is worth adding that on closer inspection even this 'technical sense' turns out to be deeply problematic[89]).

For the reasons identified above, chemistry and science in general is better considered as revealing properties that, while also meaning-constitutive, are not of any special status and should be treated on a par with those contained in the word's or concept's stereotype. Such a reformulation allows one to follow a different set of intuitions and simply classify H_2O and XYZ together as kinds of water (or kinds of the *functional kind water*, to use Fodor's [1994: 31] term) – much like brackish water and clear water are kinds of water[90].

The reanalysis along the above lines is also more productive, as it makes it possible to deal with a vast array of cases inexplicable on Putnam's account. The first type of relevant examples are the words or concepts that, despite functioning exactly like 'water', have extensions that are chemically heterogeneous (for

---

[89] But even in this specialised context water and H_2O turn out to be macro- and microstructurally heterogeneous. For example, Weisberg (2006: 343) in the conclusion of his paper writes that "…there is no single kind for water that is useful in all chemical contexts. In particular, we have seen that the set of substances with molecular formula H_2O is often not a very useful chemical kind. It fails to make distinctions among substances that both chemists and ordinary language users would want to make".

[90] Note that such a possibility is admitted by Putnam: „if H_2O and XYZ had both been plentiful on Earth, … it would have been correct to say that there were *two kinds of water*. And instead of saying that 'the stuff on Twin Earth turned out not to really be water', we would have to say 'it turned out to be *the XYZ kind of water*" (1997: 241).

This, in turn, highlights another problem with Putnam's argument, namely the totally arbitrary restricting the 'referential world' of a person to the *planet* they inhabit, rather than, say, the whole universe (embracing both Earth and Twin Earth), or a single continent or country (excluding parts of the planet Earth).

instance, Chomsky [2000: 149] mentions ‘fire’, ‘earth’ and ‘air’). Another type are ‘empty’ words/concepts, such as those of fictional entities, that seem to lack any traditionally conceived reference (Segal [2000: 34–36] mentions ‘ether’, ‘phlogiston’, ‘ghost’).

2.3. Summary and conclusion

The second chapter of this thesis was devoted specifically to defending the main theoretical decision – the choice of the mentalistic perspective on the topic of concepts – and dealing with the consequences of such a step. I introduced the division into I-semantics and E-semantics as two equally legitimate but largely incommensurable research perspectives. I traced back the motivation for E-semantics to the influential work of Gottlob Frege, whose views I also briefly sketched out.

Subsequently, I turned to the debate between the philosophical positions of externalism and internalism regarding mental content. Due to intense terminological confusion it was necessary to precede the proper discussion of this topic by clarifying the notions in question.

Finally, I undertook to defend the validity of the internalistic position. This was achieved by restoring the authority of the pre-theoretic intuitions – in support of internalism – that was undermined in the wake of Putnam’s influential arguments for externalism based on the Twin Earth thought experiment. I adduced several lines of evidence as well as isolated observations that exposed the construction of Putnam’s argument as deeply deficient and fallacious.

Among the unexplored issues are those related to the precise extent of the tension between the two perspectives on concepts – that is, the precise distribution of questions that can be formulated and successfully answered within one perspective, but not the other. Especially illustrative in this context is “The Problem of Ignorance and Error” as described by Laurence and Margolis (1999: 21–23, 34–35, 47–48): a requirement that a theory of concepts should be able to explain the fact that it is possible to possess a given concept despite being

ignorant of its fundamental properties or even having erroneous beliefs about them. For example, the belief that smallpox is caused by "evil spirits or divine retribution" (1999: 21) does not rule out the very possession of the concept SMALLPOX itself. Laurence and Margolis – having already stated their mentalistic assumptions – list the above problem as a major difficulty for most theories they consider. However, "The Problem of Ignorance and Error" seems to arise directly from the *normative* aspect of concepts, i.e. is what should count as *real* smallpox, as 'fixed' by external reality independently of people's beliefs. Normativity, in turn, constitutes an aspect of 'concepthood' that does not appear to be able to be reconciled with a mentalistic perspective. Therefore, it remains to be established whether problems such as that of ignorance and error should count as genuine issues for a mentalistic theory to grapple with, or rather, should be assigned as E-semantic problems that from a mentalistic standpoint qualify as simply irrelevant. At a very minimum, some division of labour between the two perspectives is unavoidable.

PART II

THE THEORETICAL FOUNDATIONS OF THE STUDY OF CONCEPTS

Introduction and notation

In the second part of this work, I consider questions related to the object of study. In the third chapter, I introduce and concisely characterise the topic of *concepts* and the phenomenon of *categorisation* in general terms before proceeding to a discussion, mostly synchronic but also diachronic in scope. In this way, I relate the present work to historical as well as contemporary research on the above theoretical objects, placing my work in the necessary context against which it can be assessed. Chapter 3 is concluded by a short discussion of another notion vital from the perspective of the whole text, namely that of mental representation. In Chapter 4, the subject is narrowed down to the viewpoint of Cognitive Science.

An important proviso is in order at this point. Arguably, a comprehensive discussion of concepts and categorisation is impossible to achieve in any finite text, much less in this modest work. This theoretical problem, when not accompanied by adequate qualifications, is one of troublesome generality: it is simply too immense to be approached academically. In particular, the issue of concepts, being fundamental to a number of linguistic and philosophical traditions, is highly heterogeneous and contains numerous veins that are only nominally relevant to the present work. For this reason, it is necessary to give priority to in-depth, rather than in-breadth, concerns, in order to narrow this topic down to manageable proportions.

One major consequence for the layout of the present work is the decision to place more extensive historical considerations in Chapter 5. This results directly from the focus of this work. In Chapter 5, where the cognitivist context for concepts and categorisation has already been established, a historical overview can be linked to, and seen as continuous with, recent theorising supported with empirical data.

Consequently, a more detailed scrutiny of concepts in the specific understanding of this work belongs in the fourth chapter of the dissertation. Another issue that deserves mentioning at this point is that of the facts of

language use. Although I set off by invoking the *Oxford English Dictionary* definitions of the relevant words, the facts of language use – in the sense of E-language – are not the focus of the present work. The semantics (*qua* E-semantics, see e.g. Jackendoff 1989: 74) of the words *concept* and *categorisation*, as well as of their counterparts in languages other than English, constitutes an altogether different theoretical problem; consequently, it will be excluded from this thesis.

At the same time, it cannot be questioned that natural language has an important role in the development of technical scientific vocabulary, namely, by providing powerful checks on the validity of the chosen terms. Generally, if there is a tension between the technical use of a given word and its use in the vernacular, it should be resolved in favour of the vernacular: when the intended technical meaning violates certain basic intuitions from everyday use of the word, it is a very strong indication that the term should be changed (rather than our intuitions reformed). Hence, I will be satisfied to conclude that my understanding of the key terms, developed further in the text, does not stand in a radical disagreement with the applications of the respective words by language users – but will not go into this issue in any depth.

The risk of confusion resulting from the discrepancies between the everyday use and the 'dedicated' technical use – defined precisely in the Chapter 4 – is aggravated by the existence of two levels of description: meta-level versus object level. The above concerns dictate the need for staking out the distinctions with care. For a large part, this goal is achieved through the notation: regular spelling for denoting objects in the extralinguistic reality (dog), upper case letters for concepts (DOG), single quotation marks for linguistic units, such as words corresponding to concepts and objects ('dog') (but note that they are also used as distancing quotes); double quotation marks are used for actual citations from other sources or for complete formulas, and *italics* are used for emphasis (but also for words of foreign origin). The word 'concept' itself normally appears in this work as a technical term with the meaning regulated by the definition

presented in the fourth chapter. Wherever the word ‘concept’ is meant in the popular sense, roughly equivalent to ‘conception’, ‘notion’ or ‘idea’, this should normally be clear from the context. Usually this distinction is also helped by the notation: compare ‘the concept of ‘x’’, e.g. ‘the concept of ‘identity’’ (‘concept’ used in the popular sense, to the effect ‘conception of ‘identity’’) and ‘the concept X’, e.g. ‘the concept DOG’ (‘concept’ used in the dedicated, technical sense).

3. Concepts, categorisation, mental representation. Preliminary definitions and discussion. Historical background.

Introduction and caveats

The character of the third chapter of the present work is largely introductory. In this chapter, the main objective is to provide an initial discussion of the key terms, that is of *concept* and *category/categorisation*, as well as of several other ancillary but important terms, notably mental representation. The considerations that figure in this part of the text are of both diachronic and synchronic nature and are designed to sketch a preliminary conceptual geography of the subject.

This introductory overview does not (as it cannot) make any pretence to comprehensiveness. It is a constructionally necessary element, offering the Reader a theoretical background against which the more specific concerns of the rest of this work can be defined. However, it should be borne in mind that neither historical nor contemporary general characterisation itself lies among the primary research objectives of this work; it is meant as a necessary starting point for the development of subsequent chapters. As remarked above, because of the size and heterogeneity of the subject, it was necessary to profile the discussion in order to retain focus and the integrity of the work.

3.1. Concepts

3.1.1. Preliminary definitions

The relevant entry in the Second Edition of the *Oxford English Dictionary* defines the English word ‘concept’ in the following way:

> **2. a.** *Logic* and *Philos*. The product of the faculty of conception; an idea of a class of objects, a general notion or idea.

b. Hence in weakened use, a general notion or idea…[91]

Whereas such a rendering strikes one as rather unspecific, it can be argued that the generality of the definition reflects the wideness of application of this word, both in common use and in the supposedly more disciplined academic discourse.

The earliest recorded use of the word *concept* in the English language dates back to the year 1556; 1663 for the sense **2.a.** quoted above. 'Concept' is of Latin origin, having been formed from Latin 'conceptum' (neuter of 'conceptus'), from past participle 'concipere', 'to conceive'. This form can be further traced back to 'capere', to take, and ultimately, to the Indo-European root 'kap', to grasp. This illustrates an interesting point about the pattern of historical meaning transfer, by which vocabulary originally describing the tactile modality extends in meaning to cover other sensory modalities, and eventually mental actions or states (first described by Joseph Williams 1976; for a discussion of implications for the formation of conceptual metaphors see Aleksander Szwedek 2000, 2002).

3.1.2. Historical note

The English word 'concept', as remarked above, was first recorded in writing in the sixteenth century, which, in the context of over twenty five centuries of the Western intellectual tradition, can almost be considered a recent coinage. It is already evident that theoretical reflection on topics related to concepts considerably predates the entering of this word into the English language. As is

[91] The Second Edition of the OED (*Oxford English Dictionary on CD-ROM, version 3.0*) represents the word *concept* with one verbal (with an obsolete meaning equivalent 'to conceive') and one nominal entry, the latter comprising three senses, further divided into the total of eight sub-senses. Sense **1**, obsolete and equivalent to *conceit*, has four sub-senses ("thought, idea"; "disposition, frame of mind"; "imagination, fancy"; "opinion"); sense **2**, in addition to **2.a.** and **2.b.** quoted above, lists the use in attributions and combinations as a distinct sub-sense **2.c.**; sense **3**, labelled as nonce use, is "an original draft or rough copy".

the case with the better part of major themes in occidental thought, the roots of this reflection can be traced to classical antiquity.

In the light of the above, it is particularly difficult to pick out any specific areas in the history of philosophy that would correspond directly to the modern understanding of concepts: most topics in the history of philosophy can be successfully argued to coincide with concepts in one way or another. Still, *some* selections are inevitable. Generally, contemporary research on concepts in the mentalistic sense can be said to be continuous with historical investigation of *ideas*[92]. In the non-mentalistic sense, it roughly succeeds to the study of word meanings[93]. I accept the above as useful simplifications and guiding assumptions for the overview below.

3.1.3. Discussion

In addition to the diverse uses of the word *concept* in popular idiom, this term has developed a wide range of more technical uses in linguistics, psychology, philosophy, and neighbouring disciplines.

Probably the most comprehensive definitions of the term 'concept' have been developed in the logico-philosophical tradition. Marciszewski (1970: 213) determines that concept in the logical sense is the meaning of a common name (which he defines separately as a name that can function as a predicate in an atomic subject-predicate sentence). Marciszewski further explains that concept in the logical sense should be distinguished from concept in the psychological sense, which is a mental experience consisting in representing something in a non-intuitive (non-sensory) way, i.e. so that sensory images do not belong to the

---

[92] Cf. *The Oxford Companion to Philosophy*: "The term ['concept' – SW] is the modern replacement for the older term *idea, stripped of the latter's imagist associations, and thought of as more intimately bound up with language." (Rundle 1995: 146)

[93] Cf., e.g., Wojtak 1998, who reviews the problem of the relation between concepts and lexical meanings from a linguistic point of view, coming to a conclusion that the two, while not identical, are nevertheless very closely related.

content of this representation, even though they may accompany it (1970: 213). Antoni Podsiad (2000: 633–635) distinguishes concepts in the psychological-epistemological sense (as distinct from concepts in logic and concepts in methodology). A concept is a "purely mental presentation of something with its immanent content, a word of thought (*verbum mentis*), or a cognitive-intellectual form (*species intelligibilis*; a picture). Such a presentation can refer to something general, or to something specific but in a general way, and is a result of the process of abstraction"[94]. Podsiad (2000: 634) further distinguishes formal concept (an act of nonintuitive or nonimagistic presentation) from objective concept (purely intentional correlate of the act of presentation). A linguistic correlate of a concept is a name, and both concepts and names have their proprietary contents and extensions (2000: 634).

A crucial distinction related to concepts is that between concepts and images. This is stressed most forcefully by Leon Zawadowski in his critique of early associationist psychologism in linguistics (1966: 234–238), where he emphasises the qualitative difference between conceptual and imagistic content. Similarly, Zdzisław Wąsik (1987: 113) underscores the contrast by juxtaposing the definitions of concept – "the meaning (connotation) of a name, a mental counterpart of a set of features typical of objects to which the name refers (its designates)"[95] – and image – "psychological process consisting in bringing to consciousness the pictures of objects and situations not presently impinging on the sensory organs of a man, based on past perceptions and fantasy"[96].

At this point, it is useful to establish several definitions and ancillary distinctions that pervade the academically informed discussions of concepts[97].

---

[94] Podsiad (2000: 633–634), transl. from Polish – SW.

[95] Wąsik (1987: 113), transl. from Polish – SW.

[96] Wąsik (1987: 113), transl. from Polish – SW.

[97] The Reader is directed for further, more extensive reference especially to:

The terms *denotation*, *designation*, and *reference* are frequently treated as synonymous or nearly synonymous; the distinctions between them do not figure prominently in the context of this work, but it is prudent to have them all spelled out. Podsiad (2000: 172) defines *designation* in logic as a semantic function that, by means of its sense, indicates an object of which it can be truthfully predicated; *designation* in linguistics, as opposed to *signification*, is the referring of the consciousness of a language user to specified classes of objects by means of the conventional signs of this language. A *designate* of a name is an object designated by this name (in this sense of that name) (Podsiad 2000: 173). Marciszewski (1970: 51) explains that "[a] designate of name *N* in language *L*, given a certain sense of that name, is every object of which this name can be predicated preserving truth".

Reference is the most general term of the three, being also common in everyday use in addition to its more technical applications. According to Marciszewski (1970: 196) *reference* is any (case of) referring of a sign to the reality that this sign is about. Reference is dually defined by Podsiad (2000: 581), as 1) (in logic) that which is meant by a sign, and 2) (in linguistics) one of the two basic functions of language, consisting in indicating that which a sentence is about, thus making it possible to identify and reidentify the object.

Denotation is frequently meant in the general sense of *reference* above. Marciszewski (1970: 47) defines denotation as 1) the extension of a name, i.e. the set of all designates of a name (including past, present, future, and possible designates), 2) the class of all presently existing designates, 3) object of any type referred to by a categorially appropriate expression (not necessarily a class).

Ajdukiewicz, Kazimierz 1949. *Zagadnienia i kierunki filozofii. Teoria poznania, metafizyka* [Issues and directions in philosophy. Epistemology, metaphysics]. Warszawa: Wydawnictwo Czytelnik.

Ajdukiewicz, Kazimierz 1960. *Język i poznanie. T. 1, Wybór pism z lat 1920–1939* [Language and cognition. Selection of papers from the 1920–1939 period]. Warszawa: Państwowe Wydawnictwo Naukowe.

Denotation is specified by Podsiad (2000: 169–169) as the semantic relation between expressions of a language and the objects described in this language, which consists in this expression referring to a specified object; it may also mean a *denotatum*, i.e. a denoted object. Denotation is contrasted with *connotation*.

Connotation in logic is the characteristic content of a name, consisting of its necessary and sufficient properties such that any person knowing those properties is able to correctly decide whether any given object is a designate of that name (Marciszewski 1970:111). According to Podsiad (2000: 454), *connotation* is the set of properties characteristic of the designates of a name, i.e. those by means of which one can assign those designates to that name's *extension*. Connotation in the logical sense was introduced by the British philosopher John Stuart Mill[98], who considered it to be synonymous with meaning (see Podsiad 2000: 968). Still, it is necessary to distinguish that notion from connotation in the psycholinguistic sense, which Gary Leech terms affective meaning: "[t]he connotations of a language expression are semantic effects that arise from encyclopedic knowledge about its denotation (or referent) and also from experiences, beliefs, and prejudices about the contexts in which the expression is typically used." (quoted in Keith Allan, 2006: 41)

A distinction similar in nature to connotation and denotation is that into *extension* and *content*. Podsiad considers *extension* to be a synonym of denotation (2000: 214, 953), and defines it as a distributive set of all designates of a name, as well as of the concept expressed by that name (2000: 953). Marciszewski (1970: 360) explains that "the extension of a name *N* (with a sense *S*, in a language *L*) is the class of such objects of which one can predicate truthfully in *L* by means of the name *N* in the sense *S*"; and also treats extension as synonymous with denotation. A term contrastive to extension is *connotation* (in the logical sense), or sometimes *content*, when it is defined as a set of

[98] John Stuart Mill (1806–1873), a British philosopher, politician and economist. Notwithstanding his contributions to logic, he was more widely known for his liberal socio-economic philosophy and the utilitarian doctrine in ethics.

properties common to all of the designates of the name (and of its corresponding concept), and only to them (Marciszewski 1970: 333, Podsiad 2000: 901). *Intension* or *intensional specification* (i.e. through enumerating a name/concept's criterial properties) can also be used in this contrastive sense, as opposed to extension or extensional specification (i.e. through enumerating all of the name/concept's designates), but it is only rarely done so.

Below, there follows a list of properties that are typically and relatively uncontroversially ascribed to concepts. In view of the breadth and diversity of literature related to this topic, the character of this reconstruction has been motivated principally by the concerns of representativeness. Consequently, I have decided to base this reconstruction primarily on a number of recent English reference works[99], as well as established Polish reference works[100], with brief complementary information based on other representative sources. As a result, I have arrived at an introductory characterisation of concepts that constitutes a minimal, and as far as possible, theory-neutral background for the further discussion of this topic; I have reserved a more in-depth treatment for the forthcoming sections of this work, where nearly all of the points raised here will be developed and revised. The Reader should also bear in mind that the items

---

[99] Brown, Keith (ed.) 2006. *Encyclopedia of Language and Linguistics, Second Edition*. Oxford: Elsevier.

Craig, Edward, Luciano Floridi (eds.) 1998. *Routledge Encyclopedia of Philosophy, CD-ROM Edition*. London – New York: Routledge.

Wilson, Robert Anton, Frank C. Keil (eds.) 1999. *The MIT Encyclopedia of the Cognitive Sciences*. Cambridge, MA: MIT Press.

Zalta, Edward N. (ed.) 2006. *The Stanford Encyclopedia of Philosophy* (Spring 2006 Edition).

[100] Marciszewski, Witold (ed.) 1970. Mała encyklopedia logiki [A concise encyclopaedia of logic]. Wrocław – Warszawa – Kraków: Zakład Narodowy Imienia Ossolińskich.

Podsiad, Antoni 2000. Słownik terminów i pojęć filozoficznych [A dictionary of philosophical terms and notions]. Warszawa: Instytut Wydawniczy Pax.

enumerated below have been isolated as distinct points only analytically, while in fact being closely interrelated and mutually dependent.

3.1.3.1. Concepts have semantic character.

Concepts are *semantic units*, that is, they are the units of *meaning*. Two things are implied by this statement. Firstly, concepts are *units*, that is, they are discrete elements, countable entities distinct from one another, items, quanta. What follows, the totality of a conceptual repertoire is traditionally construed not as an analogue continuum but as list of separate, distinguishable items. This property would seem implicit in their description and thus normally taken for granted.

Secondly, concepts are *semantic entities* (or "*semantically evaluable entities*" – Mandik and Eliasmith 2006). Particular concepts correspond to particular and countable *contents*, or more colloquially, to particular and countable *meanings*[101]. This trait is better understood in a juxtaposition with individual nodes in a connectionist network, which generate meanings collectively, but are not individually semantically evaluable, that is they do not correspond to meanings in a one-to-one fashion. Unlike them, concepts are not 'subatomic' constituents, but rather each concept can be assigned its own meaning (at least in principle if not in practice). Such an understanding of concepts is common to mainstream linguistics, philosophy and cognitive psychology and as such will be adopted in this work.

Alternatively, concepts can be thought of as 'abilities' or 'behavioural dispositions' possessed/displayed by particular organisms. This latter understanding is largely incompatible with the former and will only marginally be referred to further in the text (this problem is discussed in Laurence and Margolis 1999: 6; see also 4.2.3.)[102].

---

[101] Alternatively, we can say that concepts *have* meanings or that they themselves *are* meanings (of linguistic expressions) (Rey 1998; Hampton 1999: 176).

[102] The dispositional accounts of meaning are beginning to be revived in embodied and enactive approaches. However, these approaches are themselves located on the peripheries of broadly

3.1.3.2. Concepts are elements of relation of signification.

In linguistic semiotics, meaning is often explained by means of the relation of signification[103]; *concepts are elements of the relation of signification*. Concepts are intermediaries between (conventional) signs and their significata, i.e. signs signify things by means of mediating concepts (Lyons 1996 [1977]: 96–97).

Typically it is assumed that signs signify broadly understood objects in the real world, including physical objects, but also abstract objects, events, relations, etc. However, especially in the context of this work it should be borne in mind that the cognitivist conceptions of language reject ontological realism (cf.

---

understood Cognitive Science (and in opposition to narrowly construed CS), precisely because of their questioning the representational assumptions that are central to mainstream Cognitive Science.

[103] The relation of signification is a paradigmatic means of explaining meaning in linguistics and semiotics. Most contemporary discussions follow Ogden and Richards (1969 [1923]) in visualising this relation as a triangle. Although the original terms used by Ogden and Richards were 'symbol', 'thought'/'reference' and 'referent' (fig. 4), most commentators (e.g. Schulte 1997: 46 – fig. 5) employ other terms, including 'concept'.

Note, however, that there are numerous conceptions of 'meaning' in linguistics and semiotics – for example, Zdzisław Wąsik (2006: 32–33) distinguishes as many as fourteen distinct conceptions of 'meaning'.

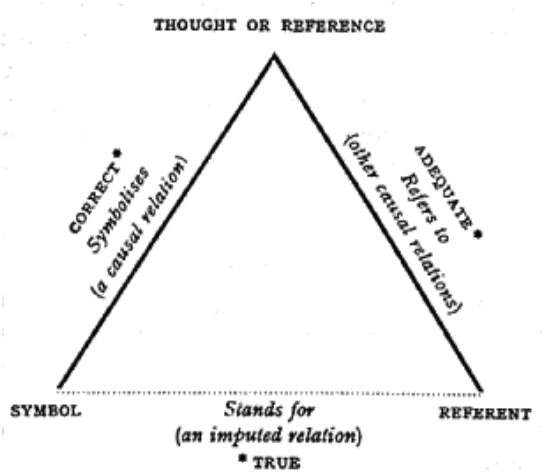

Fig. 4

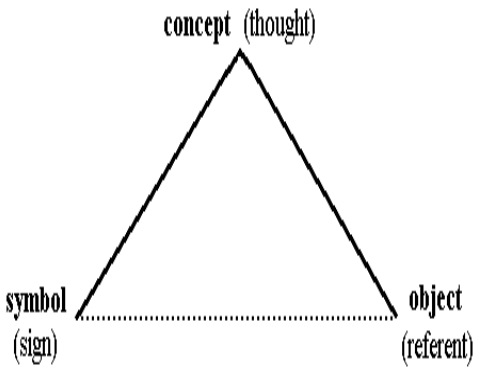

Fig. 5

Muszyński 2006: 48–51); thus in such conceptions signs (*qua* linguistic signs) must refer not to 'real' objects but instead to objects as construed by the cognitive agents (see e.g. Jackendoff 1996).

3.1.3.3. Concepts correspond to units of language.

The format of concepts is *language-like*, i.e. it is – to some degree – isomorphic with natural language[104]. Some of the aspects and consequences of this statement will be mentioned later in this section. One particular aspect that is important in the light of the relation of signification is that *concepts correspond to words* (or more precisely, lexical items) in natural languages (Rey 1998). Spelling out the precise fashion of this correspondence, however, is fraught with difficulties that will be discussed in the next chapter.

Two issues need to be shortly mentioned at this point. Firstly, the default understanding of concepts is as simple lexical concepts; that is, the linguistic units to which they correspond are single content words/lexemes (cf. James Hampton 1999: 176), which perhaps can be extended to embrace lexicalised phrases or 'entries in the mental lexicon' in general[105]. But there is at least one distinct variety of concepts that does not correspond to content words, i.e. logico-mathematical concepts that correspond to certain function words.

Secondly, most scholars (e.g. Fodor 1998, Rey 1998) distinguish between simple and complex concepts, the latter being formed by the combination of the former, thus corresponding not to words but rather to clauses and phrases. Unless explicitly stated otherwise, in this work 'concept' is meant to refer to simple lexical concepts.

3.1.3.4. Concepts are mental entities.

---

[104] One of the many ways of describing a concept is as a "word of thought (*verbum mentis*)" (Podsiad 2000: 633); see also Rundle 1995: 146.

[105] The term „mental lexicon" (analogically, „entries" in the mental lexicon) is understood here in the sense of Jean Aitchison (1996 [1987]).

Concepts are typically thought of as *mental* entities, i.e. such that exist in the minds of individual humans. However, there are important traditions in both linguistics and philosophy that treat concepts as external to and abstract from any individual human being. This recurring question, absolutely fundamental from the ontological point of view, has been discussed in detail in Chapter 2.

It may be noted that even in theories that construe concepts as abstract and non-mental, concepts must nonetheless be somehow made epistemically available to individual cognitive agents, i.e. they must at least be *grasped*, *possessed*, *accessed*, etc. by individual minds.

3.1.3.5. Concepts are building blocks of propositions.

Concepts are the elementary *building blocks of propositions* (Rey 1998; Millar 1994: 674), or the contents of complete, well-formed declarative sentences that assert or deny[106]. Therefore, concepts can be thought of as the *main constituents of knowledge*. In approaches to knowledge in which all knowledge is supposed to be propositional (c.f. the traditional, if not universally accepted, philosophical definition of knowledge as 'justified true belief'), all knowledge is also conceptual, i.e. concepts are the only kind of its constituents.

3.1.3.6. Concepts are stable and redeployable.

Concepts are *stable*, that is, they retain their identity through time. They inherit this property from their being mental representations, which have stability as a criterial feature (Nęcka et al. 2006: 26–27; Eysenck and Keane 2002: 284).

[106] As already noted, Marciszewski (1970: 261) defines 'proposition' [Polish: „sąd w sensie logicznym"] as "the meaning of a declarative sentence… that which is common to a certain class of psychological experiences".

Eysenck and Keane (2002: 245) explain that "propositional representations are considered to be explicit, discrete, abstract entities that represent the ideational content of the mind. They represent conceptual objects and relations in a form that is not specific to any language… or to any modality…".

Stability and redeployability come as relatively natural on non-mentalistic accounts of concepts, especially if concepts are considered to be eternal and immutable, e.g. in the Platonic/Fregean[107] approach (nevertheless, to the extent that concepts are expressed by the words of language, there remains the question of historical semantic change). This is also natural on those mentalistic accounts that envisage concepts as structureless (e.g. Fodor 1998 and elsewhere; but it must be made clear that to Fodor, the identity of a concept is independent of its epistemic content, which is explained in Chapter 6). Still, on the majority of developmental accounts (e.g. Carey 1999), concepts must be capable of undergoing far-reaching modifications to their constituent structure (e.g. during ontogeny) without changing their identity, i.e. without being transformed into a numerically different concept.

By virtue of being stable and static, concepts can serve as basic units in many computational models of mind (or at least higher cognitive processes, such as language processing), still dominant in today's Cognitive Science (see e.g. Pinker 1995 [1994]: 55–77). On the other hand, concepts – like other kinds of static mental representations – are seen to be unfit for the task of faithfully modelling cognition in dynamic approaches to the mind (see e.g. Port and Van Gelder 1995).

As a direct consequence of their stability, concepts are *redeployable* (Mandik and Eliasmith 2006, Rey 1998), so that the same concept can occur in different thought episodes, contributing the same 'meaning'. For example, when one considers inference, the same concept can reappear in distinct inference processes – both in the instances of the same type process and in the instances of type-distinct processes – each time with a predictable outcome[108].

---

[107] See the introductory discussion in sections 2.1.1. and 2.1.2.

[108] A special case is that of a concept being redeployed within the same thought episode; e.g. the concept STAR, underlying the word 'star', in the sentence „The little star's beside the big star" (Jackendoff 2002: 61–63); this illustrates additional constraints on the models of the neural realisation of concept activation.

3.1.3.7. Concepts are shareable.

Concepts are not idiosyncratic – a concept possessed by an individual is not a unique entity 'endemic' to this particular individual. Rather, concepts are *shareable*, i.e. a large number of individuals, ideally a whole linguistic community, must be capable of sharing *the same concept*, in a way variously described as 'grasping it', 'possessing it', 'connecting to it', 'participating in it', etc. This characteristic of concepts plays a fundamental role in explaining the ability of humans to understand others, both in linguistic communication and in folk psychology (see 3.1.3.12.).

The above property, taken at face value, leads to an immediate clash with some of the other features ascribed to concepts, principally their being mental representations. A viable way of reconciling shareability with a mental character is *via* postulating the distinction into concept tokens and concept types. This strategy is developed in 4.2.6.1.

3.1.3.8. Concepts are non-imagistic.

An extremely important aspect of concepts is that they are abstractions over individual acts of experience, a property that makes them *distinct from percepts* in that their content is non-sensory, non-imagistic (Podsiad 2000: 633–634, see also the discussion in section 3.1.3.). Perception should not be identified with sensation (or, to be more exact, sentition, or sensory stimulation) since the latter is basic and continuous signal reception, whereas the former is categorised and discrete, with individual percepts available to consciousness as units in introspection; the process of perception involves a considerable degree of active, top-down re-construction which draws on the agent's goals and background knowledge[109]. Perception, however, is still directly dependent on the immediate

[109] For an overview of the 'active' character of perception, manifest in perceptual constancies, gestaltive phenomena, etc., see e.g. Nęcka et al. (2006: 295–296). In general, this aspect of

sensory data. Percepts are concrete and can be assigned individual spatiotemporal indexes: changing the time, place or person to which a percept occurs results in a different percept. The content of percepts is sensory. In contrast, the content of concepts is abstract; e.g. the concept DOG does not depend on any specific mental image of a dog, nor any other particular (bit of) sensory experience.

Of course, perceptual content is not by definition excluded from participating in conceptual content in various ways. Indeed, such participation may be common, as indicated by the way in which the human brain works, e.g. the fact that parsing sentences with sensory contents causes activations in the relevant parts of the sensorimotor cortex[110]. Still, it must be emphasised that imagistic content cannot be *on its own* constitutive of concepts[111].

3.1.3.9. Concepts are general.

Concepts are *general* in that they do not correspond to particular events, instances, or individuals. Traditionally, they have been usually taken to correspond to *classes of things*, rather than particular, specific things (Chlewiński 1999: 12); *bona fide* concepts are of types, not individuals. Thus, the concept DOG will correspond to a class of dogs – either all actually existing (present, past and future) dogs or all existing as well as possible dogs – while the representation of an individual Fido would not have the status of a *true* concept. Alternatively, individuals can be considered to form classes with exactly one member, which would legitimise a more inclusive construal of concepts to

---

perception is emphasised by the so called „New Look" theorists – see a critical discussion in Fodor 1983: 66–90)

[110] For instance, Vittorio Gallese and George Lakoff report activations in hand areas on the motor and sensory homonculi in the cerebral cortex in parsing sentences containing the word 'to grasp' (Gallese and Lakoff 2005).

[111] See e.g. Arthur Markman (1999: 190–191), and especially Leon Zawadowski (1966: 236–238).

incorporate singular representations as well (Prinz 2006). This, however, is a slightly nonstandard understanding of the term in question.

3.1.3.10. Concepts enable categorisation.

From the abovementioned relation between concepts and classes of things it follows that concepts play a role in *categorisation*. As it stands, this statement is rather vague; exactly how it should be interpreted closely depends on the use of the terms 'categorisation' and 'category' in a given text. Concepts can be said to enable humans to categorise, they can be said to be categories themselves, or to be the intensional[112] specifications of categories (detailed discussion in 4.2.4.). The mutually defining relation between concepts and categorisation in Cognitive Science is best illustrated by the fact that in this field they are almost invariably discussed together, an issue that will also recur in the later parts of this dissertation.

3.1.3.11. Concepts support inferences.

Since concepts are the units over which the process of inferencing is defined, they make it possible for the cognitive agents to draw *inferences* (e.g. Hampton 1999: 177, Haman 2002: 18). One can speak about concepts supporting inferences in two distinct ways.

Firstly, concepts enable drawing inferences in a trivial, content-independent way, merely in virtue of being the units on which the process of inferencing runs. For example, the classic syllogism 'if x is a y, and y is a z, then x is a z' utilises the corresponding concepts X, Y, Z, even though its correctness does not hang on what those concepts are; it is thus supported by these concepts in a way fully independent of their respective contents.

More interestingly, concepts – construed as 'capsules of knowledge' – support inferences in a content-dependent way that brings into play the details of

[112] Cf. the definitions of intension, extension, connotation and denotation discussed in 3.1.3.

the semantic structure of the concepts in question. For instance, if the object x is a knife, then the knowledge of the semantics of the corresponding concept KNIFE ('knives can be used for cutting') makes it possible to draw the inference that 'x can be used for cutting'.

3.1.3.12. Concepts are units of folk-psychological explanations.

One specific area of interest where both inferences and propositional attitudes are productively employed is supposed to be the so-called 'theory of mind'[113], that is the ability to perceive others as self-governed, intelligent agents and make sense of and predict their behaviours. In other words, concepts *underlie folk-psychological explanations* (see e.g. Cain 2002; discussed in more detail in 4.2.3.2.).

3.1.3.13. Concepts combine in productive and systematic ways and underlie compositionality of though and language.

Conceptual structures – in the 'marked' sense of 'structures built from concepts' (rather than the default sense 'internal structures of concepts') – have logical properties: concepts are invoked to explain the logical properties of thought. These are mirrored by, and visible in, linguistic structures, which are assumed to inherit them from thought. The logical properties in question are *productivity* and *systematicity*[114].

Productivity[115] permits both the production and comprehension of complex expressions without previous exposure to at least one token of a given

[113] 'Theory of mind' is a misleading, but extremely well established name for this cognitive ability (at least in ethology, evolution of language, and comparative and developmental psychology). See also footnote 159 in section 4.2.5.1.

[114] Brifely explained in, e.g., Fodor 2001.

[115] Productivity is meant here in a general rather than specialised sense of productivity in word formation (see Plag 2006).

complete complex expression. Because of productivity, it is enough to know the simple expressions and the rules of their combination.

Systematicity is characteristic of a system of signs to the degree that the elements in the system combine to form complex expressions in ways that are regular and predictable. For example, when knowing the meaning of 'a y' (e.g. 'angry man') and 'b z' (e.g. 'nice girl') one is normally able to predict the meanings of 'b y' ('nice man') and 'a z' ('angry girl').

Those two properties constitute very strong arguments for the *compositionality* in human languages, supposedly derived from the underlying compositionality of human conceptual thought (cf. Fodor 2001). It must be noted that especially in natural languages, as opposed to artificially constructed languages, compositionality admits a great variety of counterexamples (e.g. Pinker and Jackendoff 2005). Still, the received view (with the possible exception of the proponents of extreme pragmatics-oriented accounts) is that no theory of language and conceptual thought can be made to work without somehow accounting for their compositionality.

3.1.3.14. Concepts are normative.

A property that many scholars, especially philosophers, deem constitutive of being a concept is *normativity* (e.g. Peacocke 1995, Chapter 5). The requirement that concepts be normative means that each concept must carry with it the criteria for its correct application. The very notion of 'correctness', in turn, presupposes a possibility for error, i.e. for misapplying the concept, as well as a viable way of judging correctness. Consequently, normativity implies the social dimension of concepthood: it is argued that the standards for correct application cannot be purely subject-internal (e.g. following the well-known Wittgensteinian 'Private Language Argument'[116]), but must instead be externally and intersubjectively constituted.

[116] Perhaps best illustrated by the 'beetle in the box' example and the following paragraphs (Wittgenstein 1953: 100–102).

From the above review of the properties ascribed to concepts by diverse theoretical perspectives it should already be clear that many of them are implicitly or even explicitly conflicting. As it has been repeatedly stated, arriving at a single, unified, fully theory-neutral technical construal of the term 'concept' is probably impossible.

3.2. Categories, categorisation

3.2.1. Preliminary definition

In general terms, categorisation[117] is the act or process of classifying, of placing something in a category or ascribing something to some group of similar entities, according to a non-random underlying principle. In a broader sense, categorisation may refer to the ability or function of performing this act or executing this process, or to the area of study concerned with the systems of categories, their formation, and application.

The Oxford English Dictionary (Second Edition) lists the word 'categorization' under the entry 'categorize', which is defined as follows: "To place in a category or categories; to classify… Hence **categorization**, the action of categorizing; classification." 'Categorisation' is of course derived from the verb 'to categorise', itself a derivate from 'category'; this last word is so defined in OED:

> **1.** *Logic* and *Metaph*. A term (meaning literally 'predication' or 'assertion') given to certain general classes of terms, things, or notions... **c.** Hence in more general use... **2. a.** A predicament; a class to which a certain predication or assertion applies. **b.** A class, or division, in any general scheme of classification...

[117] Throughout the work, I adhere to the traditional British spelling ('categorisation' with an 's') rather than to the American alternative, also recently gaining foothold in Britain: 'categorization' with a 'z'. This seemingly trivial difference has one rather crucial consequence, namely that both spellings should be remembered about during database searches.

The dictionary also offers a short description of the use of 'category' by Aristotle (sub-sense 1. a.) and Kant (1. b.). The etymology is given as: adaptation of Latin *categoria*, adopted from Greek κατηγορία: accusation, assertion, predication; abstract noun from κατήγορος: accuser.

3.2.2. Categories

The term 'category' in its theoretical sense has an ancient origin, having been first introduced by Aristotle. *Categories*, as reflected by its later classification in *Organon*, is a text on logic, and consequently, 'category' (also called 'predicament') is thought of as a primarily logical notion (but see Rijk 1988). It referred to the most abstract kinds; anything that could be either the subject of an assertion or its predicate, when considered on the highest level of generality, could be assigned to one of the categories. In more modern terminology, 'category' can be explained by reference to the notion of inclusiveness – categories can be understood as classes with maximum possible inclusiveness. The method of linguistic analysis – extracting information about the structure of the world through the analysis of language used to describe it – led The Philosopher to distinguish ten categories[118]:

a) substance
b) quantity
c) quality

[118] "Expressions which are in no way composite signify substance, quantity, quality, relation, place, time, position, state, action, or affection. To sketch my meaning roughly, examples of substance are 'man' or 'the horse', of quantity, such terms as 'two cubits long' or 'three cubits long', of quality, such attributes as 'white', 'grammatical'. 'Double', 'half', 'greater', fall under the category of relation; 'in a the market place', 'in the Lyceum', under that of place; 'yesterday', 'last year', under that of time. 'Lying', 'sitting', are terms indicating position, 'shod', 'armed', state; 'to lance', 'to cauterize', action; 'to be lanced', 'to be cauterized', affection." (Aristotle [*Categories* 1b25])

d) relation
e) place
f) time
g) position
h) state
i) action
j) affection

The logical context notwithstanding, the application of the term 'category' was metaphysical: categories provided the fundamental taxonomy not so much of logical description, as of reality itself. In other words, the world was supposed to have an intrinsic categorial structure, the discovery of which was an essentially metaphysical endeavour. This position is known as 'categorial realism'.

Since misunderstanding is rife, an important note is in order, namely, that categorial realism must not be confused with realism on Universals. In *Categories*, Aristotle directly opposes Plato's extreme realism on Universals by stating explicitly that only primary substances (individuals, e.g. a particular man like Socrates) truly exist, while both secondary substances (species, e.g. *Homo sapiens*) and other categories are metaphysically dependent; if primary substances did not exist, neither secondary substances nor other categories could exist.

Aristotle's *Categories* was an inspiration for a number of discussions, developments and revisions by ancient and scholastic authors. However, the only other equally influential system of categories was that expounded in Immanuel Kant's *Critique of Pure Reason* (Kant 2003 [1781]) in the late eighteenth century.

Kant's account, although explicitly drawing on Aristotle, was also radically different. In Kant – much like in Aristotle – the categories were not mere instruments of description, but Kant differed from the Stagirite in that he did not see categories as classes inherent in reality itself. Instead, he proposed

that the system of categories was intrinsic to the cognitive agent, being part of its cognitive equipment[119]. Thus, categories were not present in experience or somehow derived, abstracted or constructed from experience by the agent. Rather, their status was transcendental. They belonged to the cognitive faculty of the understanding (*Verstand*), being one of the *a priori*, formal conditions for cognising any possible objects of experience. It was only through the application of a conception of the understanding that the manifold given in intuition (*Anschauung*) could be synthesised into a coherent object of thought, and thus experienced[120]. Categories were the 'pure' concepts of the understanding, being the most general and devoid of any sensory content.

Kant (2003 [1781]: 68–69) singled out twelve categories which could still be grouped in four superordinate classes:

a) of Quantity
– Unity
– Plurality
– Totality
b) of Quality
– Reality
– Negation
– Limitation
c) of Relation
– Of Substance and Accident (*substantia et accidens*)
– Of Causality and Dependence (cause and effect)
– Of Community (reciprocity between the agent and patient)

[119] Cf. also the discussion in Tatarkiewicz (2003: Vol. 2. 129–132).

[120] More accurately, categories and other conceptions of the understanding were requisite for whatever cognition could be present in all '*finite*' (human-like) minds. This limitation did not hold in relation to the hypothetical 'infinite' (god-like) mind, which could rely purely on intuition (e.g. Rolewski 2002: 83).

d) of Modality

– Possibility – Impossibility

– Existence – Non-existence

– Necessity – Contingence

The Kantian system of categories retained Aristotelian spirit in so far as it remained a system of fundamental distinctions, even though the distinctions no longer pertained to the subject-external world, but rather to the necessary (but not subject-external) forms of all possible experience. It must be borne in mind that – however elegant – the Kantian classification remains purely stipulative, with no pretence to either psychological reality or experimental validity. Several commentators (e.g. Władysław Tatarkiewicz, 2003: 128) have noted that the orderly arrangement of the categories into four triples[121] itself hints at the motivation behind such a classification being mostly intra-theoretical.

More contemporarily, the notion of (ontological) 'category' has not found extensive application. One influential discussion of the notion of category, in the light of the so-called category mistakes, is that by Gilbert Ryle[122] (1951 [1949]: 22–25). While considering the different types of existence that can (indeed, should) be predicated of material and mental objects, he proposes what effectively amounts to a linguistic method of testing for two concepts belonging to distinct categories. The criterion is the presence of "absurdities" in sentences such as "she came home in a flood of tears and a sedan-chair" (1951: 22) or "…three things are now rising, namely the tide, hopes and the average age of death" (1951: 24)[123]; in the second example, the tide, hopes, and the average age

[121] Kant (2003: 68–69).

[122] Gilbert Ryle (1900–1976), a British philosopher and leadig proponent of philosophical behaviourism.

[123] This effect is described in contemporary linguistic theory as *zeugma*.

of death belong to three distinct ontological categories, and it would be a category mistake to consider them to have the same status in this respect.

In the Cognitive Sciences, two distinct contemporary applications of the notion of ontological category could be found. The one involves the field of cognitive developmental psychology (Soja, Carey, Spelke 1993 [1991]), referring to the repertoire of apparently inborn, fundamental distinctions that the infant draws upon to bootstrap word learning (thus overcoming the difficulty illustrated by the famous Quineian 'Gavagai' example[124]). The other is Jackendoff's (1992 [1987]: 149–160) linguistic analysis leading this scholar to distinguish "elements that serve as primitive <parts of speech> of conceptual structure", which "…include at least [OBJECT], [PLACE], [PATH], [ACTION], [EVENT], [SOUND], [MANNER], [AMOUNT], and [NUMBER], as well as possible others such as [PROPERTY], [SMELL], and [TIME]."

3.2.3. Categorisation

In discussing categorisation and categories one is immediately faced with a question of logical priority. The word 'category' is morphologically simpler and also seems to be conceptually prior ('categorisation' is possible only when one possesses a category to which the categorised entity can be assigned). However, in accordance with the cognitivist perspective presented in Chapters 1 and 2, my deliberate decision is to prioritise categorisation as a cognitive task accomplished by cognitive agents. The consequence of such a standpoint is an understanding of categories as tools for achieving this task, rather than abstract objects of theoretical description.

---

[124] A classic argument to the effect that a word in an unknown language (e.g. „Gavagai!") exclaimed by a native on seeing some object (e.g. a rabbit) cannot be conclusively shown to be a name for that object in that language, since an infinite number of alternative interpretations are always possible, at least in point of logic. Originally by the American logician Willard Van Orman Quine (1908–2000); discussed in the word learning context e.g. by Pinker (1995: 417–419).

In this work, categorisation in its core sense is understood, rather broadly, as *the act or process of assigning some stimulus to a given category*. One terminological remark is necessary. The interpretation of the word 'categorisation' based on popular use admits conflicting intuitions that result in its questionable application to *category formation*. 'Categorisation' is indeed sometimes used in the sense of category formation even in specialised literature (e.g. Nęcka et al. 2006: 101–103, Aarts 2006). Such a conflation may result from speaking about 'categorisation' to denote *a general area of theoretical interest*, rather than a specific psychological process. In the context of an individual cognitive agent, e.g. a human child, this leads to the question of *ontogenetic* priority between categorisation and categories, as it seems to suggest, incorrectly, that (mental representations of) categories are formed by the process of categorisation. While category acquisition/learning/formation is clearly different from categorisation, it is also clearly related. On some accounts category formation might be the reverse of categorisation; for example – as noted by Laurence and Margolis (1999: 11) – if categorisation is assumed to consist in checking for the required criterial features as posited by the classical approach, category learning may "run in reverse", consisting in assembling the features that form the category.

Typically, categorisation is presented in discussions as a high-level phenomenon, that is, a conscious phenomenon whose elements are accessible to introspection and readily verbalisable (e.g. Kalisz et al. 1996: 37–42). The above is true for *conceptual categorisation* as opposed to perceptual categorisation; and is especially prominent under the so-called classical approach, where the entire process is usually open to verbal reconstruction (e.g. 'the physical entity in front of me is a biped, and is featherless, therefore I can classify it as a man'). At the extreme end of the spectrum are legal verdicts or taxonomical decisions, which

are instances of a laborious, time-consuming and fully explicit categorisation process resting on the overt application of formal, codified rules[125].

Linguistics is the area for which categorisation so construed – i.e. as a high-level, consciously accessible process – is particularly important. An oft-quoted passage from William Labov illustrates the centrality of the notion in question (2004 [1973]: 68):

> [i]f linguistics can be said to be any one thing it is the study of categories: that is, the study of how language translates meaning into sound through the categorization of reality into discrete units and sets of units. This categorization is such a fundamental and obvious part of linguistic activity that the properties of categories are normally assumed rather than studied.

In his paper, Labov is concerned with *lexical categorisation* in the sense of staking out the 'boundaries' for word meanings (which Labov understands denotationally), the sense compatible with the subject matter of this work. Nonetheless, Bas Aarts (2006) points out that within linguistics in general, the term 'categorisation' has a range of diverse meanings. John R. Taylor in the introductory section of his book *Linguistic Categorization* (1995 [1989]: 1) highlights one prominent alternative – a common strictly linguistic understanding of 'categorisation' is, on the metatheoretical level, concerned with establishing the repertoire of terms of the metalanguage.

By contrast, categorisation as studied in psychology, for instance *perceptual categorisation* of abstract shapes, need not involve in any essential way either language or language-dependent units such as concepts (a representative example of such a construal of this notion is perceptual

[125] But even there the actual decisions are mediated by informal factors and are much more impressionistic than it might at first appear. Consider for example the notorious disagreements with respect to assigning particular fossil specimens of extinct hominids to biological taxa; as well as establishing the distinctions into taxa themselves (e.g. as reflected in the cladists versus phenetists debate). See also section 5.2.4.

categorisation of abstract shapes in humans and baboons – cf. Delphine Dépy et al. 1997). In an even more fundamental sense, any instance of principled (non-random) reduction of information complexity, or any instance of systematic 'many-to-one' mapping, is an instance of categorisation. Lakoff and Johnson (1999: 18) give the example of the human eye, where the ratio of light receptors on the retina to the neural fibres connecting the retina to the brain is of the order of 1:100. Since many inputs are 'classified' together, this counts as an example of categorisation.

An interesting case in point is the relatively recent discovery in primates of the so-called mirror neurons[126]. In the stereotypical example of a grasping action, the repertoire of possible moves is continuous along more than one dimension – for example with respect to the speed of movement, the trajectory of the arm, the trajectories of the fingers and thumb, and so on. Still, the reaction of a mirror neuron is categorical – the neuron's firing or not is a binary 'zero or one' decision. This is an example of a direct neural implementation of a high-level semantic distinction; note, too, that it does not involve either language or implicit languagelike representations.

Thus, categorisation is an absolutely fundamental, universal and pervasive low-level phenomenon that not only permeates all cognition, but also provides a basis for all behaviour in general[127]. Harnad (2002) in particular stresses its essentially sensorimotor nature. Categorisation is our intrinsic capacity:

---

[126] Mirror neurons are a relatively recent but extremely important discovery in neuroscience. The function of mirror neurons can be interpreted as performing the classification of actions as 'the same' or 'different': they fire both when an animal does a certain action and when it witnesses 'the same' action performed by a conspecific. See e.g. Rizzolatti and Craighero 2004 for a review.

[127] Such a conclusion might initially strike one as much too strong. However, in drawing it I follow numerous authors important to Cognitive Science, as is illustrated by the quotations below:

> ... to sort the blooming, buzzing confusion that reaches our sensorimotor surfaces into the relatively orderly taxonomic kinds marked out by our differential responses to it -- including everything from instrumental responses such as eating, fleeing from, or mating with some kinds of things and not others, to assigning a unique, arbitrary name to some kinds of things and not others... It is easy to forget that our categorisation capacity is indeed a sensorimotor capacity. In the case of instrumental responses... what we tend to forget is that these nonarbitrary but differential responses are actually acts of categorisation too, partitioning inputs into those you do this with and those you do that with.

But categorisation is an overarching concept with a very broad meaning and far-reaching implications for all domains of human life. Viewed from the everyday rather than cognitive-psychological perspective, categorisation is implied by such common operations and phenomena as cataloguing, stereotyping, social classes, etc. It is a relatively recent realisation that categorisation as visible in language is

---

Stevan Harnad (2005: 21): "[t]o put it most simply and generally, categorization is any *systematic differential interaction between an autonomous, adaptive sensorimotor system and its world*." [italics in the original]

George Lakoff (1990 [1987]: 5): "There is nothing more basic than categorization to our thought, perception, action, and speech."

George Lakoff and Mark Johnson (1999: 17) "[e]very living being categorizes. Even the amoeba categorizes the things it encounters into food or nonfood, what it moves toward or moves away from. The amoeba has no choice as to whether to categorize; it just does. The same is true at every level of the animal world. Animals categorize for food, predators, possible mates, members of their own species, and so on."

Eleanor Rosch (1999: 61): "One of the most basic functions of living creatures is to categorize, that is to treat distinguishable objects and events as equivalent."

Ray Jackendoff (1990 [1983]: 77): "More generally, the ability to categorize is indispensable in using previous experience to guide the interpretation of new experience… Thus an account of the organism's ability to categorize transcends linguistic theory. It is central to all of cognitive psychology."

in fact only one aspect of a much more basic phenomenon. This realisation of the continuity between language and the rest of cognition, both in the matter of categorisation and outside it, was what has since underlain fast-developing research in cognitive conceptions of language (see esp. Lakoff 1990 [1987]).

Note that categorisation is not the *only* basic, low-level cognitive process. It should be distinguished from recognition, i.e. classifying[128] two instances as instances of numerically the same individual, without necessarily assigning the individual to a broader class, and is dependent on discrimination (individuation), i.e. being able to detect discrete chunks in the input that could function as units in further processing. In the case of learned (not innate) categories, it is also dependent on memory. On the other hand, memory and discrimination are equally dependent on categorisation, and recognition, too, always seems to involve categorisation of the recognised individual at some level.

Categorisation is what introduces *discreteness* into any cognitive system. Although it is well known that categories allow diverse kinds of 'fuzzy' effects, such as hesitation, inconsistent classification of the same object by different agents[129], or by the same agent at different occasions, etc., at any given time a categorisation decision is a binary decision: something either is a member of the category or is not. A good illustration of the discreteness of category borders is the phenomenon of *categorical perception* (described by Stevan Harnad, e.g. 1987), whereby the inherently continuous spectrum of variation of a stimulus is perceived as forming a set of relatively discrete categories with only very small regions of the spectrum perceived as intermediate between any given two categories. For instance, almost any point on a colour spectrum is likely to be classified by subjects as an exemplar of some particular colour, while very few will tend to be classified as unspecified (Harnad 1987). Similarly, most vowel

[128] 'To categorise' and 'to classify' have the same meaning when 'to classify' is understood as 'to assign [something] to a class' rather than 'to form a class or a system of classes'.

[129] E.g. McCloskey and Glucksberg (1978); see also Chapter 5 in general.

sounds will be perceived as a particular vowel from the subjects' native language, with only very few sounds perceived as genuinely indeterminate.

Categorical perception further enhances the principle of exaggerating within-category similarity and between-category *dis*similarity: the subjects in an experiment typically judge two exemplars of colour shades to be much more similar to each other when they belong to the same colour category than when they belong to distinct categories, *even though* in both cases the pairs of exemplars are 'objectively 'equidistant, in the sense of being equidistant on the colour spectrum (Harnad 1987).

A historical overview of the topic of categorisation constitutes a separate major part of this work, and is presented in Chapter 5.

3.3. Mental representation

Mental representation is another key notion of the work; it will loom large in the remaining chapters of this text.

In most general terms, a representation is something that represents, i.e. *is about things other than itself*; by virtue of this fact, representations have meaning, or *content* (Egan 2006: 553). Mental representation is, *ipso facto*, a kind of representation: accordingly, it is realised by a physical vehicle (*representation bearer*) and has content, that is, represents something distinct from itself, is *about* something: an object, event, or 'state of affairs' in the world. In addition, as a result of being 'mental', mental representation requires a cognitive system to which it is internal[130].

David Pitt (2006) observes that mental representation, despite being a proprietary notion of the philosophy of mind, has had particular importance to Cognitive Science ever since the conception of this (super)discipline. It is worth reiterating that the transition from a behaviouristic to a cognitive viewpoint

[130] Cf. Żegleń (2003: 26): „…representation is information about something, encoded in a system".

resulted from a general conviction that human behaviour must be explained, not exclusively in terms of externally observable stimuli and responses, but rather in terms of underlying mental representations (see Chapter 1). Today, the representational theory of mind (in cognitivist literature often referred to as RTM[131]) provides the unifying framework for mainstream Cognitive Science, as it was explained in Chaper 1.

It has also been argued (most influentially by Hilary Putnam 1981) that representations presuppose interpreters, following the claim that nothing can be a representation 'intrinsically' purely by virtue of its physical properties, such as a structural homomorphism with the represented object, some similarity to it, etc. On this view, the relation of representing requires an external observer – an interpreter – whose presence is constitutive of this relation: 'representation is in the eye of the beholder'. This, however, gives rise to severe difficulties. If the relation of representing requires an interpreter, then mental representations require an internal, personified interpreter located, as it were, 'inside' the mind (the problem of Homunculus[132]). More generally, to the extent that interpretation itself depends on representations, the relation must postulate an endless chain of interpreters, resulting in infinite regress.

The traditional philosophical accounts of representation have resolved the above difficulty by appealing to the notion of *intentionality*[133]: an intrinsic,

[131] E.g. Margolis and Laurence 2006.

[132] Exposed by e.g. Daniel Dennett (1998: 224–225), and also Vilyaneur Ramachandran in his second Reith Lecture, accessible from: http://www.bbc.co.uk/radio4/reith2003/ram/lecture2.ram

[133] The term 'intentionality' was introduced into modern philosophy of mind by the German philosopher Franz Clemens Brentano (1838–1917), being adopted by this thinker from the works of mediaeval scholastics. Intentionality was a central notion in the philosophy of many of his students, notably the German philosopher Edmund Gustav Albrecht Husserl (1859–1938), and the Polish philosopher and logician Kazimierz Jerzy Adolf Twardowski (1866–1938). See a discussion in Żegleń (2003: 151–158).

inherent power of the mind to be directed at something, hence being about it, representing it. However, intentionality is a semantic notion. Following the naturalistic consensus in Cognitive Science and philosophy of mind discussed in Chapter 1, whereby all valid descriptions must remain within the bounds of a naturalistic vocabulary, intentionality cannot enter explanations, but must be naturalised (see theories of content below)[134].

Bearing all of the above points in mind, one may follow the strategy of Urszula Żegleń (2003: 26) who defines representation as "information about something, encoded in a system". Given that 'information', 'encoding' and 'system' are notions applicable in the natural sciences, such a construal allows one to avoid presupposing intentionality, and thus produce a naturalistic notion of representation suitable for Cognitive Science.

As is generally agreed, mental representations are *physical* entities, being realised by physical bearers. This issue and the related issue of the 'reality' of mental representations will be examined in detail in Chapter 4, in the context of discussing concepts as a kind of mental representations.

Mental representations have *contents*, by which they are type-individuated (i.e. representations having the same contents are necessarily representations of the same type). The notion of content is closely related to the notion of meaning in linguistics, and it too is notoriously elusive to definition, for precisely the same reasons.

Contents, or 'meanings', of representations must also be somehow 'grounded' or 'fixed'. At least three major naturalistic theories (in philosophy of mind called *semantics*) have been developed regarding the determination of

---

Despite having identical pronunciation and almost identical spelling, *intentionality* must not be confused with *intensionality*. The relevant distinction is clarified in a short text by William J. Rapaport: http://www.cse.buffalo.edu/~rapaport/intensional.html

[134] More radically, Jackendoff (2002: 20–21) considers the term 'representation' itself to be "intentionality-laden" and postulates its replacement (together with other similar terms, such as 'symbol', or even 'mind') by neutral terminology, such as 'cognitive structures'.

content: informational, teleological, and functional-role (summarised in Botterill and Carruthers 1999: 161–190):

a) *informational* – on the point of view of informational semantics, the content of a mental representation is determined by the causal relations between the mind and the world: a mental symbol A, with the content *A*, is reliably triggered by exemplars of the category 'a' in the world; the coincidence of instances of A and a (e.g. representations of cars and real cars) is causally correlated and supported;
b) *teleological* – according to teleo-semantics, the content of a mental state A (mental state is the relation of an agent towards a complex representation, typically a propositional attitude such as *believing that* x, *desiring that* x, etc., where x is a propositional representation) is determined by the function of the mental state which has been selected in evolution. This can be enhanced by recourse to counterfactuals (the function that would have been selected, e.g. representing cars, if a given organism had had an evolutionary history in a given environment);
c) *functional-role* – functional role semantics proposes that contents of mental representations be determined by their actual or potential causal interactions with sensations, mental states, and behaviour; in other words, content of the representation depends on the function that it has in the system. A mental representation of a car would be this particular representation that would be reliably tokened in the system's mental processes leading to appropriate behaviour in relation to cars.

The above account is to some extent biased towards linguaform mental representations that enter propositional attitudes, in concord with the traditional application of representational terminology in folk psychology. However, theories within Cognitive Science make use of numerous other sorts of mental representations, including images, impressions, schemas, image-schemas, scripts,

frames, or computational rules. Not all kinds of mental representations have to be consciously introspectible – some of them may be subdoxastic and therefore not accessible to introspection.

Finally, it must be underscored that the notion of mental representation as used in Cognitive Science departs from the minimal definition stated above in rather fundamental ways. Mental representations is used very broadly, to cover all kinds of stable postulated cognitive structures, including ones that do not have externalised *repraesentata* (do not correspond to any entities in the world external to the system). This terminological problem is addressed at length in section 4.2.4.

Summary

In the third chapter of my work, I described in general terms and then discussed three theoretical notions central to the present work: 'concept', 'categorisation', and 'mental representation'.

Firstly, I characterised the notion of 'concept' descriptively, by identifying a range of general properties that are uncontroversially predicated of concepts and supplying a short comment on each of these properties. A more detailed consideration of concepts in the particular sense most relevant to the scope of the present work constitutes a larger theoretical task and, as such, was left for Chapter 4.

Secondly, I focused on the notions of 'category' and 'categorisation'. After a short historical discussion, I established a more specific and more contemporary understanding of the latter notion, in line with contemporary Cognitive Science: as 'the act/process of grouping distinguishable stimuli together by treating them as equivalent'.

Finally, I addressed the notion of 'mental representation' as one having a major role in this work. In this section, the more traditional – intentional – understanding of 'mental representation' was highlighted, but one that was transferable from the philosophy of mind to Cognitive Science. In particular, the

main theories of content (i.e. ways of giving *meaning* to mental representations) were shortly reviewed.

4. Concepts in Cognitive Science[135]

In the third chapter, I provided a possibly general, if not fully theory-free, introductory characterisation of concepts and categorisation. But it has already been remarked that it is impossible to investigate either concepts or categories in any depth while remaining neutral on a number of philosophical issues, particularly in philosophy of mind and philosophy of language. The goal of this chapter is to clarify these basic commitments.

In this chapter, I stipulate and briefly discuss the minimal requirements for a theory of concepts in Cognitive Science: understanding concepts as mental (internal) representations that are capable of serving a number of cognitive functions. I also take two important definitional decisions that are not likewise uncontested. I propose that categories are most fruitfully approached when regarded as mental representations and that concepts are a subset of so understood categories: namely, concepts are categories with lexical correlates. These two issues are accordingly offered a more detailed treatment, with the latter claim being given a sound footing in a broad range of empirical work.

4.1. Scope of study

The scope of this work is delimited to *concepts that underlie everyday words*, that is, roughly, to *categorematic* concepts. Categorematic concepts can be defined as each having a meaning of its own and corresponding to words that can function as subjects or predicates of propositions[136].

The focus of this thesis is on 'standard' simple lexicalised semantic concepts, i.e. content concepts that correspond to commonly used open-class

[135] The content of an earlier version of this Chapter served as a basis for (Wacewicz 2010).

[136] 'Categorematic' is defined by Podsiad (2000: 434–435) as „[in traditional logic] a name of terms that fulfil an autonomous semantic function, such as e.g. nouns" [trans. SW].

lexical items, especially monomorphemic lexemes; typically, nouns, verbs, and adjectives.

A further bias of this work is towards the analysis of concrete object concepts expressed by single-morpheme common nouns (e.g. CHAIR or BACHELOR). Such is the case for two reasons. Firstly, concepts for concrete objects are the prototypical kind of concepts and seem to be psychologically and ontogenetically (e.g. Landau et al. 1998) primary to other kinds. More importantly, as a consequence of its theoretical character, this work is largely reliant on existing literature which has itself been focused mostly on concrete object concepts. This is partly the result of methodological difficulties related to the study of other kinds of concepts, as acknowledged by e.g. Rosch (1988a) or Douglas Medin et al. (2000).

Syncategorematic[137] concepts are ones that lack meanings of their own and are correlated with words that cannot serve as subjects or predicates. Concepts that correspond to function words are excluded from the present study. This concerns especially concepts underlying words that express grammatical relations (e.g. articles, pronouns, prepositions) and, most importantly, logico-mathematical concepts (AND, OR, etc.).

The distinction into content (lexical, open class) words and function (closed class) words is a sound and widely accepted one, but because of the character of this work, it can be relied on only insofar as this distinction is also *cognitively real*. Indeed, very strong empirical evidence confirms the reality of this qualitative difference. Function words are processed differently from content words, being accessed as unitised wholes rather than composed from graphemes, giving rise to a range of psycholinguistic effects, including the *missing-letter effect* (first described in a study by Alice F. Healy, referred to by Greenberg and Koriat 1991 – the reader finds it difficult to detect a particular letter when it appears in a function rather than content word). What is more, function and

[137] 'Categorematic' is defined by Podsiad (2000: 851) as „[in traditional logic] a name of terms that fulfil a semantic function only when appearing together with other terms…" [trans. SW].

content words have a very different cerebral status: they involve different patterns of activation (much more localised for function words, e.g. Pulvermüller 2002: 115–119), different types of activation (Brown et al. 1999), different patterns of breakdown in aphasic patients (largely mirroring the agrammatic/anomic distinction – Pulvermüller 2002: 69–73), access to function words betrays a right visual field bias (i.e. left hemisphere bias), etc[138].

Highly schematic words that might or might not be classified as syncategorematic (e.g. VERY, GOOD) receive only peripheral treatment. Similarly, metalinguistic (VERB, SENTENCE) or highly abstract theoretical concepts in general (INCOMMENSURABILITY, PRIMOGENITURE) stand outside the primary focus of this work. *Ad hoc* concepts (as described by Laurence Barsalou, 1983) and goal-derived concepts, as well as larger, decomposable concepts in general (RELATIONS TO NOTIFY IN CASE OF DEATH, COLOURLESS GREEN IDEAS) are likewise excluded in so far as they are clearly compositional and do not correspond to individual lexical items.

'Lexical items' – construed as the entries in the idealised mental lexicon – rather than 'words', are the linguistic unit of choice to correlate with concepts. Firstly, this does justice to the intuition that a set of different inflectional word forms ('bake', 'bakes', 'baked') can rest on one and the same concept. Secondly, this preserves the possibility of simple concepts consisting of more than one free morpheme (e.g. BLACKBOARD, RED HERRING) when they correspond to compounds or longer idiomatic expressions that have a largely noncompositional

[138] Some of the differences can be explained in terms of word frequency, but most effects remain robust even when word frequency is controlled for. Many researchers (e.g. Friederici et al. 2000) suggest that the distinction into function/content words may largely derive from a more fundamental distinction into abstract (highly schematic) and concrete (imagistic) words. This in turn seems to be in line with the position of cognitive grammar (e.g. Langacker 1987). Regardless of whether this is the case, the distinction into function and content words remains very well motivated.

semantic structure and are stored as complete units in the mental lexicon rather than being compiled online.

This division might be surprising at first sight. The status of 'concepthood' may seem to depend on a decision of a lexicographer, which not only looks arbitrary, but also refers to an externalised individuation criterion – rather than a mentalistic, system-internal criterion called for in a cognitivist approach. I hope to have convincingly resolved this difficulty by the end of this chapter.

4.2. Concepts in Cognitive Science. Concepts as lexical categories

4.2.1. Introductory remarks

Deciding on a given perspective inevitably means committing oneself to a set of assumptions, some of which will be contestable. Rather than leaving such assumptions covert or, worse still, being equivocal on them, I would like, firstly, to state them explicitly and, secondly, to shortly justify their selection in order to provide a sound foundation for the rest of my work. Some of the assumptions, such as a mentalistic rather than semantic understanding of concepts, follow directly from the main methodological choice, i.e. the fundamental commitment to the cognitivist perspective. Others result from explanatory economy and convenience, and a few issues are merely identified, but left as open questions.

As stated above, the overarching perspective assumed in this work is that of Cognitive Science. The exposition of the cognitivist perspective and the motivation for it has been offered in Chapter 1, which is devoted to these issues in full.

As a direct result, the understanding of concepts is cognitivist: the central assumption about the mental and individual, rather than mind-external and abstract character of concepts has been discussed in detail and defended in Chapter 2 (esp. 2.2.1., and 2.2.2.). This is not to suggest that inside Cognitive Science this understanding is unitary, clear and unproblematic.

The source of the difficulty is much the same as in the discussion of concepts in general: Cognitive Science is broad and diversified enough to have itself generated a richness of distinct interpretations of the term 'concept'[139]. These problems are further aggravated by the natural practice of many authors[140] not to address explicitly what exactly their use of this term should amount to, but rather to leave it implicit. Also, similar problems arise for categories and categorisation in their mutual relation with concepts. In Cognitive Science, and in cognitive psychology in particular, concepts and categorisation are routinely discussed together (examples are Barsalou 1992, Eysenck and Keane 2002 [2000]; Medin et al. 2001 [1992]; Medin 1998; Hampton 1999; and many others) to the extent of almost forming a unitary theoretical problem; still how exactly concepts are distinct from, dependent on, or basic to categories and categorisation – is hardly ever spelled out in detail (a notable exception is Barsalou 1992: 170–172). Such widespread practice of use of the term 'concept' in a non-technical, intuitive sense illustrates the fact that even within Cognitive Science, a unitary definition that could be universally embraced is impossible to achieve.

4.2.2. What is 'a concept'? Conditions on theories of concepts

Since 'concept' is a theoretical term (thus, a 'postulated entity'), the question of what it is to be a concept is in principle translatable into a set of conditions on a theory of concepts. One clearly articulated set comes from Fodor (1998: 23–34), who lists five "non-negotiable" conditions for a mentalistically oriented theory of concepts:

---

[139] One illustration might be the concluding paragraph of the entry "Concepts" in *The MIT Encyclopedia of the Cognitive Sciences* (Hampton 1999: 178): "The proliferation of different models for concept representation reflects the diversity of research traditions, the many different kinds of concepts we possess, and the different uses we make of them."

[140] Pinker and Prince 1999 [1996] is just one example.

1. concepts are mental particulars,
2. concepts are categories,
3. concepts are compositional,
4. many concepts are learned
5. concepts are public (shareable)[141].

The alternative – or complementary – way of characterising concepts is in terms of functions to whose fulfilment they are typically invoked (e.g. in Solomon et al. 1999; Medin and Smith 1984; Medin et al. 2001; Rey 1998; Prinz 2006; cf. also footnote 138 above). The most important ones, being also the most frequently listed, are *categorisation*, *inference and reasoning*, and *communication*/*word meaning*. Others include *reference determination*, *representation*, *learning*, *understanding*, *explanation*, *planning*, *prediction*; they can be seen as secondary in that they largely follow from the previous set, and also seem to overlap one another.

Another widely embraced constraint on 'concepthood'/concept possession is the *generality constraint*, first formulated as a criterion by the philosopher Gareth Evans (1982: 100–105). The generality constraint stipulates that if a subject is in possession of a concept A, then they should be able to meaningfully combine this concept with all other (semantically relevant) concepts in their repertoire; it is a condition of the possession of the concept A that the subject be able to entertain all (sensible) thoughts that are comprised of the concept A together with any other concepts possessed by this subject. It is assumed that the thoughts in question are limited to thoughts that have truth conditions; in any

[141] A few remarks are in order to prevent possible misunderstandings. Firstly, the 'non-negotiable' criteria are so called because compromising them would amount to a departure from the representational theory of mind, and hence to a departure from what Fodor assumes to be *the* point of view of Cognitive Science. Secondly, by 'concepts' Fodor means both simple and complex concepts, and (4) concerns complex concepts, thus preserving the possibility of most or all simple concepts being innate.

case, the constraint should be qualified to bar inappropriate combinations that would result in thoughts that are not well formed or nonsensical, e.g. because of category mistakes. Evans himself concedes that the generality constraint might be merely "…an ideal to which our actual system of thoughts only approximately conforms" (1982: 105), but it remains an evaluative criterion nonetheless.

It has been suggested, especially in reference works (e.g. Prinz 2006, Hampton 1999), that the requirements on a theory of concepts are usually too extensive (even within the mentalistic perspective). That is, it might be unreasonable to expect one and the same type of cognitive structure to support all of the above functions, and consequently, that it would be more fruitful to study some of the functions in separation, with the application of different theoretical models. Undoubtedly, particular theories of concepts necessarily involve certain trade-offs, i.e. theories that excel at explaining one aspect will struggle with another (e.g. prototype models might be good at explaining rapid categorisation but poor at explaining reflective rule-based inference, and *vice versa* for classical models). However, it appears to be the most theoretically interesting to assess particular accounts of concepts on the basis of their overall performance 'across the board'.

Below, I develop the criteria sketched above, addressing the most important aspects as well as ensuing problems. I rely on Fodor's enumeration as the main guideline, treating other listed issues as subsidiary.

4.2.3. Concepts are mental representations.

The practitioners of Cognitive Science are agreed on viewing concepts as mental representations (c.f. Fodor 1998; Laurence and Margolis 1999; Margolis and Laurence 2006; Solomon et al. 1999). Note that this is a nonempirical statement, prejudged by the assumed theoretical outlook. As a general statement, it leaves much room for qualifications and particularisation, but in its core does not appear to be contestable. When it is questioned, it is either done from a position with

different large-scale theoretical commitments and therefore incompatible with Cognitive Science, or the reservations are merely terminological.

Thus, for example, speaking of the "representations of concepts" from the standpoint of Cognitive Science must be seen as either indicative of the semantic perspective or actually referring to the specific implementation of conceptual structure, usually on the physical (cerebral) level, i.e. to the representation of concepts *in the brain* (this is done by e.g. Van Loocke 1999, or Shanks 2000 [1996]: 302–305). Otherwise, this would embody terminological confusion. As understood in Cognitive Science, concepts themselves *are* mental representations.

Understanding concepts as 'abilities' in any form betrays behaviourist leanings and thus also seems to go against the grain of Cognitive Science. This is certainly the case if this position leads to the elimination of the mental. Alternatively, if the mental is retained, the 'concepts as abilities' view need not conflict with the view of Cognitive Science, and given an appropriately broad definition of mental representations (see 1.3.2., 3.3.), could in principle be accommodated within it: abilities would then depend on the underlying mental representations[142].

Finally, criticisms might be targeted at the notion of mental representation itself. For example, Ray Jackendoff is unsympathetic towards 'mental representations' as well as other similar terms, finding such vocabulary to be hard to reconcile with a naturalistic outlook, and to be suggestive of a

[142] Such is the opinion of Laurence and Margolis (1999: 6, footnote 3): "Yet another alternative is the view that concepts are not particulars at all, but are, instead, behavioral or psychological abilities. We take it that behavioral abilities are ruled out for the same reasons that argue against behaviorism in general (see, e.g., Chomsky 1959). However, the view that concepts are psychological abilities is harder to evaluate. The chief difficulty is that more needs to be said about the nature of these abilities. Without a developed theory, it's not even clear that an appeal to abilities is in conflict with the view that concepts are particulars. For example, such abilities might require that one be in possession of a mental particular that is deployed in a characteristic way."

misleadingly realist semantics[143] (Jackendoff 2002: 20–21); he proposes to speak of 'cognitive structures' instead. However, this problem, although not trivial, again seems to be nominal rather than substantial: for example, Henryk Kardela (2006b) has suggested that this difficulty can be addressed by employing the term 'functional representation'.

4.2.3.1. Concepts are physically instantiated.

If concepts, as defined above, exist only internally to a cognitive system, and if cognitive systems must be realised physically, it would follow that concepts must have some kind of physical reality behind them. This question is a version of the single most important problem in contemporary philosophy of mind, namely, the mind-body problem. Following the naturalistic consensus in Cognitive Science (see Chapter 1), Cartesian mind-body substance dualism is universally or nearly universally rejected, and philosophers and cognitive scientists are agreed that the mental is dependent on (or 'grounded in', or 'realised by', or 'implemented in') the physical. In other words, the mind must have a physical substrate on which it is closely dependent. A number of proposals on the nature of this dependency are discussed in literature, ranging from strong reductionism to epiphenomenalism to forms of supervenience (see 1.3.3., 4.2.3.1., 4.2.3.2.); however, that the mind as a whole is physically instantiated is a dictum foundational to Cognitive Science, as well as mainstream philosophy of mind.

Still, the above claim does not logically imply microreductionism, that is a claim that *particular concepts* are physically instantiated. The latter statement, i.e. that individual concepts have their specific physical correlates within a cognitive system, is a different and much stronger one. 'Concept' does not belong to the physical/physiological vocabulary, but is patently a mental/psychological term. Accordingly, concepts must be individuated

[143] Gallese and Lakoff (2005) seem to be sceptical of the very term 'concept' for essentially the same reasons.

primarily on the psychological level[144], and finding a parallel mechanism of concept individuation on the physical/physiological level, although clearly desirable, is not strictly necessary. Thus, the two above statements: ('*the cognitive system is a physically realised system*' versus '*individual concepts have their specific, distinguishable physical correlates*') should not be confused, because they have a different status: the one is a theoretical dogma, the other an empirical statement that cannot be prejudged in advance. What is important is that even though Cognitive Science is obligatorily committed to the former claim, it does not require being likewise committed to the latter claim and may remain neutral in this respect.

However, the existence of a relation linking concepts to some distinct physiological units does appear plausible[145]. Such a speculation is based on a wealth of experimental observations, including data from aphasiology (selective damage to particular conceptual domains or selective disruption of semantic hierarchy, e.g. Eysenck and Keane 2002: 302–303) and animal studies (e.g. Wessberg et al.'s successful isolation of the neural correlates of representations of particular motor actions in owl monkeys [cited in Prinz and Clark 2004: 66])[146].

Finally, the very notion of 'independent individuation on the physical level' requires a comment. The most intuitively appealing possibility: that (the activations of) concepts correspond to (the firings of) individual neurons (the so-

---

[144] Cf. the distinction into three levels of description reviewed in 1.3.4.

[145] For instance, the latest version of George Lakoff's theory explicitly proposes to understand concepts in terms of neuronal structures: "*The job done by what have been called "concepts" can be accomplished by schemas characterised by parameters and their values*. Such a schema, from a neural perspective, consists of a network of functional clusters." (Gallese and Lakoff 2005: 467; italics in the original).

[146] Admittedly, the examples are very far from conclusive: conceptual domains are not individual concepts, and findings related to motor representations need not generalise to conceptual representations. What they indicate, however, is that the search for the neuronal correlates of concepts is a viable enterprise.

called 'grandmother neurons'[147]) is in fact the least viable one. Concepts as described on the psychological level do not have to correlate with static, *topological* properties of the brain, such as individual neurons, synapses, or individual neural networks. Instead, as pointed out by Jean-Pierre Changeux (1997 [1983]: 137–138)[148] 'independent individuation' might involve other variables, including *dynamic* properties, such as frequency of impulses, etc.

4.2.3.2. Concepts are real.

Admitting that the independent individuation of concepts on the physical level is an open question might be perceived as detracting from concepts: allowing a possibility that they are in some sense 'not real'. Such a line of thought stems from our everyday intuitions about criteria of existence: the primary understanding of something that exists is as a physical object with a definite shape and measurable properties. While it is customary to readily attribute existence to other kinds of entities, such as states, events, and abstractions, their existence is not usually considered to be similarly sound.

In reply it could be pointed out that the term 'concept', in so far as it is part of a theoretical vocabulary of a scientific discipline, does not need any additional motivation: its criteria of reality would be set up entirely by the philosophy of science. As is commonly agreed, these are most importantly predictive power, explanatory power, and formal 'elegance' (parsimony).

---

[147] An informal expression capturing the simplistic but not totally mistaken idea that there are individual neurons that fire when, and only when, there is a perceptual or conceptual process involving a particular person, e.g. one's grandmother – cf. e.g. Vilyaneur Ramachandran's Reith Lectures: http://www.bbc.co.uk/radio4/reith2003

[148] More generally, a given concept could be traced down to a particular activation state in n-dimensional state space of the activation states of the whole brain or one of its functional units, or with a distinct spatiotemporal pattern of brain activation – as long as the correlates are simple enough to be themselves reliably individuated on the physical level.

Thus, concepts are real precisely to the extent that they are postulated by successful scientific theories[149].

From this point of view, finding a pattern of reliable implementation of concepts in the physical substrate would support concepts simply by additionally extending their predictive and explanatory application in theories. But I think it safe to conclude that even without it, the notion of 'concept' is extremely well founded in (cognitive) linguistics, psychology and philosophy (of mind), to the extent of being practically indispensable for those disciplines. At the very least, I propose to view concepts as *postulated entities of extremely high explanatory and predictive power*, and therefore legitimate objects of study.

One particular existing theory that is strictly dependent on concepts is so called 'folk psychology'. Folk psychology is, roughly, commonsense explanations or predictions on which humans base their everyday interactions with other human beings. This theory employs generalisations of the following form:

> *if X desires p and if X believes that (if X does q, then p), then* – ceteris paribus – *X will do q*

For example:

- if Peter wants to see a lion
- and if Peter believes that if Peter goes to the zoo, then he will see a lion,
- then – all other things being equal – Peter will go to the zoo.

The proponents of folk psychology, most prominently Fodor (1994 and elsewhere; for a summary see Cain 2002), remark that folk psychology is not only successful, but is also *the only* known method for predicting and accounting for human actions.

---

[149] In this respect, the status of concepts would be no different from the status of the notions employed by paradigmatic empirical sciences, such as physics.

Folk-psychological explanations rely on concepts, because they have a propositional format, and propositions are composed of concepts (cf. section 3.1.3.12.). By this token, the success of folk psychology lends support to realism about concepts. It is important to note, however, that this relation is not symmetrical, and the reverse need not necessarily be true. A commitment to the reality of concepts does not by itself entail a similar commitment to the reality of folk psychology. For example, Botterill and Carruthers (1999, Chapter 2) note that one may still reject on other grounds such as doubtful criteria of individuation of attitudes (e.g. 'hoping' from 'expecting', etc.).

4.2.4. Concepts are categories.

In cognitive-scientific literature, there is no clear consensus on the application of the terms 'concepts' and 'category' and on their mutual relation; as a consequence, they tend to be used rather loosely and to a large extent interchangeably (see e.g. Żegleń 2003: 234; Chlewiński 1999: 47–48; Medin et al. 2001: 367; Mervis and Rosch 1981). This last case might result from the authors' intentional decision to treat these terms as synonymous, but in other cases this might be due to their treating the distinction as irrelevant and smoothing over it, or due to a failure to make the distinction in the first place. In short, the distinction between 'concept' and 'category' is often principally nominal, with the corresponding distinction in substance either missing or highly obscure.

There appear to be three main lines along which to differentiate between concepts and categories[150]. The first option is to decide that a concept is the intensional specification of a class of entities, whereas a category is its extension

[150] See also Barsalou (1992: 170–172) for a slightly different list of possibilities: 1. a category is the extension (denotation) of a concept; 2. a category is the representation of a concept's exemplars in memory; 3. a category is the representation of a concept's *kinds* of exemplars.
The last two options are different from, but consistent with, the view that I defend in this chapter.

– the class itself – the set of all individual entities in the 'real world' that 'fall under' that concept. If so, then the relation between them is that of representation: concepts are representations (mental or otherwise) of categories. Note that this proposition makes no commitments either regarding the status of concepts (abstract versus mental) or their structure. Although the classical structure of necessary and sufficient conditions is the most natural one, it is by no means the only option; for example, Hampton (1998) underscores the fact that a prototype representation can in principle be as efficient and precise in determining extension. The status of categorisation as a cognitive process could still be salvaged by the assertion that cognitive agents form *mental representations* of the categories in the external world.

On the radical version, it might be postulated that categories are inherent in the structure of the real world: the world comes as 'preformatted' into categories that exist independently of the cognising subject. This preexisting structure would be the same for all cognitive agents, and the role of those agents would be limited to discovering this structure. Categorisation decisions could be easily divided into correct and incorrect on the sole basis of being in concord with this structure.

I take it that such an extremely realistic view is untenable on a number of philosophical grounds and has long been discredited as a non-contender (e.g. Schulte 1997: 49–51). This view is also impossible to reconcile with the empirical results of linguistic and psychological research on categorisation (e.g. fuzziness and typicality effects; see Taylor 1995). It is also generally discarded in Cognitive Science[151].

[151] For example, Douglas Medin (1998: 94) opines that "[m]ore generally speaking, concepts and categories serve as building blocks for human thought and behavior. Roughly, a concept is an idea that includes all that is characteristically associated with it. A category is a partitioning or class to which some assertion or set of assertions might apply. It is tempting to think of categories as existing in the world and of concepts as corresponding to mental representations of them, but this analysis is misleading. It is misleading because concepts need not have real-world

However, the view of categories as sets of real world entities can be reformulated in a less radical and more commonly advanced version. Categories would still be groups of external (objective) entities, but the principles of assigning them to those groups would be subject-dependent. Categories would then be constructed in the process of interaction of the mind and the world. For example, the world could be seen as naturally divisible into entities objectively characterised by certain *qualities*, but categories – as the ways of grouping those entities – would only arise by the application of certain rules by the cognising agents.

But such a position is still not immune to criticism. Firstly, the problem of subject-independent reality, far from being solved, is simply moved one level down the hierarchy, from categories to qualities. Just as the existence of objective, predefined, subject-independent categories is problematic, similarly problematic is the existence of objective qualities, or attributres, on which categorisation could be based[152].

Secondly, it is unclear whether anything is gained by such a redescription. The wording 'category representation', on the face of it, suggests the strong version of the view, with categories having some existence independent of their representations. Thus, categories would have to be relatively stable, both inter-subjectively (between different subjects), and temporally (for the same subject on different occasions). Experimental data (discussed in 5.2.3.) show that this is not the case, and people not only differ in their categorisation decisions, but also are inconsistent through time.

Thirdly, a direct, unmediated epistemic access to the world in general, and hence to the 'objective' qualities of things in particular, is in principle impossible. As Jackendoff (1990: 78) puts it:

---

counterparts (e.g., unicorns) and because people may impose rather than discover structure in the world. I believe that questions about the nature of categories may be psychological questions as much as metaphysical questions."

[152] Such also seems to be conclusion of Eleanor Rosch (1988a: 319).

> [s]ince there can be no judgment without representation, categorization cannot be treated simply as the organism's comparison of some component of reality "a" to a preexisting category of dogs. Rather, the comparison must be made between the internal representations of *a* and of the category of dogs.

The second option is to place concepts and categories on the same ontological level, but consider them to be somehow differing in scope. For the reasons just sketched, cognitive linguistics follows exactly this strategy, viewing both categories and concepts as mental representations. Jackendoff (1990: 77–106) treats categories as mental structures synonymous with concept types and subsuming particular concepts tokens. For example, a representation of a particular dog would count as concept token, and it would exemplify the concept type (i.e. category) DOG. The problem with this terminology lies with 'concept tokens' which involve a counterintuitive use of 'concept'[153].

Another, slightly different, suggestion is offered by (Tabakowska 2001: 32–33). In the terminology used there again both concepts and categories are mental representations. However, concepts represent both individuals and classes, whereas categories represent classes only. Thus, concepts are more inclusive than categories: a category would be a subtype of concept. This is a well motivated proposal; nevertheless, I will not follow it for two basic reasons. Firstly, there is a strong intuition, consistent with the traditional use of the term 'concept', behind considering concepts to represent *classes*, or *types*, rather than individuals. More importantly, it appears that it is concepts that are a subset of categories, not *vice versa*. There are numerous categories that are clearly not conceptual by any usual standards – notably perceptual categories, which are typically contrasted with conceptual ones (see below).

[153] This difficulty disappears when 'concept token' is substituted by 'mental representation of an individual'. However, this would then prevent the analysis of categorisation in terms of the type-token relation.

The third option – the one chosen in this work – is to do justice to this intuition by 'tying' concepts to language. On this stance, again, both concepts and categories are treated as mental representations, but with the former being a subset of the latter. In other words, all concepts are categories, but not all categories are concepts – only those having a lexical correlate. The infinite number of nonconceptual categories can be best illustrated by the classes of stimuli used in the studies of perceptual (usually visual) categorisation tasks. For example, one may consider the practically infinite number of the *representations of the classes* of abstract shapes (e.g. Sigala et al. 2002 – see Fig. 6; Bruner et al. 1999 [1956]; see Eysenck and Keane 2002: 281), walking styles (Davis 2001), etc. that can be formed by either adult humans, infants, or animals. Such representations of classes of perceptual stimuli lack most of the properties that theories typically require concepts to have: for example, they lack a rich inferential potential and the ability to enter propositions. As a result, they cannot be called 'concepts' without seriously violating our intuitions. Consequently, they constitute examples of categories that are not (true, fully-fledged) concepts.

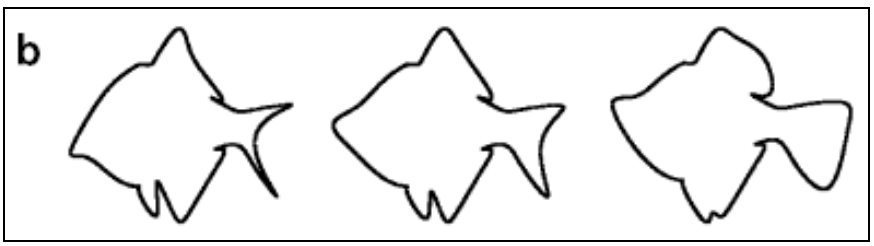

Fig. 6. Sample stimulus used in visual categorisation tasks – Sigala et al. 2002: 188.

Admittedly, this last definitional decision also has its drawbacks. However, it seems to preserve the intuitions related to the understanding of categorisation that are most relevant from the cognitive-scientific perspective.

4.2.5. Concepts have lexical correlates.

The question of the lexical correlates of concepts coincides in various ways with the broad-ranging 'language and thought' dispute: whether, and if so, in what ways and to what extent, language influences thought. This debate looms large in today's Cognitive Science, but shows little promise of being conclusively resolved; indeed, it takes so general forms that it is far from clear that this debate is so resolvable even in principle. I would like to avoid a larger-scale entanglement in this theoretical problem and focus on the specific question of the postulated constitutive relation of language (i.e. lexical items) to concepts. Before doing so, however, it will be useful to take stock of the debate, since the highlights of how language impacts thought will turn out to be relevant in subsequent sections in reinforcing the target argument.

4.2.5.1. Influence of language on general cognition

The inescapable general conclusion is that language enhances and influences cognition in a number of ways, including both the obvious and the more interesting ones. Recently, many of the threads in this discussion have been backed up with strong experimental evidence.

a) (quite trivially,) in modern societies there exist a range of cultural constructs that can only emerge – and be acquired – within social reality that itself relies on a framework laid out by language (e.g. Pinker and Jackendoff 2005: 206)[154]. It follows that certain concepts are in principle unattainable by nonlinguistic organisms. In contrast, language users can

[154] "Vast domains of human understanding, including the supernatural and sacred, the specifics of folk and formal science, human-specific kinship systems (such as the distinction between cross- and parallel cousins), and formal social roles (such as "justice of the peace" and "treasurer"), can be acquired only with the help of language." (e.g. Pinker and Jackendoff 2005: 206)

draw upon the pool of collective experience of their community in ways that allow them to guide their behaviour more successfully[155].

b) at least some (but by no means all) of human problem solving is dependent on the presence of 'inner speech'; it is not only *describable by* words and sentences of one's natural language, but is actually *rendered in* such *stricte* linguistic structures. Whether or not this phenomenon can be implicit, i.e. operating below the threshold of consciousness, is a matter of controversy. Still, there are many familiar examples where problem solvers rely on explicit use of language, such as 'talking to oneself'. One widely-cited experimental example of a situation where being a language user (as opposed to a nonlinguistic creature) predicts performance on a simple spatial cognition task comes from Elisabeth Spelke's laboratory (Hermer-Vasquez et al. 1999)[156].

c) more generally and fundamentally, language helps 'discretise' otherwise continuous cognition: divide it into distinguishable quanta. Abstract entities, relations, events, and processes become objectified[157], thus providing the basic units – corresponding to linguistic units – on which computational processes can operate. These units remain stable in the absence of a co-occurrent perceptual stimulus and are easy for retrieval

[155] For developed arguments, see in particular Daniel Dennett 1994, and Stevan Harnad 2002.

[156] The experimental design required the subjects to integrate two kinds of information: geometric (long versus short wall) and colour (blue or white wall) in order to succeed on 100%of trials; relying on only one kind of information ensured success on 50% of trials. The rates of success of both rats and prelinguistic children were close to the latter figure (suggesting no integration), whereas adults were successful on almost all trials (suggesting integration). However, when the adults were engaged in a concurrent verbal shadowing task that placed heavy demands on their linguistic processing, their performance dropped to levels indicating the lack of integration of kinds of information (Hermer-Vasquez et al. 1999).

[157] 'Objectified' not in the sense of becoming objective, but in the sense of Szwedek (e.g. 2002), that is in terms of the metaphor of concrete, physical objects.

from long term memory, and maintenance, combination and manipulation in short term memory.

This insight, historically reaching back at least to Enlightenment philosophers such as John Locke[158], has been more recently developed by a number of researchers in Cognitive Science, e.g. Jackendoff (1997)[159]. Empirical evaluation is hard to achieve, but relevant experimental data include studies such as that of enculturated chimpanzees faced with the task of grasping the relation "the same": a prerequisite for success seemed to be the ability to assign a (quasi-)lexical label to this relation (Thompson et al. 1997).

d) language enables or at least enhances metacognition ('thought about thought'): it makes it possible to form the (explicit) representations of representations (e.g. Jackendoff 1997: 202–205). This has ramifying consequences, of which the most important is perhaps the link to Theory of Mind: representing the representational states of others[160]. An

---

[158] Locke analyses this aspect of importance of language for mental processes in his *Essay Concerning Human Understanding* (1999 [1690]).

[159] „Language is the only modality of consciousness that makes perceptible the relational (or predicational) form of thought and the abstract elements of thought. Because these elements are present as isolable entities in consciousness, they can serve as the focus of attention, which permits higher-power processing, anchoring, and perhaps most important, retrievable storage of these otherwise nonperceptible elements." (Jackendoff 1997: 205)

[160] Theory of Mind (ToM) is a somewhat unfortunate but extremely well established name for the socio-cognitive ability of humans and certain higher primates to perceive others as self-governed (non just self-propelling) and rational agents capable of entertaining their own, independent mental states, i.e. beliefs, desires, fears, hopes, etc. Developed Theory of Mind implies the ability to *meta-represent*, i.e. to have mental states about the mental states of others (e.g. 'I *know* that John *thinks* that beer is in the fridge').

Although in reality it describes an extremely complex and graded set of cognitive abilities, 'Theory of Mind' serves as a useful, if simplistic, label that can be ascribed in a binary all-or-none way. In such cases, the traditional litmus test for the presence of ToM in a creature has been the creature's performance in false-belief tasks. A success in a false-belief experiment

experimental indication of a strong relation between Theory of Mind and language is the pattern observed in false-belief studies, with language competence predicting performance on false belief tasks rather than the opposite way (Hale and Tager-Flusberg 2003; Lohmann and Tomasello [quoted in Jordan Zlatev, 2007]).

e) developmentally, lexical labels facilitate the formation of categories: consistent use of a given word to describe examples of a certain salient aspect of reality helps to highlight commonalities between such elements and leads to the formation of an appropriate category. Without lexical labels acting as catalysts, category formation is delayed or not achieved at all. (This is different from points a) and c) where the 'referents' in the world are nonsalient and may be in principle unattainable without language).

   For example, Sandra Waxman and Dana Markow (1995) show that in 12-month human children, words serve as inducing stimuli to form categories of simple objects; the observed effect was the strongest with nonbasic-level categories.

   Even more strikingly, a recent study by Gary Lupyan (2006) identifies the same kind of effect in human adults.

f) the scope of at least some categories (the range of perceptual inputs that are categorised together) appears to depend partly on the way in which the corresponding word functions in a given language community. That is, the discontinuities on which categorisation is based depend – in addition to more fundamental perceptual, etc. factors – also on 'Whorfian'[161] factors related to the use of *a given natural language*.

---

requires being able to realise that another individual has an incorrect piece of information, and being able to predict this individual's behaviour that results from its acting on this incorrect information. For a review of ToM in nonhuman primates see e.g. Heyes 1998.

[161] From the American linguist Benjamin Lee Whorf (1897–1941). The so-called 'Whorfian perspective' becomes relevant in the following section, where it is addressed.

> Colour categorisation has traditionally been the area most intensively researched in relation to the Whorfian 'language influences thought' hypothesis. While no significant effects have ever been experimentally established, there is convergent evidence from a large number of recent studies that a person's natural language does in fact exert some influence on their colour categorisation (for instance, Davidoff 2001; Gilbert et al. 2005)[162]. At the same time, one must admit that the effects are rather subtle.

Far from being an exhaustive overview, the above is simply a possible classification oriented towards the implicit influence of individual lexical units. An example of an alternative classification of the ways in which language "augments human computation" is the frequently quoted one developed by Andy Clark (1998):

- *memory augmentation*
- *environment simplification* (language facilitates categorisation),
- *coordination and the reduction of on-line deliberation* (language facilitates explicit planning),

[162] Davidoff reviews, among other experiments, three cross-cultural ones (perceived subjective similarity, category learning, and recognition memory) and concludes (2001: 386): „Put together, these three new cross-cultural studies suggest that categorical perception shows the influence of language on perception. At the very least, our results would indicate that cultural and linguistic training can affect low-level perception... However, more than that, the results uphold the view that the structure of linguistic categories distorts perception by stretching perceptual distances at category boundaries".
Gilbert et al. (2005) found that colour discrimination was affected by the difference in the names of contrasted colours when the colours were presented in the right half of the visual field (thus processed by the 'linguistic' left hemisphere), but not in the left visual field. This effect was diminished when the subjects were assigned a verbal tasks related to verbal processing, but lasted when the task given engaged non-verbal working memory to a similar extent.

- *taming path-dependent learning* (learning from others in addition to from personal experience),
- *attention and resource allocation* (written and spoken words as external memory enhancements),
- *data manipulation and representation* (explicit manipulation of written text to structure one's argumentation).

4.2.5.2. Concepts as dependent on lexical correlates

The target question of the necessary relation between concepts and their lexical correlates, though dependent on the general problem sketched above, is a different and much more specific one. It must be emphasised that this question is not an empirical one that would lend itself to experimental verification or falsification. Rather, the issue is essentially analytical, being a matter of the *coherence of terminology, its concord with intuitions from natural language, and – indirectly – of agreement with larger bodies of experimental results*, by setting up a sound framework for their interpretation.

As remarked above, *I propose that concepts are only those mental representations that are correlated with lexical items* (separate entries in the mental lexicon). Roughly, mental structures that are not readily expressible in single words (or, less typically, short idiomatic constructions) are thus not considered to be concepts – although they still may play significant roles in conceptual structures. Representations that have a nonverbal nature, that is sensorimotor representations such as images, maps, sketches, sensations, motor schemas, proprioceptive schemas, etc. can of course enter conceptual structure in various ways and configurations (or at least are not in principle excluded), but they cannot be concepts, in that without some linguistic component they cannot exhaustively form complete structures of individual concepts.

This definitional decision is directly motivated by the uncontroversial statement that words (or more precisely: lexical items) express the meanings of concepts. However, its implications are much broader: it equals a strong claim

that *the two subsets of mental representations: those with lexical correlates and those without, are in some way qualitatively different*.

Such a view is controversial, and faces a strong opposition in mainstream Cognitive Science. For example, it is common for linguists and psycholinguists (such as Pinker 1995 [1994], Jackendoff 1997) to assume that conceptual structure is prior to language, and though possibly differing in range and richness, does not differ dramatically in kind between linguistic and nonlinguistic creatures. Most influentially, a similar position was argued by the philosopher Jerry Fodor (1975), who insisted that a preexisting structure of (innate) concepts was a logical requirement for the process of language acquisition to take place at all.

The arguments against lexical correlates as being constitutive of concepts proper hinge on two main points. Firstly, nonlinguistic creatures can nevertheless have advanced cognitive systems whose mental operations one is naturally inclined to describe as '*conceptual*'. Secondly, the possibility of successful cross-cultural communication is most readily explicable in terms of people of different cultural/linguistic backgrounds having the same '*concepts*' that only differ in their overt phonological realisations (and perhaps in marginal aspects of meaning).

4.2.5.3. 'Concepts' in nonlinguistic organisms

Increasingly many studies in comparative psychology show that numerous aspects of animal cognition must be given interpretations that rely upon complex mental representations. The most prominent examples – alongside ones such as avian navigation or sophisticated principles of food caching, recaching and retrieval by food-storing birds – come from primate social cognition. For instance, the primatologists Robert Seyfarth and Dorothy Cheney (2001) report that highly gregarious baboons can mentally represent and update both the relations of kinship and hierarchical social status. Due to the problem of combinatorial explosion, in groups of as many as eighty individuals this is

impossible to achieve by means of one-by-one associations, i.e. storing each relation as a separate memory entry; the monkeys must instead employ sophisticated computations defined over 'concepts' such as FAMILY or DOMINANCE.

Similarly, human infants are cognitively complex creatures who from their earliest days display advanced understanding of the surrounding world (in ways roughly consistent with the Piagetian[163] account, but at even younger ages). Consequently, infant cognition is routinely described in conceptual terms: it is customary in developmental psychology to speak of a child's 'concept' of objecthood, number, identity, self, animacy, causation, force, etc. (examples are: "infant's concept of occlusion", Renée Baillargeon 2001; "infant's concept of *twoness*/number", Karen Wynn 1993 [1992]).

It is indeed undeniable that both prelinguistic children and at least some animals do possess advanced mental representations. Still, my reply to the above examples is to deny that these representations have the status of *bona fide* concepts, and to describe them as 'proto-concepts'. The reason behind such a decision is that the presence of a lexical label, especially one frequently used in communication, exerts a profound effect on the functional characteristics of a given mental representation.

Thus, I argue that *bona fide* concepts differ from proto-concepts possessed by animals and infants – as well as from non-lexicalised representations in human adults – in the following profound ways (some of which are direct results of the correlations with lexical labels):

a) concepts become available for explicit inference (in a way described in 4.2.5.1. b),

[163] Jean Piaget (1896–1980), Swiss psychologist. Piaget was author of an extremely influential constructivist account of human cognitive development in which this process was seen in terms of successive domain-general stages.

b) concepts become stabilised, which facilitates their storage, retrieval, and manipulation (in a way described in 4.2.5.1. c),

c) concepts become further stabilised by their frequent activations caused by the use of the corresponding lexical items in communication,

d) concepts are even further stabilised by their frequent pre-activations caused by the use of semantically or phonetically related lexemes (as evident in the phenomenon of semantic and phonetic *priming*; see e.g. Aitchison 1996 [1987]: 109),

e) concepts support compositionality in that they can enter very elaborate complex representations as their building blocks. While it can in principle be imagined that the mental representations in animals and children might have this quality as well, there is no evidence in favour of such a conjecture. On the contrary, what evidence there is points to the opposite. The very first linguistic utterances by children are holophrastic (Bancroft 1995: 64) in a way suggesting limited compositionality of the underlying representations. In contrast, in normal human speech only very rarely do individual words form complete utterances: they almost invariably come as elements of larger compositional structures.

f) most concepts are domain-general in the sense that they can enter reasonings related to any subject. In contrast, the mental representations in infants and animals seem to be domain-specific, i.e. limited to a particular, narrow range of contexts. There is no evidence whatsoever that the complex computational processes invoked to explain e.g. avian navigation or primate social cognition generalise beyond their proprietary domains.

This mirrors the generality constraint (Evans 1982 – see 4.2.2.), which does not seem to be met in the case of infants/animals.

g) concepts are of a palpably different level of generality than some of the mental representations ascribed to animals and prelinguistic children. The latter are more general, lacking the rich inferential content of concepts; on the other hand, they are more closely dependent on sensory imagery. It

> seems that the mental representations of 'causality', 'objecthood', 'occlusion', etc. in children and animals can be more appropriately explained, not in terms of *concepts*, but rather in terms of *image schemas* (for such suggestions concerning infants see Mandler [cited in Jordan Zlatev: 2007], and animals – Marc Hauser 1997 [1996]).

In short, the overall changes to a mental representation resulting from its becoming correlated with a lexical 'tag' have a qualitative dimension and make the ascription of a different status fully legitimate[164].

Also, it is important to spell out an additional condition for 'concepthood', already implicit in the above argument; namely, that the lexical label correlated with a mental representation be part of a richer communicative system. While even a very simple organism can easily internalise an association between an arbitrary sign and a class of inputs, such a mechanical stimulus-response pairing is something very different from a word of human language[165]. Most significantly, as remarked in point e), lexemes are not autonomous, holistic utterances, but units functioning almost exclusively as parts of larger structures; both the existing paradigmatic and the potential syntagmatic relations between lexemes are meaning-constitutive. In addition, the frequency and stability of the communicative use of a lexical label seems to be necessary to support some of its influences, listed above, on the associated cognitive structure (entrenchment effects).

---

[164] Some thinkers, such as Daniel Dennett (1996) and Euan MacPhail (1998), argue for a still much more profound divide between linguistic and language-less creatures. According to them, language is absolutely foundational in the development of the central notion of self. Hence, consciousness as we know it cannot be meaningfully attributed to nonlinguistic creatures. Such a stance meshes with the argument presented above, but does not follow from it as a necessary consequence.

[165] See especially Terrence Deacon (1997: 65–67), who takes paradigmatic and syntagmatic relations between signs to be constitutive of their symbolic character.

In short, although one cannot ignore extant research that documents the richness and sophistication of cognition in nonlinguistic creatures, one can still qualify cognition in such creatures as nonconceptual. Very characteristically, this is precisely what many theorists of animal/infant cognition do. The chief proponent of the notion of 'nonconceptual content', José Luis Bermúdez (e.g. 2003), identifies advanced cognitive operations available to nonlinguistic creatures and to an extent downplays the significance of language for thought. Still, he endorses what he calls "the Priority Principle", that is the linguistic criterion in the divide between concepts and 'nonconcepts'[166].

There remains the final objection of a practical nature: the proposed line of demarcation between concepts and nonconcepts would disallow many of the intuitively correct uses of the term 'concept'. This objection could be valid if the term in question was consistently used in cognitive-scientific literature in a deliberate and disciplined fashion. Since it is not, the objection can be addressed by invoking the loose/strict distinction, characteristic of many technical terms. The practice of using a term in a loose sense does not itself rule out the possibility of using it fruitfully in a strict sense.

Indeed, if one wants to preserve the possibilities of employment of the term 'concept' rigorously in academic discourse, there appears to be no alternative to the distinction suggested above. *The 'lexical' way of distinguishing concepts from nonconceptual mental representations seems to be the only principled way of doing so*, remaining the only way of making this term manageable and suitable for disciplined theoretical use.

4.2.5.4. Concepts are not language-specific.

[166] This also serves to show that 'the lexical criterion' does not suggest the radical and heavily criticised 'linguistic model of thought', whereby all (higher) mental processes are defined over word-like units; quite the opposite, the very presence of nonconceptual (i.e. non-linguaform) content directly points to the existence of the other kinds of cognitive mechanisms.

One of the traditional roles of concepts has been to serve as building blocks of propositions in an essentially language-neutral way (which is to say, independently of any given ethnic language). Also, the possibility and everyday practice of effective translation as well as effective communication across different natural languages appear to be founded on a common repertoire of concepts in the speakers of those different languages. Assuming a more direct, 'Whorfian' link between concepts and words of a natural language would seem to run counter to these ideas, and thus such a decision demands additional support.

I offer what I take to be a successful reply in three steps. Firstly, it is an empirical fact that the particular natural language that a person speaks does impact that person's cognition (see 4.2.5.1. point f)). Even though the magnitude of these effects should not be overestimated, they are nevertheless impossible to dismiss completely.

Secondly, the implicit assumption that provokes the inconsistency is that in order for successful communication/translation to exist, the source and target concepts must be identical. While this assumption does not find any empirical support, there is substantial evidence to the contrary. For example, data from developmental psychology show unequivocally that communication between children and their caregivers takes place in spite of very far-reaching differences in their concepts[167].

Thirdly, and most importantly, the apparent clash between the facts of cross-linguistic communication/translation and the 'linguistic' character of concepts holds only given an extremely strong reading of this linguistic character. Most rebuttals of the 'Whorfian' relation between language and thought, including the influential criticism by Pinker (1995), focus largely on strong linguistic determinism[168]. In fact, unless one assumes the patently

[167] See Carey (1999 [1991]).

[168] The doctrine of linguistic determinism has been traditionally illustrated by the famous quotation from the American linguist Benjamin Lee Whorf (1956: 213): "We dissect nature

indefensible interpretation on which thinking is literally performed in words, there appears to be no direct contradiction between the possibility of communication/translation and linguistically influenced concepts. (See Wacewicz 2008 for a discussion).

Still, the definitional requirement that a mental representation correspond to a lexical item in order to count as a concept does not equal a commitment to any such strong views. As mentioned above, the actual structure of a particular lexical concept can include various other kinds of representations. What forms of representations participate in concepts, and in what ways, is a different empirical question[169]. On such a view, a substantial overlap between the internal structures of concepts (counterparts) in the members of two linguistic communities is sufficient for effective communication/translation to take place. What is more, the assumption of a moderate degree of non-overlap has the virtue of explaining frequent cases of irremovable difficulties in communication/translation. Alternatively, one may construe the differences in the concepts of speakers of two distinct natural languages simply as magnified interpersonal differences between speakers of the same language – a conceptualisation which also naturally leads one to the conclusion defended in this section.

---

along lines laid down by our native languages. The categories and types that we isolate from the world of phenomena we do not find there because they stare every observer in the face; on the contrary, the world is presented in a kaleidoscopic flux of impressions which has to be organized by our minds – and this means largely by the linguistic systems in our minds. We cut nature up, organize it into concepts, and ascribe significances as we do, largely because we are parties to an agreement to organize it in this way – an agreement that holds throughout our speech community and is codified in the patterns of our language."

It might perhaps be noted that, although indisputably effective, the quotation is very hard to interpret in any way that would allow for some more precise definition, as is the doctrine itself.

[169] Indeed, the definitional decision linking concepts to lexical items (or otherwise establishing precisely how they should be understood) seems to be itself a prerequisite for this question to be answerable.

4.2.6. Concepts are shareable/concepts subserve communication.

According to a consensus in Cognitive Science, concepts are *aggregates*: they are, roughly, bodies of 'knowledge'; they are comprised of and, in principle, are decomposable into, smaller or more primitive elements. A direct consequence is that concept possession/mastery is gradable, rather than being a matter of binary all-or-none decision. This, in turn, constitutes grounds for employing the notion of concept overlap/similarity to account for the concepts' role in communication in precisely the manner sketched out in the previous section.

The above consensus is contested by the minority camp of conceptual atomists, notably Jerry Fodor (esp. 1998). Indeed, the shareability of concepts – fundamental to linguistic communication, folk psychology, etc. – is Fodor's chief motivation in taking this controversial stance. Fodor's atomistic position will receive detailed treatment in Chapter 6.

4.2.6.1. Concepts as types and concepts as tokens

One specific worry related to Fodor's influential argument should be addressed at this point. As pointed out by that researcher (e.g. Fodor 1998: 30–39), the notion of 'concept similarity/overlap' is by nature a dependent notion, having a logical prerequisite in the form of 'concept identity'. In other words, it is impossible to judge similarity between two concepts without knowing what it would be like for those concepts to be *identical*. The notion of *the same concept* must be present at some level, if only as a yardstick for judging the similarity between particular individual concepts.

This, however, is clearly at variance with the mentalistic commitment. Since concepts are supposed to be mental representations possessed by individuals, each concept is instantiated in a particular, individual mind, and realised by the 'hardware' of a particular brain. Quite obviously, it is impossible for two distinct individuals to share a *numerically identical* concept (cf. e.g. Aydede 1998).

Laurence and Margolis (1999: 5–8, 75–79) see this difficulty as superficial and readily solvable by the application of type-token distinction. To them, while individual concepts are necessarily different as concept tokens, they might still exemplify the same concept type. Consequently, to deny that two distinct individuals could possess 'the same' concept "…would make as much sense [as] to say that two people cannot utter the same sentence because they cannot both produce the same token sentence. Clearly what matters for being able to utter the same sentence, or entertain the same concept, is being able to have tokens of the same type" (1999: 7).

Whereas I agree that the type/token relation is ultimately the right solution, I suggest that it is far less clear that its functioning is as unproblematic as implied by the analogy offered by Laurence and Margolis. The case of sentences is relatively unproblematic because they are type-individuated in virtue of their physical, nonsemantic characteristics (shapes and composition of their constituents). Concepts, however, are ontologically different. Type-individuation of concepts is something non-obvious; rather, it is precisely what a successful theory of concepts strives to describe.

The problem appears to have its roots in the admission of the public character of concepts, which might not be a viable option for a cognitively oriented researcher. The existence of *bona fide* 'public' concepts suggests the independence of the abstract level of concept types. It suggests that the public, abstract concept types are something 'over and above' the collections of their token instances, and they cannot be reduced to such collections. Consequently, the abstract level is not merely epiphenomenal, but must necessarily come first: either ontologically or at least logically, i.e. in the order of explanation. This in turn would threaten to discredit the autonomy of the I-semantic perspective relative to the E-semantic perspective; it might also prove difficult to square with the naturalistic approach.

I propose to resolve this final difficulty in the following way. What is relatively uncontroversial is the truly 'public' status of the *words of a natural*

language. Words are symbols and as such, despite being semantic in the sense of having meanings, can at the same time be individuated on the basis of their nonsemantic, physical properties (phonemes or graphemes). 'Public' (*sensu* 'shareable') concepts are dependent on public (*sensu* 'truly public') words – dependent not as a kind of artificial, abstractly specified relation, but rather in an actual, causal manner. As has already been established, on the present account concepts emerge developmentally, in the process of infants becoming 'linguistic creatures', that is, becoming proficient users of their native languages. The internalisation of lexical labels exerts real, causal influence on children's mental representations: children's concepts are shaped – among other things – through repeated witnessed instances of meaningful use of corresponding phonological words (in ways partly described in 4.5.2.3.).

Consequently, on the present account, concepts as individual mental representations are ontologically primary, and they do not have an inherent public character. The 'publicity' (*qua* shareability) of concept types is only derived, being dependent on the public status of the lexical labels, *the correlation with which is constitutive of concepthood*. On this account, which can be thought of as moderately neo-nominalist, the abstract concept types are purely epiphenomenal and have no additional properties over and above those directly resulting from concept tokens. A resulting additional virtue is the preservation of a naturalistic framework of explanation.

4.3. Cocnlusion

In this chapter, I considered concepts as seen from the standpoint of Cognitive Science. I made a decision to define 'concepts' as 'mental representations having lexical correlates'; such a definition makes it possible to treat 'concept' as a technical term across the Cognitive Sciences, while also preserving most intuitions from a looser use of this word in the literature. I narrowed down the subject of the present study to content concepts (prototypically, concepts for common words), having found the corresponding distinction into content and

function words well motivated, both theoretically and in terms of its psychological and neuronal reality. I also addressed several relatively less central terminological points.

Later in the chapter, I adduced arguments for the qualitative difference that is made by mental representations' being correlated with lexical items of a natural language (both as a direct result of such a correlation, and indirectly through its functioning within a linguistic cognitive system). In this context, the process of language acquisition in the child is seen as directly, causally responsible for the formation of fully-fledged concepts on the basis of more primitive proto-conceptual mental representations.

In the final sections of this chapter, I addressed the problem of the shareability of concepts, a characteristic being particularly difficult to reconcile with a mentalistic ontology. I proposed that this could be achieved by ascribing shareability to concept types, but only derivatively: resulting entirely from the correlation with truly shareable and public lexical labels that – *via* language acquisition – exert causal influence on the nature of concepts. Thus, concept types should be seen purely as idealisations exploiting an assumed high degree of isomorphism between individual conceptual repertoires of the members of a linguistic community.

I conclude that, on the present analysis, construing concepts as mental representations with lexical correlates may be considered the optimal solution from a cognitivist perspective. It has the two crucial advantages of, firstly, preserving most of the pre-theoretical intuitions connected to the word 'concept', and secondly, solving several theoretical problems, most importantly the problem of shareability.

PART III

CONTEMPORARY APPROACHES TO CATEGORISATION AND CONCEPTUAL STRUCTURE

5. Classical approach to categorisation and conceptual structure

In Chapter 5, I provide the background for discussion of similarity-based (prototype and exemplar) models of categorisation by reviewing and evaluating the most influential traditional approach to the topic of categorisation/concepts, that is the 'classical approach'. Both the review and, especially, the evaluation are carried out from the cognitivist perspective that characterises the entirety of this work.

I start from the problematic issues of terminology and highlight the pitfalls related to this broad subject matter. After (re)establishing the crucial terminological distinctions, I provide the required historical review. Rather than being a detailed historical study, it brings to the surface and then discusses the most important characteristics of the particular approaches.

5.1.1. Theories of categorisation or theories of concepts? Review of terminological problems.

As has already been described in section 4.2.4., any philosophically informed discussion of the topic of concepts and categorisation runs into terminological problems which, although seemingly trivial, turn out to be insurmountable in practice. What some academic works (e.g. Barsalou 1992, Medin et al. 2001) review as the 'theories of categorisation' is essentially identical in scope and structure to what others (e.g. Medin and Smith 1984, Nęcka et al. 2006) discuss under the heading of 'theories of concepts'. Even if some principled distinctions can be specified (such as, for instance, 'categorisation' being only one of many functions of concepts, however prominent), they are rather unsystematic and none seems to be well established in the literature.

The use of the term 'category' (hence, too, 'categorisation') is particularly problematic. Category can be construed extensionally as a class of beings, i.e. as the sum of all actually or potentially existing exemplars of a given set. Alternatively, it can be construed as an intensional specification of such a set. Nevertheless, on closer inspection this would presuppose extreme ontological realism, assuming the existence of an already categorised reality independent of its cognisers. What is more, 'intensional specification' already strongly suggests a specific method of reference determination, i.e. the 'classical' one. Also, such a construal encounters severe problems with dealing with non-concrete categories that do not have objects as their exemplars (note, too, that while the term 'exemplar' is popularly used to designate a particular individual, it can also designate a concept one step down the hierarchy, such as SPARROW being a hyponym and therefore an *exemplar* of BIRD[170]). All the reasons enumerated above serve to show that the traditional nonmentalistic understanding of 'category' as 'class of beings', although dominant, has its difficulties and cannot be treated as the only valid option (a broader discussion of this point has been provided in 4.2.4.). Finally, 'category' has a still different traditional philosophical meaning: one of the most basic and broad general classes of entities.

In this work, I propose to use to use the term 'category' in the broad but mentalistic sense (unless the context makes it very clear that the nonmentalistic and more traditional reading is intended): a category is any mental representation regardless of its level of abstraction. Thus, categories can be conceptual but also perceptual or even sensory or sensorimotor, and can be exemplified even in quite basic cognitive agents. A concept is a mental representation expressible by a lexical item and correlated with it, in the sense that the lexical item in question is present in the mental lexicon of a given agent (see 4.2.5.2.). 'Categorisation' has a dual – most likely irredeemably dual – meaning. Firstly, it is classification: the

[170] This is common in exemplar approaches to categorisation, e.g. Storms et al. 2000.

act/process of assigning some inputs to a category; secondly, it is a general theoretical topic related to the issues of categories, including category boundaries, category formation, conceptual structure, etc. It is this latter, more inclusive, sense that is intended in this work in the expression 'theories of categorisation'.

5.1.2. Mentalism, psychological reality, and theoretical goals.

In this chapter, as was so often the case in the preceding parts of this work, the overall research perspective will play a decisive role. This was already visible in the previous point, which highlighted the matters of terminology. The problem extends even further since the topic of categorisation/concepts transitions seamlessly into such topics as the form and structure of knowledge representation[171], and the encoding, storage and retrieval of items in semantic memory.

Looking from the cognitivist perspective, at the forefront of interests are the topics related to the way in which natural or artificial cognitive agents process information. Hence, cognitive scientists will be relatively less interested in aspects such as reference determination or even *ideal* cognitive economy[172], but at the same time much more interested in such subject areas as drawing inferences or concept acquisition. As nearly always when the perspective of Cognitive Science is involved, *psychological reality* remains a fundamental (probably the single most significant) criterion in evaluating particular models or larger theoretical approaches.

5.2. Classical approach

5.2.1. Exposition

[171] Such is the observation of Frank C. Keil et al. (1998), who concede that no principled, non-arbitrary distinction can be drawn between the study of categorisation and knowledge representation.

[172] The notion of cognitive economy is discussed by Eleanor Rosch (1988a).

The so-called classical approach to the issue of concepts and categorisation is both the intuitively natural and historically dominant one. Notwithstanding its prevalent criticism and apparent retreat, this model of understanding categorisation (speaking broadly) still pervades everyday thinking and continues to exert a very strong influence on Cognitive Science – as well as science in general. I agree with the opinion of George Murphy, who suggests that the classical view may be thought of as a default, pre-theoretic position on categorisation[173].

The key premises of the classical approach can be briefly summarised in the following way. Categories have internal structures that can be adequately captured in the form of conjunctions of features, and most importantly, at least some of those features are essential. Essential features are 'necessary and sufficient', i.e. they are jointly sufficient and individually necessary for something to be a category exemplar. The conjunctions of essential features form category cores that decide about category identity; these cores may – but do not have to – be accompanied by other, non-necessary (accidental) features carrying additional but non-criterial information about the category.

The classical approach is also sometimes referred to as the *traditional theory* (Laurence and Margolis 1999) or the *definitional approach* (Murphy 2002). The former term reflects the historical predominance of this view (see section 5.2.2.: History). The latter term is based on the fact that this perspective on categorisation shares its precepts with the method of intensional definition widely employed in describing the meanings of lexical items. Still other denominations for this general intellectual perspective include the *Aristotelian*

[173] „…[A] reading of the most cited work on concepts written prior to 1970 reveals its assumption of definitions. I should emphasize that these writers did not always explicitly say, <I have a definitional theory of concepts.> Rather, they took such an approach for granted and then went about making proposals for how people learned concepts (i.e., learned these definitions) from experience." (Murphy 2002: 12)

approach (Taylor 1995), the *scholastic approach* (Aarts 2006) and the *categorical view* (Labov 2004 [1973]).

Taylor (1995: 79–80) has summarised the main assumptions of the classical view of categorisation[174]:

a) all members of a category have equal status
b) all non-members of a category have equal status
c) there is a fixed set of necessary and sufficient conditions defining membership to each category
d) all necessary and sufficient features defining a category have equal status
e) category boundaries are fixed

A further point, f), could be added to the list:

f) categories form a hierarchy with transitive category membership

Points a) and b) indicate that no gradability is allowed: category membership status is binary, with exemplars simply classified as 'members' versus 'non-members' rather than being rated on a more complex membership scale. Categorisation decisions are all-or-none, with no possibilities for finer-grain distinctions or evaluations.

Point c) states the method of categorisation – through the set of necessary and sufficient conditions – but further implies that the set is 'neat': it is finite and probably very small, and the features themselves are simple in the sense of being (relatively) easy both to verbalise and to determine in the tested exemplars.

[174] A different summary is provided by William Labov (2004 [1973]: 68), who in his critical address of the classical view lists its "implicit assertions": "all linguistic units are categories that are: 1) discrete, 2) invariant, 3) qualitatively distinct, 4) conjunctively defined, 5) composed of atomic primes".

Point d) indicates the lack of weighting, that is of endowing certain features with greater relative significance than others. This indirectly results from points a) and b). If no gradability is allowed in the first place, it follows that no particular feature can carry more importance than any other, since the failure to meet any of the individual criteria has exactly the same influence on the outcome of the categorisation decision. Any given exemplar can only either meet all of the criteria and thus be classified as a member, or *not* meet all of the criteria and thus be classified as a non-member, with no other outcomes possible.

Point e) reasserts the binary character of the system of categories (member or non-member), which should admit no haziness or uncertainty about category membership. Also, it stresses the system's *universality*. The division of the world into categories is stable both across subjects and across time, so that for any two people, or for two moments in time for the same person, there should in principle be only one correct way to categorise. In such cases, different categorisation decisions would automatically imply the incorrectness of at least some of them.

Point f) expresses the organisational aspect of the system of categories: they form a hierarchy, with the more specific categories being well-defined subsets of the more general ones, progressing recursively down the hierarchy to the most specific ones. Since the subordinate category is characterised by all of the features of its superordinate category, this enables the straightforward inheritance of inferences. This in turn enables the convenient intensional definition by *genus proximum* and *differentia specifica*.

The list quoted above, rather than being an arbitrary stipulation, appears to be a manifestation of an underlying deeper ontological commitment. It seems that the points a) through f) should be complemented with an additional assumption of a more general kind, one that serves as a philosophical foundation for the classical approach:

A) categorisation (*qua* category formation) captures the pre-existing, objective structure of the world

In other words, the objective reality has a ready-made structure and comes pre-categorised even before it is experienced by the cognitive subject. The system of categories is inherent in the reality, it is single, universal, and "carves the nature at its joints". The division of the world into categories is discovered by the cognisers rather than being the product of their interaction with the world. As will become evident, it is this realist ontological position lying at the foundations of the classical approach that has become ultimately untenable and resulted in the criticism of this theoretical outlook.

Recapitulating, the classical approach is founded on the intuitively appealing idea of categories having clear-cut conditions of membership, expressible as (relatively short) lists of (relatively explicit) necessary and sufficient features.

Furthermore, a closer examination of the above points reveals additional more basic philosophical commitments. I propose that the following underlying assumptions are presupposed by the classical view:

– *essentialism* – in the world there are essences – usually *hidden*, i.e. not manifest in the surface properties – common to all members of the class (e.g. genomes are sometimes taken to constitute the hidden essences of biological species);
– *analyticity* – the internal structure of a concept (and its corresponding word) mandates a number of infallible inferences, which are true *a priori* and in all contexts, and their truth results trivially and necessarily from the concept's internal structure (e.g. "a bachelor is a man");
– *determinacy* – a given entity's membership status for any given category is always determinate, and any uncertainty in determining whether it is a member or a nonmember is always owing to the cognitive agent's epistemic limitations;

– *composition* – concepts are built from (relatively few) primitive elements that can be added to form gradually more complex compounds;
– *decomposition* (*reversibility*) – the process of acquisition is theoretically fully reversible, so that the structure of a concept can be analytically traced back to the list of all of its individual, basic component parts (with no interfering emergent qualities);
– *discreteness* – the individual component parts are distinctive, discrete wholes, separable from one another; when combined, they retain their identities rather than e.g. blending; they are also usually taken to be expressible in a verbal format (see 5.2.5);
– *reductionism* – the lists of necessary and sufficient features are nonredundant and relatively short (ruling out all noncriterial information); this translates into maximum reduction of complexity in category information, which in turn maximises the economy of coding that information.

The above theses will be progressively discussed in some more detail in 5.2.2. below as they recur in particular historical conceptions. The historical review, for obvious reasons, will have a selective nature, while still preserving the representativeness and highlighting the essence of the classical view.

5.2.2. History

5.2.2.1. Antiquity

The spirit of the classical theory of categorisation is easily spotted already in the works dating back to antiquity. One classic example are Plato's Socratic dialogues. Many of them were characterised by Socrates' elenctic method of argumentation: Socrates aimed at laying bare the shortcomings of his opponent's provisional description of a phenomenon, leading to a closer examination of the phenomenon in question so that the provisional characterisation could be

superseded with an infallible *universal definition*[175]. Such definitions were intended to be objective – i.e. to describe real phenomena rather than, for instance, people's conceptions or the meanings of words – which highlights Plato's position as an extreme ontological realist. Socratic definitions were expected to admit no counterexamples and to capture the very essence of a given phenomenon, that is its 'true nature'. The defined objects were usually highly abstract moral or philosophical qualities, such as piety (*Euthyphro*), virtue (*Meno*), or knowledge (*Theaetetus*).

Nevertheless, this philosopher is most famously associated with another definition, that of man (the human being) as 'featherless biped', the anecdotal case of which serves to demonstrate that problems with definitions have already started to be evident in Plato's time. Allegedly, after Plato defining man as 'featherless biped', Diogenes of Sinope (Diogenes the Cynic) countered this definition with the example of a plucked chicken. The source of the report is the much later work of the biographer Diogenes Laertius (VI, 40; 1696: 414)[176], and it should perhaps be noted that Plato's actual commitment to the abovementioned definition is unclear. In Plato's extant works it is found in *Statesman* 266e (transl. by Benjamin Jowett; Jowett 1902: 550), where it is expressed by Stranger – the polemist of Young Socrates[177].

---

[175] This method is very close in spirit to conceptual analysis that forms the cornerstone of the contemporary field of analytic philosophy, the main distinction being perhaps that the latter is interested in the analysis of word meanings.

[176] "*Plato* having defined a Man to be an Animal with two Legs, without feathers, and having gain'd great applause thereby, he [Diogenes of Sinope] stript a Cock, and brought him into his School, and said, here is *Plato's* Man for you: which occasioned him to add to his Definition, *With broad Nails*" [italics in the original].

[177] "I say that we should have begun at first by dividing land animals into biped and quadruped; and since the winged herd, and that alone, comes out in the same class with man, should divide bipeds into those which have feathers and those which have not, and when they have been divided, and the art of the management of mankind is brought to light, the time will have come

However, it was not Plato but rather his successor Aristotle who was the more central figure to the advance of this line of thought. Aristotle developed three rather crucial theoretical principles: the law of noncontradiction (LNC), the law of excluded middle (LEM)[178], as well as the distinction of attributes into necessary and accidental (see also Taylor 1995: 22–24, who comments on the influence of those logical principles). The seeds of the two laws are already incipient in Socrates' elenctic method of argumentation, and the seeds of the essence-accident distinction – in Socratic definitions; still full credit is due to Aristotle for their explicit formulation. Quite obviously, their importance to the Western intellectual tradition extends much farther, but here only their significance with respect to categorisation will be reviewed.

The law of noncontradiction, held by Aristotle to be a self-evident and absolutely fundamental axiom of all logical thought, decrees that the same thing cannot be simultaneously asserted and negated. This principle has been stated in several slightly different variants in *Metaphysics*, for example:

> Evidently then such a principle is the most certain of all; which principle this is, let us proceed to say. It is, that the same attribute cannot at the same time belong and not belong to the same subject and in the same respect; we must presuppose, to guard against dialectical objections, any further qualifications which might be added. (*Metaphysics* 4.3).

> But we have now posited that it is impossible for anything at the same time to be and not to be, and by this means have shown that this is the most indisputable of all principles. (*Metaphysics* 4.4).

---

to produce our Statesman and ruler, and set him like a charioteer in his place, and hand over to him the reins of state, for that too is a vocation which belongs to him".

[178] These two „laws" are sometimes called „principles", and abbreviated accordingly to PNC and PEM. The law of non-contradiction is often referred to, somewhat misleadingly, as „the law of *contradiction*".

The law of excluded middle, although very often confused or conflated with the former law, is something distinctly different from it. Aristotle writes:

> But on the other hand there cannot be an intermediate between contradictories, but of one subject we must either affirm or deny any one predicate. (*Metaphysics* 4.7)

> While the doctrine of Heraclitus, that all things are and are not, seems to make everything true, that of Anaxagoras, *that there is an intermediate between the terms of a contradiction, seems to make everything false*; for when things are mixed, the mixture is neither good nor not-good, so that one cannot say anything that is true. (*Metaphysics* 4.7) [italics added – SW]

This law requires that the truth value of every proposition (with the possible exception of those dealing with future contingencies) be *determinate*: not merely that it cannot be both true and false, but that it must be either one or the other, with no third option allowed.

With respect to categories, the law of noncontradiction and the law of excluded middle, taken together, ensure the binary, all-or-nothing nature of categorical divisions. The total universe of entities is always neatly divided into two sets, members and nonmembers of the category, that are nonintersecting and complementary. No instances are inherently undetermined regarding their membership status: even when a person is unsure or erroneous in their classification, this is the result of his epistemic limitations, not of any haziness or indeterminacy in the objective structure of the reality.

It is the distinction into the essence of a being and its accidental characteristics that most directly reflects the spirit of the classical approach to categorisation. To Aristotle, entities are characterised by innumerable traits that might be truly asserted of them, but the truth of such assertions is contingent and not vital to the identity of the thing[179]. Such traits are called *accidents*, and their

[179] *Metaphysics 4.4.*

elimination would not change the identity of the thing in question. On the other hand, entities also have their *essences* that can be stated in the form of a conjunction of necessary and sufficient (defining) features (see 5.2.1.c)).

> The parts which are present in such things, limiting them and marking them as individuals, and by whose destruction the whole is destroyed, as the body is by the destruction of the plane, as some say, and the plane by the destruction of the line; and in general number is thought by some to be of this nature; for if it is destroyed, they say, nothing exists, and it limits all things… The essence, *the formula of which is a definition*, is also called the substance of each thing. (*Metaphysics* 5.8.) [italics added – SW]

It is important to understand properly the role of Plato, Aristotle, and their contemporaries in the development of the classical approach to categorisation/concepts. Their work conveyed a general worldview that was highly convergent with this approach and, especially in the case of Aristotle, introduced a number of devices useful in its development. Still, it would perhaps be an overstatement to credit any individual philosopher with the name of the 'founding father' of this theoretical outlook. Again, it is probably more productive to conceptualise the classical approach as the default, natural position, certain crucial aspects of which were brought to the surface already in the writings of the Greek masters.

5.2.2.2. Modernity

The default, natural character of the classical approach ensured the continuation of its historical dominance for the following several centuries. Notably, it formed the implicit foundation for the empiricist doctrine of knowledge, which is aptly observed by Laurence and Margolis (1999). Accounting for the origins of knowledge was the main theoretical objective of empiricism, so the most central aspect of the classical approach in this context was the elegance with which it

explained concept formation (concept acquisition). This was achieved by the idea of *combination* of simple elements into progressively more complex structures, and ultimately into concepts ("Ideas" in the original terminology).

On the generalised empiricist account, the acquisition of concepts consists in assembling their internal structures from more basic elements which in turn are assembled from even more basic elements, and ultimately from sense data. At least on the level of ideas, this process is characterised by discreetness: no blending between components occurs, so that particular components remain distinct. Thus, the process has the property of being theoretically reversible – through the analytical reversion of concept acquisition one is able to deconstruct concepts into their primitive component parts or at least extract their essences.

As is usually the case, paradigmatic examples come from the concepts of physical objects. John Locke envisages the process of the formation of concepts of physical objects in the following way:

> …[I]deas of Substances are such combinations of simple ideas as are taken to represent distinct particular things subsisting by themselves; the supposed or confused idea of substance, such as it is, is always the first and chief Thus if to substance be joined the simple idea of a certain dull whitish colour, with certain degrees of weight, hardness, ductility, and fusibility, we have the idea of lead; and a combination of the ideas of a certain sort of figure, with the powers of motion, thought and reasoning, joined to substance, the ordinary idea of a man. (Locke 1999 [1690]: 149).

> …[T]he greatest part of the ideas that make our complex idea of gold are yellowness, great weight, ductility, fusibility, and solubility in aqua regia, etc., all united together in an unknown substratum. (Locke 1999: 300)

This last quote, although clearly illustrative of the classical-like assumptions behind Locke's system, is in fact part of a larger section explaining the ontological status of such properties. In an interesting way, this is a deviation from the principle A listed above – categories as mirroring the objective reality –

towards a considerable degree of mind-dependence, a theme later developed especially by George Berkeley. Once again it needs to be stressed, however, that the chief motivation behind the classical mindset of the empiricists was the idea of explaining the acquisition of all knowledge by reference to the relatively simple processes of *combination* of units (of ultimately sensory origin), which made it possible to postulate minimal inborn cognitive machinery.

5.2.2.3. The twentieth century

In one form or another, the implicit assumption of the classical view prevailed in our intellectual tradition in most, if not all, of the past century. Many of the most pervasive intellectual currents of the first part of the twentieth century can easily be shown to implicitly embrace the classical precepts. Similarly, early mentalistic research into concept acquisition in developmental psychology was founded on a classical-like general outlook.

For the logical positivists such as Rudolf Carnap[180], one of the guiding ideas was that of reducing all statements to atomic statements about basic observational facts; the very possibility of such a reduction was a criterion of classifying a statement as truly meaningful. This can be diagnosed as reflecting classical-like intuitions, albeit on a propositional level rather than on the level of individual words/concepts. These, however, too were postulated to be reducible, first to other words, and then progressively to "primary" or "protocol senses" reporting basic experienced qualities (Carnap 1959 [1932]) – at least in case of 'meaningful' as opposed to 'metaphysical' words. Another caveat is that the interests of the logical positivists were largely limited to scientific discourse.

Behaviourists, researching category learning with abstract stimuli saw it essentially as discovering the common property possessed by all members and no non-members (cf. the works by Clark L. Hull and Kenneth L. Smoke, as

[180] Rudolf Carnap (1891–1970), a philosopher and logician of German origin, later an American citizen, a leading logical positivist and a member of the intellectual group known as Vienna Circle.

discussed e.g. by Murphy [2002: 12–16] and Storms [2004]). Although the objects of study were not so much 'concepts' or 'word meanings' but rather artificial categories, the general framework was classical in the sense of relying on the necessary and sufficient features. This line of thought essentially agreed in this respect with the later work on learning theory conducted within a representational paradigm, represented notably by Jerome Bruner (e.g. Bruner et al. 1999 [1956]).

5.2.3. Criticism

Serious opposition against the broadly understood classical view did not emerge until the later half of the twentieth century; since then, however, the classical perspective has faced very strong criticism on many grounds, leading to its systematic decline in Cognitive Science. At a minimum, it has gradually lost its appeal as a theoretical approach giving any promise of eventually describing accurately the mentalistic range of phenomena related to categories, such as categorisation decisions or concept acquisition. The growing criticism has been backed up by a broad and increasing range of objections having both conceptual and empirical nature.

Murphy (2002: 16–49), who takes an extremely critical position against the classical approach, groups the objections into "in-principle arguments" and "empirical problems". Medin and Smith (as quoted in Medin 1998: 95–96) enumerate three main reasons: the failure to specify defining features, the existence of goodness of example effects, and the existence of unclear cases. Medin and Rips (2005: 37–72) focus on the problems with hierarchical arrangement of categories, such as the failure of the hyponyms to inherit all of the defining properties of the hypernym. Laurence and Margolis (1998: 14–27) further specify the list of reservations, adding the problems with psychological reality, analyticity, ignorance and error, conceptual fuzziness, and typicality effects. Partly drawing on these lists, below I discuss, comment, and summarise

the main lines of objections, suggesting a general philosophical reason for the inadequacy of the theoretical perspective in question.

5.2.3.1. Ludwig Wittgenstein

It is widely accepted, at least by linguists and philosophers (somewhat less by cognitive psychologists, who tend to put more stress on the issue of typicality effects, as described below), that an important, possibly decisive blow against the classical approach was dealt by Ludwig Wittgenstein[181] in a work published posthumously as *Philosophical Investigations* (Wittgenstein 1987 [1953]). Such legendary status of Wittgenstein's work in this particular respect is rather astounding. It must be made clear that – contrary to a popular belief – the book never explicitly touches on the problems of categorisation, and its focus is entirely different. Instead, the relevant sections of the text have the character of comments made in passing by way of addressing the question of the definition of the book's central term, "language-game" (*Sprachspiel*). Wittgenstein's observation runs as follows (1987: 31–32):

> For someone might object against me: "You take the easy way out! You talk about all sorts of language-games, but have nowhere said what the essence of a language-game, and hence of language, is: what is common to all these activities, and what makes them into language or parts of language..."
>
> And this is true.—Instead of producing something common to all that we call language, I am saying that these phenomena have no one thing in common which makes us use the same word for all,— but that they are *related* to one another in many different ways. And it is because of this relationship, or these relationships, that we call them all "language"…
>
> …Consider for example the proceedings that we call "games". I mean board-games, card-games, ball-games, Olympic games, and so on. What is

[181] Ludwig Josef Johann Wittgenstein (1889–1951), an Austrian philosopher whose work exerted enormous influence on the philosophy of language in the twentieth century, precursor of the intellectual movement known as *the linguistic turn* which emphasised the study of natural language in its relation to philosophical problems.

> common to them all?—Don't say: "There *must* be something common, or they would not be called 'games' "—but *look and see* whether there is anything common to all.—For if you look at them you will not see something that is common to *all*, but similarities, relationships, and a whole series of them at that. To repeat: don't think, but look!—Look for example at board-games, with their multifarious relationships. Now pass to card-games; here you find many correspondences with the first group, but many common features drop out, and others appear. When we pass next to ballgames, much that is common is retained, but much is lost.—Are they all 'amusing'? Compare chess with noughts and crosses. Or is there always winning and losing, or competition between players? Think of patience. In ball games there is winning and losing; but when a child throws his ball at the wall and catches it again, this feature has disappeared. Look at the parts played by skill and luck; and at the difference between skill in chess and skill in tennis. Think now of games like ring-a-ring-a-roses; here is the element of amusement, but how many other characteristic features have disappeared! And we can go through the many, many other groups of games in the same way; can see how similarities crop up and disappear[182]. [italics in the original]

It should be noted that in the context of theories of categorisation, where this argument is usually applied, it is remarkably weak[183]. What it does establish is an original case for explicit definitions' *sensitivity to counterexamples*, a trait that has since been underscored by many researchers. One popular example is the often quoted case of the inadequacy of the definition of 'bachelor' as 'unmarried

[182] Wittgenstein then extends his analysis to the concepts of 'number', 'proposition', 'derive', 'guide' and 'reading', with a similar conclusion, i.e. that it is impossible to specify the definitions of the relevant terms. His best known quotes that deliver this conclusion are „[a]nd the strength of the thread does not reside in the fact that some one fibre runs through its whole length, but in the overlapping of many fibres" (1987: 32), and „[i]n order to find the real artichoke, we divested it of its leaves." (1987: 66).

[183] The very idea of over two millennia of intellectual tradition being destroyed by an implicit argument contained in almost a single passage of text must be felt as deeply disturbing. The status of Wittgenstein's observations in this respect might in a large part result from his general status as a major figure in twentieth century philosophy.

man', another – Fodor's "painting one's paintbrush" analysis (Fodor 1981a), in which each of a number of successive attempts at decomposing the verb 'to paint' is countered by a counterexample.

The sensitivity to counterexamples, while indeed posing considerable difficulty to the classical approach, is still very far from being conclusive evidence against it. The incontrovertible fact that explicit definitions are sensitive to counterexamples translates into a claim that it is very difficult to provide lists of necessary and sufficient conditions for any given concept (analogically, word meaning) by means of introspective linguistic analysis, where the conditions would be expressed by other words – but not into a strong refutation of this general view.

The challenge of sensitivity to counterexamples invites at least three possible replies. Firstly, the difficulty in producing successful definitions does not itself constitute a proof that this is impossible in principle (see e.g. Medin 1998). Secondly, there may be other methods of arriving at the lists of conditions than introspective analysis drawing on the linguistic intuitions of a researcher or researchers (see especially Harnad 1990, 2002, 2005). Thirdly, and most importantly, the problem might lie in the implicit assumption of the verbal format of the features, that is that the features can be adequately captured in terms of words (cf. section 5.2.5.).

A final point concerns the *positive* side of the Wittgensteinian argument, that is the idea of *family resemblance*. This is the idea that category members, on analogy with members of a family, do not universally share any single property, but rather are distinguished by each exemplifying some of the set of features characteristic of the family as a whole: "I can think of no better expression to characterize these similarities than <family resemblances>; for the various resemblances between members of a family: build, features, colour of eyes, gait, temperament, etc. etc. overlap and criss-cross in the same way" (Wittgenstein 1987: 32). Again, this has not been developed by Wittgenstein in any greater detail, but has proven to be almost equally seminal as the negative part of his

argument. Notably, family resemblances were a leading inspiration behind the work of Eleanor Rosch (see below).

5.2.3.2. Willard van Orman Quine

Another line of philosophical reflection widely taken to constitute evidence against the classical view was inspired by the logician and philosopher Willard van Orman Quine. In particular, his influential text "Two dogmas of empiricism" (Quine 1961 [1951]) shed light on the problems with the notion of *analyticity*. The essence of Quine's argument in this text was that there could be no principled and nonarbitrary dividing line between 'the empirical' (fact, observation, 'pure' experience) and 'the linguistic' (the description, the coding, the 'semantic component'); between the sentences confirmable synthetically and analytically. Thus, analyticity was downgraded from something qualitatively and categorically different to being simply an idealised extreme on an uninterrupted continuum with syntheticity. Once again, Quine's paper was not meant as a critique of a theory of categorisation – it was an important text in the philosophy of science questioning the possibility of dividing *the propositions of a science* into two distinct sets: ones whose truth needs to be demonstrated by reference to experience (synthetic) and ones whose truth can be demonstrated on purely conceptual grounds (analytic). The relevant problematic consequence of Quine's text – analysed in Laurence and Margolis (1999) – resulted from the questioning of the overall validity of the notion of analyticity understood generally. The classical view requires that there be a precise, *qualitative distinction* between a small class of inferences mandated by the core structure of a concept, which should be true *analytically* (necessarily in virtue of the concepts meaning), and all other inferences, contingently true or false. In the opinion of Laurence and Margolis, with such a qualitative difference removed or even made a matter of degree the classical approach loses its fundamental rationale.

While undeniably philosophically influential, for the cognitive scientist this line of criticism carries only limited impact. It must be kept in mind that

Quine's original motivation and context was related to the philosophy of science, not theories of concepts or categorisation. Thus, the argument applies to a cognitivist study of categorisation only on the strength of analogy of science to cognitive processes[184]. From the perspective of Cognitive Science it qualifies simply as a valuable insight, and more conclusive arguments should come from empirically oriented studies.

5.2.3.3. Psycholinguistic experiments (typicality effects)

Outside of philosophy and traditional linguistics, it was empirical work by cognitive psychologists that has given most momentum for discrediting the classical approach. Of particular import was the research of Eleanor Rosch and her establishing the phenomenon of *typicality effects* (although it is worth noting that the philosophical insight of family resemblances was a guiding inspiration behind her proposals, e.g. Rosch and Mervis 1996 [1975]: 442–443). More specifically, this concerns the pervasive and wide-ranging effects that the phenomenon of *typicality* has on the performance of subjects in tasks based on processing category-related information.

Rosch's point of departure was the realisation that categorisation is a ubiquitous phenomenon and a fundamental process in nature, vastly exceeding the area of 'high-level', conceptual categorisation (cf. section 3.2.3.). She understood categorisation as a real process carried out by living organisms constrained by their environments and resources such as time and energy. There followed a shift from an 'abstract'[185] to a *functional* understanding of categories, as organised around the structure of goals of the cognitive agent. Perhaps the most prominent consequence was the prediction of a *differential status* of

---

[184] For other arguments related to the controversial status of the general analogy between cognitive processes and science see Botterill and Carruthers (1999: 69–73).

[185] "My own work on categories did *not* originate in learning theory, in concept identification, in formal linguistic semantics, or in semantic memory; it might have taken a very different course had it done so." (Rosch 1988b: 374; italics in the original).

exemplars as members of a category relative to *what is important* from the perspective of a given agent or – especially in the context of linguistic categories – community of agents.

In a series of experiments, Rosch (prior to 1973 publishing under the name Heider) and her co-workers conclusively established a range of effects that were related to the item's typicality as a member of a given category. How typical an item was for its category was found to reliably influence (reviewed e.g. in Rosch 1988a, 1988b):

- *reaction times*; in responses to questions of the structure 'is an X a Y', where X is an exemplar and Y is a category, response times are inversely correlated with the item's typicality, i.e. subjects respond faster for X's that are typical for the category Y; so that e.g. they react faster to 'is robin a bird' than to 'is turkey a bird';
- *speed of learning*; children are faster in forming categories based on exposure to 'good' (typical) than to 'bad' (atypical) exemplars of a category;
- *spontaneous generation probability/order*; subjects asked to generate a list of members of a given category tend to produce typical members with higher probability than atypical members and usually before them;
- *use of hedges and sentence substitutability*; sentences of the structure 'an X is a Y' are readily qualified with hedges such as 'technically' (e.g. as 'a turkey is technically a bird') when X is an atypical category member but not when it is a typical member; in contrast, only typical members can figure in constructions as 'an X is an Y *par excellence*'.

Typicality effects are a basic cognitive phenomenon that is extremely well-established and extremely pervasive[186]. Worth appreciating is the fact that they are by no means restricted to lexical-semantic categories, but appear in all areas of cognitive processing, for all types of subjects. They are found for artificial as well as natural categories (Posner and Keele, referred to by Murphy [2002: 28–31]), in perceptual as well as conceptual categorisation and in animals as well as in humans (Dépy et al. 1997).

A point that requires clarification is the distinction between *typicality effects* and the related but separate topic of *prototype theories* or *models* of concepts/categorisation (in this work, the former topic is discussed in this section, the latter – in the next chapter). This is best illustrated by the position of Rosch herself, who warned against equating those two subject matters (1988a [1978]). Although her earlier texts (e.g. Rosch and Mervis 1996 [1975]) might have encouraged different conclusions, in a later paper, Rosch (1988a: 319–320) made it clear that her general discussion of the prototypes should not be taken as a proposal of a specific prototype theory of concepts/categorisation. Her own point, as clarified in the quoted text, was more modest, i.e. merely that typicality effects are a robust and important phenomenon that must be accommodated by any successful theory with serious aspirations to psychological reality: "In short, prototypes only constrain but do not specify representation and process models"[187].

---

[186] Such is, for example, the opinion of Zdzisław Chlewiński (1999: 176) and Laurence Barsalou (1992: 175–176), among many others.

[187] Rosch (1988a: 319–320): "2. Prototypes do not constitute any particular processing model for categories... What the facts about prototypicality do contribute to processing notions is a constraint - process models should not be inconsistent with the known facts about prototypes...

3. Prototypes do not constitute a theory of representation of categories. Although we have suggested elsewhere that it would be reasonable in light of the basic principles of categorization, if categories were represented by prototypes... such a statement remains an unspecified formula until it is made concrete by inclusion in some specific theory of representation…

It is precisely this constraint that is violated by the classical approach. Models based on the classical premises lack the explanatory power to account for typicality effects. The classical approach has no resources to capture the gradability of structure and thus cannot differentiate between the typicality statuses of the items under classification. These either match the set of criterial features, thus counting as category members, or fail to match the set, thus counting as nonmembers, with no further distinctions possible (cf. 5.2.1. a) "all members of a category have equal status", and b) "all non-members of a category have equal status".

Much like in the two previous cases, it is easy to misinterpret the exact power and scope of critical argument against the classical view resulting from the phenomenon of typicality effects. It is therefore necessary to underscore that – strictly speaking – this phenomenon does not overtly *contradict* the classical view; it is simply *not accounted for* in this tradition, and so does not constitute a conclusive argument in and of itself. While typicality must be acknowledged as a cognitively real phenomenon, categorisation (at least in principle) can still be maintained to be a psychological process independent of typicality effects. Indeed, in case of at least some types of categories, such dissociation between categorisation and typicality was shown experimentally (e.g. Armstrong et al. 1999 [1983][188]). This leads to the proposals that the classical view can be salvaged by upgrading it to a 'hybrid view'[189], on which concepts would

---

4. Although prototypes must be learned, they do not constitute any particular theory of category learning... In short, prototypes only constrain but do not specify representation and process models."

[188] In this experiment, certain concepts such as ODD NUMBER proved *both* to be well-defined (accordingly, to have clear boundaries) *and* to have graded typicality structure. Typicality structure, however, did not influence categorisation, i.e. how typical an item was rated for the concept had no effect on whether it was appropriately classified.

[189] Also termed the *Binary Model* (Hampton 1995), the *Dual Theory* (Laurence and Margolis 1999) or the *Revised Classical View* (Murphy 2002). Nęcka et al.'s (2006) use of the term *The Probabilistic View* in this context seems to be incorrect.

incorporate 'classical' cores as well as more superficial 'identification procedures' characterised by graded typicality.

To a degree, the question of the validity of 'hybrid models' reflects the distinction between the philosophical and the cognitivist stance. Although there is some experimental evidence specifically favouring the prototype to the "Binary" view (provided by e.g. Hampton 1995) on empirical grounds, that latter stance also has the major drawback of being in principle unappealing to a cognitively minded researcher. Even assuming the possibility that the classical part of the model could handle the task of categorisation, the notion of categorisation would be construed very narrowly. To a cognitive scientist, categorisation is the most interesting as a reliable shorthand for a whole set of associated phenomena, in particular *differential behaviour* and *inferencing* (cf. Barsalou 1992: 25: "…[t]he primary purpose of making categorizations is to support inferences relevant to the perceiver's goals: Categorization is usually not an end in itself."). From such a standpoint, one categorises an entity as a dog not just for the sake of categorising itself, but rather in order to gain access to a rich body of inferentially available information associated with this category; and he can even be reliably expected to behave in certain ways appropriate for encounters with dogs. Since the 'classical cores' have been shown to be deeply deficient in explaining this extended cognitive dimension of categorisation, their theoretical value is minimal[190].

5.2.3.4. Context sensitivity and vagueness

A specific phenomenon of particular interest, also related to typicality, was first identified in a study by Michael McCloskey and Sam Glucksberg (1978). They found that – especially for atypical members of a category, that is items neither

[190] Cf. also the highly critical opinion of Murphy (2002: 39–40), who complains that the 'revised classical theory' starts from the position of attempting to mitigate the criticism based on typicality effects in order to preserve certain preexisting theoretical goals rather than starting from the psychological data itself.

clearly belonging nor clearly not belonging to a category, e.g. egg for ANIMAL – categorisation decisions could differ not only intersubjectively (between different people) but also *intra-subjectively* (for the same person across time). As shown by McCloskey and Glucksberg, in some cases the subjects were inconsistent in some of their classification decisions, so that one and the same person could classify the same item differently on different occasions. This is unexpected and hard to explain if one assumes the classical view, since if categorisation decisions are reached *via* a property-matching process, it is difficult to see how it could yield results unstable through time for the same subjects.

A classic example of vagueness/fuzziness from the field of linguistics is discussed in the widely quoted text by William Labov (2004). In Labov's experiment, the subjects were requested to name the objects presented to them on drawings (one at a time, in randomised order) – see Fig. 7. The most interesting finding was related to the "Contexts" trial, in which the subjects were asked to imagine the objects in one of four context, e.g. with someone drinking coffee from the object, or the object standing on a table filled with mashed potatoes. The imagined context turned out to have a substantial impact on the categorisation of the objects on drawings as evidenced by their naming.

Such a result stands in direct conflict with the classical view (specifically, its essentialist premises), which does not allow contextual – and therefore noncriterial – features to influence categorisation decision outcomes. In theory, this objection could be addressed by re-qualifying the contextual variables of such sort as criterial features and thus including them into the extended set of necessary and sufficient features. This, however, would lead to further problems. Firstly, it would be extremely difficult to constraint the set of relevant contextual variables, and secondly, their differential importance would still contradict the classical tenets (cf. point d) in 5.2.1.); thirdly, they could turn out to be very difficult to formalise and operationalise.

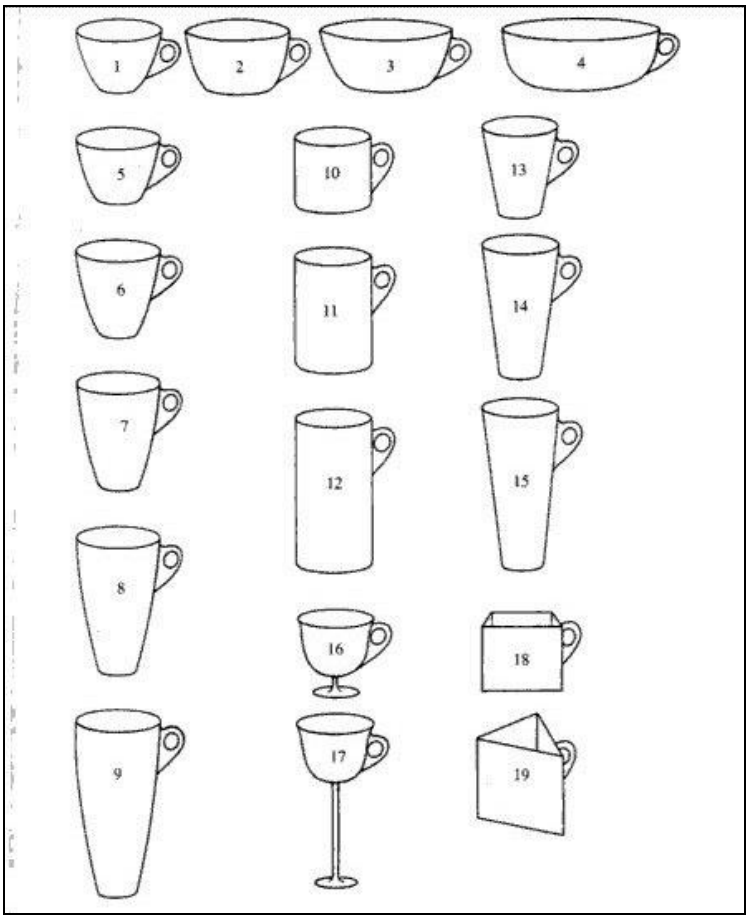

Fig. 7. Visual stimulus used in the study by Labov (2004 [1973]).

A natural way of dealing with the cases such as those in McCloskey and Glucksberg 1978 and Labov 2004 is parallel to a popular philosophical response to the so called 'the Sorites paradox', or 'the paradox of the heap'. This paradox illustrates the difficulty in establishing the exact number of grains of cereal needed to constitute a heap. Although a heap consists of a finite number of grains greater than one, when adding grains one by one it seems to be impossible to identify any particular turning point at which the next grain would complete the qualitative transition after which the entity in question could finally be described as 'a heap' (reviewed e.g. in Dominic Hyde 2005).

A frequently offered solution rests on resorting to the division of the indeterminacy pinpointed by the paradox into *semantic* and *epistemic*[191]. If the indeterminacy is qualified as merely *epistemic*, it is a viable option to claim that

[191] Cf. Hyde 2005.

there indeed exists a specific borderline number of grains necessary and sufficient for making a heap. The apparent indeterminacy would result purely from the epistemic limitations of the subjects who would be normally unable to determine this number with the requisite accuracy[192].

By the same token, it could be proposed that in reality the meaning of words/concepts from the studies in the vein of (McCloskey and Glucksberg 1978) and (Labov 2004) quoted above is fixed and not vague. It would follow that for every (conceptualised) object such as an individual cup or an instance of a disease it either does or does not belong to a classically construed category, understood here as a class of entities in the world. Whereas the subjects might not always be able to adequately determine the categorial status of the olive or a receptacle, *in reality*, the olive either is or is not a fruit, and a given receptacle either is or is not a cup. It is already obvious, however, that from a cognitivist perspective such a reply is deeply uninteresting for the same reasons as were mentioned in some of the previous objections. It is precisely this epistemic, cognitive dimension that a psychologically real theory of categorisation should be able to explain; when it is missing from the account, the theory can no longer count as a theory of *categorisation* (the broad subject area of categorisation) but rather as an ontological theory.

5.2.4. Evaluation

As can be seen from the previous section, the criticism of the classical approach is overwhelming. In particular, in Cognitive Science, it is considered a non-contender (although in fairness it should be noted that this has never been the primary area of application of the work in the spirit of the classical view). Very clearly, the classical way is *not* how humans go about categorisation tasks, and if

[192] It could be observed that some 'soritic' predicates are more amenable to such an explanation than others. While seemingly valid for 'is red' or 'is bald', such a strategy feels drastically less palpable for predicates like 'is fast', 'is shy', or 'is fairly big'.

the criterion of psychological reality is violated, the theoretical approach in question must automatically be discarded.

While such a strong conclusion is well founded, one should also remember that the scope of the refutation of the classical view is not absolute, and that several theoretical areas can be identified in which it appears to be the most adequate description of phenomena. One such area is specialised scientific vocabulary, which is often thought to be classical *par excellence*: specialised scientific terms are not so much described by their definitions as called into being by them, so theoretically they should be fully reducible to their definitions. Another area is categorisation as a conscious, rule governed process. Humans are fully capable of acting in ways compatible with the classical view in that they can follow a set of explicit rules in sorting tasks, e.g. in assigning objects to specific categories on a basis of predefined feature checklist. For example, if a person has mastered the rules of chess, he or she can categorise the moves into legal and illegal one by the process of conscious application of explicit, linguaform rules.

Still, even such cases are not without their caveats. Firstly, technical vocabulary is often found to be sensitive to counterexamples, which demonstrates that at least some technical terms have other stabilising elements over and above their definitions[193]. As to conscious, rule governed categorisation, experimental data show that it is always to some degree influenced by nonclassical effects. Even when one is given a very clear set of explicit rules, noncriterial factors still tend to have an impact on categorisation outcomes, and certainly have a significant impact on the categorisation *process*. This is shown unequivocally by Edward Smith et al. (1998), who discuss and then replicate a study in visual categorisation of artificial stimuli by Scott W. Allen and Lee R. Brooks (quoted and analysed in Smith et al. 1998: 172–180). They have established that the subjects, although presented with an unambiguous rule, also

[193] Cf. the example of a ‘straight line’ in non-Euclidean geometry.

invoke implicit 'similarity' procedures at the same time, as evidenced by the error rate and reaction times[194]. Somewhat similar effects were shown for natural conceptual categories in (Hampton 1995), where the categorisation decision of subjects still depended on the presence of the "characteristic features" even though the subjects had been explicitly instructed to rely exclusively on the "defining features" of a category (cf. also the further comments by Smith et al. 1998: 180–181).

A much better candidate for the area where the classical view can be at least partly vindicated is that of people's *beliefs* regarding the categorial structure of the world. While their concepts, and as a result, their actual categorisation decisions are patently nonclassical (as discussed above), people appear to be convinced that the external world, especially in the biological dimension, is partitioned into classes of beings, each of which has clear boundaries and has an identity defined by its hidden *essence*. This is known as *psychological essentialism*. The term in question comes from Douglas Medin and Andrew Ortony (1989), who dub it "the psychologically plausible analog of the logically implausible doctrine of metaphysical essentialism". The authors explain: "[t]his would be not the view that *things* have essences, but that people's *representations* of things might reflect such a belief… We think there is evidence that ordinary people *do* believe that things have essences" (Medin and Ortony

[194] In this experiment, the subject were asked to categorise pictures of imaginary animals as Diggers versus Builders. The criterion for categorisation was the presence of at least 2 of the three attributes in the test object. There were two groups of subjects – the Rule group (who were told the rule for categorisation) and the Memory group (who did not know the rule and were supposed to make guesses on the basis of previous exemplars). The most interesting results concern the so-called Negative Matches - examples that were technically Builders (they matched the rule), but were more perceptually similar to Diggers. The Rule group was better at categorising normal Builders than Negative Matches, which suggested that apart from rule-based strategy they also resorted to some similarity judgements. If they had relied solely on rules, they should have categorised all builders, including Negative Matches, with the same accuracy.

1989: 183, italics in the original). Psychological essentialism has subsequently received sound empirical confirmation, although mostly with respect to biological kinds (e.g. the study by Woo-kyoung Ahn et al. 2001). Despite this last limitation, psychological essentialism might turn out the only major aspect of categorisation relevant to Cognitive Science that is highly compatible with the classical approach.

5.2.5. Specific problem: feature format

The problem of the format of the *features* of concepts has reappeared throughout the present chapter. Here, it is singled out as a separate point, a decision that is substantiated by the need to highlight its utmost importance.

Much of the study on categorisation, especially in linguistics and philosophy, appears to have rested on the unspoken underlying assumption that adequate and relevant categorisation features are overt and can be given in a verbal representational format (single words, 'primes' roughly equivalent to single words, "subject-generated verbal predicates", etc.). The identification of such features requires analytical scrutiny of the studied concept, involving selective attention and drawing on working memory resources, and the subsequent encoding of the features in a symbolic format (in the process of describing a feature verbally). The cognitive resources used in such a task are thus characteristic only of controlled, rule-based processing.

It is crucial to bear in mind that there is no guarantee that the features which one uses in this analytical, post-factum, symbolic description of a particular category are exactly the features upon which one actually bases her categorisation decisions. It might be useful to revert to Wittgenstein's (1987) well-known example of family resemblance – but in a *literal* way. On encountering a (known) human face, one is instantly able to categorise it as a particular familiar face; one completes this categorisation task using low-level cognitive processes that does not enter his consciousness. The face can relatively easily be described verbally, employing such predicates as 'has bushy eyebrows',

'has a long nose' or 'has a pointy chin', but there is little doubt that even a very detailed description of this kind will be highly unsatisfactory; the features actually involved in the categorisation process must be of a rather different nature. The true features are below the level of our verbal access[195].

Particularly relevant here is Jackendoff's (1996: 545) comment on the character of semantic primitives of word meanings (which on most views are equivalent to 'essential features' of concepts):

> [I]t is not necessarily the case that all (or any) semantic/conceptual primitives are independently expressible *as words*. Just as the smallest isolable speech sounds (phonemes) are composites of distinctive features that cannot appear independently, so it appears to be the case that all word meanings are composite, made up of semantic/conceptual constituents that cannot appear in isolation. That is, word meanings are "molecular" entities in the "chemistry of concepts", while semantic/conceptual primitives are subatomic or even quarklike. This being the case, the ultimate decomposition of a lexical item cannot be expressed in terms of a linguistic paraphrase. [Italics in the original.]

When properly considered, this turns out to have interesting consequences for the evaluation of the generalised classical view. Consider the influential Wittgensteinian objection reviewed in 5.2.3.1. Rather than the proper interpretation to the effect of 'necessary and sufficient features are very difficult if not impossible to state verbally', it is often given an overly strong interpretation to the effect of 'there is no feature whatsoever, of whatever character, that would be exemplified by all category members and no nonmembers'. Harnad (2002) takes issue with such a view, pointing to the

[195] Jackendoff [1990: 42–43] introduces the example of face recognition to reach a similar conclusion, although in a slightly different context.

logical necessity of the existence of *some* invariances, as implied by the very phenomenon of successful categorisation[196]:

> (1) Successful Sorting Capacity Must Be Based on Detectable Invariance.
>
> The theorist who wishes to explain organisms' empirical success in sorting sensorimotor projections by means other than a detectable invariance shared by those projections (an invariance that of course need not be positive, monadic and conjunctive, but could also be negative, disjunctive, polyadic, conditional, probabilistic, constructive -- i.e., the result of any operation performed on the projection, including invariance under a projective transformation or under a change in relative luminance -- indeed, any complex boolean operation) has his work cut out for him if he wishes to avoid recourse to miracles, something a roboticist certainly cannot afford to do. (Harnad 2002)

Harnad's point is that for each category there simply must – as a matter of logical necessity – exist an invariance that is responsible for supporting the category's identity, *and that the difficulties for pinpointing such invariances stem from their elusive nature rather than their nonexistence*. Firstly, the invariances in question may consist, and probably do consist, in something much more complex than the simple process of binary feature checking, as assumed in the basic classical models. One example of a plausible invariance is the distance from the category prototype (that is: the invariant element shared by all category members is the distance from the prototype on the scale of similarity, computed over the matrix of weighted shared features), which is precisely the mechanism underlying the similarity-based (that is, prototype and exemplar) approaches. However, the other point to bear in mind is that the features involved in

[196] This, in a way, is already acknowledged by Wittgenstein: „But if someone wished to say: <There is something common to all these constructions—namely the disjunction of all their common properties>—I should reply: Now you are only playing with words. One might as well say: <Something runs through the whole thread—namely the continuous overlapping of those fibres>." (1987: 32)

establishing the invariances might be present on a deeper, sublinguistic (using Jackendoff's metaphor quoted above, "quark" rather than "atomic") level.

5.2.6. Natural Semantic Metalanguage

An illustration of the above problem might be the original approach to lexical/conceptual semantics known as NSM (Natural Semantic Metalanguage), developed over several decades by a Polish linguist Anna Wierzbicka. Natural Semantic Metalanguage is an approach to semantic analysis of word meanings in natural languages, the central element of which are the so-called 'primitive concepts' (Wierzbicka 1996: 35–110), also variously referred to as 'indefinabilia', 'atomic expressions', 'semantic primes', or the 'alphabet of human thought' (cf. Kalisz 1998). While Wierzbicka's approach cannot be strictly classified as 'classical in spirit' – at least not without caveats and reservations – it shares with it at least one crucial general assumption, i.e. that of decomposition of wholes (word meanings) into features that themselves have a semantic character and can be expressed lexically[197].

Semantic primitives (primes) as construed in NSM are the ultimate units of semantic analysis. The meanings of other words in a natural language can be expressed in short paraphrases involving only semantic primes and/or intermediate expressions that could themselves be stated in terms of semantic primitives. In contrast, the primes are atomic, and even though in principle they could be broken down into sets of arbitrary features, they are not amenable to being further decomposed in the way described above (Wierzbicka 1996: 28). Each of the postulated semantic primes needs to have a lexical exponent; 'lexical exponent', however, "is used in a broad sense to include not only words, but also bound morphemes and phrasemes (fixed phrases)" (Goddard 2002: 406–407). Semantic primes are arrived at through "in depth analysis of *any* natural

[197] Wierzbicka herself seems to subscribe to the widely understood definitional view in her rebuttal of Fodor's arguments against the posibility of defining lexical concepts (Wierzbicka 1996: 237–257), and also, to a degree, in her criticism of the prototype theory (1996: 148–169).

language" (Wierzbicka 1996: 13, italics [SW]): for each postulated primitive, there is (hypothesised to be) a lexical exponent of that primitive in any natural language. This implies *universality*, but not only on the level of linguistic sings, but also conceptual repertoires of the speakers of all human languages. This universality across human languages is suggested to be founded on a strong innate component (Wierzbicka 1996: 16–19).

An example of an NSM paraphrase could be the definition of the concepts 'game' supplied by Wierzbicka (1996: 158–159) as a reply to the influential Wittgensteinian argument quoted above:

games

(a) many kinds of things that people do
(b) for some time
(c) "for pleasure" (i.e. because they want to feel something good)
(d) when people do these things, one can say these things about these people:
(e) they want some things to happen
(f) if they were not doing these things, they wouldn't want these things to happen
(g) they don't know what will happen
(h) they know what they can do
(i) they know what they cannot do

The NSM approach, however, is not without its problems. Symptomatically, the analysis of 'game' (or, strictly speaking, the concept GAME) just cited is accompanied by a footnote explaining that "metaphorical extensions, ironic or humorous use, and the like" are excluded from being covered by this definition; Wierzbicka (1996: 159) resorts here to the distinction into "playful extensions" and "basic meaning", which she does not further substantiate. The difficulties with semantic primes are concisely reviewed by Roman Kalisz (1998), who groups them into the following categories:

- incompleteness and incomprehensibility

- generality
- gestaltive effects
- grammaticality
- awkwardness

In short, Kalisz (1998) points out that firstly, Wierzbicka's paraphrases in fact to a large extent rely on intermediate expressions, and when these are substituted with their analyses into primitives, the paraphrases become incomprehensible. Secondly, the paraphrastic explications are usually too general (unrestrictive), extending to cover examples not falling under the relevant concept. Thirdly, the paraphrases only approximate the original meaning, being unable to capture some of the emergent subtleties that are specific to the unreduced forms. Fourthly, closed class words, although frequently appearing in the paraphrases, are underrepresented in the proposed repertoire of primitives; as a result, the ideal of building explanations *fully* reducible to the set of primitives is difficult to achieve in practice. Finally, in practice the explanations often turn out to be clumsy, calling for increased processing effort in comprehension.

Despite Wierzbicka's ingenuity – acknowledged by commentators including Kalisz himself – it is clear that complete representations of word meaning can only very rarely (if at all) be reconstructed from semantic primitives. I suggest that this follows directly from the main commitment of the NSM approach (to the primitives having lexical exponents), thus constituting another example of the feature format problem.

5.3. Summary and conclusion

In Chapter 5, I reviewed the so-called 'classical approach' to the problem of (broadly understood) categorisation and conceptual structure. After clarifying the potential terminological difficulties, I summarised the main theoretical tenets, which was followed by a brief historical review. In later sections, I discussed the main objections against the classical view, most importantly the problems with

typicality effects and with irremovable context sensitivity. Recapitulating on the strengths and weaknesses of the classical approach, I determined that no attempt to salvage this general stance was convincing and, at least in Cognitive Science, this approach offers no promise for a successful application.

In the final section, I considered the crucial problem of establishing the features relevant for categorisation. Specifically, I focussed on the *format* of the features. I concluded that the assumption about the possibility of capturing the features verbally, i.e. the assumption that the features can be adequately described by words, might be a major stumbling block to developing more accurate models of categorisation. What is more, even those decompositional accounts of conceptual content or word meaning that explicitly avoid identifying conceptual components with words must ultimately refer to the descriptive categories offered by a natural language. This is a pervasive and possibly even irremovable problem resulting directly from the qualitative rather than quantitative nature of linguistic and philosophical study.

The above treatment of the classical approach leads to several additional considerations that might turn out to be worth exploring in more detail in further philosophically-oriented research. Firstly, the failures of the classical view with respect to the cognitive dimension seem to reflect a consistent pattern. In view of the division of labour between the mentalistic and externalistic perspectives presented in Chapter 2, it is possible that the classical approach meshes much more naturally with the latter alternative. For example, the normativity of concepts appears to be an essentially E-semantic subject area, with very limited potential for exploration within a mentalistic framework. Still, in the applications related to normativity – especially where the requirement for an explicit statement of a set of criteria results from the very nature of the task, e.g. in normative definition – models based on classical assumptions may turn out to be indispensable.

Another, and a possibly surprising reflection stems from a juxtaposition of the abovementioned pattern of failures of the classical approach with the pattern

of failures of the general project of classical Artificial Intelligence. It has been repeatedly observed (perhaps most pungently by Jerry Fodor [1983], [2000]) that AI has so far failed to live up to its early promise. Considering that since the influential paper by Alan Turing (1950) the standard measure of success of AI has been the capability for intelligent conversation with a human, it is the mastery of the concepts of natural language that seems to be the foremost obstacle to the successful progress of AI. One may note that if the idea of meanings-as-lists-of-necessary-and-sufficient-conditions was in fact a feasible one, and if concepts for standard lexical items were in fact implementable in the classical spirit, then the problem of concept acquisition in machines could be reduced to the implementation of a set of primitives plus simple computational rules. Accordingly, the as yet insurmountable difficulties of classical AI may partly reflect the troubles with the classical theory of conceptual content.

6. Conceptual atomism and its refutation

6.1. Introduction

In the previous chapter, I provided a review of historically the most influential theory of conceptual structure in linguistic, philosophical and cognitive-psychological literature on the topic of categorisation, broadly construed (as defined in 3.2.3.). The review has been accomplished from a recent historical perspective, and has been supplemented with a critical evaluation from a cognitivist standpoint represented in this work.

All of the theories – or more appropriately, theoretical approaches – considered in Chapters 5 and 7 can be grouped together as 'decompositional' views on concepts/categorisation. Such a name is pertinent, because the common ground between such views consists in the intuitively compelling guiding assumption about the structural complexity of simple concepts resulting from composition of more basic parts. More precisely, by this assumption simple concepts are posited to have internal structures that can – at least potentially – be analysed into sub-components, and that are the main factor deciding about their individuation.

Conceptual atomism is an alternative view, distinct from all of those characterised above in claiming that concepts display no internal structure. On this view, it is concepts themselves that are the most primitive building blocks of cognitive meaning, with no possibilities for further decomposition. Although this stance – arguably a minority view in Cognitive Science – has several contemporary adherents (e.g. Ruth Millikan, 2000), the most vocal of them is arguably Jerry A. Fodor (especially 1998), whose formulation of conceptual atomism has come to serve as its paradigmatic construal.

Reasons for a special treatment of the Fodorian position start from its originality, the difficulties it has posed for the attempts at its accurate reconstruction (cf. Laurence and Margolis 2002), and the extremely influential status of its author. However, one motivation to address and convincingly refute it stands out prominently. As suggested by the subtitle of his 1998 classic (*Concepts: Where Cognitive Science Went Wrong*), Fodor's conceptual atomism, rather than being yet another view of concepts, constitutes a challenge for entire mainstream Cognitive Science in this respect. If one is prepared to follow Fodor to the full extent of his argument, and consequently, if one accepts that concepts are indivisible atoms of (propositional) thought, then no legitimate form of conceptual analysis can be exercised, and analytic philosophy and lexical semantics face the need of a radical reformulation of their fundamental goals (cf. Fodor 1998: 162–163; Fodor 2000: 350; and also Wierzbicka 1996: 257)[198].

That Fodor is not in fact correct seems to be uncontested (consider the converging overall verdicts of, among many others, Murat Aydede [1998], Kent Bach [2000], Eric Dietrich [2001], Alex Levine and Mark H. Bickhard [1999], as well as the authors of reviews in the March 2000 issue of *Mind and Language*[199]). Still, much less unanimity is achieved regarding the exact motivation of this dismissal, other than the extreme counterintuitiveness of some of his proposals. Focussing their comment on another aspect of Fodor's doctrine, namely extreme *nativism* (*innatism*)[200], Laurence and Margolis (2002: 26) write:

---

[198] This consideration may be largely irrelevant to non-mentalistic approaches to language and concepts (broadly characterised as E-linguistics in chapters 3 and 4), where the main commitment is to descriptive power. However, in the present work as well as all other cognitively oriented projects, in which psychological reality is a major desideratum, Fodor's argument must be taken seriously.

[199] Dietrich's (2001: 94) quip: "[Fodor] wouldn't be happy if anyone agreed with him" perfectly captures the general spirit of the debate, both with respect to its content and its form: the characteristically waggish style of Fodor's later writings is often adopted by his opponents.

[200] A position somewhat relaxed in Fodor 1998. See below.

> Not surprisingly, Fodor has had few supporters. Philosophers seem to have taken the conclusion to be so patently absurd that they think the argument behind it barely needs to be addressed…
>
> As will become clear, we think that these reactions are deeply problematic. Apart from anything else, responses like these have encouraged a superficial understanding of Fodor's argument. This is unfortunate since, in spite of the near universal rejection of its conclusion, the dialectic that Fodor's argument generates remains extremely influential.

Laurence and Margolis's observation, despite the abovementioned qualification that it pertains to the nativistic rather than atomistic aspect of Fodor's theory, readily transfers to the present context as well. This is due to the fact those aspects represent different facets of the same comprehensive and carefully constructed theory of cognitive meaning. Thus, the argument for radical concept nativism can be treated as the argument for conceptual atomism run one step further. Conceptual atomism, in turn, links very closely to informational semantics. These have been described as "natural allies" (Fodor 1998: 156), but in fact this relation generalises to most other building blocks of Fodor's theoretical edifice. Especially when commenting on elements that individually strike one as strongly counterintuitive, it is important to have before one's eyes the complete picture of mutually supporting and illuminating theses.

6.2. Jerry Fodor's theory of concepts

The roots of Jerry Fodor's general theoretical outlook regarding meaning in thought and language can be traced to two rudimentary but far-reaching commitments: to *naturalism* and to (propositional) *folk psychology*[201]. These can be supplemented by the observation about the systematic nature of human

[201] Cf. M. J. Cain's diagnosis in the introduction to his monograph on Jerry Fodor: "Fodor has two basic commitments: one is to folk psychology and the other is to physicalism" (Cain 2002: 1). Contrary to Cain, I opt for *naturalism* instead of the latter term. Rather than any changes in essence, it involves slightly different (less reductionistic) connotations.

thought processes. The Fodorian project can be seen almost in its entirety as advances in spelling out the consequences of those two guiding desiderata. Naturally, over the span of several decades some aspects of Fodor's philosophy have undergone certain modifications, and in what follows I focus on the more recently defended positions.

6.2.1. Naturalism

A tenet that is central to Jerry Fodor's theoretical programme, as it is to contemporary mainstream Western philosophy and science in general (see 1.3.3.), is that there exist no supernatural properties and reality is at bottom uniformly physical; hence, mind and meaning are purely natural phenomena whose properties are, at bottom, strictly physical properties. Such an approach requires that all relations and phenomena that are postulated in science must at least be made compatible with a physical explanation, actual or at least possible[202]. In particular, intentional/semantic relations must be suitable for an ultimate reformulation in a naturalistic vocabulary involving non-intentional and non-semantic phenomena and relations, such as causal or nomic relations[203]. Naturalism requires no separate substantiation, since it receives practically unanimous endorsement within Cognitive Science, counting among its most basic precepts.

Note that this is not equivalent to a strong reductionistic position that would prophet or even advocate the dissolution of all other sciences in physics. Fodor defends the legitimate status of special sciences in general and psychology

[202] "It's a methodological consequence of our conviction - contingent, no doubt, but inductively extremely well confirmed - that everything that the sciences talk about is physical. If that is so, then the properties that appear in scientific laws must be ones that it is possible for physical things to have, and there must be an intelligible story to tell about how physical things can have them." (Fodor 1994: 5)

[203] "I want a *naturalized* theory of meaning; a theory that articulates, in nonsemantic and nonintentional terms, sufficient conditions for one bit of the world to be about (to express, represent, or be true of) another bit." (Fodor 1993 [1987]: 98; italics in the original)

in particular (e.g. Fodor 1980 [1974]); according to him, the ultimate viability of strong reductionism, while possible, is not likely, much less necessary.

6.2.2. Folk psychology

Folk psychology has been concisely introduced in section 4.2.3.2., where it was quoted as a possible supporting argument for the reality of concepts. What is most relevant for the present purposes is that firstly, folk psychology is extremely successful, and secondly, its operations are routinely couched in propositional terms.

Although the spontaneity and efficiency of folk psychological explanations are universally familiar, Fodor's celebrated "lecture" example (1993 [1987]: 3–8) serves as a convenient illustration. From the telephone conversation in which Fodor is asked to lecture in a distant city and accepts the invitation, one can effortlessly and accurately predict Fodor's consequent arrival in that city on a given date. Such predictions concerning people's behaviour can be achieved exclusively by means of folk psychology and, despite the 'ceteris paribus' clause, they tend to be very reliable – especially when compared to predicting the course of arguably less complex natural phenomena, such as the weather. Fodor takes these considerations to provide evidence for the reality of folk psychology as a genuine mental process, whereas the majority of philosophers, notably Daniel Dennett (e.g. 1998: 81–94) and Patricia Churchland, remain much more sceptical about such an interpretation.

The other important point regarding folk psychology is that it exploits generalisations not unlike those functioning in regular sciences. In order for its sentences to generalise across contexts and people, its basic component parts – propositions and, in turn, their component parts, concepts – must remain 'public', i.e. invariant for different subjects. This bears the vital consequence that the identity of concepts (qua concept *types*) must be preserved across individuals, which motivates Publicity Constraint as a "non-negotiable" criterion for a theory

of concepts (Fodor 1998: 28–34; see section 4.2.6.): to Fodor, such a theory must account for different people sharing the same concepts.

6.2.3. Systematic nature of human thought (compositionality)

One additional corollary of the central status of folk psychology and, especially, of *propositions* serving as its vehicles, is that thought is implied to have at least certain logical properties. But the arguments for compositionality are not dependent on the above; rather they are founded on independent observation. In short, a system is compositional if it is productive in predictable ways. The pithiest definitions are probably those supplied in Fodor 2001 (p. 6)[204]:

> Both human thought and human language are, invariably, productive and systematic; and the only way they could be is by being compositional. (Productivity is the property that a system of representations has if it includes infinitely many syntactically and semantically distinct symbols. Systematicity is the property that a system of representations has (whether or not it is productive) if each of the symbols it contains occurs with the same semantic value as a constituent of many different hosts).

Productivity and *systematicity* constitute very strong intuitive criteria for concept possession, as it indeed seems legitimate to credit a cognitive agent with a possession of a given concept only if she can satisfy the Generality Constraint (i.e. be able to recombine a given concept with all other semantically and syntactically appropriate concepts in her repertoire[205]), and also apply the concept systematically (i.e. her being able to grasp the meaning of xRy, e.g. 'John loves Mary', is itself sufficient to guarantee that she is able to grasp the

[204] Towards the end of this text, Fodor apparently contradicts himself, claiming that "language is not compositional" (2001: 14). This should properly be read as an assertion that the compositionality characteristic of language is secondary and imperfect, and it is only thought that is primarily and perfectly compositional.

[205] As postulated by as Gareth Evans 1982; see also more extensive comments in section 4.2.2.)

meaning yRx, e.g. 'Mary loves John'). To Fodor, systematicity and productivity are fundamental properties of the human conceptual system (and, derivatively, of natural languages), and their successful explanation is a *sine qua non* criterion for a viable account thereof – hence the vehement attacks on connectionist architectures as models of the human mind (e.g. Fodor and Pylyshyn 1988), which are at best capable of approximating systematic/productive behaviour, but have so far been unable to implement fully general systematicity/productivity.

6.2.4. Consequences

Throughout his career, Jerry Fodor has remained dedicated to the core tenets outlined in 6.2.1., 6.2.2. and 6.2.3. The positions reviewed below are most fruitfully seen as the consequences stemming from the above foundations and, when Fodor's anti-relativistic leanings are also taken into consideration, are at least partly predictable from them. Below I present a very brief account of the key consequences that together form a framework into which conceptual atomism then fits in naturally. Conceptual atomism itself, the position central to the interests of the present work, receives a more detailed exposition in the next major section (6.2.5.).

6.2.4.1. Language of thought

Despite being introduced as early as in 1975, the idea – or the hypothesis of – the *language of thought* (LOT or *mentalese*; Fodor 1975) has continued to exert a powerful influence on Cognitive Science, with such prominent researchers as Steven Pinker (e.g. 1995: 55–82) considering it one of its mainsprings. In crudest terms, LOTH (language of thought hypothesis) stipulates that (propositional) "thought and thinking are done in a mental language, i.e., in a symbolic system physically realized in the brain of the relevant organisms" (Aydede 2004). By this hypothesis, LOT is a linguaform computational system with combinatorial syntax and compositional semantics that generates complex, sentential representations from simple wordlike representations (i.e. concepts), the

'sentences' being identical to complex physical brain states in particular organisms, the 'words' – to distinct component parts of those states. Whenever an organism tokens a propositional attitude with the content P, it does that by virtue of tokening an isomorphic sentence S in LOT, which happens by virtue of tokening the corresponding brain state (cf. Cain 2002: 55–56)[206].

All sources of motivation for LOT derive in one form or another from the logical-cum-linguistic properties exhibited by human thought. Firstly, other representational formats, such as images, are fundamentally inadequate for implementing linguistic meanings with their potential for compositionality, inference, etc. Note that Fodor does not question the reality of picturelike and other sensory representations, but he follows the later Wittgenstein (e.g. 1987 [1953]: 131–132)[207], and ultimately Frege, in demonstrating imagistic representations to lack context independence, stability, generality, and other features necessary for supporting propositional, publicly shareable meanings. Secondly, a discrete symbolic medium is required to implement full systematicity and productivity that, as has been shown in 6.2.3., Fodor takes to embody the most important attributes of human thought.

---

[206] Strictly speaking, either 'identical' or 'brain states', as well as 'organism' need to be qualified, since LOT is multiply realisable. In addition to the possibility of LOT being differently physically encoded for different people, this also means that LOT is at least in principle possible for non-human and/or artificial beings. Note, too Fodor's very strong realistic stance on the ontology of LOT: it is an actually existing, cerebrally realised system, not a mere descriptive heuristic.

[207] E.g. remark 449: "We do not realize that we *calculate*, operate, with words, and in the course of time turn them sometimes into one picture, sometimes into another.—It is as if one were to believe that a written order for a cow which someone is to hand over to me always had to be accompanied by an image of a cow, if the order was not to lose its meaning." (Italics in the original.)

Incidentally, and ironically, it is Wittgenstein's (1987 [1953]) argument for the primacy of society to rules (and to rule-governed systems such as languages), and hence against the possibility of private language, that most forcefully militates against LOTH.

The multiple realisability of LOT notwithstanding, in humans it is supposed to be a universal, innate piece of cognitive equipment. Its innateness allows Fodor to explain concept acquisition without resort to learning (see 6.2.4.3.).

6.2.4.2. Criticism of inferential role semantics (holism)

Much of Fodor's support for his favoured conception of semantics – understood in the sense of a method for individuating meanings – relies on a negative argument against the rival views, especially inferential role semantics. *Inferential role semantics*, or *conceptual role semantics*, [IRS/CRS] is a type of functional role semantics mentioned in 3.3. subsection c), where the adequate roles are restricted to inferential roles (see e.g. Block 1998).

From Fodor's perspective, the most powerful objection against inferential/conceptual role semantics is that they are seen by to be irreparably *holistic*. If, as stipulated by IRS, a concept is individuated by its inferential relations, then – since there can be indefinitely many inferences involving numerous other concepts – its individuation is necessarily dependent on other concepts. This would be acceptable if that dependence could be shown to be 'manageable', i.e. if one could reliably differentiate the important inferences (meaning-constitutive, analytic; for example from *is a dog* to *is an animal*) from the unimportant ones (contingent, synthetic; for example from *is a dog* to *wags its tail*). Unfortunately, since Willard van Orman Quine's (1961) extremely influential attack on the analytic–synthetic distinction, it has been almost universally agreed that no principled way of such a differentiation is attainable – a point that is reinforced by the general failure of the classical views on categorisation (see 5.2.3.2.). Thus, the individuation of a concept depends to some degree on all inferences (and hence all concepts)[208], if not immediately,

[208] "To put it in slightly other terms, it seemed to us [Fodor and Ernest Lepore – SW] likely that either translation is an *atomistic* relation, so that what translates an expression of L [a language – SW] is independent of what, if any, other expressions L contains; or translation is a *holistic*

then at least indirectly, by depending on some concepts that in turn depend on other concepts, etc. IRS's producing holistic consequences is inescapable (cf. Cain 2002: 125–127).

The reason why Fodor finds holism objectionable follows directly from his commitment to folk psychology (6.2.2.), and more precisely, to folk psychology's dependence of publicly shareable, stable conceptual contents. If concept individuation is holistic, then given even the slightest interpersonal differences, concepts are non-identical (non-*type*-identical, *token* non-identity being a truism) from person to person, and folk-psychological generalisations cannot be salvaged. "…[I]f the individuation of concepts is literally relativized to whole belief systems, then no two people, and no two time slices of a given person, are ever subsumed by the same intentional generalizations, and the prospects for robust theories in intentional psychology are negligible." (Fodor 1998: 114)

6.2.4.3. Nativism and modularity

Fodor's nativism regarding concepts (radical concepts nativism) is mostly dependent on the argument for informational semantics and consequently, conceptual atomism. It should be noted that in his 1998 book, Fodor has to some extent backtracked on his rather extreme position. Still, contrary to Cain's (2002: 73–80) opinion, it seems that his stance continues to qualify as rather nativistic, if to a somewhat lesser degree[209]. A more extensive treatment of this issue is offered in the next section.

---

relation, so that what translates an expression of L depends on *all* the other expressions L contains. We saw no stable middle ground short of wholesale appeals to the analytic/synthetic distinction, which, following Quine, we took to be a Very Frail Reed." (Fodor 1994: 74; italics in the original)

[209] "Likewise, it used to seem to me that atomism about concepts means that DOORKNOB is innate. But now I think that you can trade a certain amount of innateness for a certain amount of mind-dependence. *Being a doorknob* is just: striking our kinds of minds the way that doorknobs do. So, what you need to acquire the concept DOORKNOB <from experience> is just: the kind

There is also a nativistic ring to Fodor's prominent theory of the modularity of mind (Fodor 1983, 2000; see also 1.4.6.1.), in that the developmental paths of the modules are supposed to be relatively inflexible. This topic is relatively less relevant for the present considerations, but two remarks can be of interest. Firstly, contrary to the popular reception of this theory, only peripheral systems, i.e. the processing systems for the sensory modalities and language, have modular properties, while the central system for conducting propositional operations, such as the fixation of beliefs, does not; on the contrary, it is informationally unencapsulated and domain-general. This means that 'central' cognition is non-modular. Secondly, this non-modular character of central cognition is what renders it hopeless for any productive cognitive-scientific study, and this is the case precisely because of the problem of holism (Fodor 1983: 101–129)[210].

6.2.5. Fodor's conceptual atomism and informational semantic

The two elements of Jerry Fodor's rather impressive theoretical construction that are of the most immediate interest from the perspective of the present work are *informational semantics* and *conceptual atomism*, also jointly referred to as *informational atomism* (IA) due to their interdependence. The most comprehensive treatment of those positions was given in Fodor's 1998 monograph on concepts, *Concepts: Where Cognitive Science Went Wrong*. Accordingly, in my review and the subsequent critical discussion of this

---

of mind that experience causes to be struck that way by doorknobs." (Fodor 1998: 162; italics in the original) Note that such a rephrasing does not envisage the process of concept acquisition as substantially more environmentally driven, but still presents it as rigidly dependent on the presumably innate cognitive equipment. See 6.2.5.

[210] More accurately, on Fodor's account central cognition is *doubly* holistic, being both isotropic and Quinean. This means that the level of acceptance of any belief may be sensitive, respectively, "to the level of acceptance of any other belief in the system", and "to the global properties of the system's beliefs taken collectively" (Fodor 1983: 107–108).

philosopher's defence of informational semantics and conceptual atomism, I will rely mostly on the arguments provided in this book.

The first theoretical element, conceptual atomism, can be defined as follows:

> *Conceptual atomism: concepts*[211] *are atoms, i.e. indivisible units with no internal structure.*

Thus, each of the concepts in one's conceptual repertoire constitutes a self-contained primitive that cannot be further analysed into component parts. Furthermore, relations between concepts do not play any role in content individuation: what relations a given concept may have to any other concepts is entirely immaterial to establishing its content. Similarly, the acquisition of each individual concept proceeds in isolation from the rest of the conceptual system, and the possession of any given concept is likewise independent of the possession of any other(s). On the above account it is (theoretically) possible to possess only one concept, or, for example, to possess the concept FOREST without simultaneously possessing the concept TREE (Fodor's own vivid examples include BACHELOR and UNMARRIED [1998: 14], and TUESDAY and WEDNESDAY [1998: 74]).

The unorthodox view that concepts cannot be individuated by means of interconceptual relations compels its proponents to identify an alternative method of successful content individuation. According to Fodor, such a method, i.e. a *semantics*, is informational:

---

[211] For the sake of accuracy it should be reemphasised that here, as elsewhere in the text, "concepts" are meant in their default understanding – in line with 3.2.3. c) and 4.1. – i.e. as *simple* concepts. The considerations below, quite obviously, do not apply to complex concepts. To a first approximation, conceptual atomism claims that what is morphemically unstructured at the level of language must also be unstructured at the level of thought (although the reverse need not hold, cf. Fodor's standard example of DOORKNOB).

Informational semantics: contents are individuated by the nomic relations that link concepts to appropriate elements of the mind-external world.

This is to say that the instances of a certain class of entities in the world (e.g. dogs) reliably cause the tokenings of a given concept (e.g. DOG), and the stable and lawlike (nomic) character of this relation of tokening is meaning-constitutive for this concept[212]. Dogs ("instances of doghood", to use Fodor's wording, who himself borrows it from James J. Gibson) causally co-vary with the tokenings of the concept DOG, and the variety of possible causal chains from dogs to DOG reflects the variety of the means of *semantic access* (Fodor 1998: 75–80). The basic mode of semantic access is perceptual – when a person simply sees/hears/touches a dog – but may take much more complex forms for more abstract concepts or in non-standard cases. Fodor (1998: 29, 76, 79) repeatedly cites the example of the deaf-blind American author Helen Keller, whose disabilities did not prevent her from the acquisition of English, testifying to her grasp of the underlying concepts.

Conceptual atomism and informational semantics, mutually supportive but considerably less intuitively appealing and consequently less readily acceptable than the available rival views, require an elaboration. Fodor's argumentation builds upon what Alexander Levine and Mark Bickhard (1999) call a "what else?" argument: the systematic and comprehensive criticism of the alternative views, in the light of which his preferred position emerges as the only viable one.

---

[212] "[T]he fact that DOG means *dog* (and hence the fact that <dog> does) is constituted by a nomic connection between two properties of dogs; viz. *being dogs* and *being causes of actual and possible DOG tokenings in us*." (Fodor 1998: 73, italics in the original)

"<[D]og> and DOG mean *dog* because <dog> expresses DOG, and DOG tokens fall under a law according to which they reliably are (or would be) among the effects of instantiated *doghood*." (Fodor 1998: 75, italics in the original)

Fodor divides inferential role semantic (IRS) accounts of concepts into two main groups – definitional and similarity-based ("statistical", as he labels them) – and addresses them with counterarguments of both empirical and conceptual nature.

6.2.5.1. Criticism of definitional accounts

Definitional accounts of conceptual structure are such that assume most concepts to be composed of lists of features (which are usually taken to be other concepts) in a way that the content of the given concept is fully inherited from the contents of its components; such a view is equivalent to the Classical View characterised in section 5.2.1. From this standpoint, concepts *are* definitions (strictly speaking, they are *definienda* or structural descriptions; Fodor 1998: 41, footnote 1), or feature bundles. Note that their being feature bundles is a stronger claim than the possibility of predicating features of concepts (Fodor 1998: 63), since it requires that concepts be composed of those features in a way deciding about their identity. Another vital hedge is that such a description pertains to most, but not *all* concepts, because there must be at least some primitive concepts that avoid the problem of circularity.

Definitional accounts are a species of IRS, since the structural description that is supposed to constitute a given concept can be readily recast in terms of inferences: for example, the fact that the concept BACHELOR comprises the concept MALE trivially licences the inference that if someone is a bachelor, he is a male. Nevertheless, *bona fide* definitional accounts are immune from Fodor's two foremost criticisms against IRS, that is the problem of holism/publicity and the failure to support compositionality (Fodor 1998: 44). As to the first, the identity of a given concept is decided exclusively by the contents of its definition, which is finite and exactly stated, and no holistic sensitivity to the rest of the conceptual system arises; furthermore, different people can have identical concepts (concept types) in virtue of the sameness of the definitions of these concepts. As to compositionality, definitions readily compose by contributing all of their contents to the content of the complex expression. In short, with respect

to holism and compositionality, definitions behave exactly like complex concepts. BACHELOR remains the same from person to person, because it is constituted only by UNMARRIED and MALE, independently of any idiosyncratic and contingent beliefs about bachelors that particular people might entertain; also, it contributes all of its contents to complex mental representations such as HUNGRY BACHELOR which then has the same content as HUNGRY UNMARRIED MAN.

Accordingly, the objection that Fodor raises against the definitional accounts of concepts is ultimately of empirical nature. Firstly, Fodor points out that it is only 'true' definitions, i.e. ones that fully reduce the given concept to its component parts, that qualify as safe from his reservations against IRS. In order to avoid the discussed problems with publicity and compositionality (as well as with concept acquisition), the structural descriptions proposed as definitions must indeed achieve full synonymy with the corresponding concepts (Fodor 1998: 48–49). Fodor's summary denies the existence of such *bona fide* definitions on the ground that empirical research within Cognitive Science has repeatedly failed to identify any incontrovertible cases.

The core of Fodor's argumentation developed over two chapters need not be reviewed here, considering that his central conclusion about the status of definitional accounts is uncontroversial and was independently reached and substantiated in the previous chapter of my work. However, Fodor proceeds to lay a much stronger claim, one regarding the impossibility of even partial (non-complete) decomposition of concepts[213]. Being indivisible atoms, concepts lack

[213] Such a claim is tangential to the conclusion just reached: from the point of view of publicity and compositionality, it suffices that there exist no genuine definitions of the sort described above. Still, admitting the possibility that (at least some) concepts might (at least in part) be constituted of some identifiable components would undermine both conceptual atomism and informational semantics: the former because concepts could no longer be considered indivisible, and the latter because the mind-world nomic relation would cease to be the only factor relevant to establishing contents.

not only well defined internal structures, but any structure whatsoever. This, however, clashes with commonsensical data, such as the intuition of the polysemy of 'keep' in "Susan kept the money" and "Sam kept the crowd happy" (Fodor 1998: 49–50; examples originally from Jackendoff 1992)[214]. The meaning of the word 'keep', and hence of the concept KEEP, seems to be partly identical and partly different across the sentences, an explanation that questions the indivisibility of 'keep'/KEEP and assumes structural complexity. Fodor insists that, since no more primitive level of semantic analysis exists than that of concepts themselves, 'keep'/KEEP must be univocal, and the only relation that the word 'kept' has in common in both quoted sentences is simply the relation of *keeping* (Fodor 1998: 55).

6.2.5.2. Criticism of similarity-based accounts

Fodor's argument against the similarity-based accounts of concepts is a mirror image of his critique of definitional accounts. In his analysis, Fodor focuses heavily on prototype accounts, which he takes to be representative of the whole class in all relevant respects, and which are discussed in the seventh chapter of my work. In this case, empirical evidence for the reality of prototypes – *sensu* typicality effects – cannot be questioned, a fact that Fodor acknowledges (1998: 93) and, as will be explained, whose consequences he is prepared to take very seriously. Nonetheless, in Fodor's opinion, probabilistic theories are critically vulnerable to his chief conceptual objection, that is the problem of compositionality. According to this researcher (Fodor 1998: 100), "the status of the statistical theory of concepts turns, practically entirely, on this issue".

The attack against the possibility of prototypes being concept-constitutive is launched in Fodor 1998 along the lines familiar from his earlier work (e.g. Fodor 1995: 14–19, Fodor and Lepore 1996). The key reservation, termed here

[214] Following Gilbert Ryle's analysis quoted in section 3.2.2., one could compose a zeugmatic „Susan kept both the crown happy and her money", diagnosing a category mistake that would indicate that each example of 'keep' belongs to a different category.

'the Standard Objection', is succinctly captured in the two initial sentences of Fodor and Lepore's text (1996: 253–254): "[t]here is a standard objection to the idea that concepts might be prototypes (or exemplars, or stereotypes): because they are productive, concepts must be compositional. Prototypes aren't compositional, so concepts can't be prototypes".

As was illustrated in 6.2.3., compositionality is a basic fact about human language and thought. The truistic observation that people's understanding and production of meaningful sentences is not limited to the set of previously encountered examples, but is readily extended to an infinite number of novel ones cannot be explained without the recourse to compositionality, i.e. "the derivation of the content of a complex concept *just* from its structure and the content of its constituents" (Fodor 1998: 104). Even though the precise way in which the content of a complex concept is derived may at times be rather intricate (cf. e.g. Jackendoff's [2002: 378–394] distinction into "simple" and "enriched" composition), the status of compositionality remains beyond doubt.

The problem of prototypes with compositionality is twofold. Firstly, indefinitely many complex, i.e. composed, concepts (e.g. "GRANDMOTHERS MOST OF WHOSE GRANDCHILDREN ARE MARRIED TO DENTISTS" [Jerry Fodor's example cited in Laurence and Margolis 1999: 36], CARS THAT HAVE BEEN SCRATCHED BY CHINESE WOMEN, etc.) lack prototypes. Secondly, and more importantly, for those complex concepts that do have prototypes, their prototypes do not seem to be constructed from the prototypes of their constituents. PET FISH serves as Fodor's main example: the prototype of a pet fish (i.e. the goldfish) is totally unpredictable from the prototypes of pets and of fish, which presumably are respectively dogs and cats, and e.g. carp or mackerel, depending on the specific socio-cultural context. Fodor (1998: 107) emphasises that PET FISH/'pet fish' is *not* idiomatic, but is formed compositionally, as attested by the inferences from *is a pet fish* to *is a pet* and *is a fish;* this can be compared with the idiomaticity of 'hot dog' and resulting fallible inferences from *is a hot dog* to *is hot* and *is a dog*. To support

compositionality, prototype theories would have to propose a reliable mechanism specifying, for any given complex concept, exactly *which features* and exactly *to what degree* are contributed by each of the simple concepts. Fodor is satisfied that no principled account of this sort is attainable.

The above simple but still powerful argument does not, however, exhaust the relation Fodor's theory bears to the topic of prototypes. The existence of typicality effects such as prototype enhancement, decreased reaction times, increased rates of spontaneous generation, etc.[215] constitutes a conclusive proof that the formation of prototypes is a pervasive trait of the organisation of human knowledge. As has already been remarked, Fodor does not attempt to question this finding, and endorses the existence of prototypes. The resolution of this *prima facie* contradiction is instrumental in shedding light on Fodor's theory in a more general context, helping one to avoid a simplistic treatment of this philosopher's position.

Central to the understanding of Fodor's theory of concepts is the fact that – in virtue of its being an informational theory, as opposed to inferential role theories – it allows one to make distinctions, on one hand, between *concept identity* and *content identity*, and on the other, between *concepts* and the *epistemic factors* related to them; distinctions that are not fully available to the proponents of IRS. This rather crucial insight tends to be easily overlooked when one intuitively accepts the standard perspective of the more 'natural' inferential role theories. On the different versions of IRS[216]:

- *concepts* are individuated by their contents,
- *contents* are individuated by the inferences they license (*is a bachelor* licenses *is unmarried* in virtue of the internal structure of the concept BACHELOR which has UNMARRIED as one of its constituents);

[215] Phenomena discussed in the section 5.2.3.3.

[216] E.g. Block 1998.

– *inferences*, in turn, are epistemic, which, put simplistically, means that content individuation depends on one's *knowledge*; in short, one can possess the concept BACHELOR only on condition that one knows that bachelors are unmarried males.

On the view sanctioned by Fodor, however,

– *concepts* are individuated syntactically (as symbols of the language of thought/LOT),
– *contents* are individuated by their nomic mind-world relations (lawlike causal co-variation of BACHELOR tokenings and bachelors)[217], and
– any *inferences* one may be able to draw or any *knowledge* one may happen to possess with regard to their concepts – are entirely irrelevant to the individuation of either.

In other words, on Fodor's atomistic-informational account, *no epistemic factors are content-constitutive*. This, however, should not be taken to mean that epistemic factors have been totally eradicated. As observed by Laurence and Margolis (1999: 65): "[l]ike any other theorist, the atomist holds that people associate a considerable amount of information with any concept they possess. The only difference is that whereas other theorists say that much of the information is collateral (and that only a small part is constitutive of the concept itself), atomists say that all of it is collateral."

[217] "Given my view that content is information, I can't, as we've just seen, afford to agree that the content of the concept H_2O is different from the content of the concept WATER. *But I am entirely prepared to agree that they are different concepts*. In effect, I'm assuming that coreferential representations are *ipso facto* synonyms and conceding that, since they are, *content* individuation can't be all that there is to *concept* individuation." (Fodor 1998: 15; italics in the original, underline mine [SW])

Accordingly, since the nature of prototype structures and prototype effects is epistemic – e.g. what prototype one has for BIRD depends on what one explicitly or implicitly *knows* about birds – it does not affect content individuation. Prototypes, neither being concepts nor entering their structures, are merely associated with concepts in a way tangential to the individuation of the latter's contents. In much the same way, inference is still possible on an atomistic account: it suffices that the inferences that are drawn in virtue of any information associated with a given concept are collateral and not content-constitutive. From Fodor's standpoint, that red is a colour does not contribute to the meaning of RED (nor 'red'); nevertheless, a reliable inference from *is red* to *is a colour* is still operational (in Fodor's [1998: 110] words, "RED entails COLOUR"). Still, such an elimination of epistemic factors from the account of concepts has important ramifications that make it an open target for criticism, a consequence that will be developed further in the discussion.

6.3. Criticism of Fodor's conceptual atomism

Numerous lines of criticism have been advanced against Fodor's view of concepts, with scant agreement on the precise reasons of either the failure or at least inadequacy of his proposals. In my opinion, most of this criticism is misplaced: the arguments, while essentially correct, fall short of addressing the actual view of that author in its full sophistication, and so fall short of questioning it. In the present section, I shall argue that the most critical objection actually turns out to be the most trivial one, namely, that when given the appropriate exegesis, *Fodor's theory does not qualify as a legitimate theory of concepts*. The underlying reasons, however, appear much less trivial, and can be traced down to a fault regarding one of his "non-negotiable criteria of a theory of concepts" (1998: 23–34), namely, the requirement of "publicity".

6.3.1. Radical concept nativism

Among the most far-reaching consequences of Fodor's contention that concepts are atomic are those for the theory of concept acquisition. The natural account of concept acquisition available to the various versions of IRS involves the process of *learning* in the form of compiling the target concept from its more basic constituents. On pain of infinite regress, there must also exist primitive concepts, which, lacking any more basic elements to be compiled from, cannot be learned and must therefore be innate. Still, their number can be assumed to be relatively small. In contrast, conceptual atomism, by its commitment to the thesis that all (simple) concepts are structureless, must reject learning altogether, with the corollary that all (simple) concepts are primitive and thus innate.

What is controversial, therefore, is not the very postulate of the existence of innate concepts itself; rather, it is the scale of this nativism that has provoked severe criticism. To remain consistent, a conceptual atomist seems to be committed to the innateness of e.g. DEMOCRACY, CARBURETTOR, as well as indefinitely many other concepts, including those for *future* cultural and technological inventions. As noted by Laurence and Margolis (2002: 26; cf. the quote in 6.1.) such *radical concept nativism* appears to be preposterous and its extreme intuitive implausibility alone seems to merit its rejection.

It should be observed, however, that the above objection exploits our IRS-informed presupposition that concepts are *constituted* by at least some of the knowledge related to them, and the implausibility actually regards the amount of assumed innate *knowledge*. The distinctions available to Fodor make the situation more complex, allowing him to claim that while concepts are indeed (in a way) 'innate', their contents are not, and neither is their proprietary 'epistemology'. This equals a departure from radical concept nativism for a more moderate position, one on which our strongly antinativist intuitions have no bearing, since – although remaining distinctly nativistic – Fodor's present stance does not hold concepts to be innate in the *radical* sense. This is so because, on the present proposal, concepts are neither innate nor learned, but are activated by our minds 'locking to' particular properties – a relation that cuts across the traditional

dichotomy, dividing the labour between 'the innate' and 'the learned' in a nonstandard way[218]. What is innate ultimately turns out *not* to be concepts *per se*[219]. Rather, it turns out to be a sort of species-specific system of category formation.

Central to the understanding of the issue of concept nativism is the doorknob/DOORKNOB problem (d/D problem; DOORKNOB being Fodor's standard example of a concept), phrased by Fodor in the following way: "why is it so often experiences of doorknobs, and so rarely experience of whipped cream and giraffes, that lead one to lock to *doorknobhood*?" (Fodor 1998: 127) Despite its trivial appearance (and whimsical statement), the d/D problem is real, but its treatment by Fodor is obscure and calls for a careful explication. The d/D problem should best be understood by transforming it into the related question of 'why do humans categorise in the way in which they do, and not in any of the countless other logically possible ways?'

Firstly, note that the d/D problem does not have an explicit linguistic component, i.e. it is not the d/'d' or D/'d' problem. Thus, it is not the banal problem of why it is the word 'doorknob' that, in English, happens to be the name for doorknobs or for the corresponding concept; as is well known, in any natural language this is decided by arbitrary convention. Rather, it is a problem of why, given that there are infinitely many ways of sorting things into categories, humans use precisely this and not other small subset of those possible ways.

---

[218] „The natural, appalled, reaction to radical concept nativism is: <But how *could* you have a concept like DOORKNOB innately?> To which the proper answer is: <That depends a lot on what it is to have a concept.> According to the present proposal, to have a concept is to be locked to the corresponding property." (Fodor 1998: 141; italics in the original, underline mine – SW)

[219] „What has to be innately given to get us locked to *doorknobhood* is whatever mechanisms are required for doorknobs to come to strike us as such… [T]he kind of nativism about DOORKNOB that an informational atomist has to put up with is perhaps not one of *concepts* but of *mechanisms*." (Fodor 1998: 142; italics in the original)

At this point, it may help to consider a quote from Jorge Luis Borges as commented on by Eleanor Rosch (1988a: 312):

> The following is a taxonomy of the animal kingdom. It has been attributed to an ancient Chinese encyclopedia entitled the Celestial Emporium of Benevolent Knowledge:
>
> > "On those remote pages it is written that animals are divided into (a) those that belong to the Emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (i) those that tremble as if they were mad, j) innumerable ones, (k) those drawn with a very fine camel's hair brush, (l) others, (m) those that have just broken a flower vase, (n) those that resemble flies from a distance (Borges, 1966: 108)."
>
> Conceptually, the most interesting aspect of this classification system is that it does not exist. Certain types of categorizations may appear in the imagination of poets, but they are never found in the practical or linguistic classes of organisms or of man-made objects used by any of the cultures of the world.

In other words, human categories, understood as the extensions of concepts, strike us as homogenous. Although humans could in principle classify things-presently-categorised-as-doorknobs and things-presently-categorised-as-giraffes together under a single concept corresponding to a single lexeme (e.g. FLURG, 'flurg') – in practice we do not. Such a pattern of categorisation, absolutely arbitrary to human beings, could in principle be entirely natural to a hypothetical but logically possible non-human mind. Conversely, it is only for human minds that doorknobs can come to form a coherent category. The fact that it is possible for humans to naturally find all *these things and not others* to be members of a single, uniform category is *a fact about the innate, species specific makeup of our minds*. Human minds, in virtue of being human minds, license certain patterns of categorisation while proscribing indefinitely many others.

It needs to be underscored that humans have an innate *spectrum* of viable categorisation patterns rather than just one, rigid, universal pattern. As is well known, 'the same' experiential world can be partitioned into linguistic classes in distinct ways in different cultures. Compare the English 'doorknob' (similarly, DOORKNOB) and the Polish 'klamka' (KLAMKA), which covers both 'doorknob' and '(door)handle'. As is evidenced by their functioning as lexemes, both those categorisation patterns are valid possibilities, latent in the innate equipment of all human beings. Again, this is opposed by 'arbitrary' categories. To give an extravagant example, a category could be formed composed of metal objects, trapezoids, inequities, and some (but not all) trips; it is a logically possible category that is nevertheless unnatural to human minds and consequently never used and never lexicalised.

To sum up, on Fodor's account, the innate concepts should be construed not as concepts in the usual meaning, but rather as *'the humanly accessible ways to divide the reality into coherent classes'*.

6.3.1.1. Role of ontology

The crux of Fodor's argument lies in transferring a large part of the theoretical burden from the problem of concept acquisition to problems of ontology. Fodor's ontological claims exactly complement the innatist claims, being their mirror image. Such a strategy also allows Fodor to remain consistent with his realistic leanings.

Despite sounding intricate, Fodor's claims originate at bottom in a relatively basic Kantian constructivist insight about the inherently two-factor nature of cognition[220], in which the contributions of the subject are inseparable from the contributions of the object – if the two can be distinguished, it is only possible by means of abstract conceptual analysis, but not in practice. Similarly, on Fodor's account reality exists not as pre-structured into ready-made entities

[220] Briefly mentioned in the discussion of categories in section 3.2.2.

and classes of entities, but the structure in reality is only found relative to human, or humanlike, minds. Note that the metaphor is neither that of the mind *discovering* preexisting structure (naïve realism), nor that of the mind *imposing* structure on originally unstructured reality (antirealism). Doorknobs really do exist (ontological realism), but can be perceived and categorised *as doorknobs* only by human, or humanlike, minds – in virtue of their innate cognitive equipment, which enables human minds, but not other hypothetical minds, to cognise properties such as *doorknobhood* and form concepts such as DOORKNOB[221].

The situation is further complicated by the fact that in addition to mind-dependent properties, Fodor distinguishes natural kind concepts (based on natural kind properties). This distinction, however, is made on unclear grounds and is deeply problematic. For example, water *sensu* H_2O is classified as a natural kind, but DOG is not (cf. Fodor 1998: 150–151; 137 [footnote 11]) – a decision that might strike as somewhat arbitrary. In any case, natural kind concepts form a clear minority of concepts in the human repertoire and are not in any sense crucial to the present analysis. Furthermore, the general idea of natural kinds (as having a special status secured by the authority of the modern natural sciences) is itself highly questionable, as was shown in 2.2.3.

6.3.2. Problem of elimination of epistemic factors

The elaborate and philosophically refined construction of Fodor's argument commands respect and merits a careful analysis. This scholar's account might

[221] "The basic idea is that what makes something a doorknob is just: the kind of thing from experience with which our kind of mind readily acquires the concept DOORKNOB. And, conversely, what makes something the concept DOORKNOB is just: expressing the property that our kinds of minds lock to from good examples of instantiated *doorknobhood*." (Fodor 1998: 137; italics in the original)

"*Being a doorknob* is just: striking our kinds of minds the way that doorknobs do. So, what you need to acquire the concept DOORKNOB <from experience> is just: the kind of mind that experience causes to be struck that way by doorknobs." (Fodor 1998: 162; italics in the original)

turn out to be defensible and valuable on philosophical grounds. However, more doubts arise as to whether it remains similarly valuable from the cognitivist perspective. The gravest problem seems to be the simple fact that Fodor proposes basically no explanation for the phenomena of central interest to Cognitive Science – a criticism very close to those already raised against the classical view in sections 5.2.3.3. and 5.2.3.4. While Fodor has never claimed to have such ambitions, such a slant of his arguments results directly from other underlying assumptions whose importance can be shown to be overrated.

As it has already been discussed, for Fodor the epistemic factors (one's 'knowledge') should be excluded from a theory of concepts. This is possible because he proposes that firstly, concepts are not individuated based solely on their contents, but chiefly based on their nonsemantic properties, i.e. physical properties as symbols of LOT. Secondly, unlike in mainstream Cognitive Science, concepts are not 'capsules of knowledge'. The contents of concepts are not constituted by knowledge associated to concepts, but rather from nomic mind-world relations (lawlike causal co-variation of things in the world and the tokenings of the concepts they fall under).

In Fodor's view, the adoption of such a semantics is made necessary by the threat of *holism* (6.2.4.2.). That author perceives meaning holism not simply as a phenomenon to be explained, but as a severe problem to any functional role semantics – and especially inferential/conceptual role semantics (IRS/CRS), where contents are determined by the available inferences, which in turn depend on the knowledge associated with concepts. Since it is in principle impossible to show which inferences count as meaning-constitutive and which are merely subsidiary, meaning (content) depends to some degree on all possible inferences. To Fodor this is unacceptable, for the reason that the identity of any single content is distributed over the whole conceptual network and thus impossible to establish conclusively.

A naturally convincing reply is that there exist numerous examples of other systems whose elements are highly interdependent in the relevant respects

but nevertheless preserve their identities and are easily singled out from the rich web of interrelations because most of those interrelations can safely be ignored. Consider for example the distribution of gravitational forces in the universe, which is in principle influenced by all existing bodies with mass; however, one can neglect almost all of the influences and effortlessly identify the proprietary gravitational fields of individual celestial bodies (essentially the same point is made by Keil and Wilson [2000: 314][222]). To Fodor, however, such a solution cannot be made to work, since it is too 'loose': with the phenomenon of holism taken into account, the identification of contents in IRS – and thus, assuming IRS, of concepts themselves – is always only *approximate*. Consequently, any two contents can only be determined to be *almost* identical (very similar), whereas Fodor insist that a successful theory of concepts must spell out strict criteria for two contents to be *exactly* the same; to him, the notion of content-similarity is secondary to and dependent on the notion of content-identity (e.g. Fodor 1998: 30)[223].

---

[222] "All of the moons of our solar system are influenced in their orbits by all other masses in our solar system; but each planet and its moons form a coherent system distinct from any other one, constituting a system that can be almost completely understood at that level of analysis. Explanatory beliefs are not distributed evenly in the web of understanding. They form tight, richly structured clusters that then have sparse links to other clusters. Beliefs about the mechanics of solid objects, for example, are richly structured and tightly interconnected, but their connections to the cluster of beliefs about minds are comparatively very few. (It is, admittedly, awfully hard to know how to count, but by any metric that is devised, the difference would be huge.)" (Keil and Wilson 2000: 314)

[223] "In fact, however, the idea that content similarity is the basic notion in intentional explanation is affirmed a lot more widely than it's explained… On one hand, such a notion must be robust in the sense that it preserves intentional explanations pretty generally; on the other hand, it must do so *without itself presupposing a robust notion of content identity*. To the best of my knowledge, it's true *without exception* that all the construals of concept similarity that have thus far been put on offer egregiously fail the second condition". (Fodor 1998: 30; italics in the original)

Still, the rationale for such a strict demand can be easily questioned. As has been remarked several times: Fodor's original motivation in this respect can be traced to a specific and overly restrictive condition of the publicity/shareability of concepts, motivated in turn by the efficiency of human communication and folk-psychological reasoning. In section 4.2.6.1. I proposed a different understanding of the public character of concepts, as logically secondary to and ontogenetically derived from truly public words of a natural language: it is the lexical labels of a natural language that are the only type of entity that is truly public, being type-identical for all subjects. With this solution accepted, the stabilising invariant dimension to human concepts is seen to reside in language (*qua* E-language) and no longer has to depend on the eccentric requirement of the (type-)identity of everyone's mental representations. This allows for the reintroduction of some natural degree of variety between 'the same' concepts in different people, and so neutralises Fodor's main objection against functionally based accounts of meaning.

To recapitulate, Fodor's position is undeniably original and sophisticated. By rethinking and modifying his former theses, Fodor actually manages to rescue his theory from the most frequent accusation, namely that of radical concept nativism that would lead to absurd and unacceptable consequences. At the same time, Fodor's position becomes effectively a philosophical account, of only limited relevance to the concerns of Cognitive Science. The cognitivist criteria of adequacy of a theory of categorisation are essentially epistemic, i.e. how *knowledge* associated with concepts is acquired and put to use, e.g. in abductive inference (this point has already been heavily emphasised, especially in section 5.2.3.3.). As is rightly observed by Eric Dietrich (2001), even if one accepts Fodor's policy of divorcing epistemic factors from the theory of concepts, one still has to complement the theory of concepts with a successful theory of those 'epistemic factors', which then inherits nearly all of the problems interesting to a cognitive scientist.

6.4. Recapitulation

The sixth, and penultimate, chapter of this book, treated the atomistic stance on concepts/categorisation, as expounded by its leading proponent, Jerry Fodor. The motivation for addressing Fodor's claims in considerable detail was, beside this philosopher influential status, the challenge that it created for the basic tenets of mainstream Cognitive Science. I discussed Fodor's position holistically, in the sense of tracing and reconstructing the interdependence of his views regarding particular problems such as naturalism, nativism, or language of thought. After a reconstruction of this philosopher's intricate theoretical stance and the correction of possible misinterpretations, there followed a critical examination.

Although I defended Fodor from the most common line of criticism – that of radical concept nativism – showing his position to be philosophically well-founded, I pointed out a different problem with his view on concepts. The rejection of functional accounts of meaning such as IRS, which are so central to Cognitive Science, turns out to result from a peculiar choice of his theoretical priorities – specifically, an overly restrictive commitment to the 'public' nature of concepts.

I hope to have demonstrated that firstly, such a commitment is ill-founded, and secondly, that it gives rise to the main weakness of Fodor's account, namely, that the attempt to solve this mostly philosophical problem is made at the cost of the neglect of the cognitively interesting issues related to concepts.

7. From prototype to exemplar models in nonlexical and lexical categorisation

7.1. Preliminary remarks

In the seventh, and final, chapter, I conclude my work by considering the possibilities of application of the exemplar view to the issue of lexical-semantic categorisation. Having established the inadequacy of the major competing approaches – the classical view in Chapter 5 and Fodor's atomism in Chapter 6 – I turn to their similarity-based alternatives, that is the prototype and exemplar views. As a necessary starting point, I briefly examine the relevant theoretical foundation, that is the very notion of similarity in which these approaches are grounded. My specific aim in this chapter is to demonstrate why the exemplar approach can be legitimately thought of as qualitatively different from other approaches under consideration, and secondly, to provide a strong footing for the claim – hopefully to be borne out by future interdisciplinary empirical research – that the exemplar approach shows promise for overcoming the diagnosed shortcomings of the rival views. Accordingly, the convention of this closing part of my work is largely that of research postulates and sketching out further prospects.

As should already be clear, this concluding chapter is not intended as a comprehensive analysis or even detailed presentation of the similarity-based accounts of categorisation. It should be kept in mind that especially the general topic of prototypes constitutes a diversified subject area, extending beyond the topic of concepts and important to many fields of Cognitive Science[224]. In particular, the prototype theory of categorisation, broadly construed, can be taken

[224] E.g. low-level perception (see Dépy et. al. 1997).

to constitute the paradigm approach adopted in cognitive linguistics (though arguably not in Cognitive Science at large).

Because of the abundance and popularity of existing theoretical proposals built around the notion of prototypes (e.g. Taylor 1995, Lakoff 1990) [225], I do not discuss this topic in detail – only to the extent to which it is necessary for making the distinction from the closely related exemplar view. Note, too, that some versions of prototype theory developed in cognitive linguistics differ in important points from the basic description stated below – e.g. on George Lakoff's extended version of prototype theory the prototype is no longer the organising element of the category and, crucially, similarity to prototype does not govern category membership[226]. Therefore, the focus of the following sections is on the general assumptions of the exemplar approach, the reasons for its limited popularity, the similarities to as well as the differences from the prototype view, the possible advantages, and the implications it carries for lexical categorisation and for Cognitive Science in general.

7.2. Similarity as theoretical notion

The prototype and exemplar approaches are frequently grouped together under the rubric of *similarity-based* approaches (e.g. Hampton 1998, Smith et al. 1998)[227]. In concord with this denomination, both rest on the general assumption that categorisation is a function of similarity, i.e. that categorisation decisions

---

[225] Important critical discussions of this topic are offered in: Osherson, Daniel, Edward E. Smith 1999. [1981]. "On the Adequacy of Prototype Theory as a Theory of Concepts". In: Margolis and Laurence (eds.), 261–278. [*Cognition* 9, 35–58], and Wierzbicka, Anna 1999. *Język - umysł - kultura* [Language – mind – culture]. Edited by Jerzy Bartmiński. Warszawa : Wydawnictwo Naukowe PWN.

[226] Summarised and discussed in Bombor (2005: 20–22).

[227] I employ the term 'similarity-based approaches' rather than '*probabilistic* approaches' because it seems to be used more uniformly in the literature. The term 'probabilistic views' is used mostly in the same sense, to include both prototype and exemplar theories (e.g. Medin et al. 2001: 387); however, some (e.g. Murphy 2003, Nęcka et al. 2006) use it differently.

causally depend on the judgement of how similar the categorised item is to a stored representation (whatever the exact form of this representation – see below). The concept of similarity, or *resemblance*, thus becomes of central significance and requires a closer examination.

7.2.1. Problems with similarity

Similarity is a notion that has had an immense impact on the discussions of issues related to categorisation (Medin and Aguilar 1999). While constituting a potentially useful or sometimes even indispensable theoretical tool, it also evokes concern about the possibilities of its successful application in particular models. This is owing chiefly to the dilemma regarding the broad versus narrow construal of similarity, with the apparent lack of a productive middle ground. A natural way of understanding similarity narrowly is to treat it as perceptually driven, a construal patently too limited for most of interesting applications to categorisation (see below)[228]. The only viable alternative is to understand similarity broadly, with the inclusion of abstract, relational, functional and contextual aspects. Similarity so construed becomes more useful for theoretical applications, but immediately becomes sensitive to the difficulties with operationalising it with the rigour requisite for scientific discourse, or "constraining" it[229].

A famous reservation very often referred to in the discussions of the validity of the concept of 'similarity' is the one voiced and developed by Nelson

---

[228] The default construal of similarity as *perceptual* similarity is very common in everyday explanations of categorisation, consider e.g. the popular „a whale is a mammal despite its being more *similar* to fish".

[229] "…[S]imilarity can be a notoriously unconstrained variable." (Hampton 1998: 139)

"The main criticism has been that the notion of similarity is too unconstrained to be useful as an explanatory principle." (Medin and Aguilar 1999: 104)

"[s]imilarity is too flexible and unconstrained to serve as a grounding explanation for categorization… [if] similarity is overly flexible and context-dependent, then similarity would be in as much need of explanation as categorization." (Goldstone 1994: 126–127)

Goodman[230] (1992 [1970]). This apparently multifarious problem proves to be fundamentally that of underspecification: any two entities can be construed as similar in some respects, e.g. in virtue of sharing an infinite number of features (be they irrelevant ones, such as 'is smaller than a galaxy'), and judging what respects and what features are the *relevant* ones must circularly refer back to the notion of similarity. By the same token, Medin et al. (2001) observe that for every pair of entities A and B, they can be estimated as similar (or dissimilar) in one respect or another, and "a zebra and a barber pole could be seen as more similar than a zebra and a horse, if the feature striped is given sufficient weight" (Medin et al. 2001: 387).

Similarity, even in a general sense based on the natural language understanding, is a notion considerably more complex than it could at first appear. As is well known – but contrary to pre-theoretical intuitions – similarity is not symmetrical, but directional. Pairs of entities will be judged as more or less similar depending solely on the direction of comparison, i.e. A can be judged as more similar to B than B to A, sometimes considerably. Such a directional character of similarity was originally shown by the economist and cognitive scientist Amos Tversky (2004 [1977]) and later confirmed by a number of other studies (e.g. Medin and Goldstone 1995). Likewise, actual similarity judgements show considerably less transitivity (if A similar to B and B similar to C, then A similar to C) than could be naturally expected.

While similarity is most often conceptualised in terms of a two-argument relation (similar [A, B]), such a formulation leaves this notion grossly underspecified. This problem is discussed by Medin and Goldstone (1995: 83–87), who – also building on Goodman's (1992) insight – conclude that "[s]pecifically, the statement that *A* is similar to *B* in respect *C* is an incomplete, misleading analysis of similarity. At a minimum, similarity statements need to be expanded to include <according to comparison process *D*, relative to some

[230] Henry Nelson Goodman (1906–1998), an American philosopher.

standard *E*, mapped onto judgments by some function *F*, from perspective *G*>" (note that this comment concerns similarity *statements*). Such a strategy of constraining similarity, however, has the price of introducing dangerously 'loose' variables: variables such as 'perspective' might in practice turn out to be very difficult to operationalise.

Furthermore, although the correlation between similarity and categorisation is both intuitively obvious and has been firmly established experimentally, still – as is well known since at least David Hume – correlation alone is no proof of causation, and the direction of the causal link remains difficult to show conclusively. In other words, similarity can give rise to categorisation, but the relation could run in reverse; and researchers such as Philippe Schyns (1997) or Douglas Medin (Medin et. al. 2001: 389) have repeatedly pointed out that not only can an item be categorised on a basis of similarity, but two items can be viewed as (more) similar precisely because they share a category label.

The above points illustrate the main challenges for the attempts of spelling out an approach to categorisation/conceptual structure based on similarity. Firstly, similarity has to be stated in some more tractable format. Precisely because the tools of classical logic are not naturally fit for the task of capturing the nuances of the notion of similarity, models based on statistical procedures must instead be employed. Secondly, if a feasible way of judging the direction of dependence between similarity and categorisation could be devised, that would strengthen the approach philosophically.

7.2.1.1. Note: similarity versus 'shallow' perceptual similarity

There is a natural propensity to construe similarity in terms of merely *perceptual similarity*, especially outward visual appearance, carrying the implication of it being a 'shallow' heuristic. This is especially visible in frequent juxtapositions of 'surface similarity' with 'deeper' organising principles of a category (often suggesting a division of labour between the similarity-based 'identification

procedures' and the rule-based 'conceptual core', as discussed in 5.2.3.3.). A prime example is the principle of "original sim"[231], whereby young children categorise – and frequently 'miscategorise' – items based exclusively on their outward appearance. Later in development children gradually abandon this strategy, being driven instead by other kinds of information, such as causal relations. For example, circus lions dressed up in tigers' costumes are judged to be tigers by younger children ('original sim'), but to be lions by older children.

Nevertheless, the tendency to constrain similarity in this particular way is unfounded and overly limiting. Similarity between two items can hold in virtue of all types of respects: there are no reasons why more abstract features or dimensions should be excluded from being taken into account (cf. Hampton 1998: 138, 142–143). Particularly illustrative is the issue of categorisation of items other than concrete, material objects – consider for example detecting the 'instances of beauty' or issuing verdicts in legal cases. These categorisation decisions cannot rest on 'outward appearance' and patently require other forms of similarity. Therefore, in this respect similarity needs to be understood in a general sense, as a candidate for a more universal categorisation mechanism.

7.2.2. Ways of constraining similarity

The theoretical value of similarity hinges on the way in which it can be formalised with the introduction of appropriately strong constraints. When this has been achieved, similarity can be successfully applied as a direct tool for modelling categorisation. Minimally, similarity could be a vital component in the models, if still having to be complemented with other principles; ideally, similarity alone could predict categorisation decisions. Presented below are the most influential approaches – of varied specificity – to harnessing similarity so that it becomes adequate for grounding categorisation. Note that except for overall similarity, the other methods listed below assume the breakdown of the

[231] "Original sim" is a term invented by Frank C. Keil (cited in Hampton 2001: 16); see also Keil et al. 1998.

stimulus to be categorised into feature bundles, which makes them sensitive to the feature format problem (5.2.5.).

7.2.2.1. Overall similarity

It has been argued that judging category membership by *overall similarity* is an often used categorisation procedure, at least for perceptual categorisation. It is accomplished by a rapid response to a stimulus by essentially *holistic* processing, without – or prior to – breaking the stimulus down into separate features or aspects. Robert Goldstone and Laurence Barsalou (1998: 240–241) make a case for a wide application of undifferentiated overall similarity especially in rapid visual categorisation. A related strategy is *categorisation by blurring*, whereby one classifies novel instances without forming the precise schema of a category, but purely by comparing the novel instance to a known member of the category and blurring over the irrelevant differences. In Goldstone's and Barsalou's words (1998: 250): "[o]ne does not need to know what makes something a dog in order to categorize the neighbor's poodle as a dog, as long as one knows that a beagle is a dog, and is able to ignore (blur over) the differences between poodles and beagles".

Overall similarity has the major advantage of avoiding the feature format problem sketched out in the section 5.2.5. Categorisation by overall similarity is rapid and automatic, and takes place without recourse to explicitly coded features, so the problem of the linguistic format of the features does not arise. Still, the trade-off appears to be that such a strategy has little use in nonperceptual categorisation; consequently, the prospects of applying it to the study of lexical categorisation seem to be very limited.

7.2.2.2. Typicality

In apparent contrast to the predictions of the classical view – cf. 5.2.1. point a) – different members of a given category are consistently judged as differentially representative of the category to which they belong. This level of perceived

representativeness is extremely robustly correlated with several variables of subjects' performance in category-related tasks (see 5.2.3.3.) and is measured as *typicality*, or *goodness of example*. In other words, category members are gradable in terms of 'how well' they exemplify their category. In the prototype view, the phenomenon of typicality gradient is taken as a clue to the actual category structure: typicality is assumed to be the similarity of an item to the category prototype(s) and, at the same time, to be the sole variable underlying categorisation decisions (the exemplar view sees the phenomenon of typicality as emergent rather than reflecting the actual structure of mental representations). Note that typicality so conceived remains an intentional rather than formal measure, itself in need of further specification or explicit coding.

James Hampton (1998: 139) proposes to treat typicality as "similarity in respect of those attributes which form the intensional representation of the prototype concept"; typicality thus becomes "…a constrained form of similarity, in which the respects (and their relative importance) are determined by the conceptual representation itself." (1998: 139) Vital is the fact that while typicality assumes similarity as dependent on the comparison of feature bundles, the nature of those features is not prejudged: they may encode not only perceptual but also abstract characteristics. The exact mechanism of quantifying over feature overlaps may vary, as particular models will adjust the specific computations to best reflect the performance of human subjects.

7.2.2.3. Family resemblance

The concept of family resemblances was originally suggested by Ludwig Wittgenstein (1987 [1953]) as an alternative to the understanding of meanings of words in the traditional, essentialist spirit (see 5.2.3.1.). Rather than there being any properties common to all designata of a word, there can be found only overlaps of properties: "the various resemblances between members of a family: build, features, colour of eyes, gait, temperament, etc. etc. overlap and criss-cross in the same way", hence the name (Wittgenstein 1987: 32). In the 1970s,

developing this inspiration, Eleanor Rosch hypothesised that the patterns of feature overlaps might be what explains both typicality ratings and categorisation decisions: the 'best' category members tend to have the most attributes in common with other category members while at the same time having the least attributes in common with category nonmembers (in modelling, attributes may also be given differential weighting according to their relative importance). Rosch and Mervis (1996 [1975]) found empirical support for this hypothesis, with later empirical work generally confirming the overall soundness of this conclusion[232].

One rather important difference should be reemphasised between Rosch et al.'s treatment of family resemblances and the intuitive understanding of this notion. This difference consists in enhancing the basic intra-category dimension with the addition of the *inter-category* dimension: the exclusion of the non-overlapping attributes, i.e. ones that are uncommon for other category members. Thus, high family resemblance scores, predicting both typicality and categorisation, reflect both minimised within-category differences *and maximised between-category differences*.

7.3. Prototype and exemplar models of categorisation

The prototype approach to categorisation/concepts arose as a reaction to the classical approach early in the 1970s. The main philosophical inspiration had been that of the Wittgensteinian 'family resemblances', while the decisive empirical contributions came from the work of Eleanor Rosch and her associates, already mentioned in 5.2.3.3[233]. Since the general topic of prototypes (prototype representations) is very extensively covered in extant cognitive-linguistic literature, I will not treat it at length here. Rather, I will examine its main

[232] See especially the discussion in Murphy 2002

[233] Other key influences, especially the contributions from the logician Lofti Zadeh and the anthropologists Brent Berlin and Paul Kay, are discussed in Lakoff (1990 [1987]: 12–57).

theoretical underpinnings, and turn to highlighting the crucial difference from the exemplar view, focussing in particular on the advantages of the latter.

The exemplar approach to concepts/categorisation appears to have been somewhat neglected in Cognitive Science, at least until recently. Although exemplar-based models also have a considerable history, dating back to as early as the 1970s (e.g. Medin and Schaffer 1978; cf. Medin et al. 2001: 379), the influence of the exemplar view on Cognitive Science has been relatively limited. This is especially striking when compared to the tremendous impact that the closely related prototype view has exerted in the areas of linguistics and philosophy; many important discussions (e.g. Chlewiński 1999, Fodor 1998) either ignore the exemplar view or fail to acknowledge it as a contender distinct from the prototype view. Reasons for such a situation include, but are not limited to, the smaller tractability of exemplar models as well as their natural inclination towards *stricte* perceptual categorisation, and will constitute a topic of discussion in the later sections of this chapter.

7.3.1. What is 'a prototype'?

When reflecting on the notion of 'a prototype', one is faced with ontological problems along the lines already known from the discussion of the terms 'concept', 'category', etc. That is, do prototypes exist as real entities in the subject-external world, as abstracta, or as subject-internal mental representations? If the notion of a prototype as a mental representation is accepted, there arise additional difficulties with the particular format of such a representation.

The idea of category prototypes as specific exemplars in the world (e.g. a particular individual sparrow for BIRD) must be discarded due to immediate absurd consequences: for example, it is unclear how one would establish a particular individual to be the best exemplar for each category, or how people could gain epistemic access to that particular individual. The externalistic understanding of prototypes as abstract beings is also ruled out in a work where

the fundamentally mentalistic perspective to meanings has been assumed and defended – for reasons discussed in 4.2.4.

The mentalistic reading of the notion of prototype again admits two interpretations: as a singular representation versus a summary one. Early work by Rosch favoured the first construal, i.e. of prototypes as individual best examples (as represented in the cognitive system): "[p]rototypes appear to be just those members of a category which most reflect the redundancy structure of the category as a whole." (Rosch et al. 1976: 433) However, an interpretation both more popular and more useful for modelling takes a prototype not to represent any particular individual, but to be an abstract summary representation – formed by a statistical generalisation over the features of the category members and usually encoded as a list of features[234].

7.3.2. Categorisation by prototype

The foundation of the prototype approach to concepts is the discovery of typicality effects taken one step forward – to the assumption that concepts are mentally represented as prototypes[235] (or, more consistently with the terminology advocated in this work: concepts are encoded as prototypes, or concepts are mental representations taking the form of prototypes). Accordingly, the process of categorisation consists in measuring the similarity of the categorised item to the category prototype, typically achieved through detecting feature overlaps: sufficient similarity triggers the 'member' categorisation decision, while insufficient similarity triggers the 'nonmember' decision.

---

[234] Accordingly, a trout can be said to be the prototype of the category FISH only in so much as it is likely to be judged a *prototypical member* of this class of objects, i.e. of fish. Still, the prototype representation of the category FISH is an abstract representation consisting of such (weighted) features as 'lives in water', 'has fins', 'is slimy', and so on.

[235] This is a 'strong' assumption. Rosch (Rosch 1988a, Rosch and Mervis 1996) stressed repeatedly that the existence and pervasiveness of typicality effects merely *suggests*, and by no means *proves*, that categories are represented as prototypes.

James A. Hampton, one of the main proponents of prototype models of categorisation, describes the basic meta-model of categorisation by prototype as having three essential components: an *intensional representation*, a *metric of similarity*, and a *threshold criterion* (Hampton 1998). The intensional representation is a set of features – either an unstructured list, or a structured form such as a frame, tree, or schema – often with weights dependent on their relative importance. The metric of similarity is a specific mathematical mechanism for calculating the similarity of the categorised item to the above standard, i.e. the intensional representation, based on the number and weights of matching and non-matching features; the computations may also take into account the inter-category aspect (see 7.2.2.3. above). The threshold criterion can be conceptualised as the category border in the similarity space: it is a binary mark that, when reached or exceeded by the computed similarity rating, produces the 'member' categorisation decision, and otherwise produces the 'nonmember' decision.

7.3.3. What is 'an exemplar'?

The notion of 'an exemplar' in the exemplar view is a non-obvious one and requires specification – much like the notion of 'a prototype', but in slightly different ways. Intuitively, exemplars are real-world countable objects or samples; however, not surprisingly, cognitivist approaches to categorisation assume exemplars to be mental kinds of beings: exemplars as representations that are *stored in memory*. Although there suggests itself the simple interpretation of an exemplar as a representation of an individual, this issue is more complex, as illustrated by the quote from Murphy (2002: 58):

> Suppose that I know a bulldog that drools a great deal named Wilbur… How do I decide, now, whether a friend of mine, who is complaining about her new dog's drooling, has a bulldog? According to the exemplar view, I would have to retrieve all the dog exemplars that I know that drool (no small number), and then essentially count up how many of them are bulldogs. But in retrieving

> these exemplars, how do I count Wilbur? Does he count once, because he is only one dog, or does each encounter with Wilbur count separately? Put in more formal terms, do I count types (Wilbur) or tokens (Wilbur-encounters)?

It seems that 'exemplars' are best made sense of in terms of distinct *memory traces*. It can be assumed that every occurrence of an item leaves a separate memory trace; alternatively, people store only individuals, so that each occurrence of the same individual would not leave a separate trace, but update and strengthen the present one for this individual. In any case, instance-level information should be preserved in the system in addition to individual-level information rather than excluded, since it has been shown to exert influence on categorisation (Barsalou et al. 1998). Another key point is that each exemplar must appear with its associated category label.

A common spontaneous objection against the exemplar fashion of processing category-related information is the concern that it could place too much strain on long term memory resources; simply put, there may be no 'room' in memory to remember all encountered exemplars of all known categories. The idea of people being able to 'remember' every single exemplar of every single object and other conceptualised entity that they come across might indeed seem counterintuitive. However, such a scenario does not contradict what is known about the workings of the human central nervous system. Crucial here is the understanding of 'remembering': the exemplar view does not require that the experienced exemplars or instances be remembered individually in full detail, in the form accessible to later conscious retrieval; it stipulates merely that 'storing' an exemplar means *each of them* having *some* impact on the structure of the category. The increase in stored information takes place at an arithmetic rate, therefore sidestepping the risk of combinatorial explosion: the number of stored exemplars is *large*, but – given what is known about the human brain – not *unmanageably large*. The synaptic plasticity, both in terms of the creation of new synapses and adjusting existing connections' strengths, is sufficient to

accommodate the consequences of an exemplar model of categorisation (see also Barsalou 1992: 27).

7.3.4. Categorisation by exemplars

The foundational commitment of exemplar models (such as Exemplar-Based Random Walk – Palmeri 1997; Generalised Context Model – Medin and Schaffer 1978; or various exemplar predictors in Storms et al. 2000)[236] is that of *resisting* the default ideas:

- of abstraction of criterial features,
- of reduction of complexity in category information, and
- of creating categories as nonredundant representations.

On this view, people store in memory all encountered instances of a given category to form an emergent representation that is *distributed*, *collective*, and *heavily redundant*. According to exemplar-based models such as the ones mentioned above, categorisation decision process consists in retrieving from memory a certain (potentially huge) number of candidate exemplars that are similar to the test stimulus, and classifying the test stimulus on the basis of its relative similarity to the exemplars from the various categories. This means that on coming across e.g. some animal, subjects invoke a number of similar exemplars, choose the best match (or a few closest matches[237]), extract the category label attached to the exemplar(s), and assign that label (e.g. 'cat') to the encountered animal. Needless to say, just as in other models of categorisation, this is assumed to be a split-second process performed in the cognitive unconscious.

---

[236] Also based on the discussions by Storms (2004), Murphy (2003: 50–53, 73–114), and Medin et al. (2001: 378–392).

[237] As in EBRW (Palmeri 1997, Nosofsky and Palmeri 1997).

Consider for example the Exemplar-Based Random Walk model (EBRW; developed over Robert M. Nosofsky's Generalised Context Model, GCM) of speeded *perceptual* categorisation (Nosofsky and Palmeri 1997). In the model, exemplars are conceptualised as points in multidimensional psychological space, with their similarity (inversely) reflected by their distance in that space, i.e. the distance in the psychological space *decreasing* with growing similarity. Note that rather than feature checks, multidimensional scaling is used as a similarity metric, so that exemplars' similarity on any dimension may vary continuously (exemplars being continuously more or less similar in any given respect, e.g. more or less 'red') rather than being a matter of a binary match/mismatch (+/- 'red'). On presenting the test stimulus (the item to be categorised), *all* exemplars stored in memory 'race' to be retrieved one by one, and their category label is read. More similar exemplars, being less 'distant', are thus more likely to 'win the race'. The retrieval process continues until enough evidence is build – a sufficient number of sufficiently similar exemplars are retrieved – for the item to be assigned to a given category. Among the strengths of EBRW is its computational flexibility, which allows it to successfully incorporate the influence on categorisation of such cognitive factors as context, selective attention, and the recency of presentation, resulting in a wider scope of predictions related to performance in category-based tasks and thus yielding greater psychological reality.

7.4. From prototype to exemplar models in lexical categorisation

In the literature on categorisation and concepts (e.g. Taylor 1995, Chlewiński 1999), the qualitative difference between classical and prototype models is typically heavily emphasised. With respect to this salient opposition, the exemplar view is normally grouped together with the prototype one: they both are similarity-based, and in lexical categorisation they are continuous with each other (as shall be seen, the distinction between prototype and exemplar models is not always easy to make). Both the prototype and the exemplar views are

strongly 'nonclassical'; in contrast to the tenets of the classical approach (5.2.1.), they converge on the same set of alternative theoretical assumptions:

a) category members have differential, graded status (either as members of that category or at least in terms of how representative they are),
b) category nonmembers have differential, graded status: even the entities that technically fall outside the category borders can be rated as more, or less, similar to the category prototype or to stored exemplars,
c) no single feature can be necessary for category membership,
d) features by which people categorise are weighted, that is, they differ in their relative significance,
e) category boundaries are fuzzy; categorisation decisions are a matter of judgement; they may vary across individuals and depend not only on the categorised item, but on seemingly external factors such as context,

A) categorisation (*qua* category formation) does not capture the objective structure of reality, but creates such a structure that is optimised for the cognitive agent's maximally effective functioning in the world

7.4.1. Distinguishing exemplar from prototype models

The above convergence, however, is motivated in slightly different ways for both approaches. In the prototype view, some of the assumptions may be taken as revealing the specific representational structure of categories, whereas in the exemplar view they arise as consequences of the categorisation *mechanism*. A good illustration of the difference is the crucial phenomenon of typicality effects reviewed in 5.2.3.3. In prototype models, typicality is intrinsically encoded in the structure of representation, as it constitutes the same psychological variable that is used for making categorisation decisions: typicality effects are a direct consequence of the categorial representation having a prototype format. In the exemplar models, typicality effects will also naturally hold, but for a different

reason. Exemplar models predict the efficiency of performance to increase as a function of the number of similar stored exemplars; more prototypical instances of a category are, as a rule, more similar to (a greater number of) stored exemplars of the category than nonprototypical instances, and, as such, would be categorised faster and more accurately, thereby giving rise to typicality effects. Thus, 'prototypes' will be emergent statistical artefacts of the underlying processing of category-related information.

In the context of semantic-lexical categorisation, the distinction between prototype and exemplar models becomes even more elusive. The difference between an exemplar-based and a prototype model is very much reduced to the type of computational process – calculating feature overlap with many exemplars rather than a single prototype – as is comprehensively illustrated by the following passage from Storms et. al. (2000: 53):

> To summarize, in the context of natural language categories like fruits, vegetables, vehicles, etc., three different theoretical views may be distinguished depending on the levels at which abstraction does or does not take place. The first view assumes that no abstraction whatsoever takes place and that only memory traces of particular encountered instances are stored. Any category-related judgment is based on these memory traces, as no abstract information is stored with verbal concepts. The second view assumes that abstraction may take place, but only at a level lower than the concepts studied, that is, at the level of tomatoes in case vegetables are studied. The representation of the studied natural language concepts, like vegetables and vehicles, is comprised of lower level concepts like tomatoes and bikes, respectively. Finally, the third view states that abstraction (also) takes place at the level of the studied natural language concepts and that (characteristic) features of their exemplars are directly stored at this level. The latter view can be labeled the prototype view, and the first two views are exemplar views. (Storms et al. 2000: 53)

7.4.2. Distinctiveness and advantages of exemplar models

As has been remarked before, owing to the major commonalities between the prototype and exemplar views, they are easily conflated, with the exemplar view typically losing its autonomy and being incorporated as a subtype of the prototype view. To a degree, this is understandable considering the well-established status of the prototype approach (especially in cognitive linguistics[238]), and the relative obscurity of the exemplar approach. However, it is vitally important to appreciate the critical respect in which exemplar models are sharply distinguished from both prototype *and* classical models. This qualitative difference consists in the nature of postulated representation referred to above. Central to both classical and prototype view is the postulate of what George Murphy (2002: 42–58) names a *summary representation*, with considerable reduction in the amount of preserved category-related information. In contrast, the *fundamental assumption behind exemplar models is the lack of an explicit, summary representation*, which in a way 'shifts' the level on which the categorial representation resides one step 'down' – to the level of individual remembered instances.

The key characteristic is information redundancy, and the particular resultant strength of the exemplar view is the retention of a substantial body of category-related knowledge – especially in comparison to only skeletal category-related information preserved by classical or prototype models. Thus, exemplar models best embody the cognitive mechanism postulated by Evan Heit and Laurence Barsalou (1996) under the name of *instantiation principle* – it consists in preserving rather than filtering out detailed information about individual instances. Instantiation principle has since been found to play a significant role in categorisation[239]. This is also partly equivalent to the incorporation of noncriterial, encyclopaedic information as inherently constitutive of concepts'/categories' meaning, precisely as postulated in the cognitive

[238] Cf. e.g. Kardela 2006a: 202, Kalisz 2001: 50.

[239] E.g. Storms et al. 2000; but see also Murphy 2003: 114.

conceptions of language. The exemplar stance also seems to be consistent with the way in which categories are put to use. This counts as an important advantage, in view of the repeated criticism of the rival approaches for treating categorisation as an end in itself and the consequent inability to account for category-related phenomena (5.2.3.3., 6.3.2.). For example, people possess knowledge not only of the attributes of a category, but also on its internal structure, including the correlations between the attributes; this supports within-category predictions and inferences. The canonical examples here involve people's ability to make the inferences that if a bird is *large*, it is *unable to sing* (Medin et al., 2001: 379), and if it *has large wings*, it is more likely to *inhabit regions near the sea* and *live on fish* (Medin and Smith, 1984: 125).

Exemplar models of natural language categories (e.g. Storms et al. 2000) assume the representations to be given as collections of features ("feature bundles"), making them partly vulnerable to the feature format problem (5.2.5.). However, their advantages also become clearly visible. Each categorial representation comprises *x* stored exemplars, and assuming that each of the exemplars can be broken down into a collection of *y* features, the distributed category representation effectively consists of the total of *x* * *y* features. For example, if the category A consists of 22 stored exemplars, and each of these exemplars consists – on average – of 6 features, then the total number of features in terms of which the category A is cumulatively represented is 22 * 6 = 132. Many of those features will of course overlap, being common to most of the exemplars. However, firstly, the ratio of recurrence itself conveys information about the feature's importance for the category relative to other features, and secondly, the total number of nonrepeating features will still be substantially larger than for prototype models. Idiosyncratic features of seemingly marginal relevance that would be lost in classical and prototype models still reside in the representation and, though pushed back to deep background, can nevertheless be retrieved to influence categorisation as well as category-related processes. Moreover, information about feature correlations is preserved (such as is *large* -

is *unable to sing* for BIRD – see above). Those points are particularly significant if one remembers the dynamic character of categorisation and its sensitivity to relational and contextual properties: the inclusion of a broader knowledge base helps to account for this context-dependence.

A final point concerns the relative unpopularity of the exemplar approach: as observed above, the application of exemplar-based models to semantic-lexical categorisation has so far been very limited, especially when compared with the very wide-ranging popularity of prototype models. Storms (2004: 5–7) identifies two major stumbling blocks in this respect. Firstly, instances of natural language concepts seem to have to be abstractions themselves ("what is an exemplar"), and secondly, the appropriate selection of features is problematic ("what are the relevant features").

While accepting Storms's diagnosis, I suggest that this situation can be shown to have more fundamental basis, itself being *another consequence of the refractory 'feature format' problem* (5.2.5.). As was mentioned above, exemplar models have always had their roots in the area of perceptual categorisation, where they continue to thrive. The emphasis on the statistical and distributed nature of representation leads to heavy reliance on advanced computational procedures. While features and dimensions used in perceptual categorisation are much more amenable to purely quantitative treatment, for lexical-semantic categorisation this is much more problematic. In the case of lexical categories, some level of intentional description seems to be inevitable, which may be more difficult to reconcile with the inherently mathematical-computational nature of the exemplar approach.

7.5. Summary

The seventh, concluding chapter of my dissertation addressed the so-called 'similarity-based' approaches to concepts and categorisation. I scrutinised the notion of similarity for its theoretical soundness and applicability for modelling categorisation. Sample ways of constraining similarity were concisely described.

Subsequently, I presented the basic assumptions as well as general operating mechanisms of both the prototype and exemplar views of categorisation, highlighting their commonalities, but also drawing attention to the key respect in which they differed, that is the summary versus distributed nature of representation. Briefly addressed was also the ontological/methodological question of the specific nature of both 'a prototype' and 'an exemplar'.

What I identified as the main asset of the exemplar view was the distributed, cumulative nature of categorial representation maximising 'instantiation principle' (as opposed to classical and prototype summary representations maximising 'cognitive economy'). Statistical generalisations over a large number of exemplars allow exemplar-based models to mimic the desirable characteristics of prototype models, in particular the phenomenon of typicality effects. At the same time, the principle of inclusion of large portions of noncriterial and seemingly 'collateral' knowledge brings the exemplar view closer to the way in which actual cognitive agents (humans) perform category-related operations.

Conclusion

Categorisation may be the single most basic cognitive process in organisms, and as an area of theoretical inquiry, it is certainly fundamental to Cognitive Science as a whole. In the words of Lakoff and Johnson, "every living beings categorises", treating distinguishable stimuli as equivalent in certain respects – respects important for the organism's successfully functioning in its environment. Thus, categorisation truly underlies all cognition. At the other end of the spectrum, high-level cognition is organised and permeated by language, giving rise to mental representations that count, and can function as, fully blown *concepts*.

The study of language becomes particularly attractive when it is not practised as an isolated, purely descriptive enterprise; rather, its appeal is the greatest when it can be demonstrated to have wide-ranging implications for the study of the human mind. Half a century ago, precisely such was the motivation of Chomsky as a co-founder of the emerging science of cognition. Despite its evolution over the succeeding five decades (depicted in Chapter 1), Cognitive Science has preserved its standpoint on the phenomenon of language as well as the goals and methods of its study.

Precisely such is, too, the spirit of this book. Its main commitment is to the participation in Cognitive Science, a commitment whose theoretical consequences were spelled out and defended in Chapter 2. The cognitivist nature of this work consists in making connections to a larger body of interdisciplinary research: both in the sense of drawing from this research, and in the sense of yielding conclusions that, hopefully, might prove transferable and valuable to other fields within CS.

The theoretical achievements of this work present themselves as follows. Chapter 1 was devoted to introducing Cognitive Science in both historical and

contemporary context. Emphasis was placed on the current profile of this broad field of study, and the constitutive role of interdisciplinary collaboration. Chapter 2 laid down the foundations for the mentalistic perspective that was assumed throughout this work; it also targeted polemically the influential argument against such a perspective, based on Hilary Putnam's thought experiment. Chapter 3 systematised the definitions of the term 'concept' present in the literature, arriving at a list of fourteen features ascribed to concepts in a range of theoretical outlooks. It also examined the notions of category/categorisation and mental representation.

Chapter 4 contained a crucial argument regarding the character of the relation between concepts and lexical items. It was proposed that the term 'concept' can be best made to work in the context of Cognitive Science when it is understood as a mental representation correlated with an entry in the mental lexicon. The above proposal was then substantiated by a survey of relevant evidence.

Chapter 5 consisted in a review of the historically dominant classical view of concepts/categorisation. Its subsequent evaluation generally reinforced the highly critical conclusions already present in the literature. However, in the discussion, one specific problem was suggested as lying at the heart of the problems of the classical approach, namely the frequent assumption of a wordlike format of component features. Chapter 6 brought a presentation of another influential view of conceptual structure, that is conceptual atomism. Arguments were supplied for the refutation of this view.

Chapter 7 offered a short examination of the underlying tenets of the similarity based approaches to categorisation. The discussion focussed on the less popular exemplar view; it was opined that this view emerged as best fulfilling the criteria of a cognitively oriented theory. Further research, it was proposed, should be concentrated on overcoming the technical problems with application of exemplar models to lexical-semantic categorisation.

Two specific significant findings appeared to stand in certain conflict. The exemplar view, introduced towards the end of this work, was considered to show the most potential for modelling human cognition with respect to broadly understood categorisation. At the same time, it was also found to be vulnerable from the pervasive problem of the *format of features* (at least when used to model lexical-semantic categorisation), that is of dubious psychological reality of conceptual structural elements when rendered verbally.

A particularly interesting suggestion to follow may be to aim at a certain synthesis – one in which the statistical computational mechanism underlying exemplar models could be combined with a different, 'non-featural' representation format. Considering that the recently developed cognitive tradition in linguistics has made extensive use of nonpropositional (generally, 'nonlinguaform') representations, such a direction of future research looks both viable and promising.

BIBLIOGRAPHY

Aarts, Bas 2006. "Conceptions of Categorization in the History of Linguistics". *Language Sciences* 28, 361–385.

Ahn, Woo-kyoung, Charles Kalish, Susan A. Gelman, Douglas L. Medin, Christian Luhmann, Scott Atran, John D. Coley, Patrick Shafto 2001. "Why Essences are Essential in the Psychology of Concepts". *Cognition* 82, 59–69.

Aitchison, Jean 1996 [1987]. *Words in the Mind: an Introduction to the Mental Lexicon*. Second edition. Oxford: Blackwell Publishers.

Ajdukiewicz, Kazimierz 1949. *Zagadnienia i kierunki filozofii. Teoria poznania, metafizyka* [Issues and directions in philosophy. Epistemology, metaphysics]. Warszawa: Wydawnictwo Czytelnik.

Ajdukiewicz, Kazimierz 1960. *Język i poznanie. T. 1, Wybór pism z lat 1920–1939* [Language and cognition. Selection of papers from the 1920–1939 period]. Warszawa: Państwowe Wydawnictwo Naukowe.

Allan, Keith 2006. "Connotation". In: Brown (ed.), Vol. II, 41–44.

Anderson, Michael 2005. „How to Study the Mind: An Introduction to Embodied Cognition". In: Flavia Santoianni, Claudia Sabatano (eds.) 2005. *Embodied Cognition and Perceptual Learning in Adaptive Development*. Cambridge: Cambridge Scholars Press. http://www.cs.umd.edu/~anderson/papers/bes_ec.pdf

Aristotle. *Categories*. Trans. E. M. Edghill. Online edition – The Project Gutenberg. In: http://www.classicallibrary.org/aristotle/categories/arist10.txt ED 02/2006

Aristotle. *Metaphysics*. Transl. W. D. Ross. Online edition: http://classics.mit.edu/Aristotle/metaphysics.html

Armstrong, Sharon Lee, Lila R.Gleitman, Henry Gleitman 1999 [1983]. "What Some Concepts Might Not Be". In: Margolis and Laurence (eds.), 225–259. [*Cognition* 13, 263–308.]

Aydede, Murat 1998. "Fodor on Concepts and Frege Puzzles". *Pacific Philosophical Quarterly* 79 (4), 289–294.

Aydede, Murat 2004. "The Language of Thought Hypothesis." In: Zalta (ed.). URL = http://plato.stanford.edu/entries/language-thought/ ED 05/2005

Bach, Kent 2000. "A Review of *Concepts: Where Cognitive Science Went Wrong* by Jerry A. Fodor". *The Philosophical Review* 109 (4), 627–632. http://userwww.sfsu.edu/~kbach/Fodorreview.htm ED 01/2008

Baillargeon, Renée 2001. “Infants’ Physical Knowledge: Of Acquired Expectations and Core Principles”. In: Emmanuel Dupoux (ed.) 2001. *Language, Brain, and Cognitive Development: Essays in Honor of Jacques Mehler*. Cambridge, MA: MIT Press, 341–361.

Bancroft, Dennis 1995. “Language Development”. In: Lee and Das Gupta (eds.) 1995. *Children´s Cognitive and Language Development*. Oxford: Basil Blackwell, 45–80.

Barsalou, Lawrence W. 1983. “Ad Hoc Categories”. *Memory and Cognition* 11 (3), 211–227.

Barsalou, Lawrence W. 1992. *Cognitive Psychology. An Overview for Cognitive Scientists*. Hillsdale, NJ, Hove, UK: Lawrence Erlbaum Associates, Publishers.

Barsalou, Lawrence W., Janellen Huttenlocher, Koen Lamberts 1998. “Basing Categorization on Individuals and Events”. *Cognitive Psychology* 36 (3), 203–272.

Bechtel, William, Adele Abrahamsen, George Graham 1998. “The Life of Cognitive Science”. In: William Bechtel, George Graham (eds.). 1998. *A Companion to the Cognitive Science*. Oxford: Blackwell Publishers, 1–104.

Berger, Peter L., Thomas Luckmann 1966. *The Social Construction of Reality*. Garden City, NY: Doubleday.

Bermúdez, José Luis 2003. *Thinking without Words*. Oxford: Oxford University Press.

Bickerton, Derek. 1998. “Catastrophic Evolution: the Case for a Single Step from Protolanguage to Full Human Language”. In: James R. Hurford, Michael Studdert-Kennedy, Chris Knight (eds.) 1998. *Approaches to the Evolution of Language*. Cambridge, UK: Cambridge University Press, 341–358.

Björnsson, Gunnar 1998. *Moral Internalism. An Essay in Moral Psychology*. Stockholm: Norstedts Tryckeri.
http://people.su.se/~gbjorn/morint.pdf

Block, Ned 1998. “Semantics, Conceptual Role”. In: Craig and Floridi (CD-ROM).
http://cogprints.org/232/0/199712005.html (preprint)

Boden, Margaret A. 2006. *Mind as Machine: A History of Cognitive Science*. Oxford: Oxford University Press.

Bombor, Magdalena 2005. *Teoria prototypu w wybranych wersjach i jej implikacje filozoficzne* [Selected versions of prototype theory and its philosophical implications] (Unpublished MA thesis). Toruń: Institute of Philosophy at Nicolaus Copernicus University.

Borchert, Donald M. (ed.) 2006. *Encyclopedia of Philosophy, Second Edition*. Farmington Hills, MI: Thomson Gale.

Botterill, George, Peter Carruthers 1999. *The Philosophy of Psychology*. Cambridge, UK: Cambridge University Press.

Brown, Colin M., Peter Hagoort, Mariken ter Keurs 1999. "Electrophysiological Signatures of Visual Lexical Processing: Open- and Closed-Class Words". *Journal of Cognitive Neuroscience* 11 (3), 261–281.

Brown, Keith (ed.) 2006. *Encyclopedia of Language and Linguistics, Second Edition*. Oxford: Elsevier.

Bruner, Jerome S., Jacqueline J. Goodnow, George A. Austin 1999 [1956]. "The Process of Concept Attainment". In: Margolis and Laurence (eds.), 101–123. [Chapter 3 of *A Study of Thinking*. New York: John Wiley and Sons.]

Buckner, Randy L., Jessica M. Logan 2001. "Functional Neuroimaging Methods: PET and fMRI". In: Roberto Cabeza and Alan Kingstone (eds.) 2001. *Handbook of Functional Neuroimaging of Cognition*. Cambridge, MA: MIT Press, 27–48.

Buller, David J., Valerie Gray Hardcastle. 2000. "Evolutionary Psychology, Meet Developmental Neurobiology: Against Promiscuous Modularity". *Brain and Mind* 1, 307–325.

Burge, Tyler 1979. "Individualism and the Mental". *Midwest Studies in Philosophy* 4, 73–121.

Burge, Tyler 1986. "Individualism and Psychology". *The Philosophical Review* 95 (1), 3–45.

Buss, David M. 2004. *Evolutionary Psychology: The New Science of the Mind*. Second Edition. Boston: Allyn and Bacon.

Cain, Mark 2002. *Fodor: Language, Mind and Philosophy*. Cambridge: Polity Press.

Carey, Susan G. 1999 [1991]. "Knowledge Acquisition: Enrichment or Conceptual Change". In: Margolis and Laurence (eds.), 459–487. [In: Susan G. Carey, Rochel G. Gelman (eds.) 1991. *The Epigenesis of Mind: Essays in Biology and Cognition*. Hillsdale: Lawrence Erlbaum Associates.]

Carnap, Rudolf 1959 [1932]. "The Elimination of Metaphysics Through Logical Analysis of Language". Trans. Arthur Pap. In: Alfred J. Ayer (ed.) 1959. *Logical Positivism*. New York: The Free Press, 60–81. ["Überwindung der Metaphysik durch Logische Analyse der Sprache". *Erkenntnis* 2].

Changeux, Jean-Pierre 1997 [1983]. *Neuronal Man: the Biology of Mind*. Trans. Laurence Garey. Princeton: Princeton University Press. [*L'Homme Neuronal*. Paris: Librairie Arthème Fayard].

Chlewiński, Zdzisław 1999. *Umysł. Dynamiczna organizacja pojęć. Analiza psychologiczna* [T*he mind. the dynamic organisation of concepts. a psychological analysis*]. Warszawa: Wydawnictwo Naukowe PWN.

Chomsky, Noam A. 1957. *Syntactic Structures*. The Hague: Mouton.

Chomsky, Noam A. 1959. "A Review of B. F. Skinner's *Verbal Behavior*". In: *Language* 35 (1), 26–58.

Chomsky, Noam A. 1975. *Reflections on Language*. New York: Pantheon.

Chomsky, Noam A. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger.

Chomsky, Noam A. 2000. *New Horizons in the Study of Language and Mind*. Cambridge, UK: Cambridge University Press.

Chomsky, Noam A. 2005. "Universals of Human Nature", *Psychotherapy and Psychosomatics* 74 (5), 263–268.

Chuderski Adam 2002. *Wykorzystanie metod sztucznej inteligencji w badaniach nad umysłem* [The applicability of the methods of Artificial Intelligence in the study of mind]. Wortal Kognitywistyka.net.
http://www.kognitywistyka.net/artykuly/ach-wmsiwnu.pdf

Clark, Andy 1998. "Magic Words: How Language Augments Human Computation". In: Peter Carruthers, Jill Boucher (eds.) 1998. *Language And Thought: Interdisciplinary Themes*. Cambridge, UK: Cambridge University Press, 162–183.
http://www.philosophy.ed.ac.uk/staff/clark/pubs/magic.pdf

Clark, Andy. 2006. "From Fish to Fantasy: Reflections on an Embodied Cognitive Science". In: http://www.philosophy.ed.ac.uk/staff/clark/publications.html
http://www.philosophy.ed.ac.uk/staff/clark/pubs/TICSEmbodiment.pdf ED 05/2006

Clark, Andy, David Chalmers 1998. "The Extended Mind". *Analysis* 58 (1), 7–19.
http://www.philosophy.ed.ac.uk/staff/clark/pubs/TheExtendedMind.pdf

Cohen, Henri, Claire Lefebvre (eds.) 2005. *Handbook of Categorization in Cognitive Science*. Amsterdam: Elsevier.

Conee, Earl, Richard Feldman 2001 [2001]. "Internalism Defended". In: Kornblith (ed.), 231–260. [*American Philosophical Quarterly*]

Cooper, Ric. 2000 [1996]. "Explanation and Simulation in Cognitive Science". In: Green (ed.), 23–52.

Cosmides, Leda, John Tooby. 2006. "Evolutionary Psychology: A Primer". In: *Center for Evolutionary Psychology*: http://www.psych.ucsb.edu/research/cep/index.html http://www.psych.ucsb.edu/research/cep/primer.html ED 05/2006.

Craig, Edward, Luciano Floridi (eds.) 1998. *Routledge Encyclopedia of Philosophy, CD-ROM Edition*. London – New York: Routledge.

Davidoff, Jules 2001. "Language and Perceptual Categorisation". *Trends in Cognitive Sciences* 5 (9), 382–387.

Davis, James W. 2001. “Visual Categorization of Children and Adult Walking Styles”. *Lecture Notes in Computer Science: AVBPA 2001 Halmstad, Sweden*, 295–300.

De Beaugrande, Robert–Alain, Wolfgang Dressler 1981. *Introduction to Text Linguistics*. London: Longman.

Deacon, Terrence W. 1997. *The Symbolic Species. The Co-evolution of Language and the Human Brain*. London: Penguin Press.

Dennett, Daniel 1994. “The Role of Language in Intelligence”. In: Jean Khalfa (ed.) 1994. *What is Intelligence? The Darwin College Lectures*. Cambridge, MA: Cambridge University Press.

http://cogprints.org/192/00/rolelang.htm ED 04/2006

Dennett, Daniel 1996. *Kinds of Minds*. New York: Basic Books.

Dennett, Daniel 1998. *Brainchildren. Essays on Designing Minds*. Cambridge, MA: The MIT Press.

Dépy, Delphine, Joël Fagot, Jacquess Vauclair 1997. *Categorisation of Three-Dimensional Stimuli by Humans and Baboons: Search for Prototype Effects*. Behavioural Processes 39, 299–306.

Dębowski, Józef 2000. *Bezpośredniość poznania. Spory – dyskusje – wyniki*. Lublin: Wydawnictwo UMCS.

Dietrich, Eric 2001. “Concepts: Fodor's Little Semantic BBs of Thought. A Critical Look at Fodor's Theory of Concepts”. *Journal of Experimental and Theoretical Artificial Intelligence* 13 (2), 89–94.

Diogenes Laertius 1696. *The Lives, Opinions, and Remarkable Sayings of the Most Famous Ancient Philosophers*. Transl. unknown. London: R. Bentley. [Peri bion̄ dogmaton̄ kai apophthegmaton̄ ton̄ en philosophia eudokimēsanton̄].

Donald, Merlin 2004. “The Virtues of Rigorous Interdisciplinarity”. In: Joan M. Lucariello, Judith A. Hudson, Robyn Fivush, Patricia J. Bauer (eds.) 2004. *The Development of the Mediated Mind. Sociocultural Context and Cognitive Development*. London: Lawrence Erlbaum, 245–256.

Egan, Frances 2006. “Representation in Language and Mind”, In: Brown (ed.), Vol 10., 553–556.

Evans, Gareth 1982. *The Varieties of Reference*. Oxford: Oxford University Press. (ed. by John McDowell)

Eysenck, Michael W. (ed.) 1990. *The Blackwell Dictionary of Cognitive Psychology*. Cambridge, Ma: Basil Blackwell Ltd.

Eysenck, Michael W., Mark Keane 2002 [2000]. *Cognitive Psychology*. 4th edition. Hove, UK: Psychology Press Ltd.

Farkas, Katalin. 2003. „What is Externalism?". *Philosophical Studies* 112, 187–208.

Fitch, Tecumseh W., Marc D. Hauser, Noam Chomsky 2005. "The Evolution of the Language Faculty: Clarifications and Implications". *Cognition* 97 (2), 179–210.

Fodor, Jerry A. 1975. *The Language of Thought*. Cambridge, MA: Harvard University Press.

Fodor, Jerry A. 1980. "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology". *Behaviorial and Brain Sciences* 3 (1), 63–109.

Fodor, Jerry A. 1981a. "The Present Status of the Innateness Controversy". In: Jerry A. Fodor (ed.) 1981. *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, MA: MIT Press, 257–316.

Fodor, Jerry A. 1981b [1974]. "Special Sciences, or the Disunity of Science as a Working Hypothesis". In: Block N(ed.) 1981. *Readings in Philosophy of Psychology. Volume 2.* Cambridge, MA: Harvard University Press, 120–133. [*Synthese* 28, 77–115.]

Fodor, Jerry A. 1983. *Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.

Fodor, Jerry A. 1993 [1987]. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Fodor, Jerry A. 1994. *The Elm and the Expert. Mentalese and Its Semantics*. Cambridge, MA: MIT Press.

Fodor, Jerry A. 1995. "Concepts; A Potboiler". *Philosophical Issues* 6, 1–24.

Fodor, Jerry A. 1998. *Concepts: Where the Cognitive Science Went Wrong*. New York: Oxford University Press

Fodor, Jerry A. 2000. *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. Cambridge, MA: MIT Press.

Fodor, Jerry A. 2001. "Language, Thought and Compositionality". *Mind and Language* 16 (1), 1–15.

Fodor, Jerry A., Zenon W. Pylyshyn 1988. "Connectionism and Cognitive Architecture: a Critical Analysis". *Cognition* 28, 3–71.

Fodor, Jerry A., Ernest Lepore 1996. "The Red Herring and the Pet Fish: Why Concepts Still Can't Be Prototypes". *Cognition* 58, 253–270.

Frege, Gottlob 1960a [1952] {1891}. "Function and Concept". Transl. Peter Geach. In: Geach and Black (eds.), 21–41. {"Funktion und Begriff". Vortrag, gehalten in der Sitzung vom 9. Januar 1891 der Jenaischen Gesellschaft für Medizin und Naturwissenschaft, Jena: Hermann Pohle}.

Frege, Gottlob 1960b [1952] {1892}. "On Concept and Object". Transl. Peter Geach. In: Geach and Black (eds.), 42–55. {"Über Begriff und Gegenstand". *Vierteljahresschrift für wissenschaftliche Philosophie* 16, 192–205}.

Frege, Gottlob 2001a [1952] {1892}. "On Sense and Reference". Trans. Peter Geach and Max Black. In: Martinich and Sosa (eds.), 7–18. {"Über Sinn und Bedeutung". *Zeitschrift für Philosophie und philosophische Kritik* 100, 25–50}.

Frege, Gottlob 2001b [1956] {1918}. "Thought". Trans. Peter Geach and Robert Stoothoff. In: Martinich and Sosa (eds.), 19–31. {"Der Gedanke. Eine Logische Untersuchung". *Beitrage zur Philosophie des Deutschen Idealismus* I, 58–77}.

Friederici, Angela D., Bertram Opitz, D.Yves von Cramon 2000. "Segregating Semantic and Syntactic Aspects of Processing in the Human Brain: an fMRI Investigation of Different Word Types". *Cerebral Cortex* 10 (7), 698–705.

Gallagher, Shaun 2005. *How the Body Shapes the Mind*. Oxford: Oxford University Press.

Gallese Vittorio, George Lakoff 2005. "The Brain's Concepts: The Role of The Sensory-Motor System in Conceptual Knowledge". *Cognitive Neuropsychology* 22, 455–479. http://www.unipr.it/~gallese/PCGNSIOBA9.pdf ED 03/2006

Galton, Antony 1993. "On Specification and Implementation". In: Hookway and Peterson (eds.), 111–136.

Geach, Peter, Max Black (eds.) 1960 [1952]. *Translations from the Philosophical Writings of Gottlob Frege*. 2nd edition. Oxford: Basil Blackwell.

Gilbert, Aubrey L., Terry Regier, Paul Kay, Richard B. Ivry 2005. "Whorf Hypothesis Is Supported in The Right Visual Field But Not The Left". *Proceedings of the National Academy of Sciences* 103, 489–494.

Glasersfeld, Ernst von 1995. *Radical Constructivism: A Way of Knowing and Learning*. London – Washington: The Falmer Press.

Glasersfeld, Ernst von 2008 [1984] {1981}. "An Introduction to Radical Constructivism". *AntiMatters* 2 (3), 5–20. [In: Paul Watzlawick (ed.) 1984. *The Invented Reality: How Do We Know What We Believe We Know?* New York: Norton, 17–40 (translated from German)] {Paul Watzlawick (ed.) 1981. *Die Erfundene Wirklichkeit*. Munich: Piper, 16–38}.
http://anti-matters.org/ojs/index.php/antimatters/article/download/88/81

Goddard, Cliff 2002. "Whorf Meets Wierzbicka: Variation and Universals in Language and Thinking". *Language Sciences* 25 (4), 393–432.

Goldman, Alvin I. 2001. "Internalism Exposed". In: Kornblith (ed.), 207–230.

Goldman, Alvin I. (ed.) 1993. *Readings in Philosophy and Cognitive Science*. Cambridge, MA: MIT Press.

Goldstone, Robert L. 1994. "The Role of Similarity in Categorization: Providing a Groundwork". *Cognition* 52, 125–157.

Goldstone, Robert L., Lawrence W. Barsalou 1998. "Reuniting Perception and Conception". *Cognition* 65 (2-3), 231–262.

Goodman, Nelson 1992 [1972]. "Seven Strictures on Similarity" In: Mary Douglas, David L. Hull (eds.) 1992. *How Classification Works: Nelson Goodman Among the Social Sciences*. Edinburgh: Edinburgh University Press 13–22. [In: Lawrence Foster, John W. Swanson (eds.) 1970. *Experience and Theory*. Amherst: University of Massachusetts Press, 19–29.]

Gopnik, Myrna 1997. "Language Deficits and Genetic Factors". *Trends in Cognitive Sciences* 1 (1), 5–9.

Green, David (ed.) 2000 [1996]. *Cognitive Science. An Introduction*. Oxford: Blackwell Publishers.

Green, David 2000 [1996]. "Introduction". In: Green (ed.), 1–22.

Greenberg, Seth N., Asher Koriat 1991. "The Missing-Letter Effect for Common Function Words Depends on Their Linguistic Function in the Phrase". *Journal of Experimental Psychology: Learning, Memory, and Cognition* 17 (6), 1051–1061.

Grodzinsky, Yosef 2003. "Imaging the Grammatical Brain". In: Michael C. Arbib (ed.) 2003. *Handbook of Brain Theory and Neural Networks. 2nd edition*. Cambridge, MA: MIT Press, 551–556.

Hale, Courtney M., Helen Tager-Flusberg 2003. "The Influence of Language on Theory of Mind: A Training Study". *Developmental Science* 6 (3), 346–359.

Haman, Maciej 2002. *Pojęcia i ich rozwój. Percepcja, doświadczenie i naiwne teorie* [Concepts and their development. Perception, experience and naïve theories]. Warszawa: Wydawnictwo Matrix.

Hampton, James A. 1995. "Testing the Prototype Theory of Concepts". *Journal of Memory and Language* 34, 686–708.

Hampton, James A. 1998. "Similarity-based Categorization and Fuzziness of Natural Categories". *Cognition* 65, 137–165.

Hampton, James A. 1999. "Concepts". In: Wilson and Keil (eds.), 176–179.

Hampton, James A. 2001. "The Role of Similarity in Natural Categorization". In: Ulrike Hahn, Michael Ramscar (eds.) 2001. Similarity and categorization. Cambridge: Cambridge University Press, 13–26.

Harnad, Stevan 1987. "Psychophysical and Cognitive Aspects of Categorical Perception: A critical overview." In: Stevan Harnad (ed.) 1987. *Categorical perception: The groundwork of cognition*. New York: Cambridge University Press, 1–28.

Harnad, Stevan 1990. "The Symbol Grounding Problem". *Physica D* 42, 335–346. http://users.ecs.soton.ac.uk/harnad/Papers/Harnad/harnad90.sgproblem.html

Harnad, Stevan 2002. "Symbol Grounding and the Origin of Language". In: Matthias Scheutz (ed.) 2002. *Computationalism: New Directions*. Cambridge, MA: MIT Press, 143–158. http://users.ecs.soton.ac.uk/harnad/Papers/Harnad/harnad02.symlang.htm

Harnad, Stevan 2005. "To Cognize is to Categorize: Cognition is categorization". In: Cohen and Lefebvre (eds.), 20–43. http://cogprints.org/3027/1/catconf.html

Haugeland, John 1985. *Artificial Intelligence: The Very Idea*. Cambridge, Ma: MIT Press.

Hauser, Marc D. 1997 [1996]. *The Evolution of Communication*. Cambridge, MA: Bradford Books [Cambridge, MA: MIT Press].

Hauser, Marc D., Noam Chomsky, Tecumseh W. Fitch 2002. "The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?". *Science* 298, 1569–1579.

Heit, Evan, Lawrence W. Barsalou 1996. "The Instantiation Principle in Natural Categories". *Memory* 4 (4), 413–451.

Hermer-Vasquez, Linda, Elisabeth S. Spelke, Alla S. Katsnelson 1999. "Sources of Flexibility in Human Cognition: Dual-task Studies of Space and Language". *Cognitive Psychology* 39, 3–36.

Heyes Cecilia M. 1998. "Theory of Mind in Nonhuman Primates". *Behavioral and Brain Science*s 21 (1), 101–134.

Holyoak, Keith J. "Psychology". In: Wilson and Keil (eds.), xxxix–li.

Hookway, Christopher, Donald Peterson (eds.) 1993. *Philosophy and the Cognitive Sciences*. Cambridge, UK: Cambridge University Press.

Hyde, Dominic 2005. "Sorites Paradox". In: Zalta (ed.). http://plato.stanford.edu/entries/sorites-paradox ED 12/2007

Jackendoff, Ray 1990 [1983]. *Semantics and Cognition*. Cambridge, MA – London: MIT Press.

Jackendoff, Ray 1992 [1987]. *Consciousness and the Computational Mind*. Cambridge, MA: MIT Press.

Jackendoff, Ray 1996. "Semantics and Cognition". In: Shalom Lappin (ed.) 1996. *The Handbook of Contemporary Semantic Theory*. Oxford: Blackwell, 539–559.

Jackendoff, Ray 1997. *The Architecture of the Language Faculty*. Cambridge, MA, London: MIT Press.

Jackendoff, Ray 1999 [1989]. “What is a Concept, That a Person May Grasp It?”. In: Margolis and Laurence (eds.), 305–333. [*Mind and Language* 4, 68–102].

Jackendoff, Ray 2002. *Foundations of Language. Brain, Meaning, Grammar, Evolution*. New York: Oxford University Press.

Jackendoff, Ray, Steven Pinker 2005. “The Nature of the Language Faculty and Its Implications for Evolution of Language (Reply to Fitch, Hauser, and Chomsky)”. *Cognition* 97 (2), 211–225.

Jackson, Frank 1982. “Epiphenomenal Qualia”. *Philosophical Quarterly* 32, 127–36.

Johnson-Laird, Philip N. 1980. “Mental Models in Cognitive Science”. *Cognitive Science* 4, 71–115.

Jowett, Benjamin. 1902. *The Dialogues of Plato*. New York: Charles Scribner's Sons.

Kalisz, Roman 1998. “Is It Possible to Operate with Primitives in Every Explanation?”. In: Lewandowska-Tomaszczyk (ed.), 55–64.

Kalisz, Roman 2001. *Językoznawstwo kognitywne w świetle językoznawstwa funkcjonalnego* [Cognitive linguistics in the light of functional linguistics]. Gdańsk: Wydawnictwo Uniwersytetu Gdańskiego.

Kalisz, Roman, Wojciech Kubiński, Andrzej Buller 1996. *In Serach of a Frame of Mind. An Introduction to Cognitive Linguistics and Artificial Intelligence*. Gdańsk: Wydawnictwo Uniwersytetu Gdańskiego.

Kant, Immanuel 2003 [1781]. *Critique of Pure Reason*. Transl. J.M.D. Meiklejohn. Online edition – The Project Gutenberg. In: http://www.gutenberg.org/dirs/etext03/cprrn10.txt ED 02/2006. [*Kritik der reinen Vernunft*. Riga: Johann Friedrich Hartknoch]

Kardela, Henryk 2006a. Metodologia językoznawstwa kognitywnego [The methodology of cognitive linguistics]. In: Stalmaszczyk (ed.), 196–233.

Kardela, Henryk 2006b. “Pojęcie reprezentacji we współczesnym językoznawstwie”. [The notion of representation in contemporary linguistics]. *Kognitywistyka i Media w Edukacji* 8 (1–2), 137–166.

Karmiloff-Smith, Annette 1994. “Precis of *Beyond modularity: A developmental perspective on Cognitive Science*”. *Behavioral and Brain Sciences* 17 (4), 693–745.

Keil, Frank C., W. Carter Smith, Daniel J. Simons, Daniel T. Levin 1998. “Two Dogmas of Conceptual Empiricism: Implications for Hybrid Models of the Structure of Knowledge”. *Cognition* 65 (2-3), 103–135.

Keil, Frank C., Robert A. Wilson 2000. “The Concept Concept: The Wayward Path of Cognitive Science. Review of Fodor's *Concepts: Where Cognitive Science Went Wrong*.” *Mind and Language* 15, 308–318.

Kelly, George Alexander 1955. *The Psychology of Personal Constructs Volume 1. Theory and Personality*. New York: Norton.

Klawiter, Andrzej. 2004. "Powab i moc wyjaśniająca kognitywistyki" [Charm and explanatory power of Cognitive Science]. *Nauka* [Science] 3, 101–120. http://www.staff.amu.edu.pl/~uampsych/kognitywistyka/powab.pdf

Kornblith, Hilary (ed.) 2001. *Epistemology: Internalism and Externalism*. Oxford: Blackwell Publishing.

Labov, William. 2004 [1973]. "The Boundaries of Words and their Meanings". In: Bas Aarts, David Denison, Evelien Keizer, Gergana Popova (eds.) 2004. *Fuzzy Grammar: A Reader*. Oxford: Oxford University Press, 67–90 [Charles-James N. Bailey, Roger Shuy (eds.) 1973. *New Ways of Analyzing Variation in English*. Washington, DC: Georgetown University Press, 340–373].

Lakoff, George 1990 [1987]. *Women, Fire, and Dangerous Things. What Categories Reveal about the Mind*. London – Chicago: The University of Chicago Press.

Lakoff, George, Mark Johnson 1999. *Philosophy in the Flesh. The Embodied Mind and its Challenge to Western Thought*. New York, NY: Basic Books.

Landau Barbara, Linda Smith, Susan Jones 1998. "Object Perception and Object Naming in Early Development". *Trends in Cognitive Sciences* 2 (1), 19–24.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar. Vol I: Theoretical Prerequisites*. Stanford: Stanford University Press.

Lau, Joe 2006. "Externalism about Mental Content". In: Zalta (ed.) http://plato.stanford.edu/entries/content–externalism/ ED 10/2006.

Laurence, Stephen, Eric Margolis 1999. "Concepts and Cognitive Science". In: Margolis and Laurence (eds.), 3–82.

Laurence, Stephen, Eric Margolis 2002. "Radical Concept Nativism". *Cognition* 86, 25–55.

Laurence, Stephen, Eric Margolis 2007. "The Ontology Concepts – Abstract Objects or Mental Representations?" *Noûs* 41 (4), 561–93. http://www.philosophy.dept.shef.ac.uk/papers/OntologyofConcepts.pdf

Levine, Alexander, Mark H. Bickhard 1999. "Concepts: Where Fodor Went Wrong". *Philosophical Psychology*, 12 (1), 5–23.

Lewandowska-Tomaszczyk, Barbara (ed.) 1998. *Lexical Semantics, Cognition and Philosophy*. Łódź: Łódź University Press.

Libet, Benjamin 1999. "Do We Have Free Will?". *Journal of Consciousness Studies* 6 (8-9), 47–57. http://m0134.fmg.uva.nl/publications_others/BLfreewill.pdf

Locke, John. 1999 [1690]. *An Essay Concerning Human Understanding*. University Park, Pa.: Pennsylvania State University Press. [London: Thomas Bassett].

Lupyan, Gary 2006. "Labels Facilitate Learning of Novel Categories". In: Angelo Cangelosi, Andrew D.M. Smith, Kenny Smith (eds.) 2006. *The Evolution of Language. Proceedings of the 6th International Conference (EVOLANG6)*. Singapore: World Scientific Publishing.

Lyons, John 1996 [1977]. *Semantics*. Vol 1. Cambridge: Cambridge University Press.

McCloskey, Michael E., Sam Glucksberg 1978. "Natural Categories: Well Defined or Fuzzy Sets?" *Memory and Cognition* 6 (4), 462–472.

Macphail, Euan 1998. *The Evolution of Consciousness*. Oxford: Oxford University Press.

Mandik, Pete, Chris Eliasmith 2006. "Concept". In: Chris Eliasmith (ed.) 2006. *Dictionary of Philosophy of Mind.* In: http://www.artsci.wustl.edu/~philos/MindDict/index.html ED 03/2004

Marciszewski, Witold (ed.) 1970. *Mała encyklopedia logiki* [A concise encyclopaedia of logic]. Wrocław – Warszawa – Kraków: Zakład Narodowy Imienia Ossolińskich.

Marcus, Gary F., Simon E. Fisher 2003. "FOXP2 in Focus: What Can Genes Tell us About Speech and Language?" *Trends in Cognitive Sciences* 7 (6), 257–262.

Margolis, Eric, Stephen Laurence (eds.) 1999. *Concepts: Core Readings*. Cambridge, MA: MIT Press.

Margolis, Eric, Stephen Laurence 2006. "Concepts". In: Edward N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (Spring 2006 Edition) http://plato.stanford.edu/archives/spr2006/entries/concepts/ ED 03/2006

Markman, Arthur B. 1999. *Knowledge Representation*. London – Mahwah – New Jersey: Lawrence Erlbaum Associates, Publishers.

Martinich, Aloysius, David Sosa (eds.) 2001. *Analytic Philosophy: An Anthology*. Oxford: Blackwell Publishers.

Medin, Douglas L. 1998. "Concepts and Conceptual Structure". In: Paul Thagard (ed.) 1998. *Mind Readings. Introductory Selections on Cognitive Science*. Cambridge, MA: MIT Press, 93–126.

Medin, Douglas L., Marguerite M. Schaffer 1978. "Context Theory of Classification Learning." *Psychological Review* 85 (3), 207–238.

Medin, Douglas L., Edward E. Smith 1984. "Concepts and Concept Formation". *Annual Review of Psychology* 35, 113–138.

Medin, Douglas L., Andrew Ortony 1989. “Psychological Essentialism”. In: Stella Vosniadou, Andrew Ortony (eds.) 1989. *Similarity and Analogical Reasoning*. New York: Cambridge University Press, 179–196.

Medin, Douglas. L., Robert L. Goldstone 1995. “The Predicates of Similarity”. In Cristina Cacciari (ed.) 1995. *Similarity in Language, Thought, and Perception*. Brussels: BREPOL, 83–110.

Medin Douglas L., Cynthia Aguilar 1999. “Categorization”. In: Wilson and Keil (eds.), 104–105.

Medin, Douglas L., Elisabeth Lynch, Karen O. Solomon 2000. “Are There Kinds of Concepts?”. *Annual Review of Psychology* 51, 121–147.

Medin, Douglas L., Brian H. Ross, Arthur B. Markman 2001 [1992]. *Cognitive Psychology*. 3rd Edition. Orlando, FL: Hartcourt, Inc.

Medin, Douglas L., Rips, Lance J. 2005. “Concepts and Categories: Memory, Meaning, and Metaphysics”. In: Keith J. Holyoak, Robert G. Morrison (eds.) 2005. *The Cambridge Handbook of Thinking and Reasoning*. Cambridge: Cambridge University Press, 37–72. http://www.psych.northwestern.edu/~medin/publications/Rips,%20Medin%202005%20Concepts-categories.pdf [preprint]

Mervis, Carolyn B., Eleanor Rosch 1981. “Categorization of Natural Objects”. *Annual Review of Psychology* 32, 89–115.

Millar, Alan 1994. “Concepts”. In: Ronald E. Asher, J. M. J Simpson (eds.), *The Encyclopedia of Language and Linguistics*. Vol 4. Oxford: Pergamon Press.

Miller, George A. 1956. “The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information”. *The Psychological Review* 63, 81–97. http://www.well.com/user/smalin/miller.html

Miller, George A. 2003. "The Cognitive Revolution: a Historical Perspective". *Trends in Cognitive Sciences* 7 (3), 141–144.

Millikan, Ruth G. 2000. *On Clear and Confused Ideas: An Essay about Substance Concepts*. Cambridge: Cambridge University Press.

Murphy, George L. 2002. *The Big Book of Concepts*. Cambridge, MA: The MIT Press.

Muszyński, Zbysław 2006. “Założenia Filozoficzne w Koncepcjach Językoznawczych” [Philosophical assumptions in linguistic conceptions]. In: Stalmaszczyk (ed.), 38–56.

Nęcka, Edward, Jarosław Orzechowski, Błażej Szymura 2006. *Psychologia poznawcza* [Cognitive psychology]. Warszawa: Wydawnictwo Naukowe PWN, ACADEMICA Wydawnictwo SWPS.

Nosofsky, Robert M., Thomas J. Palmeri 1997. "An Exemplar-Based Random Walk Model of Speeded Classification." *The Psychological Review* 104 (2), 266–300.

Ogden, Charles Kay, Ivor Armstrong Richards 1969 [1923]. *The Meaning of Meaning. A Study in The Influence of Language upon Thought and of The Science of Symbolism*. 10th edition. London: Routledge and Kegan Paul.

Osherson, Daniel, Edward E. Smith 1999. [1981]. "On the Adequacy of Prototype Theory as a Theory of Concepts". In: Margolis and Laurence (eds.), 261–278. [*Cognition* 9, 35–58].

Oxford English Dictionary on CD-ROM, version 3.0. 2nd edition. 2002. Oxford: Oxford University Press

Palmeri, Thomas J. 1997. "An Exemplar-based Random Walk Model of Perceptual Categorization". In: Michael Ramscar, Ulrike Hahn, Emilios Cambouropoulos, Helen Pain (eds.) 1997. *Proceedings of the Interdisciplinary Workshop on Similarity and Categorisation*, Edinburgh: University of Edinburgh Press, 181–187.

Pawelec, Andrzej 2005. *Znaczenie ucieleśnione. Propozycje kręgu Lakoffa* [Embodied meaning. The proposals of G. Lakoff and his followers]. Kraków: TAiWPN Universitas.

Peacocke, Christopher 1995. *A Study of Concepts*. Cambridge: MIT Press.

Piłat, Robert 2007. *O istocie pojęć* [On the nature of concepts]. Warszawa: Wydawnictwo IFiS PAN.

Pinker, Steven 1995 [1994]. *The Language Instinct: The New Science of Language and Mind*. Harmondsworth: Penguin Books. [New York: William Morrow.]

Pinker, Steven 1997. *How the Mind Works*. New York: WW Norton and Company.

Pinker, Steven, Alan Prince 1999 [1996]. "The Nature of Human Concepts: Evidence from an Unusual Source". In: Ray Jackendoff, Paul Bloom, and Karen Wynn (eds.) 1999. *Language, Logic, and Concepts: Essays in Memory of John Macnamara*. Cambridge, MA: MIT Press, 221–261 [*Communication and Cognition* 29, 307–361].

Pinker, Steven, Ray Jackendoff 2005. "The Faculty of Language: What's Special about it?" *Cognition* 95(2), 201–236.

Pinker, Steven, Jacques Mehler (eds.) 1988. *Connections and Symbols*. Cambridge, MA: MIT Press.

Pitt, David 2006. "Mental Representation". In: Zalta (ed.) http://plato.stanford.edu/entries/mental-representation/ ED 05/2006 http://plato.stanford.edu/archives/win2005/entries/mental-representation/ [permanent link]

Plag, Ingo 2006. "Productivity". In: Brown (ed.), Vol. 10, 121–128.

Podsiad, Antoni 2000. *Słownik terminów i pojęć filozoficznych* [A dictionary of philosophical terms and notions]. Warszawa: Instytut Wydawniczy Pax.

Popper, Karl Raimund 1978. Three worlds. The Tanner Lecture on Human Values. Delivered at The University of Michigan, April 7, 1978. http://www.tannerlectures.utah.edu/lectures/documents/popper80.pdf

Popper, Karl Raimund 1972. *Objective Knowledge: An Evolutionary Approach*. Oxford: The Clarendon Press.

Port, Robert F., Timothy van Gelder (eds.) 1995. *Mind As Motion – Explorations in the Dynamics of Cognition*. Cambridge, MA: MIT Press.

Prinz, Jesse 2006. "Concepts". In: Borchert (ed.), 414–420.

Prinz, Jesse, Andy Clark 2004. "Putting Concepts to Work: Some Thoughts for the Twentyfirst Century". *Mind and Language* 19, 57–69.

Pulvermüller, Friedemann 2002. *The Neuroscience of Language: on Brain Circuits of Words and Serial Order*. Cambridge, UK: Cambridge University Press.

Putnam, Hilary 1975 [1960]. "Minds and Machines". In: Hilary Putnam 1975. *Mind, Language, and Reality: Philosophical Papers vol. 2*. Cambridge, UK: Cambridge University Press, 362–385. [In: Sidney Hook (ed.) 1960. *Dimensions of Mind*. New York: New York University Press, 148–180]

Putnam, Hilary 1981. *Reason, Truth, and History*. New York: Cambridge University Press.

Putnam, Hilary 1995 [1981]. "Brains in a Vat". In: Hilary Putnam 1995 [1981]. *Reason, Truth and History*. Cambridge, UK: Cambridge University Press, 1–21.

Putnam, Hilary 1997 [1975]. "The Meaning of <meaning>". In: Hilary Putnam 1997 [1975]. *Mind, Language and Reality. Philosophical Papers, Volume 2*. Cambridge, UK: Cambridge University Press, 215–271. [In: Keith Gunderson (ed.) 1975. *Language, Mind and Knowledge. Minnesota Studies in the Philosophy of Science* VII, 358–398].

Quine, Willard Van Orman 1961 [1951]. "Two Dogmas of Empiricism". In: Quine, Willard Van Orman 1961. *From a Logical Point of View*. Second edition. [*The Philosophical Review* 60 (1), 20–43].

Rapaport, William J. 1996. "Cognitive Science". In: http://www.cse.buffalo.edu/%7Erapaport/Papers/ http://www.cse.buffalo.edu/%7Erapaport/Papers/cogsci.pdf ED 05/2006

Rey, Georges 1998. "Concepts". In: Craig and Floridi (eds.).

Rijk, Lambertus M., de 1988. "'Categorization' as a Key Notion in Ancient and Medieval Semantics". *Vivarium* 26, 1–18.

Rizzolatti Giacomo, Laila Craighero 2004. "The mirror-neuron system". *Annual Review of* Neuroscience 27, 169–192.

Rolewski, Jarosław 2002. *Nowa metafizyka Kanta* [The new metaphysics of Kant]. Nowa Wieś: Wydawnictwo Rolewski.

Rosch, Eleanor 1988a [1978]. "Principles of Categorization". In: Allan Collins, Edward E. Smith (eds.) 1988. *Readings in Cognitive Science, a Perspective from Psychology and Artificial Intelligence*. San Mateo: Kaufmann Publishers, 312–322. [In: Eleanor Rosch, Barbara B. Lloyd (eds.) 1978. *Cognition and Categorization*. Hillsdale, NJ: Lawrence Erlbaum, 27–48.

Rosch, Eleanor 1988b. "Coherences and Categorization: a Historical View". In: Frank S. Kessel (ed.) 1988. *The Development of Language and Language Researchers: Essays in Honor of Roger Brown*. Hillsdale: Lawrence Erlbaum Associates, 373–392.

Rosch, Eleanor 1999. "Reclaiming Concepts". *Journal of Consciousness Studies* 6 (11-12), 61–77.

Rosch, Eleanor, Carolyn B. Mervis, Wayne D. Gray, David M. Johnson, Penny Boyes-Braem 1976. "Basic Objects in Natural Categories". *Cognitive Psychology* 8 (3), 382–439.

Rosch, Eleanor, Carolyn B. Mervis 1996 [1975]. "Family Resemblances: Studies in the Internal Structure of Categories". In: Heimar Geirsson, Michael Losonsky (eds.) 1996. *Readings in Language and Mind*. Oxford: Blackwell Publishing, 442–460. [*Cognitive Psychology* 7 (4), 573–605.]

Rundle, Bede 1995. "Concept". In: Ted Honderich (ed.) 1995. *The Oxford Companion to Philosophy*. New York: Oxford University Press.

Ryle, Gilbert 1951 [1949]. *The Concept of Mind*. London – New York – Melbourne – Sydney – Cape Town: Hutchinson House.

Schulte, Gwendolyn 1997. *An Interdisciplinary Perspective on the Cognitive Meaning of Linguistic Metaphor, its Interpretation and Computational Representation*. Marburg: Philipps-Universtat Marburg.

Schyns, Philippe G. 1997. "Categories and Percepts: a Bi-directionnal Framework for Categorization". *Trends in Cognitive Sciences* 1 (5), 183–189.

Scott, Sam 2006. "Cognitive Science and Philosophy of Language". In: Keith Brown (ed.), 552–562.

Searle, John R. 1980. "Minds, Brains and Programs". *Behavioral and Brain Sciences* 3 (3), 417–457.
http://www.bbsonline.org/documents/a/00/00/04/84/bbs00000484-00/bbs.searle2.html

Searle, John R. 1995. *The Construction of Social Reality*. New York: Free Press.

Segal, Gabriel M. A. 2000. *Slim Book about Narrow Content*. Cambridge, Ma: MIT Press.

Seyfarth, Robert M., Dorothy L. Cheney 2001. "Cognitive Strategies and the Representation of Social Relationships by Monkeys". In: Jeffrey A. French, Alan Kamil, Daniel Leger (eds.) *Evolutionary Psychology and Motivation, Nebraska Symposium on Motivation, vol. 48*. Lincoln: University of Nebraska Press,145–178.

Shanks, David R. 2000 [1996]. "Learning and Memory". In: David Green (ed.) 2000 [1996]. *Cognitive Science. An Introduction*. Oxford: Blackwell Publishers, 276–309.

Sigala, Natasha, Fabrizio Gabbiani, Nikos K. Logothetis 2002. "Visual Categorization and Object Representation in Monkeys and Humans". *Journal of Cognitive Neuroscience* 14 (2), 187–198.

Skrzypczak, Waldemar 2006. *Analog-Based Modelling of Meaning Representations in English*. Toruń: Wydawnictwo Uniwersytetu Mikołaja Kopernika.

Sloman, Aaron 1993. "The Mind as a Control System". In: Hookway and Peterson (eds.), 69–110.

Smith, Edward E., Andrea L. Patalano, John Jonides 1998. "Alternative Strategies of Categorization". *Cognition* 65, 167–196.

Soja, Nancy N., Susan Carey, Elisabeth S. Spelke 1993 [1991]. "Ontological Categories Guide Young Children's Induction of Word Meaning: Object Terms and Substance Terms". In: Goldman (ed.), 461–480. [*Cognition* 38, 179–211].

Solomon, Karen O., Douglas L. Medin, Elisabeth Lynch 1999. "Concepts Do More than Categorize". *Trends in Cognitive Science* 3 (3), 99–105.

Stalmaszczyk, Piotr J. (ed.) 2006. *Metodologie językoznawstwa. Podstawy teoretyczne* [Methodologies of linguistics. Theoretical foundations]. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Stillings, Neil A., Steven E. Weisler, Christopher H. Chase, Mark H. Feinstein, Jay L. Garfield, Edwina L. Rissland 1995. *Cognitive Science: An Introduction*. Second Edition. Cambridge, Ma: Cambridge University Press.

Storms, Gert, Paul De Boeck, Wim Ruts 2000. "Prototype and Exemplar-Based Information in Natural Language Categories". *Journal of Memory and Language* 42 (1), 51–73.

Storms, Gert 2004. "Exemplar Models In The Study Of Natural Language Concepts". In: Brian H. Ross (ed.) 2004. *The Psychology of Learning and Motivation: Advances in Research and Theory. Volume 45*. New York: Academic Press, 1–39.

Szwedek, Aleksander 2000. "Senses, Perception and Metaphors (of OBJECT and OBJECTIFICATION)". In: Stanisław Puppel and Katarzyna Dziubalska-Kołaczyk

(eds.) 2000. *Multis Vocibus de Lingua*. Poznań: Dziekan Wydziału Neofilologii Uniwersytetu im. Adama Mickiewicza, 143–153.

Szwedek, Aleksander 2002. "Objectification: From Object Perception To Metaphor Creation". In: Barbara Lewandowska-Tomaszczyk, Kamila Turewicz (eds.) 2002. *Cognitive Linguistics To-day*. Frankfurt am Main: Peter Lang, 159–175.

Tabakowska, Elżbieta 2001. *Kognitywne podstawy języka i językoznawstwa* [*Cognitive exploration of language and linguistics*]. Kraków: Universitas.

Talmy, Leonard 2000. *Towards a Cognitive Semantics. Vol. I. Concept Structuring Systems. Vol II. Typology and Process in Concept Structuring*. Cambridge, MA: MIT Press.

Tatarkiewicz, Władysław 2003 [1931]. *Historia filozofii.Tomy I-III*. [History of philosophy. Volumes 1-3]. Warszawa: Wydawnictwo Naukowe PWN. [Lwów: Wydawnictwo Zakładu Narodowego imienia Ossolińskich].

Taylor, John R.1995 [1989]. *Linguistic Categorization. Prototypes in Linguistic Theory*. 2nd edition. Oxford: Clarendon Press.

Thagard, Paul. 2006. "Cognitive Science" In: Zalta (ed.) http://plato.stanford.edu/entries/cognitive-science/ ED 05/2006 http://plato.stanford.edu/archives/win2004/entries/cognitive-science/ [permanent link]

Thompson, Roger K.R., Sarah T. Boysen, David L. Oden 1997. "Language-Naive Chimpanzees (*Pan troglodytes*) Judge Relations Between Relations in a Conceptual Matching-to-Sample Task". *Journal of Experimental Psychology: Animal Behavior Processes* 23 (1), 31–43.

Todd, Peter M., Gerd Gigerenzer. 2000. "The Precis of: *Simple Heuristics That Make Us Smart*, By Gerd Gigerenzer, Peter M. Todd and the ABC Research Group". *Behavioral and Brain Sciences* 23 (5), 727–780.

Turing, Alan M. 1950. "Computing Machinery and Intelligence". *Mind* 59 (236), 433–460. http://cogprints.org/499/00/turing.html

Tversky, Amos 2004 [1977]. "Features of Similarity". In: Eldar Shafir (ed.) 2004. *Preference, Belie, and Similarity. Selected Writings. Amos Tversky*. Cambridge, MA: The MIT Press, 7–45. [*Psychological Review* 84, 327–352.]

Van Loocke, Philip 1999. "The Structure and Representation of Concepts". In: Philip Van Loocke (ed.) 1999. *The Nature of Concepts; Evolution, Structure and Representation*. London, New York: Routledge, 1–7.

Varela, Francisco J., Evan Thompson, Eleanor Rosch 1999 [1991]. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

Wacewicz, Sławomir 2008. "Language and Thinking: Motives of Pinker's Criticism of Whorfian Linguistic Relativism". *Acta Universitatis Nicolai Copernici – English Studies* XV, 111–122.

Wacewicz, Sławomir 2011. "Concepts As Correlates Of Lexical Items". In: Stalmaszczyk, P. (ed.) *Turning Points in the Philosophy of Language and Linguistics*. Frankfurt am Main: Peter Lang, 201–212.

Wacewicz, Sławomir 2012. "The Narrow Faculty Of Language: What Is It, Who Has It, And How Is It Defined?". *Theoria et Historia Scientiarum* 9, 217-229.

Wacewicz, Sławomir, Przemysław Żywiczyński 2014. "From the narrow to the broad. Multiple perspectives on language evolution". *Theoria et Historia Scientiarum* 11, 5-18.

Waxman, Sandra R., Dana B. Markow 1995. "Words as Invitations to Form Categories: Evidence from 12- to 13-Month-Old Infants". *Cognitive Psychology* 29, 257–302.

Wąsik, Zdzisław 1987. *Semiotyczny paradygmat jezykoznawstwa. Z zagadnien metodologicznego statusu lingwistycznych teorii znaku i znaczenia* [A semiotic paradigm of linguistics. From questions about the methodological status of linguistic theories of sign and meaning]. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego (Acta Universitatis Wratislaviensis No 939. Studia Linguistica XI).

Wąsik, Zdzisław 2003. *Epistemological Perspectives on Linguistic Semiotics*. Frankfurt am Main – Berlin – Bern – Bruxelles – New York – Oxford – Wien: Peter Lang. (Polish Studies in English Language and Literature, Vol. 8).

Wąsik, Zdzisław 2006. "Investigative Perspectives in the Construction of Scientific Reality: An Epistemological Outlook on the Foundations of Linguistic Semiotics". In: Edyta Lorek-Jezińska, Teresa Siek-Piskozub, Katarzyna Więckowska (eds.) 2006. *Worlds in the Making: Constructivism and Postmodern Knowledge*. Toruń: Wydawnictwo Uniwersytetu Mikołaja Kopernika, 21–35.

Weisberg, Michael 2006. "Water is Not H_2O". In: Davis Baird, Eric Scerri, Lee McIntyre (eds.) 2006. *Philosophy of Chemistry: Synthesis of a New Discipline (Boston Studies in the Philosophy of Science, Volume 242)*. Dordrecht: Springer Publishers, 337–345.

Whorf, Benjamin L. 1997 [1956] {1940}. "Science and Linguistics" In: John B. Carroll (ed.) 1997. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. Cambridge, MA: MIT Press, 207–219. [Cambridge, MA: MIT Press.] {*Technology Review* 42, 227–231, 247–248.}

Wierzbicka, Anna 1996. *Semantics. Primes and Universals*. Oxford – New York: Oxford University Press.

Wierzbicka, Anna 1999. *Język - umysł - kultura* [Language – mind – culture]. Warszawa : Wydawnictwo Naukowe PWN. (ed. by Jerzy Bartmiński)

Williams, Joseph M. 1976. "Synaesthetic Adjectives: a Possible Law of Semantic Change". *Language* 52, 461–478.

Wilson, Robert Anton 1999. "Individualism". In: Wilson and Keil (eds.), 397–399.

Wilson, Robert Anton, Frank C. Keil (eds.) 1999. *The MIT Encyclopedia of the Cognitive Sciences*. Cambridge, MA: MIT Press.

Wittgenstein, Ludwig 1987 [1953] {1953}. *Philosophical Investigations*. Trans. G. Elisabeth M. Anscombe. Third edition. Oxford: Basil Blackwell {*Philosophische Untersuchungen*.}

Wojtak, Gerd 1998. "Meaning and Concept". In: Lewandowska-Tomaszczyk (ed.), 139–158.

Wójcik, Jan 2006. "Metody badań mózgu" [Methods of the study of the brain]. In: *Kognitywistyka.net*
http://www.kognitywistyka.net/mozg/badania.html

Wynn, Karen 1993 [1992]. "Evidence against Empiricist Accounts of the Origins of Numerical Knowledge". In: Goldman (ed.), 208–227. [*Mind and Language* 7, 315–332.]

Zalta, Edward N. (ed.) 2006. *The Stanford Encyclopedia of Philosophy* (Spring 2006 Edition). http://plato.stanford.edu ED 05/2006

Zhong, Chen-Bo, Katie Liljenquist 2006. "Washing Away Your Sins: Threatened Morality and Physical Cleansing". *Science* 313, 1451–1452.
Supporting online material: www.sciencemag.org/cgi/content/full/313/5792/1451/DC1

Zlatev, Jordan [2007] "Intersubjectivity, mimetic schemas and the emergence of language". *Intellectica* 2-3, 123-152.

Żegleń, Urszula 2003. *Filozofia umysłu. Dyskusja z naturalistycznymi koncepcjami umysłu* [Philosophy of mind. A debate with naturalistic conceptions of mind]. Toruń: Wydawnictwo Adam Marszałek.

GLOSSARY OF CENTRAL TERMS

Categorisation (Categorization)

As a process/act: the act or process of treating some input as equivalent to some other inputs, i.e. assigning it to a category. Not to be confused with category formation/acquisition.

As a theoretical topic: research area concerned with the problems of categorisation, category formation, conceptual structure, and related issues (3.2.).

Category

Traditionally: a class of entities.

In this work: a discrete mental representation that makes it possible to categorise, as well as underlies a range of cognitive processes (3.2.; 4.2.4.).

Cognitive Science

Interdisciplinary science of the mind, thinking, perception, and related phenomena. Discussed at length in Chapter 1.

Cognitive system

Any information processing system of considerable complexity, either being the mind of a living creature or equivalent to a mind (in some relevant aspects)

Conceptual structure

1. A structure whose elements are concepts
2. The internal structure of a concept

In this work, usually in the sense 2. Whenever the sense 1 is intended, this should be clear from the context or, if not, is explicitly stated in the text.

Concept

Traditionally: a non-sensory, non-imagistic representation.

In this work: a mental representation that has a lexical correlate (discussed at length in Chapter 4).

The following notation is used in this work:

- (no marking) word – the extralinsuitic entity, or the 'referent'; e.g. dogs bark
- (single quotation marks) 'word' – the linguistic symbol; e.g. 'dog' is the anagram of 'god'
- (capital letters) CONCEPT – the concept; e.g. DOG is acquired by children very early

Content

The 'content' (of a symbol or mental representation) is a technical term for the equally vague common term 'meaning' (of a symbol or mental representation).

Exemplar

A singular mental representation of a category member: either of a subordinate category (e.g. APPLE to FRUIT), an individual (of a dog to DOG), or an instance of an individual as encountered by a cognitive system (a single memory trace)

Idea

A historical predecessor of the term 'concept', usually conveying more sensory and imagistic associations.

In this work: used in the broad popular sense.

Lexical category

Traditionally: any of several grammatical classes of words; a part of speech.

In this work: a concept, i.e. a category that has a lexical correlate.

Mental representation

Traditionally: literally, mental *representation*, i.e. something that both is mental and stands for (represents) some other thing.

In this work: any relatively stable mental structure that can be consistently redeployed in cognitive operations; as such, it does not need to have a repraesentatum in the reality external to the cognitive system (1.3.2; 3.3.).

Notion

In this work: used in the popular, nontechnical sense of 'concept', or 'conception'

Prototype

In this work: a concept in the form of a summary mental representation, statistically generalising over the features of the category members