2017, 3 (1)

ARGUMENTA

The Journal of the Italian Society for Analytic Philosophy

First published 2017 by the University of Sassari

Produced and designed for digital publication by the *Argumenta* Staff

Argumenta is the official journal of the Italian Society for Analytic Philosophy (SIFA). It was founded in 2014 in response to a common demand for the creation of an Italian journal explicitly devoted to the publication of high quality research in analytic philosophy. From the beginning *Argumenta* was conceived as an international journal, and has benefitted from the cooperation of some of the most distinguished Italian and non-Italian scholars in all areas of analytic philosophy.

Contents

Editorial

The first issue of the third volume of *Argumenta* opens with the Special Issue *A Hundred Year of Donald Davidson*, edited by Cristina Amoretti, Mario De Caro and Francesca Ervas. The great American philosopher was born in 1917, and the extent and depth of his reflection, together with the powerful twist he gave to the philosophical discussion worldwide, deserve to be celebrated with the due critical consideration and dispassionate interest.

To this end, some of the leading scholars of Davidson's thought active today have been gathered by the editors of the Special Issue and present their own appraisal of the aspects of Davidson's philosophy they are best acquainted with, stretching their analysis from radical interpretation to the theory of meaning, from decision theory to the theory of action, from ethics to meta-ethics, from the conception of normativity to that of rationality, passing through questions involving metaphor, holism, naturalism and triangulation. It is the conviction of the members of the Editorial Board that the profundity and originality of the articles making up this Special Issue will mark a contribution to a field of thought destined to represent a point of reference for generations of scholars to come.

After the Special Issue, we present an article on the interpretation John Searle has given of J.L. Austin's view on truth, written by Odai Al Zoubi. The criticism the author deploys of Searle's claims adds an interesting and original piece to the understanding of the prominent English philosopher of language.

The section of Book Reviews then rounds off once again the number. We are proud to offer readers three other thoughtful reviews on as many interesting books.

Argumenta 3, 1 (2017): 3-4
ISSN 2465-2334

As usual, all the articles appearing in *Argumenta* are freely accessible and freely downloadable. Heartily thanking all the colleagues who have acted as referees, we wish everybody:

Buona lettura!

Massimo Dell'Utri
Editor

Argumenta 3,1 (2017)
Special Issue

One Hundred Years of Donald Davidson

Edited by

Maria Cristina Amoretti, Mario De Caro and Francesca Ervas

The Journal of the Italian Society for Analytic Philosophy

Contents

One Hundred Years of Donald Davidson
Introduction

Maria Cristina Amoretti
University of Genova
Mario De Caro
Roma Tre University
Francesca Ervas
University of Cagliari

1. Some Hints on Davidson's Philosophy

Donald Davidson (1917-2003) is one of the few contemporary philosophers of the analytic tradition who offered significant contentious contributions to many different areas of philosophy while preserving a semi-systematic character in his writings. His output was huge, ranging from decision theory to philosophy of language, from metaphysics to philosophy of action, from philosophy of mind to epistemology. In this introduction we shall focus on a limited number of themes, which we believe better exemplify the originality of Davidson's thought.

1.1 Theory of Meaning and Radical Interpretation

According to Davidson's first theory of meaning, a satisfactory theory of meaning for natural languages (that is, a theory specifying a systematic interpretation of all the potential sentences of a specific language) must take the form of an axiomatic theory. To begin with, the meaning of the words of a certain natural language must be defined by way of a finite set of axioms. Next, the meaning of all the sentences of that language must be derived through suitable rules of inference that define a potentially infinite number of theorems of the form '*s* means that *p*'.

The most original aspect of his project is its purely extensional nature. Following his teacher Quine, Davidson rejects intentional entities, such as intentions, meanings or translations, as they would create exactly the same kinds of problems that a good theory of meaning is supposed to resolve. To pair sentences with the world without appealing to intentional entities, he then exploits the extensional notion of truth -- or more precisely, of the predicate 'is true'. He in fact believes that what a sentence of a given language means can simply be read from a theory of truth *à la* Tarski (1983), whose theorems take the form '*s* is true in *L* iff *p*'. In this way, Tarski's original project of analysing truth by way of translation and identity of meaning comes to be reversed, as Davidson assumes

Argumenta 3, 1 (2017): 7-20
ISSN 2465-2334

DOI 10.14275/2465-2334/20175.AMO

the concept of truth as primitive and then exploits it to shed light on the notion of meaning. In other words, he sees a quasi-Tarskian theory of truth as the right vehicle for deriving a satisfactory theory of meaning without running the risk of reducing meaning to truth (Davidson 1967b, 1973).

Adopting a Tarski-style theory of truth to serve as a theory of meaning, all the sentences of a natural language must be formalised in first-order logic. However, in order to formalise sentences containing certain types of adverbial modifiers into first-order logic, it is necessary to admit events into our ontology (Davidson 1967a, 1985a), conceived as unrepeatable, concrete particulars located in space and time (Davidson 1985b, 1970a). This means that, even if events play a pivotal role in Davidson's overall philosophy, they are admitted to our ontology and conceived as particulars for specific semantic reasons.

Developing a theory of meaning structured as a quasi-Tarskian theory of truth is not only important from a purely semantic point of view, it is also preliminary to shedding light on radical interpretation—that is, to understand what is necessary in order to interpret a speaker without knowing anything about her/his language. In fact, a radical interpreter should first develop a Tarski-style theory of meaning and then confirm it by appealing to the speaker's external behaviour and further empirical evidence. However, the problem is precisely that of ascertaining the empirical correctness of the theory of meaning just developed.

To begin with, a radical interpreter must discover the speaker's attitudes of 'holding true'; that is, whether she/he holds a sentence true or not in particular circumstances. Of course, holding a sentence true is already a semantic attitude, but Davidson believes that it can precede interpretation: the interpreter may know that the speaker holds a certain sentence to be true without recognising which specific truth it is (Davidson 1973). Thus, the problem is that of clarifying how radical interpretation can progress from the mere assumption that the interpreter knows that the speaker holds a certain sentence to be true. As such, a circle seems inevitable: in order to determine the meaning of a sentence, the interpreter should be able to establish what the speaker believes; however, in order to establish what the speaker believes, the interpreter should be able to determine the meaning of the sentence. In Davidson's mind, however, the interdependence of belief and meaning can be easily broken up by finding "a method for holding one factor steady while the other is studied" (Davidson 1975: 167).

Such a method is supplied by the "principle of charity", which is taken not to be an option but as a constitutive element of interpretation. In its original formulation, the principle of charity compels the interpreter to assume that, *ceteris paribus*, the speaker's beliefs are by and large true, and thus to ascribe to the speaker a great number of true beliefs (true, of course, from the standpoint of the interpreter). In this way, the circle of beliefs and meaning is broken by holding belief constant and then solving the meaning. In later works, the principle of charity is divided by Davidson into two different strands: the principle of correspondence and that of coherence. While the principle of correspondence reaffirms the idea that, *ceteris paribus*, the speaker's beliefs must be regarded as largely true, the principle of coherence ascribes attribute beliefs to the speaker so as to make her/him out to be largely rational (again, from the standpoint of the interpreter). In other words, while the principle of correspondence focuses on the empirical constraints imposed on the interpretation by the external world, the principle of coherence focuses on holistic constraints imposed on the interpretation by the basic norms of rationality. However, these two principles might

sometimes be in conflict: in certain circumstances, the overall rationality of the speaker can only be maintained by ascribing either false beliefs or beliefs very different from our own (Davidson 1984: xix), the aim being that of optimising understanding of the speaker. Thus, even if few false beliefs or very limited contradictions might sometimes be attributed to the speaker, her/his beliefs must ultimately be largely true and coherent.

1.2 Conceptual Schemes

Davidson's reflections on the principle of charity are useful to understanding his critique of the idea of a conceptual scheme (Davidson 1974). In his mind, a conceptual scheme can be identified with a set of inter-translatable languages; thus, should two languages fail to be translated one into the other, two incommensurable conceptual schemes would appear. If a conceptual scheme is identified with a set of inter-translatable languages, then it can be analysed through the notion of non-translatable languages. However, Davidson aims to show that this notion is totally empty: there is no language that cannot be translated into our own language, at least in principle.

On one hand, if our attempts to translate a speaker's language into our own language fail completely, as in the alleged case of complete incommensurability, then there would be no reason to suppose that the speaker is a rational creature endowed with a language, beliefs and others propositional thoughts. Thus, there would be no reason to suppose that she/he has a conceptual scheme. On the other hand, if our attempts to translate a speaker's language into our own language only partially fail, as in the case of incomplete incommensurability, then radical interpretation has been only partially successful, nevertheless implying that we succeeded in attributing a largely true and coherent set of beliefs to the speaker (from our own standpoint). This shared common ground is what is needed to understand local differences and errors. Because whether a particular contrast is due to a divergence between conceptual schemes or a mere difference of opinions is always blurred, the very notion of conceptual scheme loses most of its interest and plausibility, proving to be empty and senseless.

Along with the notion of conceptual scheme, Davidson also aims to get rid of the notion of empirical content, in which empirical content is taken to be some neutral content waiting to be organised and carved up by conceptual schemes. The empirical content organised by conceptual schemes is individuated in empirical experience, in the totality of sensory stimuli, in sensations, etc. The content allegedly plays a mediating role between conceptual scheme and reality, so that reality becomes related to the conceptual scheme that organises the empirical content. As every conceptual scheme organises the empirical content in a different way, "what counts as real in one system may not in another" (Davidson 1974: 183).

By rejecting the distinction between conceptual scheme and empirical content, Davidson refuses any form of conceptual relativism. Indeed, he aims to show that relativism is bound to fail on its own terms. According to Davidson, exponents of relativism who expressly theorise the existence of different and incommensurable conceptual schemes include philosophers of science like Hilary Putnam and Paul Feyerabend, historians of science like Thomas Kuhn, linguists like Edward Sapir and Benjamin Lee Whorf, and even his master Willard van Orman Quine (1951, 1960). Indeed, the rejection of the two dogmas of analyt-

ic/synthetic distinction and reductionism did not lead Quine also to eliminate the third dogma of empiricism: the dualism between the organising system and something waiting to be organised. According to Davidson, by renouncing the analytic/synthetic dualism, and consequently the distinction between sentences true by virtue of both their meaning and their empirical content and those true only by virtue of their meaning, we can still support the idea of empirical content. We could in fact affirm that *all* the utterances have an empirical content, precisely as Quinean empiricism does. According to Davidson, the *basis* of Quinean empiricism without dogmas indeed still rests in the dogma of the distinction between conceptual scheme and empirical content.

1.3 Natural Language and Communication

Davidson's first theory of meaning and his reflections about radical interpretation are intellectually deep, but they may appear far from everyday communication where the widespread presence of polysemy, metaphors, malapropisms and slips of the tongue compels the interpreter to continuously readjust her/his own hypothesis and expectations about meaning. For this reason, at some stage of his carreer, Davidson called into question the image of language he shared for a long time with most linguists and philosophers of language, concluding that "there is no such thing as what some philosophers (me included) have *called* a language" (Davidson 1994: 2).

In particular, Davidson started to believe that the theory of meaning that speakers bring to a single occasion of everyday communication (their "prior theory") differs from speaker to speaker, and also, with regard to the same speaker, over time and contexts. When the communication begins, speakers adjust their original theory of meaning to maximise their understanding of the particular interlocutors they are interacting with. The result is a transitory theory of meaning (the "passing theory"), which is eventually shared by both speakers, thus allowing mutual interpretation, but is limited to just a specific occasion of communication. This means that it is not the previous sharing of the same theory of meaning that makes communication possible (no prior theory is shared), but rather successful communication that guarantees the sharing of the same theory of meaning (the passing theory is shared on single occasions of communication).

The above reflections are also important to understanding Davidson's highly contentious claim that "there is no such thing as a language" (Davidson 1986: 174). With this claim, he attempts to argue against the idea, maintained by many linguists and philosophers, that a language is essentially characterised by a set of syntactic and semantic conventional rules gained well before the specific occasions of communications, and then shared on those occasions. The distinction between passing and prior theory shows that if we take language to be characterised by prior and conventional rules, then language cannot be shared; if we take language to be shared by both speaker and interpreter, then language cannot be characterised by prior and conventional rules. This also means that, considering the relationship between idiolects and natural languages, the former are conceptually primary while the latter are only secondary, being an overlapping of different idiolects. This, of course does not amount to saying that natural languages, conceived as abstractions built up from different idiolects, do not exist, but only that they are not conceptually primary with respect to idiolects.

To be more precise, there is also something that Davidson believes must occur prior to all occasions of communication and be shared by both interlocutors at the time of communication in order for it to be successful: that is, the speaker's intention to be interpreted as she/he actually intends and expects to be interpreted. This basic intention, without which communication would not be feasible, forces the speaker to make her-/himself interpretable as much as possible and also explains why speakers belonging to the same community tend to make their idiolects uniform (Davidson 1994).

Talking of everyday communication, special attention has been given to the meaning of metaphors (Davidson 1978). On this point, Davidson criticises the idea, shared by both interactive and conceptual theories of meaning, that metaphor is a special linguistic and cognitive phenomenon, totally different from literal expressions. Davidson paradoxically claims that the metaphorical meaning of a term is not different from its literal meaning. His paper strongly criticises conceptual reduction cognitive semantics, where literal meaning is seen as a 'superficial' consequence of conceptual, cognitive mechanisms. In light of Davidson's view, a metaphor is nothing but an image that cannot be reduced to linguistic-propositional structures. In Davidson's words, a metaphor, like a picture, "is not worth a thousand words, or any other number. Words are the wrong currency to exchange for a picture" (Davidson 1978: 46).

In recent research on pragmatics, Robyn Carston (2010) partially answers this kind of criticism by distinguishing between different processing for lexicalised and novel metaphors. In her formulation, a conceptual route to lexicalised metaphors would still exist. However, in the case of novel or literary metaphors, an alternative, 'imaginative' route is hypothesised (Carston 2010; Carston and Wearing 2011). In doing so, Carston assigns more importance to the evocative power of images in metaphor understanding, and reassesses Davidson's view in which metaphors have "no other meaning than the literal one" (Davidson 1978). The "ulterior purpose" (Davidson 2001a: 272) of a metaphor is indeed to produce an imagistic effect that is exactly due to its literal meaning (White 1996, 2001). The metaphorical interpretation would maintain the literal meaning of the metaphorically used language, which undergoes a more global pragmatic process resulting in a range of communicated affective and imagistic effects. This idea is also confirmed by experimental studies showing that in the process of metaphor interpretation, demanding attentional resources are needed in order to suppress the literal meaning (Glucksberg *et al.* 2001; Rubio Fernandez 2007).

1.4 Events, Mind, Actions

As we have seen, events conceived as particulars are primarily needed to solve some controversial aspects of natural language semantics. However, they also play a central role in Davidson's argument for anomalous monism (Davidson 1970b, 1992b, 1995a), a view about the mental that aims at offering a solution to the mind-body problem by, the one hand, preserving ontological monism and mental causality and, on the other hand, eschewing reductionism and holding mental anomalousness (this last desideratum means that contrary to orthodox anti-reductionism, such as functionalism, Davidson believes no strict psychophysical laws and no strict psychological laws involving mental events exist). The anomalousness of the mental is strictly linked to Davidson's idea that the mental is holistic and normative—two features that are obviously alien to the

physical world. If, for instance, there were strict psychophysical laws connecting mental events to physical events, then mental events could be identified without referring to their holistic and rational constraints; however, this would amount to denying the holistic and normative character of the mental.

Anomalous monism follows from the attempt to accommodate three apparently conflicting principles that Davidson sees as true: (1) the principle of (psychophysical) causal interaction; (2) the principle of the nomological character of causality; and (3) the principle of the anomalousness of the mental. According to (1), some mental events causally interact with physical events; (2) states that events related as cause and effect must be covered by strict laws, thus implying that mental events involved in causal interactions should also be subsumed by such laws; however, (3) states that there are no strict psychophysical laws and no strict psychological laws.

Despite the apparent tension, Davidson believes that the above principles can all be held simultaneously. As events are conceived as concrete particulars, located in space and time, they can admit of different descriptions. So, a distinction can be made between causal *relationships*, which connect single tokens of events, and causal *explanations*, which instead connect general types of events under certain descriptions. Now, in a monistic framework, at the extensional level every single token of a mental event must be identical to a single token of a physical event; it is actually the very same event that can then be described using both mental and physical vocabulary. Single tokens of mental events can thus be part of causal relationships by virtue of their being identical to single tokens of physical events. This means that (1) can be held at the level of event tokens. An event token that is described with a mental vocabulary can be also described, at least in principle, with a physical vocabulary; under physical descriptions, types of events can possibly be subsumed by strict physical laws, and this makes the formulation of causal *explanations* feasible. Thus, (2) is fully respected. At the same time, under mental descriptions, types of events can be subsumed by no strict psychophysical laws and no strict psychological laws, which is exactly what (3) states.

Even if anomalous monism is strongly committed to ontological monism and is compatible with the current methodologies of cognitive sciences, it is still often perceived as an anti-naturalist thesis as it denies the possibility that there could ever be a science of the mind characterized by strict psychological and psychophysical laws (even if it does not deny the possibility of non-strict laws). That being said, other authors, such as Jennifer Hornsby (1997) and John McDowell (1985), have proposed stronger readings of the anomalousness of the mental, rejecting Davidson's token identity theory and identifying the mental *only* with the level of propositional thought.

Davidson not only defends mental causation, but he also boldly maintains the causal character of reasons for action (Davidson 1963). In his view, the explanation of action through reasons is a form of causal explanation, as the reasons that truly explain actions are just causes of those actions. Looking at a specific action, there are always many possible reasons that can rationalise it in an equally suitable way. Claiming that reasons are causes makes it easy to select the real reason among the various alternatives: the real reason is that which actually caused the action.

The downside of claiming that reasons are causes is that such a claim seems in contrast with the holistic and normative character of the mental: to talk about

causes there must be strict laws, but such laws are eschewed from the mental. To reconcile causal explanation and rational explanation, Davidson again exploits his conception of events as concrete particulars. Reason and action are two events (the agent's believing and desiring such and such, and the agent's acting in a certain way) that can be differently described, admitting of both a mental and a physical description. When described with mental vocabulary, reasons can rationally explain actions but the connection between reasons and actions cannot be subsumed under any strict psychophysical or psychological law, thus preserving the holistic and normative character of the mental. However, should they be described with physical vocabulary, reasons could causally explain actions as the connection between reasons and actions physically described could be subsumed under strict physical laws. So, rational explanation can be considered a form of causal explanation inasmuch as some law-like regularities exist: even if such regularities cannot be described with the language of rationality, that is with mental vocabulary, they can still be described with physical vocabulary, at least in principle. At the same time, rational explanation remains irreducible to non-rational explanation, as no law-like regularity can be described with the language of rationality.

1.5 Triangulation

With the notion of triangulation (Davidson 1982), a situation in which two creatures mutually interact in the context of a common external world, the deep link between language and thought—which was already present in radical interpretation—possibly becomes even stronger. To explain how this is the case let us examine what conditions are, in principle, needed for having beliefs, in particular empirical beliefs—that is, beliefs concerning the external world (Davidson 1990, 1992a, 1995b, 1999, 2001b, 2001c). First, one has to understand how experience contributes to determining the content of such beliefs—content that is objective, in the sense of being true or false independently of the existence of those beliefs or the subject entertaining them. Second, the subject must be aware of the objectivity of the content, realising that what she/he believes might be false. According to Davidson, both desiderata—the determination of empirical content and the idea of objectivity—can only emerge in an intersubjective and linguistic framework that is brought about by triangulation.

In the most basic cases, what determines, at least partially, the content of an empirical belief is its *typical* cause; that is, the cause that is repeatedly associated with that content (Davidson 1999). Such a cause, however, must be accounted by with respect to its distance and width. On the one hand, the cause must be located somewhere along the causal chain from the external world to the mind; on the other, the exact portion of the world constituting *the* relevant cause must be circumscribed. According to Davidson both desiderata can in principle be obtained only within the triangular situation, where two subjects react to the same external stimuli and perceive them in a similar way. By appealing to triangulation, the distance can be accounted by putting the typical cause in the external world at the distal level where the two lines connecting each subject to the world intersect, while the width can be accounted by appealing to the sharing of each other's reaction to the external stimulus. In this way, the typical cause has been considerably narrowed down, but it still remains partially underdetermined. Only the introduction of language can solve once and

for all the underdetermination of the 'typical' cause: "A concept is defined [...] by its typical causes, given that we are already in the world of language and conceptualization" (Davidson 2001c: 124).

Exhibiting discriminatory abilities, complex as they might be, is not enough for propositional thought as the concept of objectivity is needed to say that a subject is able to classify objects and understand that what has been ascribed to a certain class may ultimately not to belong to it (Davidson 1982). In Davidson's view, the concept of error—the idea that we can be mistaken—can in principle be acquired only within the triangular situation, where two triangulating subjects can then learn to associate certain responses of the interlocutor to some relevant external stimuli and then expect the same response when the same external stimuli occur. Again, even if the triangulation is needed to create the right space for the ensuing emergence of the concept of error, and thus of objectivity, only the introduction of language can make it appear because in order to grasp the concept of objectivity, the subjects must communicate to each other the contents of their common experience (Davidson 1982, 1999).

The triangular situation not only exemplifies the social character of thought (i.e., its intersubjective dimension), but also the inextricable relationship between language and thought. Moreover, triangulation makes clear the originality of Davidson's semantic externalism, dubbed "triangular externalism". This view is quite different from other kinds of semantic externalism (Putnam 1975; Burge 1979). At first sight, it resembles a form of perceptual (or physical or causal) externalism, as meaning and content are determined by their typical external cause. However, the presence of an interpreter, a social element, is taken to be necessary to identify the relevant cause. Thus, the social element is not present in the form of social rules and convention, as in standard social externalisms, but in the form of an interpreter who enters into the process of determining the typical cause.

The notion of triangulation has received a great deal of attention in the last few years (Amoretti and Preyer 2011; Bernecker 2013; Myers and Verheggen 2016): if some authors have widely criticised it as either unable or unnecessary to determine empirical content and the concept of objectivity, others have complained particularly about the strict relationship between thought and language (Bar-On and Priselac 2011; Sinclair 2005). That being said, an interesting link between triangulation and the psychological notion of 'joint attention' has recently been highlighted (Amoretti 2013; Brink 2004; Elian 2005), yielding more naturalist readings of Davidson's triangulation.

1.6 Links to Continental Tradition

Going beyond the differences that still divide the analytic from the Continental philosophical tradition, Davidson's theses have been associated with those defended by Continental philosophers such as Jurgen Habermas (Schatzki 1986; Fultner 2011; Baynes 2016) and Hans-Georg Gadamer (Ramberg 1989; Malpas 1992; Hoy 1997). More specifically, Davidson's philosophy has been compared with that of Habermas in terms of its strong normative approach and "the rationalization of meaning and understanding" (Schatzki 1986), and is considered similar to that of Gadamer because of the relevance given to the holistic and creative nature of the interpretative act, the contextuality and flexibility of human comprehension, and the need for the constant accommodation of our theo-

ries in order to understand what the speaker means (Dreyfus 1980). Even some links with Jacques Derrida's deconstructionism have been suggested and explored, in particular his nihilism about meaning (Wheeler 2000).

Although some real affinities with certain Continental views have openly been acknowledged by Davidson himself—such as, the common dialectic method borrowed from Plato's *Philebus* (Davidson 1997)—there are also some important differences. For instance, Davidson denied the very possibility of a radical difference between speaker and interpreter in terms of conceptual schemes, which conversely is fundamental in a hermeneutic perspective such as that proposed by Gadamer. Moreover, Davidson never definitively abandoned the first formulation given to his theory of interpretation, even though in later writings (Davidson 1986, 1989) he looked at it as a "complex abstract object", a "machine" producing interpretation, and tried to provide an alternative explanation more liable to comparison with the theses of another philosophical tradition.

3. The Contributions to this Special Issue

In this final section, we offer a brief summary of the contributions to this special issue.

In his 'Truth-theoretic Semantics and Its Limits', Kirk Ludwig presents an assessment of Davidson's theory of radical interpretation on the basis of its answers to the fundamental question of the semantic notion of meaning: "What is it for words to mean what they do?" (Davidson 1984: xiii). Davidson aimed to provide the radical interpreter with the means to answer this question, i.e. a *corpus* of information necessary and sufficient to understand the speaker. A theory of meaning must give an explanation of a potentially infinite number of sentences, starting from a finite basic vocabulary and finite set of rules in order to be used by an interpreter who has finite powers. Faced with the problem of potentially unlimited linguistic productivity, Davidson turns to the 'principle of compositionality' (Davidson 1967b, 1973). Ludwig argues that the aims of a 'theory of meaning' differ from the aims of a 'meaning theory', which is instead intended to be empirically verifiable and able to give a holistic explanation of how a specific natural language works. The author concludes that for the aims of a meaning theory, this corpus of information is neither necessary nor sufficient to guarantee the comprehension of any potential utterance in the language and it is instead worth relating meaning to the roles that words and sentences play in communicative contexts (Jankovic 2014).

In their 'Davidson: Decision and Interpretation', Pol-Vincent Harnay and Pétronille Rème focus on an often neglected aspect of Davidson's work; his seminal contribution to decision theory. According to these authors, the origins and foundations of Davidson's unified theory of thought, meaning and action can be traced back to the experiments he led in Stanford during the 1950s: the 'wording effect' and the omission of meanings, which undermine decision theory as a whole, actually underlined the need to enlarge the basis of decision theory by integrating an interpretation theory that reflects mental holism more accurately. Harnay and Rème then rely on Davidson's criticisms of decision theory to shed light on the embeddedness of decision theory and interpretation theory. This approach can be particularly useful for understanding how Davidson came to develop his unified theory of thought, meaning and action as well as the overall consistency of his work.

Pascal Engel, in his 'Davidson on the objectivity of values and reasons', investigates another aspect of Davidson's philosophy that has received much less attention than others; that is, his views on ethics and meta-ethics, and in particular his argument in favour of the objectivity of moral values. His argument is based on the idea that the interpretation of desires must be holistic and presupposes a large pattern of agreement, which cannot fail to track objective truths about the values of subjects. Examining Davidson's conception of moral values in relation to what he has to say on emotions and their relations to values, Engel argues that, even assuming that the claim that values are objective can be effectively proved on the basis of constraints on interpretation, this is not enough to give us a genuine form of moral realism. Engel then suggests that a solution can be found by adopting the fitting attitude analysis of value, according to which the conditions of the correctness of emotions and attitudes must be specified, not in descriptive or factual terms, but in normative terms.

In his 'Norm and failure in mind and meaning', Akeel Bilgrami explores an apparent conflict between two of Davidson's main theses: first, the claim that normativity is constitutive of the human mind and behaviour; and, second, the idea that normativity does not constitute linguistic meaning. As human linguistic behaviour is a specific kind of human behaviour, how can the above two theses be compatible? According to Bilgrami, Davidson mistakenly understands what the first claim amounts to, as he takes normativity to enter human mentality in the form of general principles of rationality governing mental states, which are primarily causal and dispositional states. Wittgenstein, however, understands this very same claim differently, seeing mental states as *themselves* primarily normative states. The latter interpretation seems better for Bilgrami, as it can dispel the above-mentioned conflict. This can be shown by reflecting on the concept of 'failure', a concept that presupposes a norm and is thus essential to understanding the nature of norms.

Peter Pagin, in 'Radical Interpretation and Pragmatic Enrichment', reconsiders Davidson's theory of radical interpretation in light of the contemporary debate on the linguistic phenomenon of pragmatic enrichment. The focus of the paper is in fact the intermediate layer of meaning that is neither linguistically encoded nor implicated but rather explicitly said and in need of pragmatic inferential processes besides, in order to be grasped (Carston 2002; Recanati 2004, 2010). Pragmatic enrichment is due to the fact that meaning is largely underdetermined by the conventional meaning of a sentence, and as such could be a problem for radical interpretation as a speaker might hold a sentence true, not because of believing the content of the sentence in the specific context but rather as a result of a pragmatic enrichment of that content. By applying a coherence-raising account of pragmatic enrichment (Pagin 2014), the author argues that pragmatic enrichment does not constitute a problem for radical interpretation, either in upward entailing contexts because the enriched content entails the prior content, or in downward entailing contexts where enrichments tend not to occur.

In his 'Language's Dreamwork Reconsidered', Andreas Heise reconsiders Davidson's view on metaphor, according to which, in using metaphor the speaker is not conveying any message other than the literal one and metaphors are therefore not 'special' especially in terms of cognitive content. In Davidson's words, "metaphor can, like a picture or a bump on the head, make us appreciate some fact but not by standing for, or expressing, the fact" (Davidson 1978: 46). Against a traditional interpretation of Davidson's view as a defender of a non-

cognitivist theory of metaphor (Reimer and Camp 2006; Cook 2009; Lepore and Ludwig 2013), Heise argues that in later writings (such as Davidson 1986, 1993) Davidson suggests that metaphor's distinctive effect is to prompt a mental state, i.e. the mental state of 'seeing-as'. In this framework, Heise specifies the distinction between 'cognitive content' and 'non-cognitive content', and discusses the problem of whether and how it is possible to have communicative intentions toward a non-propositional content, such as the imagistic/figurative content of a metaphor.

Daniel D. Hutto and Glenda Satne, in 'Demystifying Davidson: Radical interpretation meets radical enactivism', examine two strands in Davidson's thought that apparently come into tension. On the one hand, reflections on holism and the autonomy of propositional thought seem to give way to anti-naturalism, as there would be no way for natural sciences to shed light on the relationship between the mental and the natural world. On the other, the insistence that a theory of meaning is an empirical theory pulls in the direction of naturalism, showing that the mental can be accessed by extensional tools and is thus located within the realm of nature. To solve this tension, Hutto and Satne propose to relax the conditions on how we characterise minds, and to accept the Radical Enactivist claim that minds can be intentionally directed to the world without contentfully representing it. By distinguishing *contentless* from *contentful* intentional attitudes, the connections between contentful thought and the natural world can thus become less mysterious.

In her 'Davidson's semantic externalism: From radical interpretation to triangulation', Claudine Verheggen offers a new interpretation of the journey from Davidson's early works on radical interpretation to his later works on triangulation, taking it to be continuous. In particular, she claims that Davidson's semantic externalism does not emerge first with triangulation but was already rooted in radical interpretation, and that such externalism, combining both physical and social externalism, has always been not only holistic and historical, but also social and non-reductionist. If radical interpretation establishes the broad externalist claim that the causes of an utterance play a fundamental role in determining its meaning, then triangulation explains how the relevant causes are isolated as the meaning determinants by introducing the social element. As Davidson's externalism clearly differs from orthodox ones (Putnam 1975, Burge 1979), Verheggen concludes with a brief comparison of them.[1]

References

Amoretti, M.C. and Preyer, G. (eds.) 2011, *Triangulation: From the Epistemological Point of View*, Frankfurt: Ontos Verlag.

Amoretti, M.C. 2013, "Concepts within the Model of Triangulation", *ProtoSociology*, 30, 49-62.

Bar-On, D. and Priselac, M. 2011, "Triangulation and the Beasts", in Amoretti and Preyer 2011, 121-52.

Baynes, K. 2016, *Habermas*, New York: Routledge.

[1] We warmly thank the editor of *Argumenta*, Massimo Dell'Utri, and the editorial assistants who collaborated with us in putting together this collection.

Bernecker, S. 2013, "Triangular Externalism", in Lepore and Ludwig 2013, 443-55.

Brink, I. 2004, "Joint Attention, Triangulation and Radical Interpretation: A Problem and Its Solution", *Dialectica*, 58, 179-206.

Burge, T. 1979, "Individualism and the Mental", *Midwest Studies in Philosophy*, 4, 73-122.

Carston, R. 2002, *Thoughts and Utterances: The Pragmatics of Explicit Communication*, Oxford: Blackwell.

Carston, R. 2010, "Metaphor: Ad Hoc Concepts, Literal Meaning and Mental Images", *Proceedings of the Aristotelian Society*, 110, 295-321.

Carston, R. and Wearing, C. 2011, "Metaphor, Hyperbole and Simile: A Pragmatic Approach", *Language and Cognition*, 3, 2, 283-312.

Cook, J. 2009, "Is Davidson a Gricean?", *Dialogue*, 48, 3, 557-75.

Davidson, D. 1963, "Actions, Reasons and Causes", *Journal of Philosophy*, 60, 685-700.

Davidson, D. 1967a, "The Logical Form of Action Sentences", in Rescher, N. (ed.), *The Logic of Decision and Action*, Pittsburgh: University of Pittsburgh Press, 81-95, 115-20.

Davidson, D. 1967b, "Truth and Meaning", *Synthese*, 17, 304-23.

Davidson, D. 1970a, "Events as Particulars", *Noûs*, 4, 25-32.

Davidson, D. 1970b, "Mental Events", in Foster, L. and J.W. Swanson (eds.), *Experience and Theory*, Amherts: University of Massachusetts Press, 79-101.

Davidson, D. 1973, "In Defence of Convention T", in Leblanc, H. (ed.), *Truth, Syntax and Modality*, Amsterdam: North Holland Publishing Company, 76-86.

Davidson, D. 1974, "On the Very Idea of a Conceptual Scheme", *Proceedings and Addresses of the American Philosophical Association*, 47, 5-20.

Davidson, D. 1975, "Thought and Talk", in Guttenplan, S. (ed.), *Mind and Language*, Oxford: Clarendon Press, 7-23; repr. in his *Inquires into Truth and Interpretation*, Oxford: Clarendon Press, 1984, 155-70.

Davidson, D. 1978, "What Metaphors Mean", *Critical Inquiry*, 5, 31-47.

Davidson, D. 1982,"Rational Animals", *Dialectica*, 36, 317-27.

Davidson, D. 1984, "Introduction", in his *Inquires into Truth and Interpretation*, Oxford: Clarendon Press, xiii-xx.

Davidson, D. 1985a, "Adverbs of Action", in Vermazen, B. and M. Hintikka (eds.), *Essays on Davidson: Actions and Events*, Oxford: Clarendon Press, 230-41.

Davidson, D. 1985b, "Reply to Quine on Events", in Lepore, E. and B. McLaughlin (eds.), *Action and Events. Perspectives on the Philosophy of Donald Davidson*, Oxford: Blackwell, 172-76.

Davidson, D. 1986, "A Nice Derangement of Epitaphs", in Grandy, R. and R. Warner (eds.), *Philosophical Grounds of Rationality*, Oxford: Oxford University Press, 156-74.

Davidson, D. 1989, "James Joyce and Humpty Dumpty", *Proceedings of the Norwegian Academy of Science and Letters*; repr. in French P., T.E. Uehling and H. Wettstein (eds.), *Midwest Studies in Philosophy*, Notre Dame: University of Notre Dame Press, 1991, 1-12.

Davidson, D. 1990, "Epistemology Externalised", *Análisis filósofico*, 10, 1-13; repr. in his *Subjective, Intersubjective, Objective*, Oxford: Oxford University Press, 2001, 193-203.

Davidson, D. 1992a, "The Second Person", in French, P., T.E. Uehling and H. Wettstein (eds.), *The Wittgenstein Legacy, Midwest Studies in Philosophy*, 17, Minneapolis: University of Minnesota Press, 255-67.

Davidson, D. 1992b, "Thinking Causes", in Heil, J. and A. Mele (eds.), *Mental Causation*, Oxford: Clarendon Press, 3-17.

Davidson, D. 1993, "Locating Literary Language", in Dasenbrock, R. (ed.), *Literary Theory after Davidson*, University Park: Pennsylvania State University Press, 295-308.

Davidson, D. 1994, "The Social Aspect of Language", in McGuinness, B. (ed.), *The Philosophy of Michael Dummett*, Dordrecht: Kluwer, 1-16.

Davidson, D. 1995a, "Laws and Cause", *Dialectica*, 49, 264-79.

Davidson, D. 1995b, "The Problem of Objectivity", *Tijdschrift voor Filosofie*, 57, 203-20.

Davidson, D. 1997, "Gadamer and Plato's *Philebus*", in Hahn, L. (ed.), *The Philosophy of Hans-Georg Gadamer*, Chicago: Open Court, 421-32.

Davidson, D. 1999, "The Emergence of Thought", *Erkenntnis*, 51, 7-17.

Davidson, D. 2001a, "Communication and Convention", in his *Inquiries into Truth and Interpretation*, Oxford: Clarendon Press, 265-280.

Davidson, D. 2001b, "Externalisms", in Kotatko, P., P. Pagin and G. Segal (eds.), *Interpreting Davidson*, Stanford: CSLI Publications, 1-16.

Davidson, D. 2001c, "What Thought Requires", in Branquinho, J. (ed.), *The Foundations of Cognitive Science*, Oxford: Oxford University Press, 121-32.

Dreyfus, H.L. 1980, "Holism and Hermeneutics", *Review of Metaphysics*, 34, 3-23.

Elian, N. 2005, "Joint Attention, Communication, and Mind", in Elian, N., C. Hoerl, T. McCormack and J. Roessler (eds.), *Joint Attention: Communication and Other Minds*, Oxford: Oxford University Press, 1-33.

Fultner, B. (ed.) 2011, *Jürgen Habermas: Key Concepts*, London: Acumen Press.

Glucksberg, S., Newsome, M.R., and Goldvarg, Y. 2001, "Inhibition of the Literal: Filtering Metaphor-Irrelevant Information during Metaphor Comprehension", *Memory and Symbol*, 16, 277-94.

Hornsby, J. 1997, *Simple Mindedness*, Oxford: Oxford University Press.

Hoy, D.C. 1997, "Post-cartesian Interpretation: Hans-Georg Gadamer and Donald Davidson", in Hahn, L. (ed.), *The Philosophy of Hans-Georg Gadamer*, Chicago: Open Court, 111-28.

Jankovic, M. 2014, "Communication and Shared Intention", *Philosophical Studies*, 169, 489-508.

Lepore, E. and Ludwig, K. (eds.) 2013, *A Companion to Donald Davidson*, Chichester: Wiley-Blackwell.

Malpas, J.E. 1992, *Donald Davidson and the Mirror of Meaning: Holism, Truth, Interpretation*, Cambridge: Cambridge University Press.

McDowell, J. 1985, "Functionalism and Anomalous Monism", in Lepore, E. (ed.), *Actions and Events*: *Perspectives on the Philosophy of Donald Davidson*, Oxford: Blackwell, 387-98.

Myers, R.H. and Verheggen, C. 2016, *Donald Davidson's Triangulation Argument: A Philosophical Inquiry*, New York: Routledge.

Pagin, P. 2014, "Pragmatic Enrichment as Coherence Raising", *Philosophical Studies*, 168, 59-100.

Putnam, H. 1975, "The Meaning of 'Meaning'", in Gunderson, K. (ed.), *Language, Mind and Knowledge*, Minneapolis: University of Minnesota Press, 131-93.

Quine, W.V.O. 1951, "Two Dogmas of Empiricism", *Philosophical Review*, 60, 20-43.

Quine, W.V.O. 1960, *Word and Object*, Cambridge, MA: MIT Press.

Ramberg, B.I. 1989, *Donald Davidson's Philosophy of Language*: *An Introduction*, Oxford: Blackwell.

Recanati, F. 2004, *Literal Meaning*, Cambridge: Cambridge University Press.

Recanati, F. 2010, *Truth-Conditional Pragmatics*, Oxford: Oxford University Press.

Reimer, M. and Camp, E. 2006, "Metaphor", in Lepore, E. and B.C. Smith (eds.), *The Oxford Handbook of Philosophy of Language*, Oxford: Oxford University Press, 845-63.

Rubio Fernandez, P. 2007, "Suppression in Metaphor Interpretation: Differences between Meaning Selection and Meaning Construction", *Journal of Semantics*, 24, 345-71.

Schatzki, T. 1986, "The Rationalization of Meaning and Understanding: Davidson and Habermas", *Synthese*, 69, 51-79.

Sinclair, R. 2005, "The Philosophical Significance of Triangulation: Locating Davidson's Non-Reductive Naturalism", *Metaphilosophy*, 36, 708-27.

Tarski, A. 1983, *Logic, Semantics, Metamathematics, Papers from 1923 to 1938*, Indianapolis: Hackett.

Wheeler, S.C. 2000, *Deconstruction as Analytic Philosophy*: *Cultural Memory in the Present*, Stanford: Stanford University Press.

White, R.M. 1996, *The Structure of Metaphor*, Oxford, Blackwell.

White, R.M. 2001, "Literal Meaning and 'Figurative Meaning'", *Theoria*, 67, 24-59.

Truth-theoretic Semantics and Its Limits

Kirk Ludwig

Indiana University

Abstract

This paper takes up some limitations of truth-theoretic semantics connected with the requirement that knowledge of a compositional meaning theory for a language put one in a position to understand any potential utterance in the language. I argue that associating entities, such as properties, relations, and propositions, with natural language expressions is neither necessary nor sufficient to meet this requirement. I develop an account of how a meaning theory may be formulated in terms of a body of knowledge about a recursive truth theory for a language. I consider two objections. The first is that the sort of knowledge said to suffice to enable one to use a truth theory to interpret its object language is insufficient because it fails to offer insight into semantic structure (Hoeltje 2013). I offer a response to this objection. The second is that the approach relies on antecedent competence in expressions known to be systematically related in meaning to expressions in the object language. I concede that this objection is correct and I argue that how 'that'-clauses function in explicit statements of meaning, which are our ultimate target, shows that antecedent competence in a language plays an ineliminable role in how they give us insight into meaning. I conclude that to break out of the circle of language that traditional approaches leave us in we need to relate words and sentences to the roles they are supposed to play in our communicative activities described in more fundamental terms.

Keywords: Meaning, Truth, Compositionality, Truth-theoretic Semantics.

1. Introduction

Donald Davidson was one of the most influential philosophers of the last half of the 20th century, especially in the theory of meaning and in the philosophy of mind and action. In this paper, I concentrate on a field-shaping proposal of Davidson's in the theory of meaning, arguably his most influential, namely, that insight into meaning may be best pursued by a bit of indirection, by showing how appropriate knowledge of a finitely axiomatized truth theory for a language can put one in a position both to interpret the utterance of any sentence of the language and to see how its semantically primitive constituents together with their mode of combination determines its meaning (Davidson 1965, 1967, 1970, 1973a). This project has come to be known as truth-theoretic semantics.

Argumenta 3, 1 (2017): 21-38
ISSN 2465-2334

DOI 10.14275/2465-2334/20175.LUD

My aim in this paper is to render the best account I can of the goals and methods of truth-theoretic semantics, to defend it against some objections, and to identify its limitations. Although I believe that the project I describe conforms to the main idea that Davidson had, my aim is not primarily Davidson exegesis. I want to get on the table an approach to compositional semantics for natural languages, inspired by Davidson, but extended and developed, which I think does about as much along those lines as any theory could. I believe it is Davidson's project, and I defend this in detail elsewhere (Ludwig 2015; Lepore and Ludwig 2005, 2007a, 2007b, 2011). But I want to develop and defend the project while also exploring its limitations, without getting entangled in exegetical questions.

We can distinguish two different projects in inquiries into meaning. The first is to give what I will call a theory *of* meaning, by which I mean a general account of the nature of linguistic meaning: what it is for words and sentences to be meaningful, what it is for particular words and expressions to mean what they do, and how this is bound up with our use of them. The second is to give what I will call a *meaning* theory, as opposed to a theory of meaning. A meaning theory, as I will be using the expression, is a theory for a particular language which admits of a division into a (finite number of) semantical primitives and (a possible infinity of) complex expressions, which can be stated in a finite form and which, in a sense to be made clear, enables us to specify for each of the sentences of the language what it means on the basis what its parts mean and how they are combined. A meaning theory, as I am thinking of it, is constitutively a compositional meaning theory. A further requirement on a meaning theory—the knowledge requirement—is that knowledge of what the theory states put one in a position to understand any potential utterance of a sentence of the language on the basis of understanding the contained semantical primitives and how they are combined.

Truth-theoretic semantics is sometimes said to give us no more than a translation manual does (Soames 2008). The knowledge requirement shows that the sort of theory we seek cannot be given by a translation manual, for two reasons. First, one can understand a translation manual, that is, know what it states, without understanding either language. Second, a recursive specification of a translation from one language to another need not illuminate the compositional semantic structure of either. So if truth-theoretic semantics can fulfill the knowledge requirement, its content will go significantly beyond that of a translation manual. One of the questions I raise is whether truth-theoretic semantics can meet this requirement, and more broadly whether any meaning theory can.

The two projects, that of giving a theory of meaning and giving a meaning theory, are connected. There is now a long tradition in the philosophy of language of trying to shed light on the theory of meaning by way of reflection on how to construct and to confirm meaning theories for particular languages. This is analogous to trying to shed light on the concept of truth by way of constructing truth theories for particular languages. This is exemplified in Davidson's own project in the theory of meaning, the ur-project in this tradition, in which there are discernable two elements, what I have called elsewhere the initial and the extended projects (Lepore and Ludwig 2005). The initial project focuses on how to give a compositional meaning theory for a language by way of reflection on a truth theory for the language, taking for granted knowledge of what primitives mean. The extended project aims to shed light more generally on what it is for words to mean what they do by asking how one could confirm a truth theory

for a speaker, on neutral evidence, that could be used for interpretation (Davidson 1973b, 1974, 1975, 1976).

My main focus in this paper is the initial project, that is, how to give a meaning theory for a natural language by way of constructing a truth theory for it, that is, truth-theoretic semantics. Toward the end of the paper I will come to some limitations of truth-theoretic semantics connected with the knowledge requirement, which point to the need to refocus on the theory of meaning more generally to achieve even as much insight into compositional meaning as it is the ambition of truth-theoretic semantics to provide.

The paper is organized as follows. Section 2 outlines the project of giving a meaning theory for a language and distinguishes two approaches. One takes the sentential complement 'that *p*' of the verb 'means' to refer to propositions and seeks to give a theory that directly issues in statements of what sentences mean of the form '*s* means that *p*'. The other is truth-theoretic semantics. I claim that associating propositions with every object language sentence is neither necessary nor sufficient for giving a meaning theory. That it is not necessary, for as much as can be done along the relevant lines, is to be shown in the main body of the paper. That it is not sufficient is shown toward the end of the paper in section 9, but as a kind of grace note to recognizing a limitation to the ambitions also of truth-theoretic semantics. Section 3 takes up the project of showing how appropriate knowledge about a truth theory enables one to interpret each object language sentence and understand its compositional semantic structure. The account in this section follows the main lines, with some refinements, of the account in Lepore and Ludwig (2007a), though I postpone two issues about its adequacy until sections 4 and 5, where I confront an important recent objection, and offer a response that calls for a further revision. Section 4 takes up two questions about the approach. The first is whether the sort of knowledge said in section 3 to suffice to enable one to use a truth theory to interpret its object language is sufficient (Hoeltje 2013). The second is whether ultimately (for other, more subtle reasons) the approach illuminates meaning only by relying on antecedent competence in expressions known to be systematically related in meaning to expressions in the object language. The bulk of section 4 takes up the first question. Section 5 takes up the second question and argues that there is an interesting feature of how 'that'-clauses function in explicit statements of meaning that shows that antecedent competence in a language plays an ineliminable role in how they give us insight into the meanings of the sentences they are about. Section 6 returns to the question whether assigning propositions to sentences may avoid the difficulties surveyed and argues that it rather reinforces the lesson. My conclusion is that there is a certain sense in which no propositional knowledge of a truth theory or even a more direct meaning theory suffices for understanding the object language. Section 7 is a short conclusion that suggests that to break out of the circle of language that traditional approaches leave us in we need to relate words and sentences to the roles they are supposed to play in our communicative activities described in more fundamental terms.

2. Meaning Theories

The most straightforward way to provide a meaning theory would seem to be to provide an axiom for each primitive expression in a language *L*, which (in some straightforward sense) gives its meaning, from the totality of which one could

derive formally a specification of the meaning of any sentence of the language relative to a context of utterance—a specification, moreover, grasp of which enables us to understand the sentence, while the mode of derivation enables us to understand how our understanding of the sentence rests on our understanding of its significant parts and their mode of combination. A target theorem (for a declarative sentence) could take the following form (M).

(M) For any speaker *s*, and time *t*, ϕ means in *L*, taken relative to *s* at *t*, that *p*.

We allow '*p*' to be replaced by an open sentence containing as free variables '*s*' and '*t*' which will then be bound by the initial quantifiers in (M). Call theorems of this form M-theorems. So far, so good. Now we have a divergence between at least two different approaches.

First, a natural thought is to take 'that *p*', when '*p*' is replaced by a closed sentence, to be a referring term, and see the goal as being to formulate a theory that enables us to associate with each sentence of *L* and each context of utterance a referring term of the form 'that *p*' so as to enable us to understand the sentence as uttered in that context. Let us give these referents a uniform name: propositions. For each sentence, we want to be able to derive M-theorems from axioms attaching to its primitive components. This amounts to associating a referring term with each sentence derived from axioms governing its parts. Thus, it is natural to take the axioms governing its primitive components to be relating its parts to objects in such a way that we are in a position to understand those parts, and the proposition to be a structured complex of those parts, the structure paralleling in some way the mode of combination of the primitive expression in the object language sentence. In this way we make provision for the use of the power of quantificational logic in deriving theorems from axioms and to show how the meanings of sentences depend on the meanings of their parts.

Putting aside how the technical details are to be worked out, the question arises what role the assignment of objects to expressions and sentences in the language is doing in meeting the primary goal of the meaning theory, as specified above, and in particular the knowledge requirement. The answer is that associating entities with every expression in the language is, first, not necessary in order to provide a meaning theory for a language, to the extent to which this can be done, and, second, not sufficient either.

I will take up first the claim that it is not necessary, by introducing the second approach, truth-theoretic semantics: the project of formulating a meaning theory for a language *L* (focusing for the moment on its declarative sentences) in terms of a certain body of knowledge we can have about a truth theory for the language. We will return to the second claim in a roundabout way through considering ultimately a striking limitation on the ambitions of truth-theoretic semantics.[1]

[1] Recently Greg Ray (2014) has offered a third way that takes neither Davidson's indirect route through a truth theory nor quantifies over propositions, but rather, by ascending to a meta-metalanguage, seeks to give a recursive theory that generates theorems about the truth of M-theorems (MnT-sentences, of the form '"*s* means that *p*" is true'), and then by semantic descent to arrive at M-theorems. So if one could only generate recursively the set of MnT-sentences for the language, the idea is that one would be able to grasp an explicit statement about the meaning of every object language sentence. Ray shows cleverly how to formulate a recursive theory to generate the appropriate class of MnT-sentences. But the approach, as sketched in the paper, founders on the bit of reasoning

3. Truth-theoretic Semantics

Truth-theoretic semantics aims to exploit the recursive structure of a Tarski-style truth theory in pursuit of giving a compositional meaning theory. However, a Tarski-style axiomatic truth theory is not a meaning theory. It blandly states conditions materially necessary and sufficient for the truth of sentences of an object language. Davidson's idea was that out of an extensionally adequate truth theory for a language, we can squeeze the elixir of meaning, if it has certain further properties, and we know that it does. I develop that idea in this section, in a way similar to, if not quite the same as, the way that Davidson did. For illustration, I introduce a simple axiomatic truth theory for a language *L*, [TRU], without quantifiers or context sensitivity, whose sentences are all declarative.[2] All of the central issues that concern us can be raised in connection with even this very simple theory. In the following, for convenience I use '*x* + *y* + … ' to mean 'the concatenation of *x* with *y* with … in that order'. '*N*' ranges over names and '*S*', 'S_1' and 'S_2' over sentences of the object language.

Truth theory [TRU]

1. 'Claudine' refers to Claudine.
2. 'Robert' refers to Robert.
3. For any name *N*, *N* + 'dort' is true in *L* iff what *N* refers to is sleeping.
4. For any names N_1, N_2, N_1 + 'aime' + N_2 is true in *L* iff what N_1 refers in *L* to loves what N_2 refers to in *L*.
5. For any sentence *S*, 'Ce n'est pas le cas que' + *S* is true in *L* iff it is not the case that *S* is true in *L*.
6. For any sentences S_1, S_2, S_1 + 'et'+ S_2 is true in *L* iff S_1 is true in *L* and S_2 is true in *L*.

Rules of Inference

Universal Instantiation (UI): from a universally quantified sentence any instance may be inferred.

that gets us from the MnT-sentences to the M-sentences. The reasoning is illustrated in the following:

(1) '"La neige est blanche" means that snow is white' is true.
(2) If '"La neige est blanche" means that snow is white' is true, then 'La neige est blanche' means that snow is white.
(3) Hence, 'La neige est blanche' means that snow is white.

The crucial bit in getting from what (1) states, which is output of the theory, to what (3) states, which is what we want to know, goes by way of (2), which seems simply to be an instance of semantic descent. But this presupposes that the meta-metalanguage and the metalanguage are the same, and so effectively presupposes that the theorist already understands the metalanguage. However, in general knowledge of the truth of a sentence does not give us knowledge of the language in which it is stated, and the theory that Ray presents, in the meta-metalanguage, is about sentences in the metalanguage, and could be stated in a language other than the metalanguage. Thus, knowledge of the theory (of what it states) is not sufficient for knowledge of the metalanguage in which meaning statements are expressed, which means that it does not suffice for knowledge of what (2) states. This point has been ably made now in print by Hoeltje (2016).

[2] For extensions to quantifiers and to non-declaratives see Lepore and Ludwig 2007a: Chs. 3 and 12.

Substitution (S): from any sentences of the form '*t* refers to *y*' and any sentence of the form S(what *t* refers to in *L*), S(*y*) may be inferred.
Replacement (R): S(*y*) may be inferred from '*x* iff *y*' and S(*x*).

Our criterion of adequacy for the truth theory is Tarski's Convention T, which requires that

the truth theory entails all instances of the schema (T):

(T) φ is true in *L* iff *p*

in which φ is replaced by a structural description of an object language sentence as composed out of its significant parts and '*p*' is replaced by a meta-language sentence that translates it.

A canonical truth theorem of [TRU] is any theorem derived from the axioms using only (UI), (S), and (R) whose last line is of the form (T) and in which no semantic vocabulary of the metalanguage remains (i.e., 'is true in *L*'). We call such a proof a *canonical proof*. It is clear that a canonical proof draws only on the content of the axioms in proving a canonical theorem.

What knowledge about [TRU] would enable us to know that it meets Convention T? We stipulate that

(i) in each reference axiom, the name used on the right of 'refers to' translates the name mentioned on the left.
(ii) in each predicate axiom, the predicate used in the meta-language in giving truth conditions for the object language sentence translates the object language predicate.
(iii) in each recursive axiom, the logical connective used in the meta-language to give the truth conditions for the object language sentence translates the logical connective in the object language.

We will say that if the theory meets this condition, it meets Convention A. We will say that a truth theory that meets Convention A is interpretive. It is clear that given the rules of inference and that [TRU] meets Convention A, for each sentence of the object language its canonical theorem is such that the metalanguage sentence on the right hand side translates the object language sentence for which it is used to give truth conditions, and, hence, that [TRU] satisfies Convention T.

The step to seeing how to transform these materials into a meaning theory is accomplished by restating Convention T.

The truth theory entails all instances of the schema (T)

(T) φ is true in *L* iff *p*
(M) φ means in *L* that *p*

such that the corresponding instance of schema (M) is true.

At this point, given that [TRU] satisfies Convention A, and, hence, Convention T, we can add another valid rule of inference (cf. Davidson 1970; 2001: 60):

[Transference] 'φ means in *L* that *p*' may be inferred from a canonical truth theorem of the form 'φ is true in *L* iff *p*'.

We will call any theorem so derived a canonical meaning theorem.

What this shows is that we can use [TRU] to arrive at a specification of the meaning of each sentence of the object language. This does not, however, make [TRU] a meaning theory for the object language. It is still just a truth theory. It doesn't state anything itself about what sentences in the object language mean, as opposed to the conditions under which they are true. But if it is the right sort of theory, and we know that it is, and understand it, then, it seems, we can *use it* to specify what each sentence means, in a way that makes it scrutable to us, and which, by the proof of the relevant theorem, shows how the semantical primitives in it contribute to fixing its truth condition determining meaning.

We characterized a meaning theory as a body of knowledge that puts one in a position to understand any sentence of a language on the basis of understanding the contained semantical primitives and their manner of combination in the sentence. With this in mind, we will identify the meaning theory for the object language with what body of knowledge we need to have about [TRU] that puts us in a position to use it for this purpose. In particular, if we know [K-TRU],

[K-TRU]

(i) What the axioms of [TRU] are, as stated in 1-6
(ii) What each axiom states and of each that it states what it does
(iii) That [TRU] is interpretive, i.e., meets Convention A
(iv) The rules of inference UI, S, R, and Transference
(v) What a canonical truth theorem is

then (it seems) we are in a position to infer for each sentence of the object language a meta-language sentence that explicitly states what the object language sentence means, on the basis of a proof that traces out, at each step, the contribution of each object language expression to fixing the truth conditions of the sentence to which it contributes, on the basis of reference or truth conditions given using a term synonymous with it. We can thus see what the contribution is of each semantical primitive to the interpretive truth conditions of the sentences in which it occurs on the basis of what it means. Thus, we can treat the body of knowledge characterized by [K-TRU] as a meaning theory for *L*. Here again the propositional knowledge stated in (ii) is to play the crucial role of giving us knowledge of the meaning of the axioms of the truth theory and so of the language of the metalanguage (i.e., of the truth theory). If this is correct, it shows that our meaning theory goes beyond a translation theory because one can grasp what the translation theory states without understanding either language or the compositional semantic structure of their sentences.

That completes the basic account of the goals and methods of truth-theoretic semantics. I close this section of the paper with four remarks.

(1) First, it is clear that on this way of understanding the project of truth-theoretic semantics, the truth theory is not identified with a meaning theory. The meaning theory is rather a body of knowledge about the truth theory. It is not an objection to the project then that the truth theory itself cannot do the duty of a meaning theory.

(2) Second, it is clear that the project is not to replace the investigation of meaning with the investigation of truth on the grounds that the concept of meaning is too confused for use in a properly scientific investigation of language, but rather to pursue a traditional project by means of a certain kind of indirection.

(3) Third, it is clear that it is not the goal of truth-theoretic semantics, as here conceived, to effect a kind of reduction of meaning to truth conditions, in any sense.

(4) Fourth, the approach shows (pro tem) how to achieve the aims of a meaning theory with no more ontological resources than are required for a reference theory, and, hence, it shows that the introduction of entities to assign to every significant expression of the language is not necessary in order to give a meaning theory for a language.

These are, I believe, all observations which conform to Davidson's own understanding of his project (Ludwig 2015). The distinction between the truth theory and what we can know about it that enables us to use it to interpret a language is drawn clearly in "Reply to Foster" (Davidson 1976). However, Davidson did not think that a statement of what we can know would count as a theory, as he says on the last page of "Reply to Foster", largely because of his commitment to analyzing 'states that' (or alternatively 'means that') paratactically (Davidson 1968), which is required in spelling out [K-TRU] (ii). When we put aside the commitment to the paratactic analysis, though, there would seem to be no barrier to stating what propositional knowledge would suffice explicitly. (We will return below in sections 4-6 to the question whether we have achieved everything we want.) Another departure from Davidson's own development is the introduction of Convention A as a constraint on an interpretive theory. As noted in the introduction, Davidson's project was not limited to formulating a compositional meaning theory for a language. When the project is to understand how we understand complex expressions on the basis of their components and combination, we can help ourselves to knowledge of what the primitive expressions mean. Davidson sought illumination also of what it was for primitive expressions to mean what they do. The method was to describe empirical constraints on truth theory sufficient to ensure that the theory could be used for interpretation in the form of the requirement that it be confirmable for a speaker from the standpoint of a radical interpreter—an interpreter who starts ultimately with only behavioral evidence about a speaker interacting with his environment and others. If the constraint were adequate to guarantee not merely the right outputs, but also that any theory so confirmed would provide insight into how the sentences of the language were understood on the basis of understanding of their contained primitives and mode of combination, then it would suffice for the theory to meet Convention A as well. Since my interest is in the truth theory as a vehicle for a meaning theory, I stipulate that part of what we need to know about the truth theory is that it meets Convention A and what each axiom means.

4. Is the Relevant Body of Knowledge Really Adequate?

What I have hoped to do up to this point is

(1) to explain how truth-theoretic semantics is to be understood as a pursuit of a perfectly traditional project, that of constructing meaning theories for particular languages;

(2) to show that it is mistake to suppose its work can be done by a translation theory.

Now I want to take up two important questions about it.

(i) Is the body of knowledge I have claimed to be sufficient really so?

(ii) Even if it is sufficient in some straightforward sense, and granting that it is not equivalent in any straightforward sense to a recursive translation theory, might there not yet be a sense in which the illumination of what object language expressions and sentences mean rests in part essentially not upon propositional knowledge but upon non-propositional understanding of the metalanguage, that is, antecedent competence in expressions known to be systematically related in meaning to expressions in the object language?

I address the first question in this section and the second in the next. The first question is prompted by an objection due to Miguel Hoeltje (2013), namely, that what is stated in [K-TRU] (repeated here) does not suffice to understand the truth theory.

[K-TRU]

(i) What the axioms of [TRU] are, as stated in (1)-(6)
(ii) What each axiom states and of each that it states what it does
(iii) That [TRU] is interpretive, i.e., meets Convention A
(iv) The rules of inference UI, S, R, and Transference
(v) What a canonical truth theorem is.

The meaning theory involves three levels of languages. There is the object language. There is the language of the truth theory, the metalanguage, and then there is the language in which the meaning theory is stated, the meta-metalanguage. The goal was in part to state something in the meta-metalanguage sufficient to understand the metalanguage. This is important because (a) both the canonical truth theorems and canonical meaning theorems are in the language of the truth theory and (b) the canonical proofs, which are to illuminate the compositional structure of object language sentences, are stated in the meta-metalanguage. The objection that Hoeltje makes is that (ii) in [K-TRU], which is intended to turn the trick, is not sufficient. Let's take an example of the sort of knowledge that gives us. I will vary the language of the truth theory from English to Serbian (in the Cyrillic alphabet) to bring out the point.

[*] За свако име н, н⁀"dort" је истинито у л акко оно на шта н реферира спава' means that for any name *N*, *N* + 'dort' is true in L iff what *N* refers to is sleeping.

The worry is this: how are we supposed to know anything about the relevant syntactic/semantic structure of axiom, and hence of the metalanguage sentence, from basically getting a statement of its meaning as a whole? And if we do not know that, how can this give us knowledge of the metalanguage, i.e., the language of the truth theory, which we grant is essential to using it in the way intended?

As a first remark about whether what is stated in [K-TRU] is sufficient, it is worth noting that if we have an explicit statement of the form

A means that *p*

for each of the axioms of the truth theory, then *we can state a truth theory in the meta-metalanguage for the object language*. It just consists in the list of statements that replace '*p*' in this form. We would of course also have to introduce inference rules corresponding to those for the metalanguage and define a canonical truth theorem and canonical meaning theorem for the meta-metalanguage. This is straightforward, but is not yet included in what we would know in knowing the axioms or in knowing what else is stated in [K-TRU]. This looks like it

would suffice. So just the knowledge stated in (ii) goes a long way toward what is needed, so far that just a bit more seems to gets us the rest of the way.

However, this is not how I envisioned it going. So let me return to the original idea, which was not adequately spelled out in Lepore and Ludwig (2007a). First of all, we run proofs on the truth theory as a syntactic object. So we should state the rules in the meta-metalanguage. This was not made explicit in [K-TRU]. Furthermore, the rules are stated in terms of syntactic/semantic categories that apply to metalanguage terms (names, n-ary predicates, connectives, quantificational determiners, and so on), so it presupposes a recursive syntax for the metalanguage which sorts terms into the categories necessary for the description of the structure of sentences in metalanguage in terms of the types of semantically primitive expressions in the language and how sentences are systematically built up out of them. Armed with this information, we will be able to parse the structure of the axioms of the truth theory, and given the knowledge stated in (ii), see how to interpret connectives, determiners, predicates and so on. In fact, we also omitted to include in the statement of the truth theory a statement of the corresponding information about the object language. That should be included in the statement of the truth theory, and bundled into (ii). So what we need to make explicit that we had not is that we also know a lot about the syntactic structure of the metalanguage, and that the truth theory should itself include a recursive syntax for the object language.

For example, if we know that in

За свако име н, н⁀"dort" је истинито у л акко оно на шта н реферира спава

'н' is a variable, that 'акко' is a logical connective, that 'За свако име' is a restricted quantifier, that 'За свако' is a quantificational determiner, that 'име' is a common noun, that enclosing an expression in the left and right double quotation marks forms a name, that for metalinguistic variables *v* and *u*, *v* + '⁀' + *u* is a restricted quantifier, and how the sentence is constructed out of its parts, which will determine scope relations, etc., we are in a position to interpret basically each word in it given what it means as a whole, on the basis of knowledge of its syntactic structure and the knowledge of the syntactic structure of the sentence that gives its meaning in [*] and the meaning of that sentence, which is a sentence of our meta-metalanguage. Thus, given that we know that 'За свако' is a quantificational determiner, that 'акко' is the main connective, and that 'н' is a variable, we can infer that 'За свако' means *for any*, that 'име' means *name*, that 'н' is a variable that takes names as values, that 'н⁀"dort"' means *the concatenation of N with "dort"*, that 'је истинито у л' means *is true in L*, that 'акко' means *iff*, that 'на шта н реферира' means *what N refers to* (which will be invariant across many axioms, as will 'је истинито у л акко', and 'За свако име н', which also helps us to identify the function of these phrases in the metalanguage), and finally that 'спава ' means *is sleeping*. I conclude that this challenge can be met, and [K-TRU] is to be amended in the ways indicated.

5. The Limits of Truth-theoretic Semantics

Let's turn to the second question:

(ii) Even if knowledge of the sort indicated is sufficient in some straightforward sense, and granting that it is not equivalent in any straightforward sense to a recursive translation theory, might there not yet be a sense in

> which the illumination of what object language expressions and sentences mean rests in part essentially not upon propositional knowledge but upon non-propositional understanding of the metalanguage, that is, antecedent competence in expressions known to be systematically related in meaning to expressions in the object language?

I think the answer to this question is 'yes', and that it helps to bring out something important about the limits of truth-theoretic semantics, as it has been explained here.

Let's grant knowledge of the language of the truth theory and think about how what we know about it helps us to use it to interpret object language expressions and to gain insight into the logico-semantic structure of object language sentences. First of all, knowing that the theory satisfies Convention A enables us to read off from axioms what object language expressions contribute in context to the meaning of a sentence. But the knowledge it gives us of both the content of the object language expression and its semantic role is clearly relative to our prior grasp on the corresponding content of the metalanguage expression and its semantic role. Let's take five sorts of axioms as examples. I shift to a context sensitive reference axiom and satisfaction predicate for generality.

A1. For any speaker *s*, time *t*, for any *x*, 'je' refers(*s*, *t*) to *x* iff *x* = *s*.
A2. For any function *f*, speaker *s*, time *t*, referring term *N*, *f* satisfies(*s*, *t*) *N* + 'dort' in *L* iff what *N* refers to is sleeping at *t*.
A3. For any function *f*, speaker *s*, time *t*, variable *v*, *f* satisfies(*s*, *t*) *v* + 'dort' in *L* iff *f*(*v*) is sleeping at *t*.
A4. For any function *f*, sentences S_1, S_2, *f* satisfies(*s*, *t*) S_1 + 'et'+ S_2 iff *f* satisfies(*s*, *t*) S_1 in *L* and *f* satisfies(*s*, *t*) S_2 in *L*.
A5. For any function *f*, predicate *P*, variable *v*, *f* satisfies(*s*, *t*) '(Chaque' + *v* + ')' + *P* iff every *v*-variant *f'* of *f* is such that *f'* satisfies *P*.

A1 is the most informative of these. Antecedent understanding of 'refers' tells us what the category is, and (given Conv A) we know that the clause expresses a rule of meaning for determining the referent of the expression without using an expression the same in meaning with it. But in the case of A2-A5, it is clear that understanding the content and semantic role of the object language expressions depends on our antecedent understanding of the metalanguage sentence used to give satisfaction conditions together with knowledge that the theory meets Convention A. Similarly, our understanding of the contribution of each expression to fixing interpretive truth conditions of a sentence on the basis of its meaning, as exhibited in how its axiom enters into a canonical proof of a canonical theorem for it likewise relies on our understanding of the metalanguage expression used to give satisfaction conditions for it and our knowledge that the theory satisfies Convention A.

There is a sense, then, in which the theory shows something about the object language, in light of our understanding of the metalanguage and knowledge that the theory meets Convention A, that the theory does not say. This, I think, turns out to be inescapable in giving a meaning theory of the kind we have set as our goal here. I will say more in support of this in a moment. This is not to say that the theory together with knowledge that it meets Convention A is not informative, of course, but only to say that the mode by which we gain information about the object language rests on prior understanding of metalanguage

terms and what information Convention A gives us about their relation to object language terms for which reference and satisfaction conditions are being given. This is what survives of the objection that truth-theoretic semantics could be replaced by a translation theory, namely, that it presupposes prior knowledge of a language whose terms are understood to be the same in meaning as object language terms.

But if all of this is right, if the illumination the theory provides does rest essentially in part on antecedent understanding of a language, then in what sense have we stated something in [K-TRU] that puts us in a position to understand any potential utterance of an object language sentence and to understand its compositional semantic structure? In a perfectly ordinary sense of 'body of knowledge', [K-TRU] (as modified along the line indicated at the end of the last section) does state a body of knowledge that suffices, for we can state that one needs to know that *p*, that *q*, etc., and we agree that if all of that is true of someone, then he is in a position to interpret object language sentences. But this, I think, turns out to only superficially satisfy the requirement we had in mind. For what we had in mind was that we could state a body of propositional knowledge that suffices which is essentially independent of knowledge of any language. And, on closer examination, this condition has not been satisfied.

I want to take this up in connection with the final output of the theory, which consists in explicit statements about what object language expressions mean relative to contextual parameters. For I think this is the crux of the issue, how M-sentences sentences convey to us what a sentence means, relative to a context. Consider (M-S) for example.

(M-S) For any speaker *s*, time *t*, 'Je dort' means(*s*, *t*) in *L* that *s* is sleeping at *t*.

Surely here we have got something that simply states what the sentence means! This at least does not rely on antecedent understanding of the metalanguage sentence and knowledge that it interprets the object language sentence. But *in fact it does*. We have not escaped reliance on antecedent understanding of the metalanguage.

It is easier to see this by instantiating (M) to a particular speaker and time, KL at T.

[KL] 'Je dort' means (KL, T) in *L* that KL is sleeping at *t*.

You understand the sentence used in the complement. And given that you understand it, and that you know that

> '*x* means(*s*, *t*) in *L* that *p*' is true in English iff *x* taken relative to *s* at *t* in *L* translates '*p*' in English,

you are in a position to understand the object language sentence taken relative to KL and T. But suppose that you did not understand the complement sentence. Then it is clear that you would not understand 'Je dort' taken relative to KL and T. So in fact the explicit statement of meaning relies for its effect on your understanding a sentence and understanding it to give the meaning of another sentence in a context.

To this it might be said:

> Yes, but the trouble is just that if you did not understand the complement sentence in KL, you would not grasp what proposition was expressed by the whole sentence, and so of course we would not expect you to be in a position to understand the object language sentence. This does not show at all that

> what is *stated* is not sufficient. What is stated is that the meaning is a certain proposition. And if you grasp that, you grasp the meaning of the target sentence. It does not show that the illumination comes only in the manner you sketch, i.e., via understanding of the complement sentence and knowledge that it translates the object language sentence—again, you just have to know what proposition it refers to! And while it is still true that we have to understand the complement to understand what the sentence states, that is just a reflection of the fact that to understand what someone is stating we have to understand the sentence he uses to state it.

But, as I will now argue, this response is inadequate and appealing to propositions is no help.

6. Back to Propositions

Let us take sentential complements of the form 'that *p*' to contribute a proposition to the second argument position in '*s* in *L* means *y*' (for simplicity I drop relativization to context). What kind of singular term is 'that *p*'? We seem to have three options:

1. it is a description (construed as a quantifier)
2. it is a referring term that merely introduces an object into the proposition
3. it is a term that refers via a Fregean mode of presentation.

My claim is this: no matter which of these options we take, it turns out that a true sentence of the form '*x* in *L* means *y*' enables us to understand *x* only if we can construct from '*y*' a metalanguage sentence that we understand and understand to be the same in meaning as *x*.[3]

Option 1. 'that *p*' is a description. What is the form of the description? The most natural thing to say is that, as the sentence '*p*' expresses the proposition, the description is 'the proposition expressed by "*p*" in English'. Alternatively, we could treat it as a context sensitive self-referential description: 'the proposition the speaker of this token "*p*" expresses with it'.

> φ means in *L* the proposition expressed by '*p*' in English
> φ means in *L* the proposition the speaker of this token '*p*' expresses with it.

It is clear, however, that this would not put someone in a position to understand the object language sentence unless he understands the mentioned sentence or the mentioned sentence as used by the speaker. But what if it were some non-metalinguistic description? Any proposal along these lines will be subject to the objections to option 3.

Option 2. 'that *p*' contributes only its referent to what is said. We may still, compatibly with this, take the referent to be determined by a description. We could think of it functioning like Kaplan's dthat(the *F*).

> φ means in *L* dthat(the proposition expressed by '*p*' in English).

But even so, what is said clearly does not put one in a position to understand φ independently of understanding '*p*'. For the same thing could be said using a directly referring term. Let us name the proposition expressed 'Bob'. We say the same thing then in saying:

[3] I give a parallel argument to show that reference to propositions is no help in understanding how attitude attributions help us to understand what people think in Ludwig 2014.

φ means in *L* Bob.

But clearly someone could understand what is said by this without understanding φ.

Option 3. Frege held that '*p*' in the context following 'means that' has an indirect sense that is a mode of presentation of its customary sense, i.e., the proposition expressed by '*p*'. Of course, one mode of presentation might simply be 'the proposition expressed by '*p*' in English'. But it is clear already that this will not give us understanding of the object language sentence independently of understanding '*p*' in English. What other mode of presentation might we appeal to? I venture to say that the only idea anyone has of this is given by the description of the role it is to play. But whatever it is, we know that it has to meet a certain constraint, a constraint which I think cannot be met. The constraint is that the mode of presentation must be such that one cannot present to oneself the proposition via that mode of presentation without grasping it. For otherwise, one could grasp the proposition expressed by 'φ means in *L* that *p*' without understanding φ. Grasping a proposition occurrently, as is required here, entails entertaining it. So what we require is a mode of presentation attached to 'that *p*' of which it is constitutive that one entertain its object. But the mode of presentation is distinct from what it presents, and so entertaining it is not ipso facto to entertain its object. So for no mode of presentation of an entertainable object could grasp of the mode of presentation suffice for entertaining its object.[4]

But there is an additional problem with the appeal to a non-metalinguistic mode of presentation, namely, that it would make the fact that '*p*' appears in 'that *p*' an accident of spelling. For whatever the non-metalinguistic sense that is to do the job could be attached to any arbitrary term, like 'Bob', and function in exactly the same way. But it is obvious that using a sentence we understand in the complement is crucial to the way these sentences inform us about the meaning of the sentences they are about. It is not just an accident of spelling. So even if we could make sense of a mode of presentation of a proposition that guaranteed occurrent grasp of it, this would be a fatal objection to a non-metalinguistic Fregean account: it could avoid the problem only by ignoring a central feature of the mechanism by which sentential complements do their work for us.

[4] See Ludwig 2014 for further discussion, and see especially section 5. The new cognitivist accounts of propositions championed by Soames (King, Soames, and Speaks 2014, Soames 2010, 2015) and Hanks (2015) might be thought to provide a way around this difficulty. The basic idea is that propositions are cognitive act types, of which the basic act type is that of predicating a property of an object. The relevance of this to attitude attributions is that in determining the cognitive act type to be associated with, e.g., 'Russell believed that mathematics was reducible to logic', we consider what sequence of act types are instantiated when someone understands that sentence. Since that involves understanding 'mathematics was reducible to logic', the act type is different from that involved in grasping 'Russell believed Logicism', and it explains why in grasping the former one has to entertain the proposition the 'that-clause' refers to, while that is not so with respect to the latter. But while that suffices to distinguish the propositions and explain why the former involves entertaining the proposition expressed by the sentence, it does not show that understanding the sentence in the complement is not essential to the way it functions in the language. It does not provide an account of a Fregean sense grasp of which suffices to entertain its object. It is in fact fully compatible with the view that '*s* means that *p*' conveys to us what *s* means by way of our understanding '*p*' and understanding it to translate *s*.

It is clear that as a matter of fact the way we are clued into the meaning of an object language sentence via an M-theorem for it is by way of our understanding the metalanguage sentence and the requirement that the complement sentence be the same in meaning as the object language sentence, relative to context. And it seems clear that anything you put in the second argument place in that construction will not put you in a position to understand the object language sentence except insofar as it at least codes for such a sentence. The difficulty is that understanding a sentence is not at bottom a matter of standing in a relation to an abstract object and associating it with a sentence. It is a matter of knowing how to use it for certain purposes. To convey what a sentence means we can rely on antecedent understanding of terms appropriately related in meaning, or we can explain the purpose of the sentence in the kind of activity that is central to its linguistic meaning. The second of these, however, will not take the form of relating a sentence to any object.

One could insist on a very abstract reading of mode of presentation, and allow that the mode of presentation of a proposition may simply consist in our understanding a sentence and thinking of the proposition as the one we are grasping in understanding the sentence. But this would just show that the work that sentences in the complements of M-sentences do depends essentially on our understanding them and knowing that the sentences which are the M-sentences' subjects are the same in meaning as those complement sentences.

This point extends to the various ways we have of trying to indicate the meanings of subsentential expressions by assigning them properties and relations and functions. To make meaning scrutable by assignment of an object to an expression, we must use an expression to refer to it that at least codes for an antecedently understood expression which is understood to interpret the expression whose meaning we are giving (with the except of referring terms whose referents are fully given by a rule relative to contextual parameters). This is transparent in, for example, the claim that 'rouge' in French means the property of being red, or 'aime' in French means the relation of loving.

The conclusions to reach are the following:

1. The assignment of entities to expressions in pursuit of a meaning theory for a natural language is neither necessary nor sufficient. It is not necessary because the same goals, so far as possible, can be achieved by the method of truth-theoretic semantics. It is not sufficient because assigning the correct entities alone does not put us in a position to understand object language expressions. We must assign entities using terms that code for expressions we understand and understand to interpret object language expressions to which meaning entities are being assigned.

2. A theory of meaning designed to issue in theorems of the form 'φ means in *L* that *p*' does its work inevitably in part by way of showing us something about what object language expressions mean by way of our antecedent understanding of metalanguage terms understood to interpret them. There is, thus, a limit to the depth of the illumination of meaning in the object language we can expect. We have nominally specified a body of propositional knowledge sufficient to understand any sentence in a language, but on closer examination, antecedent understanding of sentences understood to be alike in meaning with object language sentences plays a crucial role. To get greater illumination, we have to move to a different form of account. (This is obvi-

ously connected with the dispute between Davidson and Dummett (Dummett 1975, 1976) over whether a modest theory of meaning is adequate to our philosophical purposes.)

7. The Theory of Meaning

Let me conclude by returning to the question of the relation between the project of giving a meaning theory for a language and of giving a theory of meaning. The ambition of a meaning theory is to state knowledge that suffices for one to understand any utterance of a sentence in the language. Ultimately the theory should issue in explicit statements of what sentences in the language mean, as a way of expressing the sort of knowledge we want it to give us, whether we pursue the project directly or indirectly, that is to say, it should issue in M-theorems.[5] This is not a full theory of meaning, but if successful it would make an important contribution. It would provide a body of propositional knowledge independent of knowledge of understanding any particular language that would determine the meaning facts about a particular language in a way that allows understanding of the language. However, surprisingly, the very ambition to provide an explicit M-theorem for every sentence shows that what insight the theory gives relies upon prior understanding of a language together with knowledge that a sentence one understands is the same in meaning as the sentence whose meaning one wants to grasp. If this is right, then no theory of this sort can achieve its ambition.

Davidson hoped to bridge the gap between a meaning theory and a theory of meaning, as I have noted, by an account of how a truth theory could be confirmed for a speaker on the basis of evidence that did not presuppose anything about a speaker's words or any detailed knowledge of his attitudes, that is, from the standpoint of the radical interpreter (Davidson 1973b). This is a way of connecting up the structure articulated in a truth theory with the use of words by speakers of the language for which it is theory in a way, it was hoped, that would guarantee its canonical theorems were interpretive. The general idea was that we would gain insight into the content of the relevant concepts by seeing how a theory sufficient for interpretation could be confirmed on the basis of evidence that did not presuppose application of the concepts of the theory. Davidson's fundamental assumption about radical interpretation was that a correct meaning theory could be recovered ultimately from purely behavioral evidence and what we can know about agents and speakers a priori. I will not argue for it here, but I think this assumption is mistaken (Lepore and Ludwig 2005).

This still leaves us with the task of relating language—words and expressions, and ultimately sentences—to the uses to which they are put and the purposes for which they are designed. What I believe this requires is a theory of communicative institutions and the forms of collective agency that underlie them (Jankovic 2014a, 2014b). If we cannot hope for a reduction of meaning to behavioral responses to the environment, then we must relate meaning to the attitudes people have in using words and expressions in communicative contexts. This will inevitably rest on an antecedent understanding of how we are able to represent the world in thought. But it holds out the hope of breaking out

[5] These remarks therefore apply as well to Ray's account discussed in note 1.

of the circle of linguistic concepts, which the fixation on generating M-theorems in the theory of meaning does not.

References

Boisvert, D. and Ludwig, K. 2006, "Semantics for Nondeclaratives", in Smith, B. and Lepore, E. (eds.), *The Oxford Handbook of the Philosophy of Language*, Oxford: Oxford University Press.

Davidson, D. 1965, "Theories of Meaning and Learnable Languages", in Bar-Hillel, Y. (ed.), *Proceedings of the 1964 International Congress for Logic, Methodology and Philosophy of Science*, Amsterdam: North Holland Publishing Company.

Davidson, D.1967, "Truth and Meaning", *Synthese*, 17, 304-23.

Davidson, D. 1968, "On Saying That", *Synthese*, 19, 130-46.

Davidson, D. 1970, "Semantics for Natural Languages", in *Linguaggi nella società e nella tecnica*, Milano: Comunità.

Davidson, D. 1973a, "In Defence of Convention T", in Leblanc, H. (ed.), *Truth, Syntax and Modality*, Dordrecht: North-Holland Publishing Company.

Davidson, D. 1973b, "Radical Interpretation", *Dialectica*, 27, 314-28.

Davidson, D. 1974, "Belief and the Basis of Meaning", *Synthese*, 27, 309-23.

Davidson, D. 1975, "Thought and Talk", in Guttenplan, S. (ed.), *Mind and Language*, Oxford: Oxford University Press.

Davidson, D. 1976, "Reply to Foster", in Evans, G. and McDowell, J. (eds.), *Truth and Meaning: Essays in Semantics*, Oxford: Oxford University Press.

Davidson, D. 1979, "Moods and Performances", in Margalit, A. (ed.), *Meaning and Use*, Dordrecht: Reidel.

Davidson, D. 2001, "Semantics for Natural Languages", in *Inquiries into Truth and Interpretation*, first ed. 1970, New York: Clarendon Press.

Dummett, M. 1975, "What is a Theory of Meaning?", in Guttenplan, S. (ed.), *Mind and Language*, Oxford: Oxford University Press.

Dummett, M. 1976, "What is a Theory of Meaning? (II)", in Evans, G. and McDowell, J. (eds.), *Truth and Meaning: Essays in Semantics*, Oxford: Oxford University Press.

Hanks, P. 2015, *Propositional Content*, New York: Oxford University Press.

Hoeltje, M. 2013, "Lepore and Ludwig on 'explicit meaning theories'", *Philosophical Studies*, 165, 3, 831-39.

Hoeltje, M. 2016, "'Meaning and Truth' and 'Truth and Meaning'", *dialectica*, 70, 2, 201-15.

Jankovic, M. 2014a, "Communication and Shared Intention", *Philosophical Studies* 169, 489-508.

Jankovic, M. 2014b, *Conventional Meaning*. Dissertation, Philosophy, Indiana University.

King, J.C., Scott, S. and Speaks, J. 2014, *New Thinking about Propositions*, Oxford: Oxford University Press.

Lepore, E. and Ludwig, K. 2005, *Donald Davidson: Meaning, Truth, Language, and Reality*, Oxford: Oxford University Press.

Lepore, E. and Ludwig, K. 2007a, *Donald Davidson: Truth-theoretic Semantics*, New York: Oxford University Press.

Lepore, E. and Ludwig, K. 2007b, "Radical Misinterpretation: A reply to Stoutland", *International Journal of Philosophical Studies*, 15, 4, 557-85.

Lepore, E. and Ludwig, K. 2011, "Truth and Meaning Redux", *Philosophical Studies*, 154, 251-77.

Ludwig, K. 2003, "The Truth about Moods", in Preyer, G., Peter, G. and Ulkan, M. (eds.), *Concepts of Meaning: Framing an Integrated Theory of Linguistic Behavior*, Dordrecht: Kluwer.

Ludwig, K. 2014, "Propositions and Higher-order Attitude Attributions", *Canadian Journal of Philosophy*, 43, 5-6, 741-65.

Ludwig, K. 2015, "Was Davidson's Project a Carnapian Explication of Meaning?", *The Journal of the History of Analytic Philosophy*, 4, 3, 1-55.

Ray, G. 2014, "Meaning and Truth", *Mind*, 123, 489, 79-100.

Soames, S. 2008, "Truth and Meaning: In Perspective", *Truth and Its Deformities: Midwest Studies in Philosophy*, 32, 1-19.

Soames, S. 2010, *What is Meaning?*, *Soochow University Lectures in Philosophy*. Princeton: Princeton University Press.

Soames, S. 2015, *Rethinking Language, Mind, and Meaning*, *The Carl G. Hempel Lecture Series*, Princeton: Princeton University Press.

Wiggins, D. 1980. "'Most' and 'All': Some Comments On a Familiar Programme", in Platts M. (ed.), *Reference, Truth and Reality: Essays on the Philosophy of Language*, London: Routldge and Kegan Paul.

Davidson: Decision and Interpretation

Pol-Vincent Harnay and Pétronille Rème***

** Independent Researcher*
*** Université Paris-Est, AME, SPLOTT, IFSTTAR, F-77447 Marne-la-Vallée*

Abstract

Decision theory plays a central role in Davidson's work. Based on the experimentations led in Stanford during the 1950s, it is possible to track down the origins and the foundations of the unified theory of thought, meaning and action. The 'wording effect' and the omission of meanings undermine decision theory as a whole, hence the need to enlarge the basis of decision theory by integrating an interpretation theory that reflects mental holism more accurately.

Keywords: Decision Theory, Interpretation Theory, Mental Holism, Unified Theory.

> Not much attention is usually paid to Davidson's reflections on the role of decision theory in interpretation. These are, however, quintessential for anyone willing to understand Davidson's holism and the way his theory of language and that of rational action intermingle.
>
> Engel 1994: 111

1. Introduction

Very few studies have focused on Donald Davidson's work on decision theory:[1] the emphasis has rather been put on his groundbreaking work on action theory and philosophy of language.[2] Nevertheless, Davidson is, in many respects, a major author of experimental economics, which developed in the United States in the 1950s. Besides, it is worth noticing the significance of decision theory in Davidson's work—as evidenced by his numerous articles in which he deals with both his research at Stanford in the 1950s[3] and his attempt, based on Richard Jeffrey's research, to build a 'unified' theory of action and language that would overcome the weaknesses of decision theory. At least two approaches exist to gauge and put under perspective Davidson's research work on decision theory.

[1] Yet, Isaac Levi 1999 and Piers Rawling 2001 are worthy of quoting.
[2] For instance, Lepore and McLaughlin 1985, Lepore 1986.
[3] See Davidson 1980, 1985.

Argumenta 3, 1 (2017): 39-52
ISSN 2465-2334

DOI 10.14275/2465-2334/20175.HAR

A first approach, through an emphasis on economics, would consist in examining into details the axiomatizations outlined by Davidson and his team in Stanford—including Patrick Suppes. The point would be to understand the theoretical foundations of their experiments as well as their role in the debates dealing with the theory of expected utility and, as a whole, their place in experimental economics.[4] It is, however, another approach that will be tackled here, one that is conclusive with the interdisciplinarity of Davidson's writings. The approach implies relying on the lessons and criticisms formed by Davidson himself after his testing in decision theory to better understand the embeddedness between decision theory and interpretation theory. I quote: "Theory of meaning as I see it and Bayesian decision theory, are made for each other".[5] The interest of such an approach lies, first, in its purpose that is to understand the origin of the unified theory of thought, meaning and action that was ardently defended by Davidson as of the 1980s. It was indeed from the experiments led in Stanford and the criticizing of decision theory that Davidson devised a more inclusive model of decision, one putting together decision theory and philosophy of language. Secondly, the approach sheds light on the consistency of Davidson's work by offering a particular example of its ideas on mental holism.

In order to understand the interactions and imbrications between decision theory and interpretation theory, it is worth analyzing the central role of decision theory in Davidson's thought through a focus (2. Davidson and decision theory: from early experiments to the "sophisticated" theory of reason explanations) on the experiments he carried out at the beginning of his career and above all by underlining the connections pointed out between decision theory and action theory. The different failures Davidson faced when he was an experimental psychologist[6] are mostly due to skipping meanings in standard decision theory. Yet, the significance of meanings comes from the fact that choices are usually expressed verbally. As a result, when an experimenter, as often it was often the case in the 1950s, puts forward a behaviorist solution—according to which the mind is like a black box collecting data through stimuli, producing pieces of information in response—to the issue of measuring utilities and probabilities, he or she imposes a language. This language corresponds to the formulations of the options on which choices are made. A discrepancy can consequently appear between the meanings as perceived by the subjects' experiments and the ones for an experimenter devising his or her protocol. The criticism is relevant since analyzing these meanings would offer an additional mental component to the experimenter, which would allow him or her to have a better understanding of the reasons behind decision. From that base,[7] Davidson developed in the early 1980s a unified theory of action, language and interpretation (2. From decision theory to the unified theory of thought, meaning and action). Desires and beliefs are typically expressed verbally and Davidson reckons that the act of stating offers a piece of information that is directly linked to those desires and beliefs: one cannot understand what a person says if the former does not comprehend what the latter believes and desires. In other words, desires, beliefs, and meanings are connected by a strong mutual dependence that offers a significant image of the

[4] For a detailed account of this approach, see Harnay 2008.
[5] Davidson 1980: 158.
[6] Davidson 1974a, 2001c: 236.
[7] Davidson (1985) did use the term "base" to refer to his unified model.

mind. This image is both more relevant and significant than the standard behavior approach as its explanatory scope encompasses an additional datum, namely meanings: "The real problems of decision theory are problems of interpretation".[8]

2. Davidson and Decision Theory: From Early Experiments to the "Sophisticated" Theory of Reason Explanations

Decision theory was a particular component of Davidson' career. It was the first project he worked on after his PhD on Plato's *Philebus*. During the 1950s, Davidson published many articles about decision theory in which he detailed—along with Patrick Suppers—several experiments aiming at gauging the theory of expected utility's empirical validity—this theory being the cornerstone of decision theory. Furthermore, his purpose was to offer an empirical interpretation of the theory, one that is testable.[9]

Without explaining axiomatic into detail, it is relevant, though, to have a look on what Davidson learned from his experiments so as to understand why decision theory was so crucial for him. Interpreting the criticism expressed by Davidson on decision theory will be useful to fathom the emergence and the genesis of the problem of interpretation, which was his main topic of interest from the 1970s on and which fueled the afore-mentioned unified theory (2.1). In view of these analyses, it will be relevant to broadly introduce the connections between decision theory and action theory according to Davidson (2.2).

2.1 Lessons from Decision Theory

For the purpose of this article, I have decided to rely on the decision model that Davidson detailed in 1957 in *Decision Making: An Experimental Approach*. Such a choice is not trivial. After all, Davidson himself mentioned the model several times,[10] underscoring with thoroughness its weaknesses and limits. In both cases, the point was either to draw an analogy with interpretation theory[11] or to highlight mental holism.[12] The experiments resulted for Davidson in various ways of tackling his unified theory: because of different effects—including an effect of formulation—undermining the experiments' results (2.1.2), Davidson decided to promote the need to combine decision theory and communication theory.[13] He added: "A radical theory of decision must include a theory of interpretation and cannot presuppose it".[14] Likewise, the behaviorist solution systematically showcased by the experiments in Stanford—which Davidson strongly criticized—is a way to pinpoint mental holism that was a recurrent topic of discussion among Davidson's commentators (2.1.3). These two approaches will be analyzed after a brief overview of the 1957 model's results (2.1.1).

[8] Davidson 1999: 32.
[9] Davidson 1957: 4.
[10] Davidson 1980, 1985.
[11] Davidson 1974a, 1974b, 2001b: 145, 2001c: 238.
[12] Davidson 1997, 2001a: 126-27.
[13] Davidson 1974a, 2001c: 237.
[14] Davidson 1974b, 2001b: 147.

2.1.1 The 1957 Model

Davidson's input in 1957 took place as the foundations of the theory had already been laid, especially by the canonical model of Von Neumann and Morgenstern (*Theory of Games and Economic Behavior* [1947], corresponding to the second edition in which the appendix on the axiomatization of utility can be found). Within ten years, the epistemological and analytical stakes of their model were being discussed while the concepts used by Von Neumann and Morgenstern were being questioned—leading new axioms to be put forward, especially those by Friedman and Savage in 1948 and 1952.[15] In this regard, the article by Friedman and Savage (1948) can be quoted as an attempt to empirically validate Von Neumann and Morgenstern's expected utility theory. More precisely, it aims at collecting observations that deal with the behavioral choice of individuals facing risky outcomes in order to verify whether these observations are conclusive with what scholars call Von Neumann and Morgenstern's "expected utility theory". Finally, the goal is to examine the impact of the observations on the curve representing the utility function, especially in terms of aversion or appeal to risk.

Davidson's theory, in line with that trend, is complex in many respects.

The first respect has to do with its very purpose: as evidenced by the subtitle of his book written with Siegel and Suppes, *Decision Making* (1957), Davidson wanted to adopt "an experimental approach". According to them, "no satisfying empirical interpretation of decision theory has been offered, therefore it is impossible to test it".[16]

The second one is linked to its method: Davidson, Suppes and Siegel expressed a whole set of theoretical hypotheses that were testable through rephrasing Von Neumann Morgenstern and Savage's theory and that aimed at checking whether individuals maximized their expected utility by offering them several risky gambles. Based on empirical data, the scientists were able to devise a scale of utilities, evenly spaced, to verify the hypothesis of expected utility. What they did to conceive that scale was openly inspired by Frank Ramsey's works and in particular his operational method[17] allowing him to determine at the same time utilities and probabilities.

[15] Even though the influence of Von Neumann and Morgenstern on Davidson and all was significant, the Stanford team clearly followed the path of Friedman and Savage's works (1948), as proven by the attempt to obtain, thanks to experiences, utility curves similar to the ones hypothetically described by Friedman and Savage (for further details and a detailed account on the influences of the 1957 model, please see Harnay 2008).

[16] Davidson 1957: 3.

[17] Ramsey's aim in *Truth and Probability* (1926) was to underscore the link between the subjective degree of belief that one has in a proposition *p* and its current probability. Furthermore, he focused on the way a degree of belief by an agent on a given proposition could be measured. If this agent applies a certain number of norms on rationality, then the degree of belief can be represented by a measure that is conclusive with the laws of mathematics of probability, according to Ramsey. Ramsey's method consists of a ruler that represents numerically (thanks to different values) the way a person estimates the probability that something happens. The method consequently makes it possible to phrase a theory that would comprise several axioms requiring a rational behavior from the agent. From these axioms, one can calculate both the cardinal numbers of utilities (namely the subjective values of an individual on issues) and the extent to which an indi-

One major criticism of such a model expressed by Davidson is that some effects, like the formulation effect, undermine the results of Stanford team's experiments.

2.1.2 The Formulation Effect

Before analyzing into details the content and the scope of the effect, two remarks have to be highlighted.

First, the harshest criticisms against the model in 1957 and broadly speaking against decision theory were only expressed by Davidson about twenty years after Stanford's first experiments. All along his career in experimental psychology, Davidson commented upon the model merely a few times. For instance, Davidson and his colleagues discovered that winning or losing several times in a row made subjects respectively optimistic or pessimistic. As a result, this had an impact on the subjects' subsequent responses to similar offers. The increasing (or decreasing) amount of money at play also influenced choices. This means that one must identify these distortions and find a way of avoiding contamination of all the choices made.[18] Those are the kind of observations that are in *Decision Making*—observations that Davidson tried to address during the 1950s.

The second remark underscores the significance of psychologist Ward Edwards' experiments (University of Michigan) in understanding the weaknesses of decision theory pointed out by Davidson. Edwards could be arguably viewed as the first experimental psychologist to describe in details all the variables that could affect subject's choices, in one way or another. By showing how subjects systematically diverge from the objective model, Edwards opened the way to considering globally the psychological factors that influence choices. In 1954, Edwards talked about a particular effect: the "wording effect". In Edwards' mind, the wording effect corresponds to a test during which the experimenter changes the way options are verbalized by reversing the proposal that describes both gains and losses.[19] Broadly speaking, the idea is to verify whether a reversal in preference results from the reversal in wording. In other words, does the wording influence the preferences expressed by subjects? From that point on, Davidson defended the idea that including a theory of interpretation in decision theory was necessary.

In *Decision Making*, Edwards is quoted twice, though for other reasons than the "wording effect". In addition to distortions mentioned above, another issue expressed is the "recency effect": even with the special die used, subjects got attached to specific nonsense syllables, for example "if the same syllable came up three times in succession".[20] To overcome the problem, the authors postponed

vidual believes in a proposal determining issues and the way it influences his behavior on a gamble.

[18] Davidson, Suppes, Siegel 1957: 53.

[19] The options presented to the subjects were phrased as follows: "I toss a coin. If it comes up heads, I pay you $2.00. If it comes up tails, you pay me $1.00" (Edwards 1954), regarding the bets with a positive expected value. By presenting the losses before the gains while offering the same expected value, the goal was to know whether a change in the subjects' choices occurred.

[20] Davidson, Suppes, Siegel 1957: 54.

the payoffs and used three dice instead of one and "the die in use was changed after each toss".[21]

In 1974, Davidson, in his article *Belief and the basis of meaning*, quoted Edwards through referring, directly this time, to the "wording effect":

> There is not just an analogy between decision theory and interpretation theory, there is a connection. Seen from the side of decision theory, there is what Ward Edwards once dubbed the 'presentation problem' for empirical applications of decision theory. To learn the preferences of an agent, particularly among complex gambles, it is obviously necessary to describe the options in words. But how can the experimenter know what those words mean to the subject? (Davidson 1974: 147).

The argument is particularly relevant for the purpose of this article: Davidson established a parallel between decision theory and interpretation theory after criticizing the decision theory itself. He also mentioned the deadlock on meanings, whereas they are mental data necessary to reach a correct overview of the behaviors leading to deeds. This point will be developed in part 2.2.

The other main criticism of Davidson on decision theory deals with the behaviorist approach, which was systematically put into lights during Stanford's experiments.

2.1.3 The Dismissal of Behaviorism

The solution advanced by Davidson, Suppes and Siegel to the problem of measuring utilities (desires) and probabilities (beliefs) is a behaviorist approach,[22] as explained by Davidson himself: "All we had to do was to give a clear behavioristic interpretation to 'S prefers A to B' and decision theory [...] became a powerful empirical theory, eminently testable, and palpably false".[23]

According to Davidson, behaviorism is problematic in its attempt to viewing mental conditions as simple physical conditions.[24] A theory taking into account solely the physical aspect of mental conditions is indeed not worth considering, for Davidson: human behavior is part of nature (in this regard, "all mental events ultimately, perhaps through causal relations with other mental events, have causal intercourse with physical events")[25] but the idea according to which voluntary action can be applied to determinist principles, like that of physics, is to be dismissed. One reason for the dismissal of strict psychophysics' principles has to do with the holistic character of the cognitive field, which imposes to factor a growing number of components related to beliefs and motives for action. In other words, had this hypothetical physical theory of mental events rested upon such a foundation, it would have collapsed because of men-

[21] *Ibid.*
[22] Davidson, Suppes, Siegel 1957: 12.
[23] Davidson 1976, 2001c: 270.
[24] This has to do with the idea of anomalous monism: "The nomological irreducibility of the psychological means, if I am right, that the social sciences cannot be expected to develop in ways exactly parallel to the physical sciences, nor can we expect ever to be able to explain and predict human behavior with the kind of precision that is possible in principle for physical phenomena", see Davidson 1974a.
[25] Davidson 1970, 2001c: 208.

tal holism that requires to add an increasing amount of pieces of information to be understood. The temptation to narrow decision theory to match physical sciences is thus doomed—impacting concurrently the possibility of empirically testing the theory, that is to say probing its expected results.

Nevertheless, though Davidson seems to condemn decision theory for the complexity—or even the impossibility—of testing the axioms of the field's canonical models, he reaches a subtler conclusion:

> I think I have an argument to show that the main empirical thrust of an explanation of an action in decision theory, or of a reason explanation, does not come from the axioms of decision theory, or 'the assumption of rationality', but rather from the attributions of desires, preferences, or beliefs (Davidson 1976, 2001c: 273).

Studying the links between decision theory and action theory will help going deeper into this idea.

2.2 Decision Theory and Action Theory

In the 1960-70s, Davidson imagined a theory of action that had many commonalities with decision theory, especially in its structure. Both theories shared the same goals by dealing with the reasons for acting, first, and the choices available for action, on the other hand: action theory aims at explaining the motives for an isolated deed while decision theory explains why an agent chooses one action among perfectly reasonable others:

> Decision theory is a way of systematizing the relations among beliefs, desires, and actions. It does this by imposing a complex, but clearly defined, pattern on the way in which people's beliefs and desires interact (Davidson 1997, 2001a: 126).

First of all, both action theory and decision theory rely on an analysis of the roles, respectively, of desires (preferences) and beliefs (probabilities). They are, then, based on the same principle of practical syllogism as proposed by Aristotle and consequently offer a "teleological" perspective.

Davidson's decision theory, built on models and experiments from 1957-59, proved that desires and beliefs could be identified, singled out and measured—which cannot be done in action theory:

> Given the idealized conditions postulated by the theory, Ramsey's method makes it possible to identify the relevant beliefs and desires uniquely. Instead of talking of postulation, we might put the matter this way: to the extent that we can see the actions of an agent as falling into a consistent (rational) pattern of a certain sort, we can explain those actions in terms of a system of quantified beliefs and desires (Davidson 1975, 2001b: 160).

Action theory does not offer sophistication because of its structure and its underlying concepts, especially the one on rationality:

> Two ideas are built into the concept of acting on a reason (and hence, the concept of behaviour generally): the idea of cause and the idea of rationality. A reason is a rational cause. One way rationality is built in is transparent: the cause must be a belief and a desire in the light of which the action is reasonable. But

> the rationality also enters more subtly, since the way desire and belief work to cause the action must meet further, and unspecified, conditions. The advantage of this mode of explanation is clear: we can explain behaviour without having to know too much about how it was caused. And the cost is appropriate: we cannot turn this mode of explanation into something more like a science (Davidson 1974, 2001c: 233).

The principal difference between action theory and Davidson's decision theory lies in the fact that the action theory he coined is a particular simple form of explanation by reasons that does not factor the way an agent makes a choice among several actions:

> The discussion so far has been hampered, if not hamstrung, by my sticking to a particularly simple form of reason explanation, and this has prevented me from saying anything sensible about a number of problems, such as how an agent might be expected to choose among several competing actions, each of which is recommended by reasons he has. Similarly, no mention has been made of the effect of variations in the strength of desire, or degree of belief. The theory of decision making under uncertainty is designed to cope with these matters (Davidson 1976, 2001c: 268).

Likewise, action theory does not include the variations in desires' intensity or the degrees of belief—as these issues have to do with decision theory.

For Davidson, however, decision theory is a refined action theory since it moves toward "scientific respectability": "It gives up trying to explain actions one at time by appeal to something more basic, and instead postulates a pattern in behavior from which beliefs and attitudes can be inferred".[26]

Action and decision theories resort to a similar analytical reasoning. Action theory does not have a formal structure nor give the option to decide between two equally desirable actions. As for decision theory, it is static by dismissing meanings.[27] Davidson explains: "Decision theory purports to describe a static situation: the pattern of a person's attitudes and beliefs at a moment".[28] Earlier in the same article, he had shed light on the limits of a static theory on propositional behaviors: "how could we tell that subjects weren't influenced in their preferences by the experiment itself—that their preferences weren't changing as we went along?".[29]

In order to overcome these shortfalls, Davidson introduced a new theory aiming at better assessing mental holism.

3. From Decision Theory to the Unified Theory of Thought, Meaning and Action

According to Davidson, decision theory, because of its formal nature and its normative mission, is silent about worldly matters; its abstract structure does not

[26] Davidson 1974, 2001c: 235.
[27] Decision theory is static in that it fails to consider the instances when a person does not necessarily make the same choice while facing the same options, even when the circumstances leading to choosing are the same.
[28] Davidson 1976: 271.
[29] *Ibid*.

provide meaningful interpretation about the terms it uses, such as "to prefer".[30] In other words, one of the main criticisms regarding decision theory in the 1950s deals with its avoidance of providing meanings. And it is precisely by adding these elements (2.1) that the unified theory emerged (2.2).

3.1 Adding Meanings

In his article *A new basis for decision theory*[31] published in 1985 in a journal entitled *Theory and Decision*, Donald Davidson mentioned the recurrent difficulty of experimental decision theory: the experimenter acknowledges the analyses of meanings, or, more precisely, the way the subjects interpret the different objects of decision theory, such as gambles and their outcomes.

In his essay *Expressing Evaluations* (1984), which tackled directly such a difficulty, Davidson explained what the problems were:

> Bayesian decision theories have a fatal drawback: they simply assume that an interpreter can tell what propositions an agent is evaluating or choosing between, or which interpreted sentences express the agent's preferences [...] Decision theory begins with simple preferences between propositions; once these have been identified, the theory allows us to extract the beliefs and desires that went into, and explain, the preferences. But it says nothing about what determined the objects of the original simple preferences. Preferences are, of course, manifested in behavior in many ways. But this fact does not tell us how the content of preferences is fixed (Davidson 1984, 2004: 29).

This quote points out that not mentioning meanings in fact leads to at least two subsequent problems.

By leaving meanings out, decision theory overlooks an essential part of mental data that incite an agent to make a choice or a decision. By neglecting the analysis of meanings (or taking it for granted), decision theory is theory confined to a behaviorist approach whereas language is what ties the agent to people and to the surrounding world.

Another issue, correlated to the first one, can be summed up by using Davidson's own words: neglecting the analysis of meanings implies "establishing the correctness of an attribution of belief or desire involves much the same problems as showing that we have understood the words of another".[32] More precisely, decision theory and interpretation theory account for tools of measuring interdependently mental contents, by resorting to similar methods. By ignoring interpretation theory, decision theory does without an additional tool that would provide, yet, a better understanding of mental contents.

This is why, according to Davidson, "what we must add to decision theory, or incorporate in it, is a theory of interpretation for the agent, a way of telling what he means by his words".[33]

Nevertheless, in order to do so does not mean adding to the 1957 model of decision ad hoc meanings but instead trying to interpret the meanings of proposals in actors' eyes, while having access to their beliefs and desires.

[30] Davidson 1999: 32.
[31] The article is a revised version by Davidson 1980.
[32] Davidson 1974a, 2001c: 237.
[33] Davidson 1980, 2004: 155.

In other words, the aim is to conceive a theory, based on standard models of decision theory, integrating the analysis of meanings without, nonetheless, supposing in advance the existence of data to be explained: "this addition must be made in the absence of detailed information about beliefs, desires, or intentions".[34]

Analyzing meanings in decision is essential in view of the way experimenters usually assume that the words used by the subject and those used by the experimenter can be interpreted similarly. When choices are offered to the experiment's subjects, language and its meanings are somehow imposed to the individual.[35] According to Davidson, notwithstanding, imposing a language to the experimenter gets round the meanings that the subjects would give to the proposals and therefore neglects the information on the subjects' behavior influenced by the analysis of these meanings:

> To learn the preferences of an agent, particularly among complex gambles, it is obviously necessary to describe the options in words. But how can the experimenter know what those words mean to the subject? The problem is not merely theoretical: it is well known that two descriptions of what the experimenter takes to be the same option may elicit quite different responses from a subject (Davidson 1974b: 147).

For Davidson, however, the necessary interaction between decision theory and the interpretation of language goes farther. Both theories complement each other since they are tools to measure mental activities and deal with combined elements:

> Theory of meaning as I see it, and Bayesian decision theory, are evidently made for each other. Decision theory must be freed from the assumption of an independently determined knowledge of meaning; theory of meaning calls for a theory of degree of belief in order to make serious use of relations of evidential support (Davidson 1980, 2004: 158).

As we have just seen, the criticism toward decision theory by Davidson echoes the various remarks made by Tversky in 1975 and displays many forms in Davidson's several articles.

3.2 Toward a Unified Theory of Action, Language and Meaning

Integrating meanings allows Davidson to provide a broader theory where meanings, beliefs and desires are codetermined. This is in fact the unified theory of action, language and interpretation (2.2.2). In order to single out beliefs, desires and meanings, the experimenter turns into an interpreter (2.2.1).

3.2.1 When the Experimenter Becomes Interpreter

By placing the role of meanings at the very core of decision theory in its Jeffrey version, Davidson enlarges the experimenter's initial part. In fact, its part will

[34] *Ibid*.
[35] Davidson 1974a, 2001c: 236-237. In a subtler way, one could say that meanings are supposed to be known as they come from the experimenter himself.

extend to that of an interpreter. As mentioned by Davidson, "theory of interpretation is the business jointly of the linguist, psychologist and philosopher".[36] In other words, by adding meanings and by considering them as seminal in individuals' choices, the experimenter has to measure both the cardinal utilities and the subjective probabilities but also has to interpret the subjects' utterances. According to Davidson, such a procedure would be as much a mental-measuring tool as that of Ramsey—though the former would be more precise than the latter.

Turning the experimenter into an interpreter requires a profound interaction with the subject. Particularly, the interpretation mentioned by Davidson implies a comparison of the desires, beliefs and meanings of the subjects with those of the experimenter. Doing so, Davidson offers a groundbreaking view in the debate about interpersonal comparisons, those that set the standards for everyone to compare each other.

Yet, the upgrade of experimenter's role is only a detail shadowing the broader spectrum. As a whole, it is the entire empirical cornerstone—i.e. every mental information acquired—that is extended: meanings have to be processed at the same time as desires and beliefs. Decision theory has to integrate a theory of meaning.

3.2.2 The Unified Theory

More than twenty years after *Decision Making* (1957), Davidson offered a unified theory of decision and interpretation during the 1980s.

The original 1957 model described in the first part was enhanced and transformed through several articles written in the 1970s and 80s. In order to do so, this enhanced model was included into a broader "unified" theory by Davidson (1980). The term is justified by the very nature of theory's purpose, a triplet (desire, belief, meaning) that encompasses a similar theoretical reasoning as well as a distinct approach since Davidson points out complex causalities between these concepts, which induce that one concept cannot be determined without the other two being determined simultaneously too.

Decision theory and interpretation theory aim at solving two problems: pulling apart the role of beliefs and desires for the former, separating that of beliefs and meanings for the latter.

As evoked earlier, Davidson discovered that decision theory assumes that one can identify and individualize the proposals that lean toward propositional behaviors like desires and beliefs. Nevertheless, the ability to identify the propositions underpinning an agent's behavior is not to be separated from the one consisting in comprehending what the agent says.[37] Davidson adds that one usually realizes what he or she wants, prefers or believes solely by interpreting those words.[38] For the author, it is not easier to correctly establish the fact of attributing a desire to something than interpreting someone's discourse. There is, still, a need to go farther and state that both problems are identical. As a whole, one cannot determine beliefs without mastering the language of an individual; and one cannot master the language of someone without knowing what he or she believes:

[36] Davidson 1974b, 2001b: 141-42.
[37] Davidson 1980, 2004: 155.
[38] Davidson 1990: 318.

> In order to interpret verbal behavior, we must be able to tell when a speaker holds a sentence he speaks to be true. But sentences are held to be true partly because of what he believed, and partly because of what the speaker means by his words. The problem of interpretation therefore is the problem of abstracting simultaneously the roles of belief and meaning from the pattern of sentences to which a speaker subscribes over time. The situation is like that in decision theory: just as we cannot infer beliefs from choices without also inferring desires, so we cannot decide what a man means by what he says without at the same time constructing a theory about what he believes (Davidson 1974, 2001c: 238).

The unified theory of decision and interpretation suggested by Davidson can be seen as a response to the experimental issues encountered by the Stanford team:

> But stating theses mutual dependencies [between theory of meaning and decision theory] is not enough, for neither theory can be developed first as a basis for the other. There is no way to add on to the other in order to get started, each requires an element drawn from the other. What is wanted is a unified theory that yields degree of belief, utilities on interval scale, and interpretation of speech without assuming any of them (Davidson 1980: 158).

Among the expected results of such a theory, one could mention the enhancement of content due to the adding of meanings as raw data as well as the explanations available to clarify cases of potential irrationality.

Moreover, introducing meanings is a way for the theory to integrate a "dynamic" element: interpretation is built through a process of trials and errors. There are several steps before devising complete beliefs and meanings. In this respect, the principle of charity is the cornerstone that allows to maximize the agreement between the interpreter and the speaker.[39]

Guided by Richard Jeffrey's work (1983), Davidson acted on his intuitions about the connection between the decision theory and the interpretation theory by offering a slightly different version of Jeffrey's model:

> We owe to Richard Jeffrey a version of Bayesian decision theory that makes no direct use of gambles, but treats the objects of preference, the objects to which subjective probabilities are assigned, and the objects to which relative values are assigned uniformly as propositions (Davidson 1980: 160).

Though Davidson decided to use Jeffrey's model, it was not without making slight modifications. Indeed, conclusive with one of Davidson's most important statements, one of the first proposals is that one should not take meaning for granted. This amounts to using non-interpreted sentences instead of proposals because from this point of view, the latter could be assimilated to meanings themselves.[40]

Decision theory is quintessential in Davidson's work. The experiments he carried out in Stanford during the 1950s deeply influenced him on an epistemological and theoretical matter. The main problem pointed out by Davidson is that it does not include an interpretation theory of language. More specifically,

[39] For a detailed presentation, see Engel 1994.
[40] For a detailed presentation, see Harnay 2008.

the experimenter takes for granted the meanings bestowed upon the different options available by the subjects. According to Davidson, however, this is not obvious. Nothing indicates that the subject's beliefs and meanings are analogous to that of the experimenter. As a result, the different criticisms about decision theory nourished a global reflection regarding the links between decision and interpretation leading to a unified theory of action and language. The underlying central point is that one cannot analyze nor comprehend decision without resorting to language as it is both a communication tool and an indicator of mental contents. By assuming such a connection, Davidson promotes a theoretical and methodological merging of disciplines that were heretofore divided, through an emphasis on the stakes behind such a merging.[41]

References

Davidson, D. 1974a, "Psychology as Philosophy", in Davidson 2001c, 229-44.

Davidson, D. 1974b, "Belief and the Basis of Meaning", in Davidson 2001b, 141-54.

Davidson, D. 1975, "Thought and Talk", in Davidson 2001b, 155-70.

Davidson, D. 1976, "Hempel on Explaining Action", in Davidson 2001c, 261-75.

Davidson, D. 1980, "A Unified Theory of Thought, Meaning, and Action", in Davidson 2004, 151-66.

Davidson, D. 1984, "Expressing Evaluations", in Davidson 2004,19-38.

Davidson, D. 1985, "A New Basis for Decision Theory", *Theory and Decision*, 18, 87-98.

Davidson, D. 1990, "The Structure and Content of Truth", *The Journal of Philosophy*, 87, 6, 279-328.

Davidson, D. 1999, "Autobiography", in Hahn 1999, 3-70.

Davidson, D. 2001a, *Subjective, Intersubjective, Objective*, Oxford: Oxford University Press.

Davidson, D. 2001b, *Inquiries into Truth and Interpretation*, first ed. 1984, Oxford: Oxford University Press.

Davidson, D. 2001c, *Essays on Actions and Events*, first ed. 1980, Oxford: Oxford University Press.

Davidson, D. 2004, *Problems of Rationality*, Oxford: Oxford University Press.

Davidson, D., Suppes, P. and Siegel, S. 1957, *Decision Making: An Experimental Approach*, Stanford: Stanford University Press.

Edwards, W. 1954, "The Reliability of Probability-Preferences", *The American Journal of Psychology*, 67, 1, 68-95.

Engel, P. 1994. *Davidson et la philosophie du langage. L'interrogation Philosophique*, Paris: PUF.

Friedman, M. and Savage, L. 1948, "The Utility Analysis of Choices Involving Risk", *The journal of Political Economy*, 56, 4, 279-304.

[41] We would like to thank Pascal Engel and Mélaine Harnay (Université Paris III Sorbonne) for their useful comments, their expertise and encouragement. A special thought goes to Audrey, Héloïse and Mathilde who gave us the strength to complete this adventure.

Friedman, M. and Savage, L. 1952, "The Expected-Utility Hypothesis and the Measurability of Utility", *The Journal of Political Economy*, 60, 6, 463-74.

Hahn, L.E. (ed.) 1999, *The Philosophy of Donald Davidson*, The Library of Living Philosophers Series, Illinois: Open Court Publishing Company.

Harnay, P.-V. 2008, *La décision, de l'expérimentation à l'interprétation: l'apport de Donald Davidson*, Ph. D Thesis, Université Paris I Panthéon-Sorbonne.

Jeffrey, R., 1983, *The Logic of Decision*, first ed. 1965, Chicago: University of Chicago Press.

Lepore, E. and McLaughlin, B. (eds.) 1985, *Actions and Events, Perspectives on The Philosophy of Donald Davidson*, Oxford: Basil Blackwell.

Lepore, E. (ed.) 1986, *Truth and Interpretation. Perspectives on The Philosophy of Donald Davidson*, Oxford: Basil Blackwell.

Levi, I. 1999, "Representing Preferences: Donald Davidson on Rational Choice", in Hahn 1999, 531-70.

Rawling, P. 2001, "Davidson's Measurement Theoretic Reduction of the Mind", in Kotatko, P., Pagin, P. and Segal, G. (eds.), *Interpreting Davidson*, Stanford: CSLI Publications, 237-356.

Von Neumann, J. and Morgenstern, O. 1944, 1947 (2nd ed.), *Theory of Games and Economic Behavior*, Princeton: Princeton University Press.

Davidson on the Objectivity of Values and Reasons

Pascal Engel

EHESS, Paris

Abstract

Although he did not write on ethics, Davidson wrote a few papers on the objectivity of values. His argument rests on his holistic conception of interpretation of desires. I examine whether this argument can be sufficient for his objectivism about values. And supposing that the argument were correct, would it entail a form of realism about normativity and reasons? I argue that it falls short of giving us a genuine form of moral realism. My case will rest on an examination of Davidson's conception of value in relation to what he had to say about emotions and their relations to values.

Keywords: Davidson, values, reasons, moral realism, objectivity, emotions.

1. Introduction

Davidson's views on ethics have received much less attention than his views on meaning, mind and action. This is understandable, since he did not write much on ethics, although he often said that for him the most fundamental issues in philosophy are those of ethics. This concern surfaces in many of his writings, for instance in his early interest in Plato's *Philebus*, in his essay on weakness of the will, in his discussions of self-deception and in his late discussion of Spinoza (Davidson 1999).[1] And there is a field of ethics that he dealt with quite explicitly: meta-ethics. In three essays "Expressing evaluations" (1984), "The interpersonal comparison of values" (1986a) and "The objectivity of values" (1994), he drew some consequences of his conception of interpretation and rationality for the nature of moral values, and has defended an objectivist conception of these. His argument rests on the idea that interpretation of desires has to be holistic and presuppose a large pattern of agreement, which cannot fail to track objective truths about the values of agents. The argument raises several questions. First, is it correct? Can one reach the claim that values are objective on the basis of the constraints on interpretation? Second, supposing that the argument were correct,

[1] There is indeed room for developing a full-blown conception of ethics on the basis of Davidson's views in other domains, and this has been done, in particular by Bilgrami 2006, Rovane 2013 and Myers 2012.

Argumenta 3, 1 (2017): 53-68
ISSN 2465-2334

DOI 10.14275/2465-2334/20175.ENG

would it entail a form of moral realism? Third, how can it give us a realistic account of normativity and reasons? I argue that although one can develop an argument along these lines on the basis of Davidson's views on values, it falls short of giving us a genuine form of realism about reasons. My case will rest on an examination of Davidson's conception of value in relation to what he had to say on emotions and their relations to values.

2. Davidson's Argument from Interpretation to Objective Values

Before his two papers on value, Davidson had not dealt explicitly with issues about meta-ethics, although most of his work on action has close connections with moral psychology. His famous account of reasons (1963) bears clearly on the explanation of action and on what are often called motivating or explanatory reasons, although it does not deal explicitly with normative reasons in the sense of reasons based on normative beliefs about what is valuable or what one ought to do. In his article on weakness of the will (1970), however, Davidson deals with moral dilemmas and sets up a framework for the discussion of evaluative judgements. Although he does not bring this point to the fore in 1970, his account presupposes that there are genuine moral conflicts, involving real but incompatible values. He points out that Kantianism and utilitarianism, which are supposed to be objectivists about values, deny the existence of moral conflicts and dilemmas, whereas a number of philosophers, such as Williams or Foot, argued that moral dilemmas entail that values cannot be objective. In opposition to both, Davidson claims that there can be genuine conflicts between perfectly objective values. But Davidson tells us that he is not a moral realist in the sense in which this view would carry an ontological commitment to the existence of values as independent entities, a view which is open to the familiar anti-realist charge that such entities, if they existed, would be "queer" and hard to find in the natural world. Nor does he subscribe to a theory which, like McDowell's and Wiggins', insists that values are response-dependent in the way secondary qualities are so, although they track real properties in the world. But the issue, according to Davidson, is not *where* moral properties can be located in the natural world:

> Objectivity depends not on the location of an attributed property, or its supposed conceptual tie to human sensibilities; it depends on there being a systematic relationship between the attitude-causing properties of things and events, and the attitudes they cause. What makes our judgments of the "descriptive" properties of things true or false is the fact that the same properties tend to cause the same beliefs in different observers, and when observers differ, we assume there is an explanation. This is not just a platitude, it's a tautology, one whose truth is ensured by how we interpret people's beliefs. My thesis is that the same holds for moral values (Davidson 1994, 2004: 46).

His argument is not ontological, but epistemological: once we understand clearly how we can ascribe evaluative attitudes to people on the basis of their evaluative judgments, we shall be able to conclude that these attitudes are bound to track objective values.

Davidson invites us, as he does in many other contexts, to start from the necessary features of interpretation. The familiar claims are the following.[2]

(i) The task of interpretation is to ascribe to an agent propositional attitudes, such as beliefs, desires, intentions and preferences, which have certain contents. Interpretation has to start from publicly observable features of agents and of their environment and must rest on an evidential basis.

(ii) Holism: the contents of someone's attitudes necessarily depends on the contents of many other attitudes.

(iii) Charity: given that the contents of attitudes are necessarily interconnected, one must presuppose that there is at least a minimal coherence among these contents, and ascription of coherent sets of content cannot be made unless the interpreter presupposes that the agent shares a large number of true beliefs with him.

(iv) If agents are to be interpretable, they not only must share attitudes and contents that are largely similar to ours, but that are also largely correct.

Let us call this the *argument from interpretation*. On the basis of these necessary features of interpretation, which in most of his earlier writings he applies to the interpretation of beliefs and meanings, and derives from it a refutation of radical scepticism: since interpretation presupposes a massive degree of agreement on beliefs which are largely correct, these beliefs have to be about an objective world (Davidson 1981). Later on, Davidson applies this reasoning to desires, then to values, expanding (iv) into

(v) If agents are to be interpretable, they share values which are largely similar to ours, correct, and objective.

Even assuming that it works for beliefs,[3] the argument is even less straightforward for desires. *Prima facie* it should not be, for a central claim of Davidson's "unified theory of meaning and action" (Davidson 1980) is that we must interpret beliefs and desires *jointly*. Any interpretation of belief has to go through an interpretation of desires as well. Davidson insists on the fact that the pattern of desires in an agent not only is just as holistic as the pattern of his beliefs, and that they actually depend on each other when we interpret his actions:

> An interpreter cannot hope to determine the contents of a person's desires, without also determining what the person believes; and there is no way to determine the contents of either of these attitudes in a sufficiently detailed way without linguistic communication, which requires interpretation of the person's speech (Davidson 1994, 2004: 48).

But this dependency of the interpretation of desires on the interpretation of beliefs, and in turn on the interpretation of speech does nothing by itself to show that there is an argument parallel to (i)-(iv) for desires. At best, what it shows is that the argument from interpretation applies to desires *and* beliefs. But how can one apply it to those desires which are, in some sense, revealing of values held by an agent? Much here depends upon what one means by "desires". In Davidson's

[2] Here I more or less follow the very clear presentation by Myers 2012 and of Myers and Verheggen 2016; see also Myers 2004, Lillehammer 2007.

[3] Which is far from evident, as many critics have argued. See in particular Stroud 1999.

earlier writings, desires were ranked among "pro-attitudes", which can include wants, mere attractions, urges and whims (such as, to take one of his examples, the "sudden desire to touch a woman's elbow"), as well as long-standing desires, but also more elaborate attitudes of a normative kind, such as desires about what one believes to be good, worthwhile or obligatory. The latter he calls "enlightened" desires. There is a sense of "value" which applies to desires in the broad sense, which is the one that is used by decision theorists, under the name of *desirabilities* or *utilities* which, together with probabilities, determine an agent's action. But these are subjective values by definition, which do not yield any objectivity about value in general. What people individually desire can differ hugely from person to person. These are actually those gaps in having common desires which give rise to the problem of interpersonal comparisons of utility[4]. *Prima facie* what is needed for objectivity is at least some convergence of individual values susceptible to being shared by agents, possibly an agreement on these values. How do we reach such a convergence? Clearly, it is much harder to discern holistic patterns of desires within an agent than it is to discern such patterns for beliefs. If one ascribes to an agent the belief that a cloud is passing over the sun, it is natural to ascribe to him the belief that opaque objects can hide a source of light. But if one can ascribe to an agent a desire to eat an ice cream, it is hard to ascribe to her a desire, say, to eat a strawberry ice cream, or to eat an ice cream cake. Interpretation must presuppose that agents are aiming at what is *good* in general. So only "enlightened desires" about what one *believes* to be good will do. The desires that can be the basis of interpretation have to be *evaluative desires*, involving not simply an attitude of an agent towards an object, but also the belief that a certain object is valuable. In other words, they must be *normative desires*, that have propositional contents to the effect that so and so in desirable and valuable:

> To what extent do these considerations apply to the evaluative attitudes? It is possible, I think, to show that the justified attribution of values to someone else provides a basis for judgments of comparisons of value, what is called the interpersonal comparison of values. But the comparability of values does not in itself imply agreed-on standards, much less that we can legitimately treat value judgments as true or false. Now I want to go on to suggest that we should expect enlightened values—the reasons we would have for valuing and acting if we had all the (non-evaluative) facts straight—to converge; we should expect people who are enlightened *and fully understand one another* to agree on their basic values. An appreciation of what makes for such convergence or agreement also shows that value judgments are true or false in much the way our factual judgments are (Davidson 1994, 2004: 49).

Davidson makes clear here that by "values" he means three sorts of things: (a) basic values, (b) enlightened values, (c) converging values. It is not clear what the basic values could be. They could correspond to basic human needs, such as those for food, security or sex, but also to values such as justice, equality or freedom. The enlightened values are presumably those that involve what an agent considers reflectively as a value, his normative desires, on which he bases his reasons for acting. The converging values would be the end products of the process of understanding each other. But it is not clear what these would be. Would they be related

[4] Davidson actually deals with this problem in his 1986a article.

to what Bernard Williams (1985) calls "thick" concepts (such as shame or courage) or to "thin" concepts (such as *good* or *just*)?

But even if we concentrate on enlightened desires, what guarantee do we have that we shall converge on our values? It seems that what would be needed would be an equivalent of the principle of charity for desires. Sometimes, Davidson seems to suggest that there is such a parallel principle: "In our need to make him make sense we will try for a theory that finds him consistent, a believer of truths, and a lover of the good (all by our own lights it goes without saying)" (Davidson 1969; 1980: 222). But even if there were such a principle, it would function, like the principle of charity for beliefs, as an *a priori* principle of interpretation. Although such a principle is, by definition, supposed to be necessary, this would not yield a convergence on the objectivity of values[5]. For that, one needs the argument (i)-(v) above, and the holistic condition on beliefs and desires. Davidson's argument for the convergence of normative desires and values is clearly an epistemological argument, on a par with his convergence argument from interpretation of beliefs. He is quite clear that this convergence will not yield objective values in the ontological sense of separate entities, and he discards the ontological way of posing the problem of moral realism in the way Mackie (1977) and Jackson (1998) pose it: where are values? How is one to "place" them in nature? Values, for Davidson, are neither in the mind, as projections of our desires—as anti-realists and expressivists would argue—nor "out there" in the natural world or in some non-natural ideal world. If they are real, it is not in the sense of having a certain ontological status, but in the sense of being the product of a convergence in our judgements about values. When such convergence is reached, we will be able to say that our values are objective and that our judgments about them are true:

> Before we can say that two people disagree about the worth of an action or an object, we must be sure it is the same action or object and the same aspects of those actions and objects that they have in mind. The considerations that prove the dispute genuine—the considerations that lead to correct interpretation—will also reveal the shared criteria that determine where the truth lies (Davidson 1994; 2004: 47).

As Davidson is aware, the problem is that such a convergence is not guaranteed *a priori* by the holistic requirements on the interpretation of beliefs and desires. It implies that we have criteria of convergence and of divergence when we disagree on values:

> [I]f I am right, disputes over values (as in the case of other disputes) can be genuine only when there are shared criteria in the light of which there is an answer to the question who is right […] When we find a difference inexplicable, that is, not due to ignorance or confusion, the difference is not genuine […] The importance of a background of shared beliefs and values is that such a background allows us to make sense of the idea of a common standard of right and wrong, true and false (Davidson 1995; 2004: 50-51).

[5] Part of this question rests on whether the scholastic principle *Nihil appetimus nisi sub ratione boni,* which present day writers have renamed "the guise of the good" is a substantial ontological principle in moral theory or a principle of interpretation of desires. See Velleman 1992, Tenenbaum 2007, for instance.

But how can we make sure that we have such shared criteria? To say that that these are the conditions of convergence on enlightened values would be circular. The criteria of evaluation of disagreements over, for instance, the concept of "justice" are themselves evaluative, and themselves subject to disagreement. Davidson 's suggestion is obviously that such disagreements will in the end be assessable and that a core of shared values will be reached, but it is not clear that an important amount of indeterminacy will not remain, and, as Lillehammer (2007: 214-5) has remarked that "there is a uniquely fixed and determinate set of particular features of the world the positive or negative evaluation of which all agents must share if they understand each other and are otherwise well informed about the (non-evaluative) facts". The success Davidson's argument for the objectivity of moral values, however, does not turn only on his holistic interpretation argument. It turns on his capacity to be described as a theory of objective *reasons*.

3. Could Davidson Be a Reason Fundamentalist?

The recent tradition in metaethics, at least since Nagel's *Possibility of Altruism* (Nagel 1970) has accustomed us to formulate issues about moral realism and moral objectivity in terms of the concept of *reason*: can our reasons for acting, and in particular our moral reasons, be objective and is the notion of reason primitive? And traditionally two main kinds of answers have occupied the terrain: on the one hand, the Humean view, according to which reasons can be analysed as combinations or beliefs and desires, understood in instrumental terms, and on the other hand a Kantian or neo-Kantian views, according to which reasons are primitive, and considerations which make us favour certain courses of action, moral or not. This debate over the nature of reasons has many dimensions. One concerns psychology and the question of whether reasons can be causes. Another concern moral psychology and the nature of motivation. And yet another concerns the ontology of reasons.[6]

Let us, along with a recent tradition, distinguish three views about attitudes and reasons. One is that there is a distinction between motivating reasons and normative reasons: reasons for which one acts or believes (reasons one *has*), and reasons which justify an action (or a belief) and that make it rational in the eyes of the agent and of his interpreters (reasons there *are*).[7] The second concerns the nature of the attitudes: do they consist mainly in beliefs and desires, which exhaust the list of reasons to act and to believe, as Hume famously argued? Or is the notion of reason autonomous and in some sense more fundamental? The third concerns the way reasons motivate: should we accept the "Humean theory of motivation", according to which reasons have to motivate us, and must at least involve desires? The answer that one gives to these questions determine what kind of stance one takes on the ontological problem of the nature of values and norms: are these in some sense real and objective?

Davidson is notoriously a defender of the Humean view of motivating reasons: reasons are causes, and his 1963 article "Actions, Reasons and Causes" is a landmark. But he also accepts that there are normative reasons. These are the

[6] As any reader who knows the field will immediately see, I simplify excessively. Some of the main references in these debates are Nagel 1970, Williams 1979, Smith 1994, Scanlon 1998, 2014, Dancy 2000, Parfit 2011, Skorupski 2010.

[7] For an excellent overview and account of these and related distinctions, see Alvarez 2009.

reasons that an agent or a believer takes to be his best reasons and those in the light of which an interpreter must evaluate the reasons of the interpretee. These reasons are governed by the "ideal" of rationality and the principles of interpretation. The motivating reasons of an agent are never fully separated from his normative reasons: the first are not only the reasons for which the agent acted (or believed) but also those for which he *would* have acted, were he rational. Because of this rationality requirement, the answer to the second question is more complex. Like Hume, Davidson takes attitudes to be basically beliefs and desires. He always said that his main inspiration was not only Hume, but also Ramsey. His early work in decision theory[8], as well as his reading of Anscombe on intention, led him to formulate the view that reasons, as psychological states, are composed of beliefs and desires, which are the basic mental states. In many of his later writings, including in his views on interpretation, he entertained the hope of basing his whole analysis of actions and beliefs on two building blocks: a theory of beliefs on the one hand and a theory of desires, conceived *à la Ramsey* as credences and utilities, and two basic attitudes, *holding true* and *preferring true*. This minimalism permeates most of his conceptions in the philosophy of action, of meaning and of mind. He hopes to account for mental states such as intentions, hopes, regrets, surprises, and other attitudes in terms of beliefs and desires alone, and with the less possible intentional notions. The same kind of minimalism inspires his view of meaning: a theory of truth, plus the constraints on the interpretation of beliefs should be sufficient without positing meanings or senses as separate entities. Very often, as in the case of intention and of meanings, he is led to revise his minimalism, and to distinguish various levels and kinds of intention, and to introduce speaker's meanings within his initial theory. But the goal of accounting for complex notions in terms of more simple ones was always his ideal. In spite of this basic Humeanism about the nature of attitudes, which make his views seem to be close to those of functionalists in the philosophy of mind, Davidson is not a strict Humean about attitudes, since he holds that there are normative desires, desires about what we ought to desire, or about what we have reasons to desire. As we saw in the previous section, he takes these desires to track objective values and norms. Indeed, he also holds that there are normative beliefs. So he clearly has a place for normative reasons in both his psychology and his ontology of attitudes. Turning, then, to the third question: does Davidson defend the "Humean theory of motivation"—that motivation goes by way of desires (Smith 1994)? He clearly subscribes to it, in the form of what has often been called (Williams 1979) an "internalist" requirement on reasons, which he formulates in his article on weakness of the will as a "principle of continence": "perform the action judged best on the basis of all available relevant reasons" (Davidson 1980: 46) but that we can, following Myers and Verheggen (2016: 149, 159) formulate thus: "Rationality requires people always to form motivating states in line with their normative beliefs, and so always to do what they have most reason to do." In others words, there is always a "practicality requirement" on rationality and reasons: "There could be truths about what people have reason to do only if people's motivating states could be, in an appropriate sense, either correct or incorrect."[9] This requirement

[8] See Davidson, Suppes and Siegel 1957, as well as a number of essays in Davidson 1970, 1980, 2004. On these early views see in particular the interesting essay by Harnay 2010.
[9] Myers 2012: 376, Myers and Verheggen 2016: 158.

entails that although there could be cases (such as *akrasia*) where we could contingently fail to act on the basis of our best reasons, we ought, all things consider, act on the basis of the reasons that we ideally and rationality consider to be the best.

The combination of these answers by Davidson to our three questions yields a view which is hard to describe as a form of Humeanism about reasons, in spite of the fact that it involves a strong Humean basis. Davidson agrees with Hume that motivation has to go by way of desires and with the internalist or practicalist requirement. Thus, he would agree with Humeans such as David Lewis (1988) that motivation could not go by way of beliefs, and not even normative beliefs about what is good or right.[10] But Davidson would not agree with the Humeans and other anti-realists that values are not objective, and are constituted by projections of our pro-attitudes, such as desires. For a Humean, simple or sophisticated, values can never be objective in the sense that there are truths about what we have reasons to do or to believe. Davidson is in this sense clearly a realist about reasons. But could he subscribe to "reason fundamentalism" in the sense in which philosophers like Parfit, Dancy, Scanlon or Skorupski have claimed that reasons are primitive, non-psychological and normative attitudes? According to such views—and to simplify outrageously—reasons are not combinations of beliefs and desires. They are "considerations" which "favour" certain courses of action or beliefs, which cannot be analysed further. And most importantly they are not psychological states. They are facts, either as autonomous entities in the world or true propositions.[11] For Davidson, reasons cannot be fundamental in this way. Although he uses the term "reasons" in the normative sense, he still considers them as combinations of (normative) beliefs and (normative) desires, in the Humean way. And he subscribes, as we just saw, to a form of internalism and of a practicality requirement, which some reasons fundamentalist accept, but which strong moral realists like Parfit (2011) do not accept. Thus, in discussing Christine Korsgaard's (1996) version of this requirement, Parfit writes:

> We have returned to one of our main questions: how we should understand normativity. Korsgaard would be right to claim that, when realists appeal to facts about what is normatively necessary, or about what we must do in the decisive-reason-implying sense, these people do not thereby explain how we are *motivated* to act in these ways. That is an objection to normative realism if, like many Naturalists and Non-Cognitivists, we assume that normativity is, or consists in, some kind of actual or hypothetical motivating force. But realists reject that assumption. When realists claim that we have decisive reasons to act in certain ways, they are not making claims about how, even in ideal conditions, we would be motivated or moved to act. On this view, as I have said, normativity is wholly different from, and does not include, motivating force (Parfit 2011, vol. 2: 422).

Davidson could not agree with this. Although he is not a constructivist in the way that Korsgaard is, Davidson considers reasons to be essentially tied to what is *believable* and *desirable*, or with what an agent ideally would believe or desire. And he always considers his objective reasons to be capable of motivating us. And if

[10] In other words, Davidson could not accept the existence of "besires", i.e., of states which could be both beliefs (susceptible of being true or false) and desires (motivating). See my essay Engel 2015.

[11] See in particular Dancy 2000, Parfit 2011, Skorupski 2010, Scanlon 2014 and Alvarez 2009.

reasons are facts, it is not because the facts are, so to say, out there. It is because we have converged on them through a process of interpretation[12]. So, he is certainly not a moral realist in the strong sense in which this view involves the irreducibility of reasons.

A further consideration for doubting that Davidson can be a full-blown realist about normative matters concerns his reluctance to accept the existence of real norms in the epistemic domain. The principle of charity, the principle of total evidence and truth itself are norms of interpretation, but Davidson rejected strongly the suggestion that truth could be a norm for belief.[13]

4. Davidson on Emotions and Values

That Davidson is not a moral realist in this strong sense has to do with one of his other basic commitments: a commitment to a form of naturalism, albeit of a non-reductive kind. Davidson would never describe himself as a "cognitivist" about moral values, and as a "non-naturalist", both being labels that philosophers like Dancy, Parfit and Scanlon are prepared to accept. But rather than describe his views in abstract terms, it seems to me better to consider these in the light of his views about the relation of pro-attitudes and judgements about values. All his discussions of realism about values are formulated in terms of one central pro-attitude, desire. But many contemporary views about moral realism have been formulated in terms of another kind of attitude, namely *emotions*. Many forms of moral realism and of anti-realism are conceived in terms of the relation of emotions to values. Examining how Davidson's view could be placed in relation to such forms of realism can give us give us hints about how he could reply to the objections about his interpretation argument formulated in the first section above.

Why should Davidson have considered his argument for moral objectivism in terms of the relation between emotions and values, rather than between beliefs, desires, and values? Because, as we saw in the first section, his holistic argument seems insufficient to give us the appropriate objectivity, and because the psychological basis on which he rests this argument is too unspecific and narrow. Our judgments of values are not only based on beliefs and desires. Most classical and contemporary views say that they are based on other mental states, of the affective kind, namely emotions. Although Davidson never explicitly considered this possibility, we can try to reconstruct what his view was when we attend to what he has to say on emotions.

Davidson wrote very little on the nature of emotions, a topic which was, at the time when he wrote his main essays on action and mind, not as fashionable as it is today.[14] He nevertheless dealt with emotions in his essay on Hume's cognitive theory of pride (1976), on weakness of the will (1970), on paradoxes of irrationality. Although his essay on Hume is mostly exegetical, it suggests very

[12] In this sense, Davidson is probably closer to Kantian constructivism than to the form of Platonism defended by Dancy, although he is not a constructivist. Neither would he accept Parfit's kind of realism about reasons.

[13] See his reply (1999) to Engel 1999, which he reiterated in his replies to me in the volume directed by Kotatko, Pagin and Segal 2001. His reluctance to accept an ontology of norms permeates his writings, in spite of his acceptance of normative desires.

[14] He nevertheless shows that he was acquainted with the main views of his time in the Anglophone literature, e.g. those of Ryle 1949, Kenny 1963, Thalberg 1977, Solomon 1980, Gordon 1987, Pitcher 1965, Lyons 1980, Rorty 1987.

clearly views which Davidson shares with Hume. The logical space of theories of emotion is wide (and all the more so that the field of emotions is rather imprecise: there are passions, sentiments, moods, feelings, which can all be more or less occurrent or dispositional). Almost all writers agree that emotions are associated with various kinds of behaviour and bodily expressions, that most of them involve certain feelings and have a certain phenomenology, that they involve a form of appraisal and a valence, that they are related to certain informational states, on which they are based, and many writers hold that they involve judgments about values or about value properties. Many theorists accept that emotions can have reasons, but not all accept that they can be rational. Not all writers accept the idea that emotions can be objectively assessed as correct or appropriate. Depending on the weight which they put on these respective features, theories differ. The James-Lange theory insists on the behavioural component, most psychologists accept the view that emotions motivate and are associated to desires, and cognitivist theories focus on the judgements associated to emotions. The diversity of views is equally great when it comes to the relations of emotions to values. Humean theories take emotions to be the bases of evaluative judgements, which cannot be objective and true. Perceptual views of emotions and most intuitionist views hold that emotions are based on perceptions and intuitions about values. Other theories, which I shall examine below, take them to be based on the fittingness of our attitudes.[15]

Davidson's view of emotions is mostly of what is called the cognitive kind: emotions are associated to judgments. In his article on Hume on pride he tells us that Hume's theory rests on the idea that pride involves a relation to propositions:

> Hume's account of pride is best suited to what may be called propositional pride—pride described by sentences like, 'She was proud that she had been elected president'. Hume more often speaks of being proud of something—a son, a house, an ability, an accomplishment—but it is clear from his analysis that cases of being proud of something (or taking pride in something, or being proud to do something) reduce to, or are based on, propositional pride. If Hume's theory is to cope with the other indirect passions, a propositional form must be found for each of them (Davidson 1980: 277-78).

The problem is: how to conceive the relation between the state or attitude in which the emotion consists and the proposition in question? Davidson is clear that the emotion involves two sorts of things: first some perceived feature (for instance, with fear, the perception of something frightening), and second a judgment (that the perceived thing is frightening). According to Davidson Hume's theory is causal:

> The basic structure of pride and its etiology as Hume saw them is clear: the cause consists, first, of a belief, concerning oneself, that one has a certain trait; and second of an attitude of approbation or esteem for anyone who has that trait (Davidson 1980: 284).

It seems clear that Davidson himself subscribes to this analysis. But there are

[15] For good summaries on the various theories, see Deonna and Teroni 2011, Tappolet 2016.

three ways of conceiving the relation of an emotion to a judgement. The strong cognitivist holds that emotions are identical to beliefs and to associated judgements. This seems to hold for some emotions such as surprise, hope and regret, which are not clearly associated to specific behaviours, although this leaves out the feeling element. A weak cognitivist will say that emotions have a propositional component or presuppose belief, but that this component is not identical to the emotion itself. This is the case for Hume's indirect passions: pride, humility, vanity, love, hatred, envy, pity, malice, esteem, benevolence, respect, and compassion. One can also envisage a form of moderate cognitivism: emotions have "cognitive bases", which are the representations associated to them, but these bases need not be judgments (they may be representations of some sort, or perceptions which are not propositional). This seems to fit various sentiments, such as the feeling of familiarity, or the feeling of knowing. Hume is a weak cognitivist. He describes the causes of pride, and does not say that pride is constituted by one's entertaining a belief and having an attitude of approval. He just says that these are components of an emotion. Likewise, Davidson's view seems to be closer to the weak cognitivist view (Green 2013). His view seems to be summarised in the following passage of his article "intending": "If explicit judgements represent pro-attitudes, all pro-attitudes may be expressed by value judgements that are at least implicit" (Davidson 1980: 86).

This leads us to the way in which he conceives the connection between emotions and values: Davidson takes emotions to be a kind of pro-attitude, in the very wide sense which he gives to this term, and that they are at least partially associated to desires. But he does not take them to be subjective, in the sense that they would always be relative to a specific agent or circumstance. On the contrary, there are indications that Davidson takes emotions to be assessable objectively. Some signs of this can be found in his lifelong interest for Plato's *Philebus*, on which he wrote his 1949 dissertation (Davidson 1990). Plato explicitly says that some pleasures and emotions can be false. The *Euthyphro* dilemma implies that certain features of emotion, such as love, can be either response dependent or objective: either piety is being loved by the Gods or Gods love piety because it is pious. So, emotions can be, as judgments about values, the basis of values which can, as we saw in the first section, objective and true. From all this it seems clear that Davidson cannot defend another model of the relation between emotions and values, according to which values are real entities which can be in some sense perceived. There are many such views, depending on what one takes the perception in question to be. Some sort of intuition in the style of theories moral intuitionism? A perceptual judgement basing the access to the emotion? Or some non-conceptual kind of representation (Tappolet 2016)? Although he does not address this issue, Davidson cannot subscribe to any of these views, not only because he finds the causal relation which is supposed to hold between perceptions and values to be mysterious and not naturalistically explainable, but also because he never takes the ontology of values to be that of entities present in the world or in some non-natural world.

We have seen the difficulties that Davidson encounters with his interpretation argument. They would be the same if he chose, instead of judgments of value based on desires and on beliefs, another kind of analysis, to which I now turn.

5. Davidson and the Fitting Attitude Analysis of Values

This analysis is what is now called the *fitting attitude* analysis of value. It is, in a sense, a cognitivist and judgmental view, since it takes values to consist in a certain relation between attitude and judgments about the correctness of the attitude. In a nutshell, values are neither the expressions of our attitudes nor independent realities that could be perceived. They are the "formal objects" of our attitudes. Evaluative concepts have to be explained in terms of fitting or appropriate emotions. To use one of the possible formulations of this view (borrowed from Tappolet 2016, Ch. 3):

(V) X is a value if and only if there is an attitude which is fitting (or appropriate or correct) in response to x.

On this view, values are response-dependent, as they are for Humeans, but they are neither subjective nor projections out of our attitudes. They are based on our judgements about the correctness, or the *reasons* that one has to have these attitudes. In this sense, they entail a version of the "reasons first" conception of values: the main normative concepts are not axiological, but judgmental and associated to our reasons.[16]

Davidson does not explicitly discuss this view, which has roots in Brentano, but he was certainly familiar with it from his reading of Kenny's *Action, emotion and will* (1963) and of Chisholm, the main representative of Brentano in the U.S., whose views on action he both knew and discussed intensively.[17] The fitting attitude analysis starts from emotions. It does not say that they can be true or false, but that they are fitting or not, and this fittingness is itself an objective matter. The basic idea, which Davidson knew from Kenny (1963) is that emotions have a formal object, which is a value property.[18] Thus the formal object of fear is the *fearable*, the formal object of love is the *lovable,* the formal object of admiration is the *admirable*. But how can the view be made something other than a tautology? One can fear objects which are not *fearable* (little innocuous spiders) or which do not exist (monsters). How can the view yield objective values?

Davidson would probably have rejected the notion of a formal object of emotions and of other attitudes, mostly because he would have seen in it the reintroduction of an ontologically loaded notion of intentionality, which would violate his strictures on interpretation. But apart from these ontological worries, the main difficulty is: what kinds of facts can secure the fit between emotional attitude and

[16] This is why the view is sometimes called the "buck passing" account of values: the buck is passed to reasons. Scanlon 1998, 2014, Skorupski 2010 are the main contemporary defenders of such a view.

[17] See Brentano 1889, Chisholm 1957, 1976, and Davidson's essays in reply to Chisholm in Davidson 1980.

[18] When he deals with the formal object of actions, Kenny tells us that one encounters the problem of "variable polyadicity" of action verbs: how can they have a single formal object, given that actions are relative to all sorts of circumstances, such as when, how, where, by whom the action was done? Davidson's answer to this problem in "The logical form of action sentences" (1967, in Davidson 1980) is well known: he proposes to add to action predicates argument places for events, and to construe action sentences as quantifying over them. Thus, he breaks down the very notion of a formal object into a core property (expressed by the action verb), events and the properties of these events. This entails the rejection of the very notion of a formal object of actions. One can presume that Davidson would have rejected in the same way the notion of a formal object of emotions.

value? Facts about human nature? Or biological facts? Or social facts? It is hard to accept such views without falling back into a form of reductive naturalism, and without running into Moore's open question argument: any account of values or norms in terms of natural facts lose the normative character of such concepts. Another related difficulty is: how to specify the notion of fittingness? In other terms, how can we be sure that the reasons that are supposed to make an attitude correct, hence to secure the objectivity of values? Here we find again the opposition between a Humean view of reasons, which takes them to be relations of means to ends of any kind, and an objectivist notion. I may, for good instrumental reasons (say because I want a promotion or not to lose my job) find admirable something which is not admirable (*e.g.* my boss' tie, which is ugly) and desirable something which is not desirable (a saucer of mud).

Could Davidson have accepted such a view? There is an obvious similarity between his argument about the objectivity of values on the basis of convergence in judgments about values and the role which the fitting attitude analysis confers on our judgments on values. But where Davidson aims to solve the objectivity problem through his holistic conception of interpretation, the fitting attitude analysis aims to capture directly the objective values from the rightness or correctness of the emotions associated to it. But it is not clear that it can succeed better than the holistic approach. Considering the difficulties just mentioned, Ronald de Sousa argues that the appearance of tautology of the fitting attitude analysis (the formal object of love is the lovable, of fear the fearable) can be dispelled "because the attainment of success for emotions—the actual fit between the object or target of the emotion and its formal object—depends on a vast holistic network of factor that transcend my actual responses (de Sousa 2005). But if this is the case, what is this "holistic network", if not the one which makes our attitudes and beliefs holistically dependent? Holism went out by the window. It seems to have been reintroduced through the door.

6. Conclusion

The way out of these difficulties has been suggested by a number of partisans of the fitting attitude analysis of value: the conditions of the correctness of emotions and attitudes should not be specified in descriptive or factual terms, since no amount of facts could account for values and norms. The conditions have to be specified in terms which are themselves normative. (V) above should thus be reformulated:

(N) X is V iff there is an attitude which is required in response to x.

This obviously runs the risk of circularity: emotions are governed by norms which are defined by their formal objects, which in terms are defined in normative terms. What angers me must be wrong. What kind of "explanation" is that? Clearly it cannot be an explanation. But this is not an obstacle, once we understand that the fitting attitude analysis does not aim at defining the applications conditions of the norms for an emotion. It aims at stating what can be inferred from our practices. It involves necessarily an element of idealization. The attitudes that are correct are not those that are made so by a certain range of natural facts, but those that one ideally would reach if one turned one's back on those facts, and tried to adopt an impartial point of view. This form of idealization is nothing different from the objective standpoint on values and norms, which Davidson meant to be reachable from his interpretation argument. So the conclusion is that, even if he had adopted,

as I have suggested that he should have, the fitting attitude analysis, he would not have landed in a too different place from the one that he actually reached.[19]

References

Alvarez, M. 2009, *Kinds of Reasons*, Oxford: Oxford University Press.

Amoretti, M.C. and Vassallo, N. (eds.) 2008, *Knowledge, Language, and Interpretation. On the Philosophy of Donald Davidson*, Frankfurt: Ontos Verlag.

Bilgrami, A. 2006, *Self-knowledge and Resentment*, Harvard: Harvard University Press.

Brentano, F. 1889, *Von Ursrpung Sittlichte Erkenntnis*, Leipzig: Dunker and Hunblot; Engl. tr. 1931, *Our Knowledge of Right and Wrong*, revised by R. Chisholm, London: Routledge.

Chisholm, R. 1957, *Perceiving*, Ithaca: Cornell University Press.

Chisholm, R. 1976, *Person and Object*, London: Allen and Unwin.

Dancy, J. 2000, *Practical Reality*, Oxford: Oxford University Press.

Davidson, D. 1957, *Decision Making: An Experimental Approach*, Stanford: Stanford University Press.

Davidson, D. 1963, "Actions, Reasons, and Causes", in Davidson 1980, 3-19.

Davidson, D. 1969, "Mental Events", in Davidson 1980, 207-27.

Davidson, D. 1970, "How Is Weakness of the Will Possible?", in Davidson 1980, 21-42.

Davidson, D. 1976, "Hume's Cognitive Theory of Pride", in Davidson 1980, 277-90.

Davidson, D. 1980, *Essays on Actions and Events*, Oxford: Oxford University Press.

Davidson, D. 1984, "Expressing Evaluations", in Davidson 2004, 19-37.

Davidson, D. 1986, "Paradoxes of Irrationality", in Davidson 2004, 169-188.

Davidson, D. 1986a, "The Interpersonal Comparison of Value", in Davidson 2004, 59-73.

Davidson, D. 1990, *Plato's Philebus*, New York: Garland; repr. London: Routledge.

Davidson, D. 1995, 'The Objectivity of Values', in Davidson 2004, 39-57.

Davidson, D. 1999, "Reply to Pascal Engel", in Hahn 1999, 460-62.

Davidson, D. 1999a, "Spinoza's Causal Theory of the Affects", in Yovel, Y. (ed.), *Desire and Affect: Spinoza as Psychologist*, New York: Libble Room Press, 95-111.

Davidson, D. 2004, *Problems of Rationality*, Oxford: Clarendon Press.

Davidson, D. 2005, *Truth, Language and History*, Oxford: Oxford University Press.

Deonna, J. and Teroni, F. 2011, *The Emotions: A Philosophical Introduction,* Routledge: London.

De Sousa, R. 2005, "Emotional Truth", *Proceedings of the Aristotelian Society*, 76, 265-75.

Engel, P. 1999, "The Norms of the Mental", in Hahn 1999, 444-60.

Engel, P. 2001, "Is Truth a Norm?", in Kotatko, P. *et al*. 2001, *Interpreting Davidson*, Stanford: SLI, 37-51.

[19] I agree in this respect with Myers and Verheggen 2016, that in this sense Davidson's view may be close to that of Scanlon 2014.

Engel, P. 2008, "Davidson on Epistemic Norms", in Amoretti and Vassallo 2008, 123-46.

Engel, P. 2012, "Davidson and Contemporary Philosophy", in Lepore and Ludwig 2013, 588-604.

Engel, P. 2015, "Une croyance nommée désir", *Klèsis*, 31, *La philosophie de David Lewis*, http://www.revue-klesis.org/pdf/Klesis-Lewis-II-1-Engel-Une-croyance-nommee-desir.pdf (last accessed November 15, 2017).

Gordon, R. 1974, "The Aboutness of Emotion", *American Philosophical Quarterly*, 27, 11-36.

Green, M. 2013, "The Rationality of the Emotions", in Lepore and Ludwig 2013, 506-18.

Hahn, L. (ed.) 1999, *The Philosophy of Donald Davidson*, La Salle: Open Court.

Harnay, P.-V. 2010, "Une théorie du langage peut-elle fonder les comparaisons interpersonnelles? L'apport de Donald Davidson", *Revue de philosophie économique*, 2010/2, 11, 103-39.

Jackson, F. 1998, *From Metaphysics to Ethics: A Defense of Conceptual Analysis*, Oxford: Oxford University Press.

Kenny, A. 1963, *Action, Emotion and Will*, London: Routledge.

Korsgaard, C. 1996, *The Sources of Normativity*, Cambridge, Cambridge University Press.

Lewis, D. 1988, "Desire as Belief", *Mind*, 97, 323-32; repr. in his *Papers in Ethics and Social Philosophy*, Cambridge: Cambridge University Press, 2000.

Lillehammer, H. 2007, "Davidson on Value and Objectivity", *Dialectica*, 61, 203-17.

Lepore, E. and Ludwig, K. (eds.) 2013, *A Companion to Donald Davidson*, Oxford: Blackwell.

Lyons, W. 1980, *Emotion*, Cambridge: Cambridge University Press.

Mackie, J. 1977, *Ethics, Inventing Right and Wrong*, London: Penguin.

Mulligan, K. 1998, "From Appropriate Emotions to Values", *The Monist*, 81, 161-88.

Myers, R.H. 2004, "Finding Value in Davidson", *Canadian Journal of Philosophy*, 34, 107-36.

Myers, R.H. 2012, "Desires and Normative Truths: A Holist's Response to the Sceptics", *Mind*, 121, 482: 375-406.

Myers, R.H. 2013, "Interpretation and Value", in Lepore and Ludwig 2013, *A Companion to Donald Davidson*, New York: Blackwell, 314-27.

Myers, R.H. and Verheggen, C. 2016, *Donald Davidson's Triangulation Argument: A Philosophical Inquiry*, New York: Routledge.

Nagel, T. 1979, *The Possibility of Altruism*, Oxford: Oxford University Press.

Parfit, D. 2011, *On What Matters,* Oxford: Oxford University Press.

Pitcher, G. 1965, "Emotion", *Mind,* 74, 295, 326-46.

Rovane, C. 2013, *The Metaphysics and Ethics of Relativism*, Cambridge, MA: Harvard University Press.

Scanlon, T. 1998, *What We Owe to Each Other*, Cambridge, MA: Harvard University Press.

Scanlon, T. 2014, *Being Realistic about Reasons*, Oxford: Oxford University Press.

Skorupski, J. 2010, *The Domain of Reasons*, Oxford: Oxford University Press.

Smith, M. 1994, *The Moral Problem*, Blackwell: Oxford.

Solomon, R.C. 1984, *The Passions: The Myth and Nature of Human Emotions*, New York: Doubleday.

Stroud, B. 1999, "Radical Interpretation and Philosophical Scepticism", in Hahn 1999, 139-61.

Rorty, A. (ed.), 1980, *Explaining Emotions*, Los Angeles: University of California Press, 103-26.

Tappolet, C. 2016, *Emotions, Values and Agency*, Oxford: Oxford University Press.

Tenenbaum, S. 2007, *Appearances of the Good*, Cambridge: Cambridge University Press.

Thalberg, I. 1977, *Perception, Emotion and Action: A Component Approach*, Yale: Yale University Press.

Velleman, D. 1992, "The Guise of the Good", in *The Possibility of Practical Reason*, New York: Oxford University Press.

Williams, B. 1979, "Internal and External Reasons", in his *Moral Luck*, Cambridge: Cambridge University Press, 1981.

Williams, B. 1985, *Ethics and the Limits of Philosophy*, London: Fontana.

Norm and Failure in Mind and Meaning

Akeel Bilgrami

Columbia University

Abstract

The paper first gives an argument for the Davidsonian thesis that norms constitute the human mind. Then it shows that that thesis is better formulated by Wittgenstein rather than by Davidson himself. And finally, it uses the Wittgensteinian formulation of the thesis to establish why Davidson was right to further claim that linguistic meaning was not normative despite the human mind being normatively constituted. Through this entire dialectic of the paper, the concept of failure is made central to the argument.

Keywords: Mind, Intentionality, Meaning, Language, Causality, Normativity, Value, Agency, Science, Psychology, Failure, Is-ought, Fact-value, Disposition.

1. Introduction

Donald Davidson was a pioneer among philosophers in arguing that normativity was central to understanding human behaviour and that it was what set apart the understanding and explanation of human behaviour from how we understand and explain all other phenomena. Though it is true that long before him, philosophers, in resisting the overreaching claims of positivism, had claimed that the social and human sciences were value-laden, those philosophers had not (or at least not explicitly) made value or norm constitutive of the human *mind*.[1] Davidson made this last claim central to understanding human behaviour and saw in it the roots of what made the study of human behaviour and society distinctive.

Having argued this, in an extremely surprising and paradoxical turn, Davidson went on to refuse the idea that normativity also constituted linguistic meaning. How is this possible without inconsistency? How is it that human linguistic behaviour *in particular*, i.e., the utterances of sounds with meaning, are not normative whereas human behaviour, *in general*, is to be understood as nor-

[1] Davidson has presented this idea in many papers of which the most detailed in presentation perhaps is "Mental Events" (1970).

Argumenta 3, 1 (2017): 69-85
ISSN 2465-2334

DOI 10.14275/2465-2334/20175.BIL

matively constituted? Davidson never addressed this last question, indeed never so much as raised it.

I will argue in this paper that though Davidson was right in his claim that the mind is constituted by normativity, he had a mistaken understanding of what that claim amounted to. I will try and show this to be so by contrasting Davidson's understanding of the claim with Wittgenstein's earlier way of elaborating it. This, of course, implies that Wittgenstein was the real pioneer in making the claim but my aim is not to judge who gets the prize for having made it first, but rather to assess who gets the claim more right. Having done so, I will then present at the end of the paper something like an argument for the conclusion that Davidson makes about *meaning* not being normative, a conclusion for which Davidson never gave an explicit argument but which is available to be given if one has a proper Wittgensteinian and not Davidsonian understanding of how and why the *mind* is constituted by normativity.

The concept of 'failure' will play a central role in the way I approach my dialectic to present these ideas and conclusions.

2. Failure, Norms, and Norms of the Mind

Failure, the very idea, presupposes a *norm,* by the lights of which it gets counted as such. And so, failure, I will argue, is essential to understanding the nature of norms.

But I begin with a qualifying restriction. There is frequent talk of failure that presupposes something less (or other) than a norm, as when we speak of 'heart failure' or 'engine failure'. What is presupposed in these latter expressions cannot—strictly—be a norm because these are breakdowns or cessations of a *mechanism* (whether natural or artificial). Mechanisms are defined by the presence of a *causal disposition* or *tendency* of nature or artifice. Hearts and engines are disposed to or tend to behave in certain ways under certain conditions. Under these conditions, if these natural or constructed tendencies proceed without interruption or obstacle, they are said to be functioning 'well', but the term 'well' here is not—again, strictly—a normative assessment; it merely registers that the causal disposition has been triggered and that the tendency is unhampered. Hence, strictly speaking, when a heart or engine fails, these failures presuppose only a descriptive notion of what is 'normal', which by a familiar sort of alchemy gets transformed in our careless understanding into something prescriptive or normative, a transformation whose genealogy has been illuminatingly studied by philosophers such as Foucault and Hacking under the label 'normalization'.[2]

I will not be looking at this sort of understanding of 'failure', only at failure, *strictly so called*, where the lights by which it is seen to be a failure is a 'norm' in the full and irreducible sense of the term, not a norm that reduces, in the end, to a descriptive *tendency* of nature or artifice while presenting itself on the surface as a prescriptive and evaluative standard. I described this second-class idea of norm as part of a 'careless' understanding of the term but the point of Foucault's and Hacking's analysis is precisely that it is not careless (nor second-class) at all but

[2] Hacking's brilliant elaboration of this may be found Hacking 1990 and Foucault's pioneering formulation of it is presented most explicitly first in Foucault 1977 and 1978: vol. 1, but is sophisticated in subsequent works such as, especially, Foucault 2003a, 2003b, and 2006.

part of a long institutional and social construction that affects the disciplining both of subjects of study and often (via such study) the structures of social and political domination through what Foucault calls 'bio'-power. I have no wish to deny what they say. But since my interests here, like Davidson's, are narrower and more purely methodological, I am concerned to distinguish this understanding of norms from norms that are not reducible in this way to causal dispositions and tendency. I do not even want to deny what is surely frequent—that once these causal and empirical tendencies get 'normalized' and erected into norms, many of these norms, so erected, get a life of their own and are not reducible, even eventually, to their genealogical ground in these causal and empirical tendencies. If so, that is a normativity that we bestow on them and that is not reducible to their genealogical basis in tendency. So understood, they are indeed norms, strictly so-called, in the sense that is this paper's topic. What I am certain of, however, is that the examples I gave, our talk of engine failure and heart failure, cannot, just as they stand, possibly presuppose norms in the properly strict and irreducible sense. When there is failure to live up to norms, in that strict sense of norms, we are potentially subject to criticism. But though we may, of course, criticize some*one* for not maintaining a heart's health or an engine's operability, we do so only because there are other values and norms—the value of life, perhaps, or of material productivity—that the functioning of the heart or machine respectively make possible. That does not imply that a heart's or engine's failure, qua failure of a mechanism of nature or artifice, in itself presupposes any norm in the strict sense.

In the little space I have, I cannot argue for this fundamental distinction between tendency and norm.[3] I will simply take it for granted. The distinction, though it has not gone without being contested, is intuitive and plausible. It seems to be everywhere evident in human thought and action. When we say of two beliefs, p and q, that 'q follows from p' we could mean by this two quite distinct things: 1) in subjects capable of belief, the belief that q *tends* to follow the belief that p (as a matter of *causal disposition*) and 2) subjects capable of belief *ought* to (as a matter of *rationality*) believe that q, given that they believe that p. There are very familiar distinctions at play here: between cause and reason, fact and norm, is and ought; and, as I said, I will only consider failures in those episodes of human thought or action that fall afoul of the latter in each of these pairs of distinctions—failures of reason, doing what one ought not or not doing what one ought, violation of norms in the full and irreducible sense of the term.

Before I move on from these preliminary clarifications and distinctions, let me declare one more restriction that I will impose on this paper's theme. When studying failure the primary interest in the social sciences has, quite understandably, been human behaviour that runs afoul of *social*, *political*, *legal*, and *ethical* norms since those are the norms that these disciplines are most concerned with—speaking too loudly, as it might be, or driving on the wrong side of the road, committing perjury, breaking a promise, etc. Though my concern in what follows is of indirect relevance to explanations in the study of society, my direct and primary interest, again like Davidson's, is not in such failures but in the

[3] There has, as is well known, been an enormous amount of writing on the subject among philosophers both insisting on and (ultimately) denying the distinction. I give an argument for the distinction in Bilgrami 2006: Ch. 5.

failures of thought and action of individuals by *their own* lights, whether those lights coincide with the lights of social, political and other such norms or not. (Of course, an individual's lights by which she assesses her own behaviour as amounting to failure, may very often be an internalization of social norms. How could it fail to be? But that is the genesis of her lights. It does not spoil the idea that they are *her* lights. And it is, *as such*, that I am concerned with norms and our falling afoul of them in failure.)

An example of failures by our own lights may be found in the domain of psychoanalysis or psychotherapy. We may be ostracized or sent to gaol if we fall afoul of social or legal norms but we feel guilt or go to analysts or therapists because we are, by our own lights, dissatisfied with our minds and actions. When I suffer from an anxiety that prompts me to seek therapeutic attention (I am putting aside here anxieties that owe exclusively to biochemistry and for which we turn exclusively to pharmacological treatment), it is by some lights of my own that I find my behaviour to be wrong—and it need not be a moral, social… wrong. I mention psychoanalysis and therapy only as examples. The phenomenon of failing by one's own lights is far more common and far more unremarkable than is suggested by these examples. It is ubiquitous and mundane and, on the face of it, often much less interesting than the cases that are of interest in psychoanalysis. But its study may have fundamental implications for how to understand what is distinctive about the explanation of human behaviour. This is the point that has emerged with much force ever since Wittgenstein and then later Davidson put it on centre-stage.

With these two restrictions of my thematic interests declared—a restriction to failures owing to *strict and irreducible norms* and to failures within *individual* psychology—let me turn now to expounding how norms figure in individual human behaviour and psychology.

3. Norms, Causes, and Reasons

Ever since Aristotle codified the explanation of individual human behaviour in the practical syllogism, a varied terminology has been deployed to identify the states of mind that go into such explanations: 'beliefs', 'desires', and 'intentions' are the terms philosophers primarily deploy, though in the behavioural and social sciences with the decision-theoretic sophistications of Aristotle's practical syllogism to include *degrees* of belief and desire, such terms as 'preferences' or 'subjective utilities' (for desires) and 'subjective probabilities' (for beliefs) have been coined. In what follows, for the sake of simplicity, I will speak only of 'beliefs' as an omnibus term standing in for the cognitive states that go into the explanation of human behaviour, 'desires' as such a term for the conative states, and 'intentions' as such a term for the more decisional states. Thus, (very roughly) following Aristotle's practical syllogistic schema, someone may be said to desire that her thirst be quenched, believe that drinking a glass of water is the best way to satisfy that desire, and form thereby an intention to drink a glass of water. These states of mind, if all goes without hindrance, result in her drinking a glass of water. (Aristotle himself leaves the intention out and goes straight to action from the desire and belief.)

Where is normativity supposed to enter into all this? It is familiar from Davidson's writing that it is as follows: the relation between the states of mind (the belief, desire, and intention) and the human behaviour (drinking the water) is

said *not merely to be a causal* relation with the former causing the latter, but a relation that shows the latter, the behaviour, to be *rational*, given the presence of the former, the states of mind. (Here, as I said above, the notion of rationality or reason is not derived from social or legal or political norms, but from norms that govern the relations among an individual subject's own states of mind, i.e., norms of consistency, transitivity, coherence, etc., and the relations between those states of mind and her behaviour, i.e., the norms of practical syllogistic reasoning or, in its more sophisticated form, the norms of decision theory.)

Though both the element of causality and of rationality are thus in play, the relations between these two elements need much sorting out because, on the face of it, it is not clear how exactly they relate to one another; in fact, they seem to run up against each other and the rational element seems, at least at first sight, to cancel out the causal element. Failures are essential to understanding how this happens.

Take the Aristotelian practical syllogism above as a causal claim. If it is a causal claim comparable to the causal claims of natural science, it will not restrict itself to just that particular claim about that person drinking water on that occasion but aspire to greater generality. Let us transform it to a more general claim as follows:

> The desire that one quench one's thirst, the belief that drinking water is the best way, all things considered, to quench one's thirst, and so the intention that one drink water, *cause* one to drink water.

Now, like all causal generalisations, whether in natural science or about human behaviour, this one too will have to be qualified by ceteris paribus clauses to rule out spoiling conditions that have the effect of the generalization *failing* to hold. So we will have to prefix such a ceteribus paribus clause to the causal claim above: "*All things being equal*, the desire that__ the belief that__ and the intention that__ cause__".

But now there is a problem. We have no idea of a general *sort* about what is being held steady (or equal) in the ceteris paribus clause of such a causal claim. This is because we cannot really gather the different things that *spoil* such a causal claim into general categories or sorts of spoiling conditions. On one day, someone may not drink the water because he prefers just then to drink orange juice even though he knows water is better for him (for his health, say), on another day he may not drink it because he is feeling lazy, on yet another day, he may not drink it because he gets distracted…and so indefinitely on and on. There is no common element in these spoilers that can be informatively stated *ex ante*. We wait for the failures due to one or other of an indefinite number of such spoilers and the 'all things being equal' clause rules each out *only ex post*. There really is no informative thing we can say in advance about what sorts or kinds of things cause the failure of the causal generalization to hold. In advance, at best we can say something completely *un*informative such as: If she believes__ desires__ and intends__ then, *if she is rational*, she will do__.

But that really only shows that the causal claim, qua causal claim, has no empirical weight or punch, no informative, explanatory strength and power. The 'all things being equal' clause (now taking the form of an assumption of her rationality) tells us nothing that carries *empirical* information about what will spoil the causal claim and therefore what has to be ruled out. It has a 'whatever it takes' quality to it, and the mention of rationality is an admission of the com-

pletely normative nature of these allegedly causal generalisations, merely pretending on the surface to causal explanatory power. Compare such generalisations, say, to the law of falling bodies and you immediately get a sense of the contrast between the two sorts of explanations. When, it comes to the laws of natural science, we have a relatively clear *ex ante* idea of what has to be held steady by the ceteris paribus clause. That is to say, the kinds of things that would cause the generalization to *fail* to hold are things that we have a clear understanding of in advance of those failures. We know and can state in advance the *sorts* of things that spoil the generalization from holding and so we can state informatively in advance the conditions that have to be held equal. Both the terms 'ex ante' and 'sorts' are important here in understanding failure. In explanations of human behaviour conditions under which causal claims fail cannot be *sorted* into general kinds of conditions and stated *ex ante*. They can only be observed after the fact because there is nothing that they share in common that we have a grip on in advance as a general *sortal* claim that is informative. That is why we simply appeal—a waving of the hand, as it were—in the ceteribus paribus clause to the agents' rationality to save the causal claim, but in doing so it ceases to be a causal claim with any empirical import and the normative element (of an assumption of rationality) in human behaviour replaces the causal element on centre-stage. That gives a preliminary hint of the ineliminable normativity in our understanding of human minds and behaviour.

4. Wittgenstein vs Davidson on Norms of Mind

At this point, there is an interesting disagreement that might arise between two positions on how exactly to understand this normative element that so dominates the individual human mind and what constitutes failure by the its lights.

One position was made familiar by Davidson.[4] Our states of mind such as beliefs, desires, and intentions are *causal* dispositions to behave in certain ways. But unlike the causal dispositions studied by natural science (the solubility of salt, the fragility of glass, which result, under the relevant triggering conditions—the placing in water, being struck by a rock—in salt dissolving and glass shattering), such dispositions of the mind which result in human behaviour have a further feature: they are *answerable to the norms* of logic and decision theory. So, normativity enters human mentality in the form of *general principles of rationality* (the principle of non-contradiction, the principle of transitivity, etc.) to which the mental states that are disposed to cause human behaviour are answerable.

A quite different view owes to Wittgenstein. Wittgenstein claimed[5] that states such as beliefs, desires, and intentions are *themselves* primarily normative states, not causal and dispositional states. So normativity is not restricted to general principles of rationality (to which our mental dispositions such as beliefs, desires, intentions, are answerable). It is more widespread. Our beliefs, desires, intentions are, each and all, themselves normative states. What does it mean to say that mental states are themselves normative? Wittgenstein gives the following sort of example to illustrate what he has in mind. Take intentions. If I

[4] See again Davidson 1970.
[5] See his extensive discussion of both 'intentions' and 'expectations' in Wittgenstein 1953 for this view of mental states.

intend to take an umbrella when I go to work in the morning, then, *just by forming that intention*, I have generated a norm. This norm will be the basis of an assessment of my future behaviour. If I take the umbrella, I have acted in *accord* with the norm that is my intention. If I do not take it, I have *failed* to live up to that norm. And 'accord' and 'fail', as I said at the very outset, are terms that presuppose norms. To express this normativity one could, so long as one is clear that it is not a moral ought but a broader ought of rationality, say that if I intend to do something, then I ought to do it. Failing to do it is to fail of the rationality demanded by a state of mind such as an intention. We may say similar things about desires. A desire, being less decisional (or being pre-decisional in the course that practical reason takes) than intentions, may be overridden by other desires, so the norm it generates is only a prima facie norm, a prima facie ought. If I desire that I help the poor, then prima facie (unless it is overridden by other desires) it will generate a norm that will assess my future behaviour as being in *accord* with it or not. If I give money to Oxfam, for instance, I will have acted in accord with the desire that is my norm. If I do not, nor do any similar thing, I will have *failed* to live up to the norm that is—or is generated by—my desire. (About desires in particular, I should add a caveat: the term 'desires' in ordinary talk is not always used in this normative sense but rather to stand for *urges*. When it does so, it is not in the realm of normativity. Urges are precisely and primarily tendencies and causal dispositions and lack the primary normative sense that Wittgenstein has in mind. So, the term 'desire' is ambiguous. When one says, I have a desire to smoke a cigarette, I could mean either that I have an urge or I have a commitment to smoke. I can, of course, have both an urge and a commitment to smoke, but when I do, I have two distinct states of mind, not one. In what follows when I speak of desires I am restricting myself, like Wittgenstein, to the normative rather than the dispositional sense of the term. That is only to be expected since my interest in this paper is in the phenomenon of failure and, therefore, in the normatively inflected explanation of human behaviour.) Beliefs too are norms. If I believe that there is a table in front of me, that is a norm in the sense that I ought to, it commits me to, believing and not believing a range of other things: it commits me to believing that there is something in front of me, it commits me to believing that if I run very fast into it, I will likely hurt myself,[6] it commits me to not believing that there is nothing in front of me; and so on. If I do believe (and refrain from believing) these other things, then I am in accord with the norms that are generated by my belief; if not, I have failed to live up to those norms.

I have just used the word 'commits'[7] in describing the normativity of beliefs and desires. The term is a useful one and can be wielded generally to describe

[6] Here, obviously, I am committed to this only if I have some other beliefs as well, such as that tables are standardly made of a hard substance, that hard substances on impact of sufficient velocity hurt one's body, etc. All that shows is that mental states such as beliefs and desires are possessed not singly as nuggets but are holistically and inferentially linked.

[7] The idea of commitment in the study of the human mind was first elaborated in any detail by Levi 1983 in the context of a highly sophisticated and original theory of belief revision. Since then, Brandom 1994 has also deployed it in his *Making It Explicit* within a framework of 'score-keeping' and 'entitlements', and I have discussed it along quite dif-

the normativity of all such states of mind. When I have these states of mind, they are like commitments I undertake, as it were within myself (and not within the framework of some external or social contract). They are commitments in the sense of being '*internal* oughts'. Intentions commit us to certain actions, desires commit us prima facie to certain actions, and beliefs commit us to certain other beliefs.

And so, a useful way to put the difference between Davidson's and Wittgenstein's views might be this: Davidson thinks that our states of mind such as beliefs, desires, and intentions are not themselves commitments but rather (causal) dispositions which are answerable to the only real commitments we have—to the principles of rationality such as consistency, transitivity, etc. By contrast, Wittgenstein thinks that our states of mind are themselves commitments or internal norms. Our very possession of such states of mind are *commitments* to doing and thinking various things. So for Wittgenstein, the mind is cluttered with far more norms or commitments than it is for Davidson. Davidson's view of norms is an austere one (the only commitments we have are our commitments to the principles of rationality such as consistency, transitivity, etc., all other states of mind are dispositions). Wittgenstein's is a more bloated view (each and every belief, desire, and intention is itself a commitment).

Who is right? Much may turn on the answer to this question.

Both views presuppose that *norms* are irreducible to the causal and physical states of nature as the natural sciences study them. This presupposition is common ground for both of them and it is a view that I have taken for granted too, as I said at the very outset of this essay. Both views moreover claim that the *mental states* human beings possess are also irreducible to the causal and physical states studied by the natural sciences. But, on Davidson's view these (the irreducibility of norm to nature and the irreducibility of mental states to nature) are *two different* irreducibilities, whereas on Wittgenstein's view they are the *same* irreducibility. This is because for Wittgenstein beliefs and desires and intentions are *themselves* norms or commitments on a par with legal, political, ethical norms, only restricted to individual mentality. So the irreducibility of these mental states is just a special case of the general irreducibility of norms. For Davidson, on the other hand, beliefs, desires and intentions are not themselves norms. They are causal dispositions. The only norms of mind there are, are the principles of rationality. However, our mental states, despite being causal dispositions are governed by or answerable to these principles of rationality in a way that the dispositions studied by the natural sciences are not, and that is why they are not reducible to the latter. Hence, for Davidson, there are two distinct irreducibilities—of norm to nature and of mental states to nature.

One might think Ockham's razor should be sufficient to make us favour Wittgenstein's view over Davidsons, but I think there is a deeper reason to do so. And *failure* of mind is a good way to bring out the deeper reason. Let me illustrate this with an example or scenario of failure.

Suppose I believe that p. And suppose also that, in a fit of distraction, when asked, I assent to something that implies not-p. Or to keep things simpler, sup-

ferent lines in an analysis of intentionality in Bilgrami 2006: Ch. 5. But the basic idea really goes back, as I have said above, to Wittgenstein 1953.

pose that when asked I assent to not-p. I have certainly in some sense violated a principle of rationality, the principle that demands consistency, what Aristotle called the principle of non-contradiction. Now, according to Davidson that is the only failure on my part. But is it? If it were the only failure on my part, the only instruction to me would have to be: "You are inconsistent, so either give up the belief that p or withdraw your assent to not-p". But, in the scenario as I have described it, that is altogether the wrong instruction to give me. The right instruction to give me is just simply: "Withdraw the assent to not-p." Why? Because it is my belief that is a commitment whereas, just by the way the scenario unfolds, the assent to not-p is not the expression of a commitment, so not really an expression of belief in the normative sense. *It was made in distraction*, so no commitment was really made by the assent to not-p in the way that the belief that p amounts to a commitment. The example helps to bring out the sense in which beliefs themselves are commitments. In the example there is more than a failure of consistency, more than a failure of adhering to a principle of rationality that demands consistency. There is also *a failure to live up to a commitment that is present in the very existence of the belief itself*. The inconsistency involved is not between two commitments but between a commitment and an assent that does not express a commitment, an assent that is inconsistent with the only commitment that I have, which is the belief. In other words, Davidson misunderstands the nature of the failure of mind here. The scenario demands a different understanding of my failure than his instruction to me would suggest. But his position *requires* him to give me the wrong instruction to remedy my failure since his position has no other normative resources than the principles of rationality. His position therefore cannot accurately capture the real nature of my failure. It underdescribes my failure. The weakness of his position and the strength of Wittgenstein's position emerges in the fact that the latter alone would provide the right instruction in this scenario, and that would, in turn, bring out the sense in which beliefs themselves are norms or commitments. Similar examples can be run on desires and intentions.

5. Norms, Intentionality, and Explanation

There is much significance that follows from the special nature of this failure as I have just expounded it and what it reflects about the nature of the norms of our mentality.

Let me turn to that significance by first clarifying a little bit more what is meant by saying that beliefs, desires, and intentions, that go into the explanation of individual human behaviour are norms or commitments of this kind.

What is it to have such a commitment? I have said that it is not to have a causal disposition since it is normative and norms cannot be merely tendencies of nature. What this implies, then, is that one can have a commitment (a desire that one help the poor, say) and not be disposed, even prima facie, to act as the commitment requires. When that is so, one is chronically failing to live up to a commitment, but it does not mean that one lacks the commitment—if it did, we would lose the contrast between commitment (norm) and disposition.

So, a hallmark of a commitment, as of all norms, is that they do not cease to be what they are in the presence of failure. Indeed failures are essential to them in the sense that if there is no *possibility* of failure, it is doubtful that it is a norm we are speaking about. *Norms are norms only if there is a possibility of failing*

by their lights. If we were monsters of rationality and goodness, we would not have the normative concepts of rationality and morality.

It may seem that here we have a problem. If one can have a commitment and not be disposed to act on it, then it may seem that we cannot distinguish between someone who has a commitment but does not act on it and someone who does not have the commitment at all. In answering this difficulty, something important about the nature of commitments (or norms of mind) comes to the surface. The difficulty is removed only when we point out that all commitments—as a matter of definition—require *a conditionally formulated second-order disposition or dispositions*, which may be characterized as follows: when one has a commitment and fails to act on it, one is disposed to feelings of guilt, one is disposed to self-criticism, and disposed to efforts to try and do better by, for instance, cultivating the first-order dispositions to act in accord with the commitment. To define a commitment as requiring this second order disposition(s) does not reduce the commitment to a mere disposition or tendency because the second-order disposition cannot so much as be characterized except in terms of that commitment. So there is no elimination of the normative element, by bringing in a disposition of this kind at a higher order. Such a disposition is entirely parasitic on the existence of the commitment. (It is only if commitments were defined in terms of their corresponding *first*-order dispositions that we would be under the threat of reducing commitments to dispositions. But we have already said that there is no reducing commitments to first-order dispositions since the desire that I help the poor can exist without there existing any even prima facie disposition to help the poor.) And once we point out that commitments have this second order disposition built into them, we have answered our difficulty. We have a way of distinguishing someone who has a commitment and someone who lacks it. The subject who lacks the commitment lacks this second-order disposition possessed by the subject who has the commitment.

Following Wittgenstein, I have said that the states of mind (beliefs, desires, intentions…) that go into the explanation of what is distinctively (individual) human behaviour are normative states or commitments, and I have distinguished these from the more purely causal states or dispositions that human beings also possess. Does this mean that commitments have *no* causal power? Are they epiphenomenal, making no difference to the world, to the causation of human action. Are they merely relevant to answering the question "Is what someone did rational by her lights?" and not relevant at all to answering the question, "Why did someone do what she did?"

It would be an implausible limitation to impose on human subjectivity to make normative states of mind entirely epiphenomenal in this way. They cannot altogether fail to have some causal point. But if there is some causal relevance that commitments have, if they do make a difference to what occurs in the world, it had better not be the same sense of 'cause' that is present in the causal relations that are studied in the natural sciences since we know those to be causes in the merely (first order) dispositional sense. Does this mean we have two different notions of cause, one that is present in the explanations of physical behaviour studied by the natural sciences and a different notion that figures in the normatively inflected explanations of human behaviour? I cannot see any way of avoiding saying so.

If commitments have some causal effect, if they do make a difference to what individual subjects do, then statements of the form 'Her desire (understood

as commitment, not an urge) that she__caused her to__' (or more specifically, for instance, "Her desire that she help the poor caused her to give money to Oxfam") make perfectly good sense, just as much sense as statements like "Dipping the blue litmus paper in acid caused it to turn red". But they do not make the same sense. What is the difference in sense in these two uses of 'cause'?[8]

Understanding the nature and the implications of failure, as I said, is crucial to answering this question, crucial to capturing the distinctive form of causality that is present in the explanation of individual human behaviour. The natural sciences systematize the dispositions in nature, bringing them under generalizations and laws. To state what is obvious and well-known, the objects, whose causal, dispositional properties they regiment in this way are not agents and subjects in the way that the human objects of study are. Another way to put this utterly familiar point is to say that, unlike the explanations of human behaviour, they do not study phenomena which possess a richly configured 'first person point of view'.

A first person point of view is a property possessed by creatures with sentience and the property is often described with the omnibus term 'consciousness'. Being omnibus, that term is meant to capture a wide variety of phenomena. Some creatures with sentience, however, possess a first person point of view that consists not only of consciousness in the sense of what is 'given' to their senses (this is sometimes described with phrases like 'what it's like to be') but a wider phenomenology that includes the normative states we have been discussing such as beliefs and desires and intentions. These are richer (if that is the right word) since they bring within consciousness a complex element of 'self-consciousness', as was evident when I defined these normative states or commitments as requiring *second-order* dispositions to feelings of guilt, to self-criticism, and to efforts at trying to do better to live up to the commitments. There is no attributing these normative states or commitments to a subject without also attributing these higher-order dispositions. All this may seem obvious to anyone who has reflected even momentarily on what makes the behavioural and social sciences stand apart from the natural sciences.

But what we can infer, once the obvious is recorded, is something less obvious—the following distinction about causation and failure.

Suppose we have, on the basis of observation and theory, come to some causal generalization in the natural sciences—the one about acids and blue litmus, being an example. And now suppose that in future observation this generalization begins to fail to hold, we begin to observe that acids are not causing blue litmus to turn red. We cope with this failure by loosening, perhaps even eventually losing, our confidence in the initial causal generalization, in the causal power that we once thought acids to possess. In short, failure presents a certain form of crisis. It forces refutation of initially made generalizations. That, at any rate, is a model of how natural science proceeds, and Karl Popper was, of course, its most explicit philosophical theorist. As is well-known, this model was powerfully questioned first by Duhem, then Quine, and most influentially by Kuhn who described the crisis that is generated by failures of this kind in quite

[8] I am frankly claiming here that if we take normativity seriously in the study of mind we will have to introduce a distinctive notion of cause. This is denied by Davidson and also explicitly by others who have made normativity central such as John McDowell.

different terms—not as one that forces refutation, but one that forces adjustment in theory, the addition of auxiliary hypotheses, to 'save the phenomena', as Duhem put it. These theoretical adjustments were said by Kuhn and others (such as Feyerabend) to change the meanings of the terms that went into formulating the initial generalizations, so though nothing was refuted, what transpired was a change of subject. Theories about some phenomenon were not *improved* by a process of refutation of an earlier theory and the formulation of a new theory since talk of improvement requires comparability (commensurability) of the earlier theory with the later theory. But comparability, in turn, presupposes constancy of the meanings of terms in the passage from the earlier refuted theory to the later improved theory; rather, on the Kuhnian view, the later theory was no longer theorizing about the same phenomenon. Thus, 'mass' in Einstein's physics did not mean what it meant in Newtonian mechanics. This is a quite different conception of the crisis generated by failure and a quite different fate for the causal claims that fail than is suggested by the idealized model that Popper had celebrated.[9]

However, the crisis that is generated by failure in the causal claims that traffic in irreducibly normatively constituted notions of belief, desire, intention, etc., are entirely different from *either* of these conceptions of crisis because the phenomenon being studied is possessed of a first person point of view and because the normative states in question are defined in terms of second-order dispositions of the kind I mentioned earlier. Suppose, then, that we have come to see a subject as possessing a commitment—a desire, as it might be, or an intention that she help the poor—one that we expect will cause her to do certain sorts of things: give money to Oxfam, as it might be, or to panhandlers on the street, … We have this expectation because the idea of such a commitment brings with it a causal generalization, a statement of the causal power of the commitment: "commitment__causes actions such as__". Suppose, however, that (as with the failure of blue litmus to turn red when dipped in acid) she does none of those actions. What crisis does this generate? The point of these being normative states that the causal claims traffic in, is precisely to say that nothing is refuted about the correct attribution of the commitment to the agent in question. As I said, a commitment does not cease to be a commitment if it is not lived up to, that being the nature of normative states (i.e., since the *possibility* of failure is *defined* into the kinds of state they are, the fact of failure cannot cancel the idea that a commitment or normative state of mind exists). So failure refutes nothing. Does failure lead to some more Kuhnian adjustment being made by the theorist? No, that too is not what is demanded by failure. The difference between Popper and Kuhn on what transpires upon failure is a Trotskyite difference within a shared understanding that one is dealing with phenomena that possess *no* first person point of view. What does failure force, then, when phenomena with a first person point of view are the objects of study—that is, when the objects of study are *subjects* and in particular subjects with mental states not merely understood in dispositional terms but normative terms?

[9] The classic works in this familiar territory are, Popper 1959, Kuhn 1962, Quine 1953, and Duhem 1954. Feyearbend's *Against Method* (1975) is a spirited contribution to the debate, taking a sustained polemical stance against Popper's view.

The answer to this question lies in the way we have characterized the normative nature of beliefs, desires, intentions as commitments. If, as I have insisted, these states are defined in terms of certain second-order dispositions that I had elaborated above, the failure to act in accord with a commitment, disposes the subject who has the desire or commitment to be self-critical and to try and do better by way of living up to the commitment. So the point really is that in the explanation of individual human behaviour we record that commitments do *cause* actions that are in accord with the commitments—it is just that we get a sense of the distinctness of the notion of cause that is operative here, when we also record that in the cases where they *fail* to cause such behaviour, no question of the refutation of the existence of the commitment even so much as arises, since when there is failure *the phenomenon in question* from its first person point of view *itself* strives to improve the causal power of the commitment. There is simply no analogue to this in the phenomena that the natural sciences explain. Neither Popper's nor Kuhn's account of what follows upon the crisis brought about by failure are therefore relevant.

It is not as if natural scientific explanation is not norm or value laden. How many times have we heard that when we have two equally efficacious natural scientific explanations, we choose the simpler of two natural scientific theories, and simplicity is a value! And no doubt there are more interesting forms of value than simplicity that are deployed in the natural sciences. But none of this affects the nature of the causality that figures in the natural sciences. What is distinctive about the normativity of the explanation of individual human behaviour is that, for the reasons I have been elaborating, the states of mind that go into the explaining relate causally to the behaviour in a very distinctive way, and thus they explain it in a way that has no echo at all in the natural sciences.

Though my focus has been on individual human psychology, the consequences of this distinctive form of causality and explanation has and is bound to have wide consequences for how to understand social phenomena. Drawing those consequences in detail must remain a task for another occasion.

6. Norms of Mind vs Norms of Language

I have been arguing that failure illuminates the idea of a norm and I have been looking at the distinctive way it does so in the explanation of individual human behaviour, in particular. What has been key in the analysis I have offered is that the states of mind that go into the explanation of such behaviour are normative in a sense that was first illuminatingly suggested by Wittgenstein. Beliefs, desires, and intentions are themselves normative states or commitments and this makes a vital difference to the distinctiveness of how they cause behaviour and, therefore, how we understand and explain individual human behaviour.

Wittgenstein's own example of this form of normativity was given in his account of intention which has it that an intention generates a norm, dividing actions into those are in accord with the intention and those which fail to be in accord with it (accord and failure being normative notions). Failure, then, is essential to understanding norms and vice versa. The way I have put this point is to say that *there could be no norm if there was no possibility of failing to live up to it*. I want to conclude this paper with what I think is a startling and highly revealing *exception* to this otherwise impeccable claim. I want to argue very briefly that though Wittgenstein is certainly right that intentions are normative states in the

way he presents, the intentions with which individuals *mean* things with their words (in one perfectly good sense of the word 'mean' or 'meaning') *cannot fail* to be fulfilled. In short, though there can and must be failures to live up to one's intentions *in general*, there can be no failures to live up to one class of intentions, the intentions to mean something with what we say. These are a degenerate form of intention. Or if that sounds pejorative, they are a 'limiting case' of intention.

It is this last point that allows for Davidson's conclusion that *meaning* is not normative. But Davidson, who did not give any explicit argument for this quite correct conclusion, cannot really help himself to this last point since it is formulated in terms of a notion of normativity of *mind* that owes not to him, but to Wittgenstein's idea that intentions are themselves normative states, something that Davidson denies. The point, as I am making it here, is that this is correct in general of intentions, but not in particular of meaning intentions, which are a degenerate case of intentions.

Let us explore this by asking, what might amount to a failure of a meaning intention?

Perhaps something like this. I am walking down a path with a friend. I point to something and say "That's a snake" with the intention of getting him to believe that there is a snake in our path. But what is in the path is really a rope, not a snake. So I have made a mistake. A failure. But is it a failure of meaning? That cannot be right. It is a failure of other sorts. A failure of perception. A failure to utter a true statement. A failure, therefore, to communicate the facts. But did I fail to communicate the meaning I intended? No, because my meaning intention was not the intention to communicate the facts, it was not even the intention to get my friend to believe that there was a snake in our path (though I did intend that, that was not my *meaning* intention); rather my meaning intention was to use the words 'That's a snake' to *mean something in particular by my words*, viz., that there is a snake in our path, and *that* intention did not fail to get fulfilled. The fact that there was a rope in our path, not a snake, does nothing to spoil the meaning intention from getting fulfilled. The intention was impeccably fulfilled, and in fact my utterance would not have amounted to the falsehood and miscommunication of facts that it was, if my meaning intention had not been fulfilled.

And so the question arises. Can one *ever* fail to fulfill a meaning intention? What could possibly count as a failure to fulfill a meaning intention? I think, for one perfectly natural understanding of what 'meaning' is, these questions must be answered by saying 'no' and 'nothing', respectively.

Here again, as throughout the paper, it should be obvious that by 'meaning', I am interested in the actions (in this case linguistic actions) of individuals and not interested in the meanings of words as they occur in sociolects, i.e., meanings that dictionaries attempt to specify in each of their entries. If dictionaries captured all that there was to meaning, then one might contrive to say that we sometimes fail to live up to our meaning intentions. Thus, it is certainly true that when I speak a word of English, and I do not fully have a grip on what the word's meaning, *as it is given in the dictionary*, is, and my intention when I use the word is described as the intention to use that English word (*as it is elaborated in that dictionary entry*) then it can happen that I fail to fulfill that intention, so described. I do not have a full grip on what I intend. In other words I do not know what I am talking about. But even as that happens, I do have *something* in mind

when I speak those words, even if it does not coincide with what the dictionary defines them to be. And if my intention were described as meaning *that*, rather than what I do not have a grip on (the dictionary meaning of the word), then that intention is fulfilled. It is *this* notion of meaning, I am saying, that cannot fail.

The fact is that I and everybody else frequently uses words in this way, in a way that departs from dictionary definitions; and moreover frequently I (and those others) are perfectly well understood as meaning what I (they) mean by them rather than what the dictionary says about what they mean. That is to say, I mean something and I am understood to mean that, even if it does not square with what the dictionary gives as the meaning. So the notion of meaning is not exhausted by the sociolectical understanding of language. A great deal of meaning is idiolectical and is understood as such in conversation as well as in writing. It is this notion of meaning that I am concerned with since it is what is most closely tied to the intentions with which individuals speak words.

Just so as to get away from the tyranny of the dictionary, let us take an example of a slip of the tongue. Suppose I utter "I am going towndown". There is no such term as 'towndown' in the dictionary. So we have here a case of meaning that is not, not even on the surface, possessed of a dictionary meaning. Still, I mean something by that utterance because I intend to mean something by it. I intend to mean by those uttered words that I am going downtown. And suppose (plausibly, as so often happens in cases of slips of the tongue) that that is exactly what I am understood to mean by my hearers. So I have both meant something in a non-sociolectical (taking dictionaries to elaborate sociolectical meanings) sense of meaning and been understood quite correctly to mean that. And moreover my utterance is not a metaphor or a figure of speech. I *literally meant* by "I am going towndown" that I am going downtown because that is what I *intended* to literally mean by my utterance. What is true is that I did not intend to utter the *sounds* I did. I intended to utter different sounds. So it is misspeaking, but *not* a mismeaning. Such failure as there was, was a failure of vocalization not a failure of a meaning intention being fulfilled.

It would be foolish to deny that there is this notion of meaning (and literal meaning at that), tied to a speaker's intentions, which is independent of what words mean in dictionaries. Even if our words, meaning what they are intended by us to mean, coincide to a large extent (as they surely will) with what the dictionaries say they mean—this is a contingent fact. It is not conceptually required if we are to mean things with our words that we chime perfectly with the entries of dictionaries.

We have then a notion of meaning around which there cannot be any failure. That is to say when I intend to mean something by the sounds I utter—for example mean that I am going downtown, by the sounds "I am going towndown"—there cannot be any failure to fulfill the intention. I succeed in meaning just what I intend and often will be understood as meaning exactly that, even if—as in this case (which is why I picked it, to show the irrelevance of dictionaries to this notion of meaning)—the word does not even exist in the dictionary.[10]

[10] Davidson 1986 discusses examples that are in dictionaries when he discusses malapropisms. And because Davidson, unlike Wittgenstein, does not think that intentions are

This kind of intention, however, is unique among intentions. All other intentions are normative, as I have argued over the preceding sections of this paper, because there is always the possibility of our failing to fulfill them. If the possibility of failure did not exist, it would be wrong to think that we are in the region of norms. But intentions to mean something by our words cannot fail to be fulfilled. Does this reveal that we have a counter example to the thesis that intentions—being norms—presuppose the possibility of failure? No, what it reveals rather is that these intentions to mean something are a degenerate case of intentions. The Wittgensteinian claim to normativity of intentions is impeccable. It just runs up against a limiting case when it comes to meaning. Failures of meaning, in one perfectly good sense of 'meaning', are impossible, even as failures of mind are ubiquitous.

How do we diagnose this unique and peculiar property of intentions when they attach to meaning? I will not be able to answer this question at any great length as I close a paper that is already too long. But I will say just this by way of a cryptic hint of diagnosis. In the ordinary case when we intend something (when we make a commitment) and we fulfill the intention or commitment, there are two acts. The intention (the commitment) and the fulfillment of it. But in the degenerate case of the intention to mean something with what we say, the intention (commitment) and its fulfillment are not two acts, but one. These are deep waters and I regret that I cannot elaborate this diagnosis more fully in the space I have here.

References

Bilgrami, A. 2006, *Self-knowledge and Resentment*, Cambridge, MA: Harvard University Press.

Brandom, R.B. 1994, *Making It Explicit: Reasoning, Representing, and Discursive Commitment*, Cambridge, MA: Harvard University Press.

Davidson, D. 1970, "Mental Events", reprinted in his *Essays on Actions and Events*, Oxford: Clarendon Press, 1980, 207-27.

Davidson, D. 1986, "A Nice Derangement of Epitaphs", reprinted in his *Truth, Language, and History*, Oxford: Oxford University Press, 2005, 89-108.

Duhem, P. 1954, *The Aim and Structure of Physical Theory*, Princeton: Princeton University Press (first published 1914).

Hacking, I. 1990, *The Taming of Chance*, Cambridge: Cambridge University Press.

Feyerabend, P. 1975, *Against Method: Outline of an Anarchistic Theory of Knowledge*, London: New Left Books.

Foucault, M. 1977, *Discipline and Punish: The Birth of the Prison*, New York: Pantheon Books.

Foucault, M. 1978, *The History of Sexuality*, New York: Pantheon Books.

Foucault, M. 2003a, *Abnormal: Lectures at the Collège de France*, 1974-1975, New York: Picador.

themselves normative states, he does not raise the question about intentions and meaning that I am raising here.

Foucault, M. 2003b, *Society Must be Defended: Lectures at the Collège de France*, 1975-1976, New York: Picador.

Foucault, M. 2006*, Psychiatric Power: Lectures at the Collège de France*, 1973-74, New York: Palgrave Macmillan.

Kuhn, T.S. 1962, *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Levi, I. 1983, *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Chance*, Cambridge, MA: MIT Press.

Popper, K. 1959, *The Logic of Scientific Discovery*, London: Routledge (first published 1934).

Quine, W.V.O. 1953, "Two Dogmas of Empiricism", in his *From a Logical Point of View*, Cambridge, MA: Harvard University Press.

Wittgenstein, L. 1953, *Philosophical Investigations*, Oxford: Blackwell.

Radical Interpretation and Pragmatic Enrichment

Peter Pagin

University of Stockholm

Abstract

I consider a problem from pragmatics for the radical interpretation project, relying on the principle of charity. If a speaker X in a context c manifests the attitude of holding a sentence *s* true, this might be because of believing, not the content of *s* in c, but what results from a pragmatic enrichment of that content. In this case, the connection between the holding-true attitude and the meaning of *s* might be too loose for charity to confirm the correct interpretation hypothesis. To solve this problem, I apply the coherence raising account of pragmatic enrichment developed in Pagin 2014. The result is that in upward entailing linguistic contexts, the enriched content entails the prior content, and so charity prevails: the speaker also believes the prior content. In downward entailing contexts this would not hold, but I argue that enrichments tend not to occur in downward entailing contexts.

Keywords: Radical interpretation, Charity, Pragmatic enrichment, Coherence raising.

1. Two Projects or One?

H. Paul Grice and Donald Davidson shared the view that we should separate semantics from pragmatics. To this end, Grice (1975) developed the theory of *implicatures*. The main tenet was that we can separate *what is said* from *what is implicated* (Grice 1975: 25), which together make up of what is communicated. Semantics deals with the relation between a sentence and what is said *by means of* uttering that sentence, and the theory of implicatures, a part of pragmatics, deals with the relation between what is said and what is implicated *by means of* saying what is said. This relieves semantics from dealing directly with what is communicated in all cases, which avoids many complications.

Grice early on exemplified this strategy by indefinite descriptions. He notes that, for instance, "Anyone who uses a sentence of the form *X is meeting a woman this evening* would normally implicate that the person to be met was someone other than X's wife, mother, sister, or perhaps even close platonic friend" (Grice 1975: 37). He goes on to remark:

Argumenta 3, 1 (2017): 87-107
ISSN 2465-2334

DOI 10.14275/2465-2334/20175.PAG

> I am inclined to think that one would not lend a sympathetic ear to a philosopher who suggested that there are three senses of the form of expression *an X*: one in which it means roughly "something that satisfies the conditions defining the word X," another in which it means approximately "an X (in the first sense) that is only remotely related in a certain way to some person indicated by the context," and yet another in which it means "an X (in the first sense) that is closely related in a certain way to some person indicated by the context." Would we not much prefer an account on the following lines (which, of course, may be incorrect in detail): When someone, by using the form of expression *an X*, implicates that the X does not belong to or is not otherwise closely connected with some identifiable person, the implicature is present because the speaker has failed to be specific in a way in which he might have been expected to be specific, with the consequence that it is likely to be assumed that he is not in a position to be specific (Grice 1975: 38).

What is exemplified in the passage, Grice later proposed as a principle, which he called a modified Occam's Razor: *Senses are not to be multiplied beyond necessity* (Grice 1989a: 47). The strategy is to keep semantics and pragmatics separate, to keep semantics simple and to this end move tasks that can be moved from semantics to pragmatics. The two projects have separate and complementary roles.

Davidson had a similar attitude. He emphasized that a speaker always has an ulterior purpose with speaking:

> But where meaning is relevant, there is always an ulterior purpose. When one speaks, one aims to instruct, impress, amuse, insult, persuade, warn, remind, or aid a calculation. One may even speak with the intention of boring an audience; but not by hoping no one will attend to the meaning (Davidson 1984a: 9).

Davidson also insisted that linguistic meaning is *independent* of ulterior purposes:

> But the criteria for deciding what an utterance literally means—the theory of truth or meaning for the speaker—do not decide whether he has accomplished his ulterior purpose, nor is there any general rule that speakers represent themselves as having any further end than that of using words with a certain meaning and force. The ulterior purpose may or may not be evident, and it may or may not help an interpreter determine the literal meaning. I conclude that it is not an accidental feature of language that the ulterior purpose of an utterance and its literal meaning are independent, in the sense that the latter cannot be derived from the former: it is of the essence of language. I call this feature of language the principle of *the autonomy of meaning*. We came across an application when discussing illocutionary force, where it took the form of the discovery that what is put into the literal meaning then becomes available for any ulterior (nonlinguistic) purpose—and even any illocutionary performance (Davidson 1984a: 11-12).

In this passage, Davidson is speaking of the independence of the literal meaning of a sentence uttered on a particular occasion from the ulterior purpose of the speaker in uttering it. There is a further question whether the general *determination* of linguistic meaning is independent as well from the ulterior purposes that speakers have. That you cannot derive the literal meaning of a sentence from the ulterior purpose of a single utterance is compatible with the possibility that ulterior purposes have a role to play among the factors that determine meaning. In other passages, however, Davidson rejects this possibility as well:

> I agree with the hypothetical objector that autonomy of meaning is essential to language; indeed it is largely this that explains why linguistic meaning cannot be defined or analysed on the basis of extra-linguistic intentions and beliefs (Davidson 1975: 164-5).

Grice's separation of semantics and pragmatics later came to be challenged by philosophers and linguists who claimed that pragmatic processes not only generate what is indirectly communicated, but also intrude into what is directly *said*. These are known as *primary pragmatic processes*.

There has been much controversy over the alleged existence of such processes. I shall here accept, without argument, some of these claims. The main purpose of this paper is to argue that some primary pragmatic processes, known as *enrichments*, pose a prima facie problem for Davidson's autonomy claim. They in fact pose a threat as well to the adequacy of the method of *radical interpretation*.

The rest of the paper is structured as follows. In section 2, I shall briefly present the phenomenon of pragmatic enrichment. In section 3, I give a brief outline of the most relevant features of the method of radical interpretation. In section 4, I show how enrichment is a prima facie problem for the method. In section 5, I present Enrichment Theory, an account of pragmatic enrichment proposed in Pagin 2014. In section 6, I argue that if Enrichment Theory is true, then the phenomenon of pragmatic enrichment is so restricted that it is not, after all, a real problem for radical interpretation, as long as we keep to a certain type of contexts, the *upward entailing contexts*. In section 7, finally, I consider the problem induced by *downward entailing contexts*, where the situation seems to be reversed. I argue that the situation is nonetheless not reversed, because of an asymmetry in the distribution of enrichments. Enrichments do not arise, or do not have the corresponding effects, in downward entailing contexts, and hence the radical interpretation project survives the threat.

2. Free Pragmatic Enrichment

In a very wide sense of 'pragmatics', the term denotes everything that a speaker intends and a hearer interprets a linguistic utterance to convey that is not fully determined by standard (morphology and) syntax and semantics. Pragmatics, in this very general sense, covers syntactic and lexical disambiguation, anaphora resolution, ellipsis recovery, presupposition detection and accommodation, saturation, modulation, the understanding of conversational implicature, metaphor, irony, any form of indirect speech act, and more. Some of these, such as disambiguation, consist in choosing between alternatives that are made available by the semantics. Others involve adding something new. Perhaps the most basic kind is saturation.

Saturation (the term is from Recanati 2004: 7) is the process of providing values to indexicals and other context dependent expressions, including individuals to 'I' and 'her' in (1a), a quantifier domain to the quantifier 'everyone' in (1b), and a standard of height to the adjective 'tall' in (1c):

(1) a. I like her.
    b. Everyone cheered.
    c. He is tall.

These sentences also have tensed verbs, and the hearer typically assigns a time as part of interpreting their contributions to the truth conditions.

Saturation is characterized as being "linguistically mandated", as Recanati puts it (2004: 7-10). That is, there is some expression that triggers the saturation step. It is triggered because the expression needs a value in order for the sentence to express a proposition, i.e. a content that has a truth value with respect to a possible world. In case of (1a), no proposition is expressed by an utterance of the sentence except if values have been given to 'I', and 'her', and likewise a time is assigned. So, saturation is both *needed* for a proposition to be expressed, and *triggered* by a context dependent expression that is to be assigned a value.

This paper focuses on a pragmatic layer of interpretation that is characterized as being located *between* saturation and implicature. This layer of pragmatics has been named "modulations" by François Recanati (2004: 74; 2010: 5-7).[1] It has been called "explicatures" in Relevance Theory, primarily by Dan Sperber and Deirdre Wilson (1995: 182), and Robyn Carston (2002: 116), and "implicitures" by Kent Bach (1994: 126).[2] Saturation and modulation together have been called "primary pragmatic processes" by Recanati (2004: 17), intending to highlight that they occur *before* implicature.

There are important differences between these authors. Relevance theorists are contextualists; they think that the meaning of a sentence in principle underdetermines the content of an utterance made by means of it. In order to express a proposition, a contextually determined inferential process is always needed. Bach (2010: 128-29) rejects contextualism. Recanati (2004: 81-82) takes a kind of middle position. What he calls *the minimal proposition*, the result of saturation alone (after disambiguation etc.), is never, or hardly ever computed and never, or hardly ever, plays any role in utterance interpretation. On his view (Recanati 2010: 39-47), pragmatic modulations are *intertwined* with semantic interpretation: for instance, the semantic interpretation of a predicate may take as argument the *modulated* content of the interpretation of a singular term.

Recanati distinguishes between three kinds of modulation: *loosening*, *semantic shift* (or *semantic transfer*) and *free enrichment*. An example of loosening is

(2) The ATM swallowed my credit card

(Recanati 2004: 24) where the verb 'to swallow' has its application conditions *extended* to include the cash machine process referred to.

Classic examples of semantic shift come from Geoffrey Nunberg, including

(3) The ham sandwich is sitting at table 20

[1] The idea that modulation is strictly between saturation and implicature is complicated because of top-down effects: for instance, you can use the interpretation of an expected implicature in order to interpret what is said. An example from Carston 2002: 40 illustrates it:

(i) a. A: Do you want to go to the cinema?
B: I am tired.
b. B: I am tired [*enough for not wanting to go to the cinema*].

The extent of tiredness (made explicit in (b)) is inferred and entered into the interpretation of what is said because it would justify the implicature.

The mutual influence of what is said and what is implicated has been called "Grice's circle" by Stephen Levinson (2000: 186). I shall not here be concerned with this phenomenon. For further discussion of this issue, see Pagin 2014, section 2.

[2] Terminological differences are a little more complex. Officially, *explicatures* are the *results* of pragmatic operations, whereas *saturations* and *modulations* are the operations *themselves*, and *implicitures* are either the operations or the contents *added* by means of them.

as said by one waiter to another in a restaurant (Nunberg 1979; Nunberg 1995). The content of the utterance is that

(3') The ham sandwich *orderer* is sitting at table 20.

What Recanati calls *free enrichment* is any *addition* of content material that is not needed for the content to be a proposition, i.e. to have truth conditions; it is freely added in this sense. The term 'enrichment' has come to be well established in the pragmatics literature, and largely accepted across different theoretical standpoints, even though it is often employed as an alternative to more technical vocabulary that *does* differ between the positions. The phenomenon the term is used for describing is pretty much the same. A typical example is the following (from Carston 2002: 71):

(4) a. He handed her the key and she opened the door.
b. He handed her the key and she opened the door [*with the key that he had handed her*].

A normal and typical interpretation of (4a) associates with it a content that is more completely articulated in (4b), which includes additional linguistic material in brackets. The semantic contribution of this material to (4b) is the *pragmatic* enrichment of (4a). That is, when processing (4a) the hearer/reader tends to interpret it as representing a type of situation which is more completely represented by (4b). That the referent of 'she' used the key handed to her by the referent of 'he' is not semantically represented in (4a), but is "read into" it, i.e. pragmatically added during interpretation. It is semantically represented in (4b).

The idea is *not* that the hearer tacitly adds the bracketed *expression* during interpretation. It is also not the case, as in normal examples of ellipsis recovery, that there is a particular expression that *would be* recovered in any effort of making the enrichment explicit. Rather, it is the semantic content (in context) of the added phrase that matters, and often there are alternative possible linguistic additions that are semantically equivalent as far as the linguistic and extra-linguistic context goes.

One of Recanati's own official examples (2004: 8) is given in

(5) A: Do you want something to eat?
B: I have had breakfast.

According to Recanati, it is exemplified in B's utterance, where *today* is freely added to the content *I have had breakfast*. The idea is that without the enrichment, there is still a proposition, true if B has had breakfast at some time or other during his life.

Similarly, concerning enrichment, Robyn Carston says:

> It is the enriched propositions that are communicated as explicatures and which function as premises in the derivation of implicatures; the uninformative, irrelevant, and sometimes truistic or patently false minimal propositions appear to play no role in the process of utterance understanding, which is geared to the recovery of just those propositional forms which the speaker intends to communicate. The pragmatic process at work here is known as free enrichment; it is "free" in that it is not under linguistic control. So, unlike saturation, it is an optional process, in the sense that there can be contexts in which it does not take place, though these tend to be somewhat unusual (Carston 2004: 639).

I take free enrichment to be a linguistic phenomenon. Speakers *assert* propositions that are the result of free enrichment in relation to what is literally ex-

pressed (after saturation, disambiguation etc.). These are not indirect assertions.[3] I shall argue that the phenomenon of enrichment generally, and free enrichment in particular, is a problem for Davidson's method of radical interpretation. In the next section, I first present the basics of that method.

3. Radical Interpretation

Donald Davidson proposed that the proper method for testing a meaning theory for a particular language is to apply *radical interpretation* to the speaker(s) of the language. The term 'radical interpretation' is coined in analogy to Quine's 'radical translation' in chapter 2 of *Word and Object*. The intuitive idea, in both cases, is that of translation/interpretation that "starts from scratch", without any prior knowledge of the language one is interpreting, or a detailed knowledge of the attitudes of its speakers. Although radical translation/interpretation in this sense has taken place in history, for both Quine and Davidson, describing it is rather a thought experiment. The point is to identify the kind of *evidence* that is ultimately available to an interpreter and how that evidence *supports* a meaning theory, i.e. the evidential relation (Davidson 1973: 128).

> The problem of interpretation is domestic as well as foreign: it surfaces for speakers of the same language in the form of the question, how can it be determined that the language is the same? Speakers of the same language can go on the assumption that for them the same expressions are to be interpreted in the same way, but this does not indicate what justifies the assumption. All understanding of the speech of another involves radical interpretation (Davidson 1973: 125).

That my fellow speakers mean the same as I with words we both use, is not, on this view, anything we have the right to take for granted, but something that requires justification, and ultimately from evidence of the very same kind as is available to the interpreter that does start from scratch.

When interpreting a speaker who makes an assertion in a familiar language we typically infer what she *believes* on the basis of what we take the sentence to mean. If she speaks an unfamiliar language but we happen to have independent information about what she believes or wants, we can move on to a plausible guess about the meaning of the sentence she used. Ultimately, however, the radical interpreter does not have access to sentence meaning as basic evidence, and neither does he have access to information about particular beliefs, desires or intentions as data for interpretation. The meaning of sentences and the contents of attitudes will be what his theory attributes to the speaker. The evidence must be something else (Davidson 1973: 134).

Without knowing what the speaker believes or expresses the interpreter can, however, observe the speaker's linguistic utterances and reactions to utterances by others, including the interpreter himself. The interpreter can form a hypothesis about the *attitude* to a sentence that the speaker manifests. In particular, Davidson focused on the attitude of *holding-true*, or more precisely, holding true relative to a time, in order to take account of indexical sentences, like 'it is

[3] That pragmatics does "intrude" into the truth-conditional content of what is said is a controversial claim and was the subject of intense debate about ten years ago. See, for instance, Stanley 2000, Cappelen and Lepore 2005, Borg 2006. I find the claim completely convincing and the alternatives implausible.

raining' (Davidson 1973: 135; Davidson 1974: 144). Holding a sentence true is an attitude to a sentence that corresponds to *believing* the proposition that is the meaning of the sentence. Holding-true is indeed a kind of belief, but it is a belief with a very coarse-grained, impoverished content; identifying the *content* of such a belief requires no more than identifying the sentence the belief is about, for there is no need to identify the meaning of the sentence. Provided the radical interpreter can identify manifestations of the holding-true attitude on the part of the speaker, he has access to data that are independent of knowledge of sentence meaning or fine-grained individual beliefs.

At the next step, the interpreter faces a serious problem of *underdetermination*: just as the truth of a sentence depends in part on what the sentence means and in part of what the *facts* are, so the speaker's holding true a sentence depends in part on what the sentence means in the speaker's language and in part on what the speaker *believes*. So, if the interpreter knows what the speaker believes and what a sentence in the speaker's language means, he can infer what the speaker will hold true:

(INF1) 1) X believes that *p*
2) *s* means that *p*

3) Hence, X holds *s* true

Similarly, if the interpreter knows what a sentence means, and knows that a speaker holds it true, he can infer what the speaker believes:

(INF2) 1) X holds *s* true
2) *s* means that *p*

3) Hence, X believes that *p*

There is no analogous simple inference from belief to meaning:

(INF3) 1) X holds *s* true
2) X believes that *p*

3) Hence, *s* means that *p*

The simple reason why (INF2) goes through but (INF3) does not is that a (disambiguated) sentence only has one meaning, while a speaker has many beliefs, and the second premise of (INF3) does not provide the information that *this particular belief is* responsible for holding *s* true. Still, the interpreter *can* infer that among the sentences the speaker holds true, at least one means that *p* (since we are dealing with beliefs expressible in the speaker's language). Knowledge of all the beliefs of the speaker would then allow the interpreter to infer what meanings many sentences of the speaker's language must have, but not directly how these meanings are distributed over the sentences.

Initially, the interpreter knows only what sentences the speaker holds true. Different hypotheses about what the sentences mean lead to different hypotheses about what the speaker believes, and, indirectly, vice versa. Davidson refers to this as the *interdependence of belief and meaning* (Davidson 1973: 134).

How can the radical interpreter break into this interdependence? Davidson's proposal is a cornerstone in his philosophy of language. Basically, the proposal is that although the interpreter does not at the outset have any *particular* knowledge of the speaker's fine-grained beliefs or other attitudes, he does

have *general* knowledge, which he can put to use. He can know that if someone has beliefs at all, most of these beliefs are *true* (by the interpreter's lights).

> The method is intended to solve the problem of the interdependence of belief and meaning by holding belief constant as far as possible while solving for meaning. This is accomplished by assigning truth conditions to alien sentences that make native speakers right when plausibly possible, according, of course, to our own view of what is right (Davidson 1973: 137).

This is a statement of what became known as *The Principle of Charity*. In its simplest version, the idea of using Charity in interpretation is the idea that an interpretation is *better* if it leads to attributing *more* true beliefs. This works precisely because of the interdependence of belief and meaning, in particular because of the validity of schema (INF2):

(INF2) 1) X holds *s* true
2) *s* means that *p*

---

3) Hence, X believes that *p*

Premise 1) here records the evidence for the interpreter. Premise 2) states his interpretation *hypothesis*. The conclusion is used for *evaluating* the hypothesis.

Consider two global meaning theoretical hypotheses, theories T_1 and T_2. For a great number of sentences held true by the speaker, T_1 and T_2 contain substantially different hypotheses about their meaning. According to T_1, say, *s* means that *there is a hippopotamus in the refrigerator*, and according to T_2, *s* means that *there is an orange in the refrigerator* (cf. Davidson 1969: 100-101). As a consequence, the interpreter has a choice between two belief attributions to the speaker: that there is a hippopotamus in the refrigerator and that there is an orange in the refrigerator. The first belief is pretty absurd while the second may well be true. The second is preferable, in particular if it *is* true. This then speaks in favor of T_2 over T_1. This exemplifies the basic mechanism of how belief attribution influences semantics, via Charity.

A few remarks about this mechanism are in order.

1. The role of Charity is that of comparison and evaluation. In terms of the philosophy of science: Charity belongs to the *context of justification*, not the *context of discovery*. Charity is a tool for testing whether a meaning theory is acceptable, not primarily a method of selecting hypotheses to test. Other factors, such as relevance and general psychological plausibility will be important for hypothesis formation. As Davidson emphasizes as regards actual interpretation, a theory is "derived by wit, luck, and wisdom" (Davidson 1986b: 107). Neither is required for testing.

2. Charity is primarily applicable to a theory as a whole, not to individual theorems. A speaker may well have a number of false beliefs (we probably all do), and may even have some absurd beliefs as well. What is compared is the totality of belief attributions according to one theory with the totality of belief attributions according to another. That a theory generates the attribution of a belief that is *true*, gives a small positive contribution to the overall evaluation of

the theory. Conversely, that it generates the attribution of a belief that is *false*, gives a small negative contribution.[4]

3. One reason the interpreter cannot (with ordinary concepts) simply devise a meaning theory that makes the speaker hold *only* true beliefs is that the theory must be compositional; the theorems must be connected by being derived from a shared basis. If that were not the case, the interpreter could simply pick a true interpretation for any sentence held true by the speaker, because then the interpretation of one sentence would not impose any restriction on the interpretation of any other sentence. Since any belief could then come out true, the interpreter could score high on Charity. This observation shows that compositionality must be an independent requirement, because it cannot be justified from Charity (cf. Pagin 1999).

Let's go back to the content of the charity principle. A typical early formulation is

> The general policy [...] is to choose truth conditions that do as well as possible in making speakers hold sentences true when (according to the theory and the theory builder's view of the facts) those sentences are true (Davidson 1974: 152).

So, in the most basic version, interpretations should be chosen that *maximize* the rate of true beliefs among the speaker's beliefs, judged according to the standards of the interpreter (cf. Davidson 1975: 169). Theories that do *not* reach the maximum rate of truths, i.e. those that are not among the *best* theories, must be rejected.[5]

The interpreter is always to maximize the rate of true beliefs *by the interpreter's standards*, or *according to the interpreter's view of the facts*, and this comes to the same thing as maximizing agreement; it is compatible with the falsity of the beliefs of both speaker and interpreter. We get real maximizing of the rate of truth only by assuming that the interpreter's beliefs are all true (or the interpreter is dropped).[6]

In the most basic formulations, applying Charity is to make a *comparison* between theories. However, *acceptable* theories must not only be best, they must also be *good*. They must be such that speaker and hearer are rendered *largely* in agreement, i.e. such that *most* of the speaker's beliefs come out true, by the interpreter's standards:

> What justifies the procedure is the fact that disagreement and agreement alike are intelligible only against a background of massive agreement (Davidson 1973: 137).[7]

The requirement of massive agreement is typically presented in the context of an *argument* for Charity in the previous respect. However, it is clearly an aspect of Charity in its own right, since it provides an *absolute* requirement on the rate of

[4] Davidson notes that some false beliefs are more destructive than others, in case the speaker would be expected to know better. Cf. Davidson 1975: 161.

[5] Davidson accepts the consequence that two or more different theories can be equally good but better than all others. These top-ranking theories are then *all* true, despite being apparently incompatible. This is what Davidson calls "indeterminacy of interpretation", analogous to Quine's indeterminacy of translation. Cf. Davidson 1979.

[6] Davidson did assume the possibility of an omniscient interpreter in Davidson 1986a.

[7] See also Davidson 1975: 168-69.

truth (it must be high), which complements the *relative* requirement (it must be the highest).

We are now in the position to see why the enrichment phenomenon is a problem for the method of radical interpretation.

4. The Enrichment Problem

The core of the radical interpretation method is that, firstly, we can infer what a speaker X believes from data about what sentences X holds true and a hypothesis about what the sentences mean, and secondly, that we have, in the principle of charity, a filter on acceptable belief attributions. These two factors combined provides a filter on acceptable meaning hypotheses.

The problem that stems from pragmatics, and in particular free pragmatic enrichment, is that the first factor, the link between the meaning of the sentence and the content of the belief, is distorted by an additional factor: the enrichment. Schematically, instead of (INF2), we have

(INF2') 1) X holds *s* true
2) *s* means that *p*
3) *p* is enriched to *q*

---

4) Hence, X believes that *q*

Clearly, if we have no idea about the result of an assumed enrichment of *p*, then we have no idea of what *q* is, and then we have no input to our Charity test on belief attributions, simply because we have no belief attributions. If, in enrichment, the sky is the limit, then the method of radical interpretation simply delivers nothing.

On reflection, we can see that the situation is not *that* bad. Enrichment is not an operation from any content to any content, but provides an *addition* to content that is already in place. It is in principle possible to enrich (6a) to (6b) but not to (6c):

(6) a. John kissed Mary.
b. John kissed Mary at midnight.
c. Bill broke his leg.

Since enrichment does not delete conceptual components of the original content, there is a restriction on what the enriched content can be, given the hypothesis about the meaning of the sentence.

However, that restriction is not of tremendous help, since it still leaves an infinity of possibilities open, as illustrated in (7):

(7) a. I am sick.
b. [*The man*] I [*saw on the bus*] [*was doing what my aunt Augusta does when I*] am [*at her place and she believes that her neighbor is*] sick.

If (7a) can be enriched to (7b), then, although not *every* content can be an enrichment of (7a), there is no upper bound to contents that in principle can be.

That there is a potential infinity of possible enrichments of any particular content is not, however, in itself a decisive blow to the adequacy of radical interpretation. For, if all enrichments were *plausibility preserving*, it would not matter that they are infinitely many. Suppose that it is plausible that X believes that *p*. If enrichments are plausibility preserving, then for any possible enrichment *q* of *p*, it is also plausible that X believes that *q*. Then, the hypothesis that *p* is the

meaning of sentence *s* gets a positive degree of confirmation from the hypothesis, given that X holds *s* true, *whatever* enriched proposition *q* it is that X believes. Likewise, if it is not plausible that X believes that *p*, then for any possible plausibility preserving enrichment *q*, it would still be implausible that X believes that *q*, and hence the hypothesis would get a degree of disconfirmation. So, it is not the infinity of enrichments itself that is the main problem.

The problem for the method is rather that there is no a priori reason to believe that an enrichment cannot lead from a plausible belief content *p* to an implausible belief content *q*, and *vice versa*. This means the combination of a meaning hypothesis for a sentence *s*, jointly with the fact of X holding *s* true, gives neither confirmation nor disconfirmation of the meaning hypothesis, for after enrichment, the resulting belief attribution can be either plausible or implausible, depending on the enrichment assumption.

The upshot is that unless we can restrict the range of available enrichments very tightly, and ideally such that available enrichments are plausibility preserving, the method of radical interpretation does not deliver any results at all.[8]

Such a restriction on the range of available enrichments would require a systematic theory of enrichments. Is that possible? Davidson himself was clearly pessimistic about theories of pragmatics:

> I do not believe there are rules or conventions that govern this essential aspect of language. It is something language users can convey to hearers and hearers can, often enough, detect; but this does not show that these abilities can be regimented. I think there are sound reasons for thinking nothing like a serious theory is possible concerning this dimension of language (Davidson 1990: 313n).

As we shall see in the next section, however, there are reasons to be more optimistic.

5. Enrichment Theory

Much of the literature on modulations in general and enrichments in particular has been concerned with convincing readers that these phenomena exist, and with some sub-categorizing of different kinds of modulation. Little has been done in the way of explanation. Relevance Theory has indeed offered a number of principles that would help explaining why this or that enrichment occurs, but the principles offered are not sufficient for predicting any particular enrichment on their own.[9] Some predictive principles have been offered by Levinson (2000), especially with his theory of *I*-implicatures. The basic idea there is that the speaker says as little as possible, and the hearer infers as much as possible, and Levinson (2000: 117-18) offers some more concrete principles for achieving the-

[8] That pragmatic phenomena provide potential issues for the method radical interpretation, especially if contextualism is true, is noted by Kathrin Glüer (2011: 40n).

Furthermore, if we think of the conveying of an *enriched* belief as an ulterior purpose of the speaker, and the interpreter needs to arrive at a meaning theory by means of belief attributions that depend on enrichments, then Davidson's principle of the autonomy of meaning does not hold. It is doubtful, however, that conveying a belief should be counted as an ulterior purpose.

[9] This is not the place for criticism of Relevance Theory with respect to predictive capacity. For discussion, see Pagin 2014: 88-92.

se ends. These principles are a mixed bunch, however, and I think enrichments often are not motivated by his general idea.[10]

In Pagin 2014, I proposed a theory that, I claimed, does this, given only background beliefs about the world. The general idea was that free enrichments occur because they *raise the coherence*: the proposition arrived at *after* enrichment, the *enriched* proposition, has a higher degree of coherence than the prior proposition *before* the enrichment, the *original* proposition. As regards mandatory enrichments, the theory does not explain why an enrichment occurs, since some enrichment is needed irrespective or coherence, but rather (where there is enough background information) why some enrichment occurs rather than an alternative.

The theory was spelled out by means of an ordinal scale of *coherence strength*. In doing so I was building on the theory of coherence relations (rhetorical relations) of Andrew Kehler (2002), who in turn developed idea presented by Jerry Hobbs, for instance in Hobbs (1985). After a suggestion by Hobbs, Kehler used the categories of *connections between ideas* of David Hume (1748) as his basic categories of discourse relations: *Resemblance*, *Cause-Effect*, and *Contiguity*.

Kehler's cause-effect relations are *Result*, *Explanation*, *Violated Expectation*, and *Denial of Preventer*. An example of *Explanation* (Kehler 2002: 21) is

(8) George is dishonest. He is a politician.

This satisfies the *Explanation* relation insofar as it is *presupposed* that being a politician *implies* being dishonest.

Kehler's resemblance relations are *Parallel*, *Contrast*, *Exemplification*, *Generalization*, *Exception*, and *Elaboration*. An example of *Parallel* is

(9) Dick Gephardt organized rallies for Gore, and Tom Daschle distributed pamphlets for him (Kehler 2002: 16).

This exemplifies *Parallel* since organizing rallies for Gore and distributing pamphlets for Gore are subsumed under a common more general property/activity, such as *doing something in favor of Gore*. In addition, Dick Gephardt and Tom Daschle both had the property of being high-ranking Democratic politicians.

There is only one contiguity relation: *Occasion*. It is exemplified in

(10) George picked up the speech. He began to read.

This exemplifies *Occasion* since we read (10) is conveying that nothing happened *between* the picking up and the start of the reading.

Partly based on these coherence categories and coherence relations, I offered the following scale of coherence strength (Pagin 2014):

Scale of coherence strength

0) Vacuity
1) Contiguity type relations
2) Resemblance type relations
3) Possibility type relations
4) Necessity type relations

The scale runs from weakest (0) to strongest. Degree 0, Vacuity, is the measure of discourse without coherence, like

(11) My dad bought a car. Bananas are yellow.

[10] For discussion, see Pagin 2014: 92-95.

Degree 1, contiguity, could be exemplified by

(12) The table is covered with books. A cat is lying on the sofa.

On reading (12), one typically makes a so-called *bridging inference* to the conclusion that the sofa mentioned in the second sentence is in the same room as the table mentioned in the first, thus close to each other.[11]

Degree 2, Resemblance, belongs to discourses where there is a certain type of thematic unity, typically together with contiguity. The *Parallel* example above from Kehler is a good example.

Degrees 3 and 4 mark coherence between states of affairs in virtue of either, loosely speaking, making *possible* (degree 3), or making *necessary* (degree 4). Often, these relations are causal in nature. In that case, a degree 3 coherence pertains to a discourse where one fact mentioned is such as to *enable* another fact or event also mentioned. This is exemplified in (4): when we read into (4a) that the female subject opened the door with the keys that had been handed to her, we take the fact stated in the first conjunct to enable the action reported in the second, i.e. the opening of the door.

Degree 4, Necessity, concerns states of affairs related by *consequence* either causal/evidential or logical/conceptual. For instance, teleological explanations belong to this type:

(13) a. The man took out a knife. He was going to cut the rope.
b. The man took out a knife. [*He did this because*] he was going to cut the rope.

Taking out the knife is a causal consequence of intending to cut the rope, presented in a teleological manner as a consequence of the purpose itself.

The general idea of the theory, called *Enrichment Theory* (ET) (Pagin 2014), is that a free enrichment takes place if it maximally raises the coherence compared with the unenriched content, given constraints regarding the *plausibility* and *accessibility* of the enrichment. The plausibility constraint depends on general background beliefs, often called "world knowledge". For instance, the enrichment in (13b) depends on the background belief that rope-cutting is commonly done with a knife (it is commonly known to be feasible, hence commonly intended). The accessibility constraint concerns the complexity of the added content; it should be quick and easy to think.

The theory can in fact explain many of the enrichment examples in the literature, and many more examples as well. For instance, the enrichment in (13) is explained by the fact that it raises the coherence of the discourse from 1 (assuming the taking-out occurs just after the onset of the intending) to 4, Necessity, in that the taking-out is represented as a consequence of the intending.[12]

The enrichment in (4) is explained by its raising the coherence from degree 1 to degree 3, as getting the key is thought to enable the opening of the door. Degree 4 is not reachable here, unless we think that handing her the key somehow *causes* her to open the door. This is indeed possible, but the background plausibility of this assumption is not very high. The theory predicts, however, that those who *do* find it high, would also read this stronger relation into the sentence.

[11] Cf. Clark 1975, Levinson 2000: 37-38, Wilson and Matsui 1998. According to Levinson, bridging inferences are examples of I-implicatures.

[12] Compare: (i) The man took out a handkerchief. He was going to cut the rope.

Enrichment can take place within a single predication, in that it can relate a property that is *predicated* of a subject to a property that is *attributed*:

(14) a. The temperature has risen to a dangerous level.
b. The temperature has risen [*from a non-dangerous*] to a dangerous[*ly high*] level.

(15) a. A tall man picked up the book.
b. A kind man picked up the book (Pagin 2014: 83).

In (14a), two properties are ascribed to the temperature (as a changing property of some entity); that it has risen and that it is (at the time or utterance) at a dangerous level. These two facts could be unrelated without affecting the truth of the sentence, since the temperature might have been at a dangerous level before the rising. It could also be the case that the level is dangerous because the temperature is *low*, or that it is within a particular interval. Intuitively, we read into the statement the temperature was at a non-dangerous level before the rising. This is explained by the theory, since coherence is thereby raised to degree 4: the rising causes the danger. In addition, we take it that the danger depends on the temperature being high. This is not necessity for the rising to cause the danger, but it is natural to take the rising first to cause the temperature's being high, and the latter again to cause the danger.

In (15b), the property of being kind is attributed to a man and the property of picking up the book is predicated of him. Intuitively, these are taken to be connected, in that it is seen as an *act of kindness* to pick up the book. It is not completely easy to situate the example on the scale, since it would involve difficult considerations about reasons and causes, but I find it natural to say that the act of picking up the book, insofar as it is done out of kindness, is *motivated* by kindness as a trait, i.e. a disposition to perform acts that are beneficial to others. By contrast, no corresponding relation can be seen in (15a), since there is no plausible connection between being tall and picking up a book. This again can be contrasted with

(16) A tall man took down the book.

where being tall is easily seen as *enabling* the man to take down the book (from a high shelf, for instance).

For a precise statement of ET, see Pagin (2014: 76). The paper also has many more examples. We can now apply ET to the enrichment problem for radical interpretation.

6. Enrichment Theory and Radical Interpretation: The Simple Connection

We can note a central feature of Enrichment Theory:

(EEO) The enriched proposition *entails* the original proposition.

The enriched proposition is always *more specific* than the original proposition.[13] In possible-worlds terms: the set of worlds where the enriched proposition is true is a *subset* of the set of worlds where the original proposition is true.

[13] This does not hold of bridging inferences, where, according to Enrichment Theory, enrichment does not take place. Rather, the raising of coherence to level 1 occurs in *saturation*, with the assignment of time and location parameters.

Standard examples in the literature exemplify this. For instance, in

(4a) He handed her the key and she opened the door

(4b) He handed her the key and she opened the door [*with the key that he had handed her*]

the worlds where she opens the door with the key he gave her are all worlds where she opens the door (someway or other). Similarly, in

(13a) The man took out a knife. He was going to cut the rope.

(13b) The man took out a knife. [*He did this because*] he was going to cut the rope.

And similarly, again in the breakfast example of Recanati (5a), and similarly, in turn, in just about every example in the literature. I argued (Pagin 2014) that this shows that enrichment does *not* take place in order to satisfy *Charity*: if the original proposition is false, then the enriched proposition is false as well, since it entails the original one. Rather, something else is going on, and I proposed exactly coherence raising.

If Enrichment Theory is correct, and the entailment from the enriched to the original propositions projects from the examples in the literature to all (or perhaps virtually all) cases of pragmatic enrichment, we also have a result that is relevant for the prospects of radical interpretation. For then, *if* Enrichment Theory is correct, then enrichments *are* plausibility preserving, in the sense of section 4. For if it is plausible that the speaker X believes the enriched proposition, then it is (at least standardly) plausible as well that X believes the original proposition, since it is entailed by the enriched proposition. And in the case of the Charity inference, (INF2'), the original proposition is exactly the meaning of the sentence *s* (in context, if necessary).

We get the same result by taking Charity itself into account. For, in the most basic case, it is plausible that a speaker believes a proposition provided that proposition is *true*. If we have a prima facie reason to believe that X believes the enrichment proposition *q*, since *q* is *true*, then we also have a reason to believe that X believes the original proposition *p*, since *p* (entailed by *q*), is true as well.

The upshot of the investigation so far is that, if Enrichment Theory is true, then the phenomenon of pragmatic enrichment is so restricted that it in fact does *not* have any negative consequence for the adequacy of the method of Radical Interpretation. The complete picture is more complicated, however, because of downward entailing contexts.

7. The Problem of Downward Entailing Contexts

Consider the following classical example from Bach (1994: 278). A mother is talking to her child, who has had a cut in a finger:

(17) a. You are not going to die.
b. You are not going to die [*from that cut*].

In this case, on the standard analysis of the example, the salient background content is the content of the mutually shared knowledge that the child has a cut in the finger. The enrichment operates on a *part* of that content: it enriches *die* to *die from that cut* or *die because of that cut*.

On the standard analysis, the unenriched content of the mother's utterance is blatantly false: she says that the child is not (= never) going to die. The en-

richment, into saying that the cut will not cause the child to die, then turns the utterance from having a false content into one having a true content. Hence, charity may be a motivating factor of the enrichment. But for the current account of the relation between enrichment and radical interpretation, the standard analysis of the example provides an apparent counterexample. For the mother to hold (17a) true looks like indicating a false belief, which prima facie speaks against the interpretation hypothesis. This is of course a bad result.

From the coherence raising point of view, the problem seems to be that the enrichment occurs in a *downward entailing context*. A downward entailing (DE) context Φ... is such that if $p \models q$, then $\Phi(q) \models \Phi(p)$. Hence, the context in a sense reverses the entailment relation. Negation is of course the basic downward entailing operator, i.e. an operator that induces DE contexts. If p entails q, then $\neg q$ entails $\neg p$. Other DE contexts are 'Nobody...' and 'if…, p', i.e. antecedents of conditionals. Upward entailing (UE) contexts are opposite: Φ is upwards entailing just in case if $p \models q$, then $\Phi(p) \models \Phi(q)$.

The effect on the present account is obvious: if every enriched proposition within the local context entails the prior proposition, then when the enrichment occurs in a DE context, the proposition P expressed by the containing (unembedded) sentence will entail the resulting proposition P' expressed after the enrichment. In the present example, *that you will not die* entails *that you will not die from this cut*. Hence, in general, if enrichment supports charity in UE contexts, it undermines charity in DE contexts.

The result is potentially very bad for the radical interpretation approach to linguistic meaning. The question is, however, what the pattern of enrichment actually is in DE contexts. Let's reconsider the Bach example.

Suppose we make the following morbid addition to the example: unbeknownst to the mother, the child has swallowed some poison and is in fact going to die within an hour, although not from the cut. Is then the utterance (17a) true? If the content of the utterance is that of (17b), then it is indeed still true. Intuitions may be divided here. Mine is that it is in fact false in this case.

Also, from the point of view of coherence raising, the original idea would be that enriching with *from that cut* would *raise* the content *you are going to die* to the necessity level of *you are going to die from that cut*. However, if the future tense involves unrestricted quantification over future times, then the prior proposition is already at necessity level, so there is in fact no raising: it is part of folk theory that by natural necessity, everyone dies sooner or later. There would have been a raising only if the cut had made the child mortal.

The natural and intuitive correction of the standard analysis is to introduce domain restriction for the temporal domain to the *near future*. That is, what the mother says, *before enrichment*, is that the child is not going to die *in the near future*, i.e. *any time soon*. This is itself not a case of enrichment, but a case of domain restriction.[14] That is, the contents before and after enrichment are

(18) a. Not: For some time t in the near future, you will die at t.
b. Not: For some time t in the near future, you will die at t from that cut.

If this is right, then in Bach's original example, the mother's prior content (18a) is true, not false. The embedded propositions will be

(19) a. For some time t in the near future, you will die at t.

[14] I argue for this claim in Pagin 2014, section 7.

b. For some time *t* in the near future, you will die at *t* from that cut.

This enriched embedded proposition (19b) of course *does* entail the prior embedded proposition (19a). Moreover, coherence then *is* raised: on the assumption that the cut causes the child to die, what it does cause is not that the child dies *some time or other*, since that it would do anyway, but that the child dies soon after the time of getting the cut, which it would not have done without getting the cut.

Assuming that the enrichment *from that cut* does take place, in accordance with the standard analysis, because it takes place in a DE context, the result is that the *prior* total proposition (18a) does entail the *enriched* total proposition (18b). Nevertheless, we do not get a counterexample to the application of charity, simply because, unlike in the standard analysis, the prior proposition (18a) is both believed by the speaker and true, not disbelieved and false, as in the standard analysis. Hence, the Bach example, on this account, does not provide a counterexample.

However, it may seem that this result gets us out of the frying pan and into the fire. For in the morbid alternative scenario, when the child will die from having swallowed poison, the original proposition (18a) is false, while the enriched proposition (18b) is true. The speaker holds the sentence (17a) true, and believes the enriched proposition (18b). It may therefore look as if we do get the counterexample in this variant of Bach's example.

We do not however, and the reason is that it is built into the example that the speaker does *not* know, and hence (because it is very unlikely) has good reasons not to *believe* that the child will die in the near future because of having swallowed poison. Therefore, the speaker *does* believe (18a), just as in the original scenario. In this alternative scenario, the belief is false. That could be a problem for charity itself, but again is not in this case. Rather the false belief in (18a) is an explicable error. Under the epistemic circumstances, belief in (18a) is precisely what should be predicted. Hence, the standard strategy of radical interpretation deals with this case:

> Some disagreements are more destructive of understanding than others, and a sophisticated theory must naturally take this into account. Disagreement about theoretical matters may (in some cases) be more tolerable than disagreement about what is more evident; disagreement about how things look or appear is less tolerable than disagreement about how they are; disagreement about the truth of attributions of certain attitudes to a speaker by that same speaker may not be tolerable at all, or barely. […] The methodology of interpretation is, in this respect, nothing but epistemology seen in the mirror of meaning (Davidson 1975: 169).

The hardest case remains, however. For what are the contents and truth values in the even more morbid alternative case where the mother *knows* that the child has swallowed poison and will soon die? In this case, if by uttering (17a) she actually asserts (18b), not (18a), then what she asserts is in fact *true*, despite the imminent death of the child, and despite her knowledge of this fact. The prior proposition (18a) is false, and not believed. Hence, we seem to have ended up in the worst scenario for the radical interpretation project.

This is the hardest case, and I cannot discuss it thoroughly in the present paper. I believe that the radical interpretation project survives this problem as well, however, because of an asymmetry between enrichments in UE and DE contexts. To bring this out, let's go back to Carston's case of (4), repeated here as (20):

(20) a. He handed her the key and she opened the door.
b. He handed her the key and she opened the door [*with the key that he had handed her*].

Consider the scenario where he hands her the key and she does open the door, but not with the key that he had handed her. What are the intuitions about the truth value of the assertion made by means of (20a)? We must distinguish between the case where the speaker *knows* or *believes* that (20b) is false and the case where he believes that (20b) is in fact true and intends to convey this belief. In this latter case, I think the assertion is simply false: the content of the assertion is that of (20b), although it is not fully articulated, and the proposition asserted is false.

In the first case, we should again distinguish between the sub-case where the speaker understands that the enriched proposition will be conveyed, and the sub-case where he does not. In either of these sub-cases, I take the assertion to be *true* but misleading. It is true because the content that is both believed and literally expressed is true. The enriched content is not believed, and not literally expressed. Hence, the speaker does not *lie*. In the sub-case where the speaker is aware of the enrichment, the speaker intentional misleads the hearer, and in the case the speaker is unaware of the enrichment, he unintentionally misleads the hearer.

Intuitions about these cases may not be completely robust, and perhaps they get more shaky when we turn to the DE cases instead. Thus consider:

(21) a. He handed her the key but she did not open the door.
b. He handed her the key but she did not open the door [*with the key that he handed her*].

Consider again the same scenario, where she does open the door, although not with the key that he handed her. Now, in case the speaker of (21a) believes that she did not open the door in any way at all, the belief expressed and the content asserted is simply false.

The crucial case is that where the speaker has a correct belief about the scenario and still makes his assertion by means of (21a). Is the assertion true or false? Is it true but misleading?

In the UE case of (20), where the speaker has correct beliefs about the scenario, we settled for *true but misleading*. It seems to me that in the DE case of (21a), this result is ruled out. The reason is in fact rather simple: opening the door without using the key given to one is a way of opening the door, *not* a way of *not* opening the door. Compare:

(22) a. He handed her the key and she opened the door, although not with the key that he had handed her.
b. *He handed her the key and she did not open the door, although she did open it without using the key he had handed her.

Here, (22a) is perfectly fine: the default enrichment is cancelled by the additional conjunct. By contrast, (22b) is clearly (in my view) unacceptable. Accordingly, we cannot make an assertion of (21a) come out as true but misleading. It is simply false, and (22) brings out the asymmetry between the UE and DE cases.

Let's turn back to the mother's assertion of (17a) in the doubly morbid case: the child has swallowed poison and will soon die, and the mother knows this. In accordance with the discussion above, I think that the mother's assertion is false, not true but misleading. She tells the child that it will not die in the near

future, and that is simply false. This indicates that she is *not* asserting the enriched proposition (18b), which is true.

With this outcome, the radical interpretation project survives. For it is then not the case that the mother holds the false sentence (18a) *true* because of believing the enriched (18b), thereby incorrectly inducing a disconfirmation of the interpretation hypothesis of the radical interpreter. Rather, on the present analysis, she does hold (or at least would be expected to hold) (18a) false. Asserting (18a) would amount to lying.

In a way, this outcome is in agreement with Grice's first maxim of quantity:

(Quantity 1) Make your contribution as informative as is required (for the current purposes of the exchange) (Grice 1975: 26).

This is in accordance with Grice's discussion of disjunction (Grice 1989a: 45-47) and the idea of asserting *the stronger*. In upward entailing contexts, the enriched content is the stronger statement, while in downward entailing contexts, the *un*-enriched content is the stronger content.

It has seemed to many in the semantics-pragmatics literature that the weaker, enriched content (17b) is what is asserted by the mother. I think this is a mistake based on not taking the reasons for the temporal domain restriction into account. With the domain restriction in place, we can stick to the quantity principle of asserting the stronger. If this is right, then, in DE contexts, or at least under negation, enrichments do not even arise (or if they arise, they do not affect the asserted content).

Intuitions concerning the examples considered support the application of Grice's quantity principle, and also supports the project of radical interpretation. It remains an open question whether the study of other examples would yield a different result.[15]

References

Bach, K. 1994, "Conversational Impliciture", *Mind & Language*, 9, 124-62.

Bach, K. 2010, "Impliciture vs. Explicature: What's the Difference?", in Soria, B. and Romero, E., *Explicit Communication. Robyn Carston's Pragmatics*, London: Palgrave MacMillan, 126-37.

Borg, E. 2006, "Minimalism vs. Contextualism in Semantics", in Preyer, G. and Peter, G. (eds.), *Content and Context. Essays on Semantics and Pragmatics*, Oxford: Oxford University Press.

Cappelen, H. and Lepore, E. 2005, *Insensitive Semantics. A Defense of Semantic Minimalism and Speech Act Pluralism*, Oxford: Blackwell.

[15] This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 675415.

For comments and discussion of the ideas in this paper, I am much indebted to Kathrin Glüer-Pagin and to participants in the one-day conference *Donald Davidson Day*, Centenary Symposium, March 2017, at Princeton University, organized by Gilbert Harman and Ernie Lepore. I benefited from comments by in particular Ernie Lepore, Una Stojnic, and Paul Horwich.

I first pointed out the problem discussed in a talk at the Stockholm conference *Systematicity of Pragmatics*, May 2009, organized by Robyn Carston and myself.

Carston, R. 2002, *Thoughts and Utterances*, Oxford: Blackwell.

Carston, R. 2004, "Relevance Theory and the Saying/Implicating Distinction", in Horn, L. R. and Ward, G. (eds.), *The Handbook of Pragmatics*, Oxford: Blackwell, 633-56.

Clark, H.H. 1975, "Bridging", in Schank, R.C. and Nash-Webber, D.L. (eds.), *Theoretical Issues in Natural Language Processing*, New York: Association for Computing Machinery.

Davidson, D. 1969, "On Saying That", in *Inquiries into Truth and Interpretation*, Oxford: Clarendon Press, 93-108. Originally published in *Synthèse*, 19, 1969, 130-46.

Davidson, D. 1973, "Radical Interpretation", in *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press, 125-39. Originally published in *Dialectica*, 27, 1973, 313-28.

Davidson, D. 1974, "Belief and the Basis of Meaning", in *Inquiries into Truth and Interpretation*, Oxford: Clarendon Press, 141-54. Originally published in *Synthèse*, 27, 1974, 329-43.

Davidson, D. 1975, "Thought and Talk", in *Inquiries into Truth and Interpretation*, Oxford: Clarendon Press, 155-70. Originally published in Guttenplan (ed.) 1975.

Davidson, D. 1979, "The Inscrutability of Reference", in *Inquiries into Truth and Interpretation*, Oxford: Clarendon Press, 227-41.

Davidson, D. 1984a, "Communication and Convention", *Synthese*, 59, 1-15. Reprinted in Davidson 1984b, pp. 265-80.

Davidson, D. 1984b, *Inquiries into Truth and Interpretation*, Oxford: Clarendon Press.

Davidson, D. 1986a, "A Coherence Theory of Truth and Knowledge", in *Subjective, Intersubjective, Objective*, Oxford: Oxford University Press, 137-53. Originally published in Lepore (ed.) 1986.

Davidson, D. 1986b, "A Nice Derangement of Epitaphs", in Lepore (ed.) 1986, 89-107.

Davidson, D. 1990, "The Structure and Content of Truth", *Journal of Philosophy*, 87, 279-328. Reprinted in Davidson 2005.

Davidson, D. 2005, *Truth and Predication*, Cambridge, MA: The Belknap Press.

Glüer, K. 2011, *Donald Davidson. A Short Introduction*, New York: Oxford University Press.

Grice, H. P. 1975, "Logic and Conversation", in Cole, P. and Morgan, J.L. (eds.), *Speech Acts*, Vol. 3. Syntax and Semantics, New York: Academic Press, 41-58. Reprinted in Grice 1989b, ch. 2, 22-40. Page references to the reprint.

Grice, H. P. 1989a, "Further Notes on Logic and Conversation", in *Studies in the Ways of Words*, Cambridge, MA: Harvard University Press, 41-57.

Grice, H. P. 1989b, *Studies in the Ways of Words*, Cambridge, MA: Harvard University Press.

Guttenplan, S. (ed.) 1975, *Mind and Language*, Oxford: Oxford University Press.

Hobbs, J.R. 1985, *On the Coherence and Structure of Discourse*, Tech. rep. CSLI-85-37, Stanford: CSLI.

Hume, D. 1748, *An Inquiry Concerning Human Understanding*, New York: Liberal Arts Press, 1955 edition.

Kehler, A. 2002, *Coherence, Reference, and the Theory of Grammar*, Stanford: CSLI.

Lepore, E. (ed.) 1986, *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, Oxford: Blackwell.

Levinson, S.C. 2000, *Presumptive Meanings. The Theory of Generalized Conversational Implicature*, Cambridge, MA: MIT Press.

Nunberg, G. 1979, "The Non-Uniqueness of Semantic Solutions: Polysemy", *Linguistics and Philosophy*, 3, 143-84.

Nunberg, G. 1995, "Transfers of Meaning", *Journal of Semantics*, 12, 109-32.

Pagin, P. 1999, "Radical Interpretation and Compositional Structure", in Zeglen, U. (ed.), *Donald Davidson. Truth, Meaning and Knowledge*, London: Routledge, 59-74.

Pagin, P. 2014, "Pragmatic Enrichment as Coherence Raising", *Philosophical Studies*, 168, 59-100.

Quine, W.V.O. 1960, *Word and Object*, Cambridge, MA: MIT Press.

Recanati, F. 2004. *Literal Meaning*, Cambridge: Cambridge University Press.

Recanati, F. 2010. *Truth-Conditional Pragmatics*, Oxford: Oxford University Press.

Sperber, D. and Wilson, D. 1995, *Relevance. Communication & Cognition*, 2nd ed., Oxford: Blackwell.

Stanley, J. 2000, "Context and Logical Form", *Linguistics & Philosophy*, 23, 391-434.

Wilson, D. and Matsui, T. 1998, "Recent Approaches to Bridging: Truth, Coherence, Relevance", *UCL Working Papers in Linguistics*, 10, 1-28. Reprinted in Wilson and Sperber 2012, 187-209. Page references to the reprint.

Wilson, D. and Sperber, D. 2012, *Meaning and Relevance*, Cambridge: California University Press.

Language's Dreamwork Reconsidered

Andreas Heise

Institut Jean Nicod (CNRS, EHESS, ENS, PSL), Paris

Abstract

This paper offers both exegetical and systematic reconsiderations of Donald Davidson's view on metaphor. In his essay *What Metaphors Mean*, Davidson argued against the idea that metaphors have any kind of propositional content beyond the literal meaning of the relevant sentence. Apart from this negative claim, Davidson also made a constructive proposal by suggesting that metaphor's distinctive effect is to prompt a mental state of seeing-as. These two points seem connected insofar as Davidson makes the following assumptions. First, metaphors cause their distinctive effects in an a-rational way. Second, seeing-as is a non-propositional mental state. If we side with Davidson in thinking of meaning as rational and propositional, then it follows that metaphors' distinctive effects cannot be an instance of meaning. They have the wrong format and are brought about in the wrong way. Against this background, I distinguish a strong reading and a modest reading of Davidson's wrong-kind objection to metaphorical meaning. By taking into account some of Davidson's later pronouncements on the matter, this paper aims to show that Davidson did not hold on to the strong version of the wrong-kind objection. This would open up the way to conceiving of metaphorical meaning in terms of speaker's meaning, were it not for the fact that Davidson sticks to the wrong-way objection. The two concluding sections examine the cogency of the wrong-way objection as applied to metaphorical speaker's meaning, and offer a model for thinking about the a-rational mental causation Davidson thought metaphors exhibit.

Keywords: Davidson, metaphor, speaker's meaning, mental causality, rationality, seeing-as, conceptual innovation.

1. Introduction

Significantly, Donald Davidson (1978: 31) starts his seminal article *What Metaphors Mean* (WMM) with a metaphor: "Metaphor is the dreamwork of language". More significantly still, he goes on to point out the ways in which he wishes to compare metaphor to Sigmund Freud's notion of dreamwork. For Davidson, there are two relevant points of comparison. First, the interpretation of both metaphors and dreams "reflects as much on the interpreter as on the originator". Second, the act of interpretation in both cases is "a work of the im-

Argumenta 3, 1 (2017): 109-125
ISSN 2465-2334

DOI 10.14275/2465-2334/20175.HEI

agination". This second point implies, as Davidson explains, that the interpretation of metaphor is barely, if at all, "guided by rules".

Yet is Davidson's opening metaphor apt? Is using metaphor really like dreaming, and does the interpretation of metaphors and dreams really work in the same way? What would this mean for using metaphor in serious cognitive enterprises such as philosophy? Whether we are happy with Davidson's metaphor seems to depend, among other things, on how seriously we are willing to take dreams on the one hand, and metaphors on the other. Presumably, the analogy would work fine for those with a positivist leaning who think that neither metaphor nor dreams carry cognitive content relevant for the business of doing, say, philosophy. Yet Davidson (1978: 32-33) distances himself explicitly from such a view by asserting that his scepticism about metaphorical meaning does not end up construing metaphor as "confusing, merely emotive, unsuited to serious, scientific, or philosophic discourse". To the contrary, he considers metaphor a legitimate device in science, philosophy, and law, claiming that it is "effective in praise and abuse, prayer and promotion, description and prescription". Would he be willing to say the same thing about dreams? Be this as it may, the main question is whether Davidson's theory delivers his desired result for metaphor. There are two stumbling blocks that Davidson puts in the way of those, like him, who want to claim that metaphor can be used effectively in description and praise, for instance. Both of these blocks fall from his comparison between dreams and metaphors.

First, if the imagination is equally involved in both making and understanding a metaphor, and if we think of the imagination as involving image-like forms of representation, then it might seem that metaphor has the wrong format to feed into descriptions of the world. Elisabeth Camp (2013: 363) calls this *the wrong-kind objection*. Indeed, Davidson's positive proposal concerning the nature of metaphor is that it induces a state of seeing-as. Metaphor "makes us see one thing as another", as he (1978: 47) puts it. In considering a metaphor, we attend to "some likeness, often a novel or surprising likeness, between two or more things" (Davidson 1978: 33). To illustrate this effect, Davidson refers to the ambiguous image of a duck-rabbit, made famous by Ludwig Wittgenstein (1984). What are we led to see in such cases? For Davidson (1978: 46), there is a difficulty in answering this question, because there seems to be "no limit" as to what is called to our attention, and "much of what we are caused to notice is not propositional in character". This would explain the unparaphrasability and creative richness often attributed to metaphors. Yet if we accept that this quasi-perceptual quality applies to metaphor, we run into the wrong-kind objection (1978: 47): "How many facts or propositions are conveyed by a photograph? […] Bad question. […] Words are the wrong currency to exchange for a picture".

Second, if the interpretation of the metaphor "reflects as much on the interpreter as on the originator", as Davidson (1978: 31) says, then one way to spell this out would be to say that the originator lacks privileged access or first-person authority over the interpretation of the metaphor. If the author of a metaphor lacks rational control over its interpretation, and we deem it necessary for meaning to be such that the speaker, if competent, has rational control over the interpretation of her utterance, then whatever interpretation she or her interpreter comes up with, is produced in the wrong way, as it were. This is, in a nutshell, Davidson's *wrong-way objection* (Camp 2013: 364).

In the following, I discuss Davidson's wrong-kind objection in more detail from an exegetical point of view (section 2). I will offer a strong reading as well as a modest reading of this objection, arguing that Davidson is more likely to have endorsed the modest reading. Once we accept that some propositional content may be associated with metaphor, rather than none, the way seems to open up to construe this propositional content in terms of speaker's meaning—assuming that Davidson's arguments against construing it as semantic meaning in his sense are convincing. I explore this option in section 3 by drawing on some of Davidson's remarks about similes. Conceiving of metaphorical meaning this way has surely seemed attractive to a number of people.[1] What went unnoticed by some, however, is that Davidson's wrong-way objection may be construed such that it attacks not only the notion of metaphorical expression meaning but also the idea that there might be something like metaphorical speaker's meaning.[2] At the same time, Davidson does not bring up speaker's meaning in WMM. The notion shows up only in later essays such as *Communication and Convention* (CC; Davidson 2001a), *A Nice Derangement of Epitaphs* (NDE; Davidson 1986), *Locating Literary Language* (LLL; Davidson 1993a), and *The Social Aspect of Language* (SLA; Davidson 1994). It is a burden on the exegetical reconsideration proposed here to tell a plausible story about the line Davidson might have taken on metaphorical speaker's meaning. Sections 4 and 5 are devoted to this task, with section 5 raising systematic considerations against the background of the exegesis offered in section 4. In the concluding section, I shall take stock of what I consider the most plausible conception of metaphor that we may ascribe to Davidson.

2. The Wrong-Kind Objection to Metaphorical Meaning: Strong Reading and Moderate Reading

In order to give a flavour of the tension between the strong and moderate versions of Davidson's wrong-kind objection, I will quote a number of conflicting passages from WMM:

> Metaphor makes us see one thing as another by making some literal statement that inspires or prompts the insight. Since in most cases what the metaphor prompts or inspires is not entirely, or even at all, recognition of some truth or fact, the attempt to give literal expression to the content of the metaphor is simply misguided (Davidson 1978: 47).

Here Davidson emphasises that the effects of metaphor are not propositional.[3] Otherwise they would serve the purpose of recognising some truth or fact. Yet in other passages Davidson seems to waver. In fact, this wavering can already be found in the above, for Davidson says that what the metaphor prompts is "not entirely" recognition of some fact. This seems to leave open the possibility that

[1] For an overview see Reimer and Camp 2006: 855-57, and Carston 2012.
[2] Read: Implicated speaker-meant content of utterances deemed metaphorical that goes beyond the corresponding sentence's literal meaning.
[3] "Propositional effects" is shorthand for mental states or events that have propositional content. I shall say more about this in section 4.

sometimes it might be. Another telling passage that features the same kind of ambivalence is the following:

> What we notice or see is not, in general, propositional in character. Of course it *may* be, and when it is, it usually may be stated in fairly plain words (Davidson 1978: 47).

In these two passages, Davidson wavers, seemingly allowing, albeit reluctantly, that metaphor may result in propositional effects. One way to account for this ambivalence would be to say that he stresses the non-propositional side of metaphorical effects when talking, as it were, to the proponents of metaphorical meaning, and he accommodates potentially true beliefs inspired by metaphor in order to avoid falling back into the kind of denunciation of metaphor associated with positivism or empiricism.

Apart from these ambivalent passages, some statements in WMM seem to commit Davidson in a more or less straightforward manner to the idea that part of what metaphors convey are propositional thoughts:

> Metaphor does lead us to notice what might not otherwise be noticed, and there is no reason, I suppose, not to say these visions, *thoughts* [my italics, A.H.], and feelings inspired by the metaphor, are true or false (Davidson 1978: 41).

> Once we understand a metaphor we can call what we grasp the "metaphorical truth" and (up to a point) say what the "metaphorical meaning" is (Davidson 1978: 33).

Passages like these give hope to the project of contriving a theory that allows for metaphor to convey propositional contents, in some way or other, apart from possibly achieving further effects as well. Then again, we find declarations that block this route outright:

> The central error about metaphor is most easily attacked when it takes the form of a theory of metaphorical meaning, but behind that theory, and statable independently, is the thesis that associated with a metaphor is a cognitive content that its author wishes to convey and that the interpreter must grasp if he is to get the message. This theory is false, whether or not we call the purported cognitive content a meaning (Davidson 1978: 46).

A number of interpreters have commented on the conflict between the passages above (Camp 2013: 365; McGuire 2001). Does Davidson allow for the possibility that speakers might succeed in producing propositional effects in their audience by means of metaphor, or does he want to preclude this possibility? It seems that WMM lends itself to at least two readings of this issue, and I think the issue cannot be resolved by looking only at WMM. That is why I suggest casting a glance at some of Davidson's later pronouncements on the matter.

It is striking that nowhere in WMM does Davidson avail himself of the notion of speaker's meaning or "non-natural meaning" (Grice 1957).[4] Rather, Davidson's (1978: 40) use of the term "meaning" is emphatically restrictive in that

[4] I will use the labels "speaker's meaning" and "non-natural meaning" interchangeably in this paper.

paper so that it applies only to the "ordinary" or literal meaning words and sentences have "prior and independent of the context of use". As I indicated in the introduction, though, Davidson does make use of the notion of speaker's meaning and some other elements of Grice's theory of conversation in his later work.[5] For present purposes, I shall rest content with drawing attention to a section in LLL where Davidson (1993a: 300) concedes that we may decide to use the word 'meaning' also for "what the metaphor carries us to". He adds, in an endnote, that he was "foolishly stubborn" about the word 'meaning' in WMM. I take this statement to give us a prima facie reason to believe that here he abandons the strong version of the wrong-kind objection to metaphorical meaning in favour of the modest one. If true, this would be noteworthy for the strong version threatens any attempt to construe metaphor's effects in terms of speaker's meaning, at least so long as we think of speaker's meaning as propositional. This will be my starting point in the next section.

3. Systematic Reconsideration I: Similes, Metaphors, and Speaker's Meaning

While some people have sympathised with the strong reading of Davidson's wrong-kind objection to metaphorical meaning, most notably Richard Rorty (1987), a wide range of other people have found this position unconvincing (Bergmann 1982; Black 1979; Haack 1987; Moran 1989; Reimer 1996). The strong sceptical attitude towards metaphors conveying propositional content in any way whatsoever seems even less plausible given that in WMM Davidson did not take into account any construal of metaphor's effects in terms of speaker's meaning (McGuire 2001). Of course, if metaphorical effects were entirely non-propositional, this would also pose a problem for such a construal so long as we think of speaker's meaning as propositional (Camp 2013; Carston 2010). Thus, the fact that Davidson abdicates strong non-propositionalism about metaphor's effects is crucial here as well.

Now, what could such a construal of metaphorical effects in terms of speaker's meaning look like? Interestingly, Davidson himself gives us an idea by discussing simile:

> Having decided, we might then say the author of the simile intended us—that is, meant us—to notice that similarity. But having appreciated the difference between what the words meant and what the author accomplished by using those words, we should feel little temptation to explain what has happened by endowing the words themselves with a second, or figurative, meaning (Davidson 1978: 40).

It is true that Davidson speaks about simile here, not about metaphor. And he uses this comparison for his own purposes, asking rhetorically:

> [Simile] may make us think deep thoughts, just as a metaphor does; how come, then, no one appeals to the "special cognitive content" of the simile? (Davidson 1978: 45).

[5] Despite this, Davidson remained sceptical about other parts, most notably the metasemantic project of grounding meaning in non-linguistic intentions (Cook 2009; Davidson 1994: 12, n. 13).

Davidson's argument here works via modus tollendo tollens: If metaphor had a special propositional content, then similes should have a special propositional content too. Similes do not have special propositional content (or at least nobody claims they do). Hence metaphors do not have special cognitive content either.

Of course, it is open to us to question the assumption that similes do not have special propositional content. This would seem even more plausible on the assumption we may share with Davidson, and be it only for the sake of the argument, that the literal meaning of similes consists merely in pointing out that the two things compared share some property or other. If we use this minimalistic conception of what similes literally mean, it is as good as certain that similes will violate either Grice's (1989) Maxim of Relevance or his Maxim of Quantity. "If everything is like everything, and in endless ways" (Davidson 1978: 39), then pointing out this triviality does not look like a worthwhile contribution to most conversations. Davidson (1978: 42) himself recognises that conversational oddity or irrelevance might be what triggers the recognition of metaphor. But he stops short from spelling out the parallel between his remarks on simile and metaphor, and Grice's treatment of conversational implicature. Stressing the fact that, in such cases, the special propositional content does not attach to the words, is beside the point for someone who wants to propose an analysis of metaphorical content in terms of speaker's meaning along Gricean lines. For indeed in order to trigger the search for an implicated speaker's meaning, the words need to be taken in their literal meaning at first. It is precisely the resultant oddity that justifies the interpreter in looking beyond—so long as she has reason to presume the speaker to be cooperative and rational.

Interestingly enough, Davidson comments, much later, on the option of treating metaphor along the lines of conversational implicature. His comment is prompted by Oliver Scholz's (1993) paper on metaphor, to which he replies by saying, among other things, the following:

> Here I should make a remark about Gricean implicature. One important motive in Grice's treatment of conversational implicature was the same as my motive in saying what I did about metaphor: to separate those aspects of communication which can be treated only informally from those which can be given formal semantic treatment, namely, the relatively literal which underlies all the rest. I have nothing but admiration for what Grice did in that direction. It seems to me to be one of the classical defenses of the possibility of a serious theory of meaning (Davidson 1993b: 173).

This statement seems to lend further support to construing metaphor's effects as instances of implicated speaker's meaning. The assumption that implicatures can only be treated informally should not stop Davidson from considering such a construal. After all, by the time of NDE (1986: 446) Davidson goes as far as saying that even the recognition of literal or "first" meaning relies only on "rough maxims". Moreover, Davidson (1978: 46) gives further fodder to such a construal by comparing metaphorical effects to the way the Delphic Oracle communicates, according to Heraclitus: "It does not say and it does not hide, it intimates".

Against the hope however, that Davidson's account of metaphor might be easily converted into a Gricean one, it should be noted that the wrong-way objection remains valid. Indeed, there is reason to believe that it is actually Davidson's crucial argument against the idea of metaphorical meaning. This is the

argument I focus on in the following section, before assessing its cogency in sections 5 and 6.

4. The Wrong-Way Objection to Metaphorical Meaning: Mental Causality beyond Rationality

In the exegetical reconsideration of Davidson's wrong-kind objection in section 2, we have encountered reasons to believe that Davidson preferred the moderate over the strong reading of the wrong-kind objection. In the systematic reconsideration in section 3, I sketched a way in which Davidson may be turned into a simile theorist of metaphor. This construal requires that we conceive of the propositional effects of metaphor that the moderate reading of the wrong-kind objection allows for as instances of speaker's meaning. To come back to a more exegetical perspective, I want to quote an interview that Davidson gave to Kathrin Glüer that seems, at first sight, to support this project:

> Davidson: If you call somebody a rat, and you intend that as a metaphor, somebody is not getting what you mean if they think that the word "rat" means a despicable person. They have to know what a rat is in order to understand the metaphor. First meaning depends upon your past practice or at least whatever a person has to go on to figure out what the first meaning is. Now, what happens after that? In my original article on metaphor, I resisted calling it meaning because it didn't have that character. It wasn't something that you could be expected to have prepared people for in advance, something they are used to and so forth. Otherwise they would just take the word to be ambiguous or just to mean that. […]
>
> Glüer: [...] To me, it is absolutely plausible to say [...] that [...] the interpretation of metaphors requires that there is this first meaning, however ephemeral it might be, that gets the interpreter going on what the speaker is ultimately up to. The speaker has to have two (or more) intentions: one semantic that settles for the first meaning, and a secondary, "metaphoric" intention as well.
>
> Davidson: That is exactly what I would say (Davidson and Glüer 1995: 82-3).

Now the crucial question is what does this "secondary metaphoric intention" mean? For those in the business of defending metaphorical speaker's meaning, it would certainly be tempting to construe it as a reflexive intention in Paul Grice's (1957) sense. Yet while this secondary intention could indeed be construed as an instance of non-natural meaning, it could be construed equally well as a case of natural meaning, possibly in the sense of a "perlocutionary effect" (Tirrell 1991: 154), or an "ulterior purpose" (Davidson 2001a: 272).[6] Before giving reasons for thinking that the second interpretation is more likely to be true, I deem it

[6] "Perlocutionary act" or "perlocutionary effect" are technical terms from speech act theory that originated with J.L. Austin (1962). Davidson (2001a, 1993a) uses the terms "ulterior purpose" or "ulterior effect", respectively, to refer to what he considers roughly the same things. Furthermore, we may consider perlocutionary effects to be instances of what Grice (1957: 378) called "natural meaning". I will assume here that these different terminologies do in fact refer to the same phenomena, yet I do not think that any argument hinges on this assumption. In case of doubt, Davidson's terminology should be considered binding. While I deem it desirable to link these two frameworks to each other, and to speech act theory, I shall not pursue this project here.

worthwhile to recall some elements of Davidson's thinking concerning illocutionary force and ulterior purposes.

In WMM, Davidson (1978: 33) stresses that metaphor belongs to the domain of language use. In this context, it is worth noting that, for Davidson, already the illocutionary force belongs to that domain (Davidson 2001a, 2001d). Moreover, both force and ulterior purpose are, on this view, not guided by rules—in the sense of conventions—even though the identification of the illocutionary force is at the same time a necessary condition for understanding:

> There is one intention not touched on by a theory of truth which a speaker must intend an interpreter to perceive, the [illocutionary, A.H.] *force* of the utterance. An interpreter must, if he is to understand a speaker, be able to tell whether an utterance is intended as a joke, an assertion, an order, a question, and so forth. I do not believe there are rules or conventions that govern this essential aspect of language. It is something language users can convey to hearers and hearers can, often enough, detect; but this does not show that these abilities can be regimented. I think there are sound reasons for thinking nothing like a serious theory is possible concerning this dimension of language. Still less are there conventions or rules for creating or understanding metaphors, irony, humor, etc. (Davidson 1990: 312-13, n. 56).

If it were not for the fact that humour or jokes show up twice, it would be clear that Davidson is talking about illocutionary force in the second sentence, and ulterior effects in the last sentence of this footnote. Be this as it may, if he had wanted to signal that metaphors, irony, and humour belong to the same category as assertions, orders, and questions, which seem clear cases of illocutionary force, he would not have opened the last sentence with "still less".

Why does it matter whether we think of metaphorical effects in terms of illocutionary force or ulterior purposes? In some sense, it does not seem to matter much for Davidson insofar as, for him, neither is guided by rules, as the quote above shows. In that sense, both belong to the domain of language use. Yet force is different from ulterior effects to the extent that, first, it is governed by a Gricean reflexive intention ("a speaker must intend an interpreter to perceive […] the force of the utterance"), and second, that it is essential to understanding ("[a]n interpreter must, if he is to understand a speaker, be able to [identify the illocutionary force, A.H.]"). These conditions do not hold in the case of ulterior purposes. It is not a necessary condition on the bringing about of some ulterior effect like scaring ("Boo!") that the interpreter be able to detect the speaker or agent's intention to bring about that effect. Their success, the bringing about of the intended effect, does not depend on the audience's recognising this intention nor on the speaker intending for her audience to recognise her intention to bring about the intended effect. In other words, such acts are not acts of communication.

If the effects of metaphor should turn out to be ulterior effects, then this is just another way of saying that the wrong-way objection applies to metaphorical speaker's meaning, as will become clearer in the following section. Indeed, there is evidence that construing the wrong-way objection this way comes close to what Davidson had in mind. Let us have a look at his reply to a paper by Pablo Quintanilla (1999) on metaphor and conceptual innovation:

> But as Professor Quintanilla says, when metaphor affects the propositional contents of beliefs, not all that eventuates can be rationalized.[7] He describes such a process as the displacement of non-propositional thoughts by propositional thoughts. I am slightly less inclined to speak of non-propositional thoughts, but only because I (perhaps somewhat arbitrarily) restrict the word "thought" to mental states and events with propositional contents. But it hardly matters; certainly the ways metaphor, imagination, conceptual creativity, and daydreaming work their wonders in the mind are cases of mental causality which is outside or beyond the rational (Davidson 1999: 327).

In order to get a sense of this kind of mental causality that lies outside the rational, I suggest casting a glance at Davidson's (2004) essay *Paradoxes of Irrationality* (POI). There he rehearses, first, what *rational mental causality* amounts to on his account. Two elements are crucial for this type of causality, which he also refers to as *reason explanation*. First, in order for mental states, typically belief-desire pairs, to explain some other mental state, action, or event, the propositional contents of the former need to stand in "appropriate logical relations"[8] to those of the latter (Davidson 2004: 179). Second, the former needs to cause the latter.[9] In short, reasons are, for Davidson, mental causes that stand in certain logical relations to the effects they cause. It follows from this that both the cause and the effect cited in some reason explanation need to lend themselves to a description in propositional terms. For if one, or both, of the relevant states or events defy the assignment of some propositional content, no logical connection could be established in the first place. This setup allows Davidson, in a second step, to characterise cases of mental causality that are beyond the rational, meaning that they fail to meet, in one way or another, the two conditions on reason explanations just outlined.

Davidson discusses four such cases in POI: manipulation, for instance, luring a person into your garden by growing a beautiful flower (2004: 181); perception, for instance, coming to believe that a bird is flying by (2004: 179); association, for instance, humming a tune in order to recall a name (2004: 186); and self-improvement, namely changing your attitudes or desires by means of a second-order desire (2004: 186-7). As a general term for these four cases, I suggest the label *a-rational mental causation*, or *a-rationality*. Davidson makes it clear that he wants to distinguish these four cases from the cases of *irrationality* with which he is primarily concerned in POI, but that I propose to largely ignore here.[10]

[7] Even in this late pronouncement, we once again encounter the ambivalence in Davidson's formulation ("not all") that I pointed out in section 2.

[8] As Gozzano (1999: 138-39) explains, the appropriate relations that account for rationality comprise principles of decision theory and logic, as well as, for instance, the requirement of total evidence for inductive reasoning.

[9] These conditions are necessary for rational mental causality, yet not sufficient, as Davidson (2004: 173) claims. I follow Davidson in neglecting this complication in the present context.

[10] In order to justify this omission, I point to my introduction where I argued that Davidson did not consider metaphor to be an irrelevant, possibly irrational, rhetorical device that stands in the way of serious scientific inquiry, for instance. As long as we can avoid construing metaphor as irrational within Davidson's framework, I think this line should be pursued.

While these four cases fail to meet the conditions on rational mental causality, they do so for different reasons. Perception is, for Davidson, a case where the cause cannot be described in propositional terms. Manipulation, by contrast, is a case where the cause can be assigned a propositional content, namely the desire that you enter my garden in combination with my belief that I will succeed by growing a beautiful flower there. Yet while my intention may cause the desired effect, let us assume you do enter my garden, it does not, as Davidson (2004: 181) puts it, constitute the reason on which you acted. Rather, in the scenario that we are invited to envisage, it is your desire to have a look at the flower that causes you to enter. This is, thus, a case where there is no appropriate logical connection between cause and effect. Self-improvement, and incidentally the cases of irrationality that Davidson discusses in POI, are structurally similar to the manipulation case insofar as the logical relation is "missing or distorted" (Davidson 2004: 179).[11] At the same time they differ from manipulation to the extent that the causal relation obtains within the mind of a single person, instead of spanning two minds. Association, finally, seems to share with perception the deficiency that the cause lacks propositional content—at least if we think about Davidson's (2004: 186) case of "humming a certain tune". It nevertheless differs from perception, because the cause is internal, rather than external, to the individual's mind.

Against this background we can now return to the guiding question of this section: How are we to understand the "second metaphoric intention" that Davidson endorses in his reply to Glüer? If the systematic reconsideration I offered in section 3 is to go through, that is if metaphorical meaning can be thought of as speaker's meaning, then this intention needs to be reflexive—in the sense that the speaker intends for her audience to recognise it, and that it serves as a reason for the audience to retrieve the relevant propositional content. Yet in his reply to Quintanilla, Davidson argues that metaphor's effects are caused in the "wrong way", that is in a way that defies the requirements on reason explanation. If this were true, then it would undermine the idea of metaphorical speaker's meaning. In my second systematic reconsideration, hence, I shall examine the cogency of this objection. In doing so, I will also work towards clarifying which of the four cases of a-rationality mentioned above would seem to fit metaphor best. I draw together my results in the conclusion where I present what I take to be the most plausible view of metaphor that we may attribute to Davidson on the basis of my exegetical and systematic reconsiderations.

5. Systematic Reconsideration II: Davidson and Grice on Rationality in Communication

In section 3, I offered a reading of WMM that paves the way to construing metaphor's effects in terms of implicated speaker's meaning.[12] Davidson himself

[11] In "First Person Authority" (FPA), Davidson (2001b: 7) considers yet another case that seems to fit this pattern, the "noninferential" recovery of authority over a previously unconscious attitude in psychoanalytic practice. This example is telling given the comparison between dreamwork and metaphor with which Davidson opens WMM.

[12] To be sure, I do not want to argue myself that metaphorical meaning, if there is such a thing, is best cast in the mould of implicated speaker's meaning. My aim is merely to evalu-

seemed to be open to such a construal. That is at least what his reply to Scholz suggests. The reply to Quintanilla I discussed in section 4, however, stands in conflict with this interpretation. There he makes it clear that metaphor's effects are to be understood as an instance of a-rational mental causation. I suggest pinning down these conflicting interpretations with reference to the "secondary metaphoric intention" that Glüer named. In order to make sense of Davidson's toying around with the idea of implicated speaker's meaning as a model for metaphor, this intention would have to be reflexive. Reflexive intentions are a necessary condition for speaker's meaning. If, however, we take Davidson's wrong-way objection seriously, then this intention cannot be reflexive.[13] Rather, it would be an instance of mental a-rationality. Davidson thus faces a dilemma. Either he accepts that metaphor's effects may be construed as conversational implicatures in Grice's sense, but then he would have to accept that these effects are within the domain of the rational—at least on a Gricean understanding of rationality. Or else he sticks with his assessment that the effects of metaphor are a-rational, but then it seems that they cannot be construed as conversational implicatures.

We have at least two potential options to resolve this dilemma in a way that salvages as much as possible of what seems to be Davidson's conception of metaphor.[14] One option would consist in showing that Davidson and Grice simply employ different standards as to what counts as rational forms of communication. The second option I will examine is whether we find in Grice a model of a-rational verbal behaviour that can accommodate what Davidson says in his reply to Quintanilla. The goal is to assess whether Davidson's wrong-way objection holds even if we try to construe the effects of metaphor in terms of implicated speaker's meaning.[15] Since the wrong-way objection rests on the contention that these effects are produced in a purely causal, hence a-rational way, two things need to be shown in order for the objection to apply to speaker's meaning. For one, we would need to show that rationality is indeed a demand on speaker's meaning, something I will largely take for granted here. For another, we would have to establish that the Grice's and Davidson's frameworks are sufficiently similar with respect to the underlying notion of rationality. Otherwise the dispute threatens to be merely verbal.[16] Due to lack of space, I have to defer an examination of this option to resolve Davidson's dilemma to some later occasion. In other words, I shall additionally assume that the two frameworks are relevantly similar.

ate whether such a project would be compatible with Davidson's pronouncements. For an overview of the problems that such a project has to face, see, for instance, Carston 2012.

[13] The reasons backing up this contention will become more evident in the course of this section.

[14] Here I still assume that there might be a way to reconcile most of what Davidson says about metaphor, in WMM and beyond. Of course, an exegetical reconsideration of a quite different sort might simply aim for the conclusion that Davidson's pronouncements on the matter cannot be made coherent.

[15] I am thus defending Davidson, as he appears in his reply to Quintanilla, against Davidson, as he appears in his reply to Scholz.

[16] This would be a case of Davidson having his cake and eating it too. If different standards of rationality were in play, Davidson could maintain that metaphorical effects are a case of speaker's meaning, while denying that speaker's meaning has to be rational in his sense of the term.

I turn thus to the second option to resolve the dilemma. Grice considers two cases of a-rationality that may disqualify some speaker's intention from counting as an instance of non-natural meaning.[17] For one, the speaker must not think it is a "foregone conclusion" (Grice 1957: 384) that her reflexive intentions will play no role in achieving the relevant cognitive effects in her audience. If she did, this would preclude her from even forming such a communicative intention, since having an intention requires, on most conceptions, that we believe we can, in principle, realise the corresponding goal.[18] Under the same heading, Davidson (1986: 440) discusses a snippet from a dialogue between two characters in Lewis Carroll's (1994: 100) *Through the Looking Glass*. Humpty Dumpty tries to convey to Alice that he has just produced a nice knockdown argument against her position (on a topic irrelevant to our concerns) by saying "There's glory for you". He could not believe, though, that Alice would be able to interpret him correctly out of the blue, Davidson argues. And indeed, upon Alice expressing her bafflement, Humpty Dumpty rejoins "Of course you don't [know what I mean by 'glory', A.H.]—till I tell you". So, on both Davidson's and Grice's account, Humpty Dumpty did not have, and could not have had, a reflexive, communicative intention. He knew, as his confession shows, that Alice would not be in a position to get, without further information, what he was up to by using the term 'glory' in that way.

Now, why should someone who speaks metaphorically fail to have a reason for believing that her audience grasps what she is trying to convey? For one way to understand Davidson's wrong-way objection to metaphorical speaker's meaning, which I reconsidered in the previous section, would have it that indeed she must fail to have such a reason. This would explain why the act of speaking metaphorically is a-rational for Davidson.

One way to construe the wrong-way objection along these lines would consist in combining it with the wrong-kind objection. We find this line of argument most explicitly stated in Davidson's reply to Oliver Scholz who chides him for relying on an unclear notion of seeing-as. Davidson's answer comes pat:

> I thought Wittgenstein and others had made the notion of seeing as clear enough to make their (and my) point: there are important experiences that cannot be reduced to one way or another of grasping a propositional content. If it is a central function of (fresh, active, live) metaphors to induce such experiences, no theory of reference or truth can cope with what is distinctive about metaphor (Davidson 1993b: 173).

We are back to the wrong-kind objection, or so it seems. If the mental state of seeing-as is such that it does not involve grasping any proposition, and given that it is a hallmark of rationality that it operates holistically via inferential links between propositions (Gozzano 1999: 138), then these links break down in the

[17] In fact, I distinguish here two cases that for Grice (1957) were but one, namely the second one I am going to discuss. Yet I think that this second case is just a specific instance of the more general problem that sometimes the speaker's intentions do not play the role they are supposed to play in communication when it comes to producing the relevant cognitive effects. Davidson brings up another case that fits this general description, as we will presently see.

[18] According to a more modest conception, it would only be required that we do not believe that our aim is impossible to meet (Longworth 2017).

face of seeing-as. If we read Davidson that way, then he can even allow for the case that entertaining this state may result in recognising some fact or other. This would be a mere by-product, though, and the speaker may not, if she believes that the mental state of seeing-as works in the way described by Davidson, harbour any hopes that she is in a position to exert any rational control over these by-products.

This argument looks, at first, similar to the second case that may hamper the rationality of speaker's meaning, according to Grice (1957: 382). If Herod shows Salome the severed head of John the Baptist on a charger, then the communicative intentions of the agent fail to serve as reason for the audience to form the relevant belief, say, that St. John is dead. Or so Grice argues.[19] Davidson's argument in the quote above seems to be structurally similar insofar as it involves, via the mental state of seeing-as, a perceptual or quasi-perceptual element. Unlike Grice, however, he does not think that this element gives away too easily, as it were, the relevant information, without the speaker's communicative intention playing any role. Rather, the quasi-perceptual mental state of seeing-as interferes with meaning and communication in a different way. Since there is, as Davidson presumes, no proposition to grasp, no belief will be formed either, and just as little information will be gathered.[20]

This combination of the wrong-kind and wrong-way objections probably construes Davidson's argument against metaphorical content most faithfully. Yet, to some extent, I agree with Scholz's critical stance that prompted Davidson's reply above. The role that Davidson wishes to assign to seeing-as in the context of the combined argument seems to rest on assumptions that are questionable. To begin with: Is Wittgenstein's notion of seeing-as the right model for metaphor's distinctive effects, given that metaphor is a phenomenon of language or thought, but not, or not necessarily, of perception (Kemp 1991: 86)? Even if seeing-as, or related notions such as "taking-as" (Tirrell 1991: 149) or "imagina-

[19] I am sceptical about the cogency of such cases. I would argue that the information Salome can infer without referring to Herod's communicative intentions is not the same as that which she gains via recognition of Herod's intention. In the first case, the relevant proposition is, indeed as Grice suggests, something to the effect that John the Baptist is dead. In the second case, however, it would be something like the realization that Herod had John the Baptist killed. These propositions are sufficiently distinct, I take it. More generally, my tendency would be to claim that whenever people stage evidence such as severed heads, what they want to communicate is richer than the conclusion the evidence indicates by itself. Presumably, the additional element concerns the communicator's involvement or attitude regarding the relevant states of affairs. I shall not, however, pursue this line of argument here.

[20] For Wittgenstein, seeing-as seems to have been a general term for an array of mental states that range from cases bordering on conceptual thinking, at one extreme, to cases bordering on non-conceptual perception, at the other (Glock 1996: 38). While appreciating the reversible figure of the duck-rabbit arguably requires you to possess the concepts of both a duck and a rabbit, Davidson apparently thought otherwise. Either he conceived of seeing-as as a sui generis mental state that defies description in propositional terms, as his reply to Scholz seems to suggest, or else he patterned seeing-as on perception. For as we have seen in section 4, perceptual inputs also do not have propositional content for Davidson. Then again, Davidson grants that perception causes beliefs, and beliefs do have propositional content. So the a-rationality of perception could only serve as a model for metaphor if Davidson conceded that at least some of the ensuing effects may be propositional. See sections 2 and 6 for more on this.

tive seeing" (McGinn 2004: 48-55), should turn out to capture aptly what is distinctive about metaphor, is Davidson right in holding that this mental state necessarily defies the grasping of propositions?[21] Finally, even if the mental state of seeing-as were non-propositional on a conception of propositions as structured entities, would the argument still go through if we switched to a conception of propositions as functions from possible worlds into truth values (Moran 2017: 384; Stalnaker 1972)?

By appealing to the quasi-perceptual state of "metaphorical" seeing-as, Davidson's combined argument would seem to be closest to the case of perception in the fourfold typology of a-rationality discussed in section 4. The wrong format of seeing-as, that is the presumed absence of any propositional content to be grasped, would hence explain why metaphor fails to correspond to the requirements of reason explanation. For reasons sketched in the previous paragraph, however, this is not, from a systematic perspective, the most promising line of argument in defence of Davidson's wrong-way objection to metaphorical speaker's meaning. Having said that, the combined argument might well be the most faithful to Davidson's position from an exegetical point of view. I will not try to settle this exegetical question here. Rather, I will explore, by way of conclusion, if there is another way to construe the wrong-way objection that avoids the problematic notion of seeing-as.

To that effect, I submit, first of all, that the wrong-way objection to metaphorical speaker's meaning needs to be construed as a modal claim as follows: The speaker must not have any reason to expect that she exerts rational control over the effects of the metaphor. For, of course, there is also a version of the wrong-way objection that is not a question of modality, but merely an empirical matter. It may well be that sometimes speakers or writers use metaphor without having any definite content in mind that they wish to convey. Perhaps the purpose of certain poetic metaphors such as "Time is a pond in which the past bubbles to the surface" (Ransmayr 1987: 166) is not to make an assertion about the nature of time, but to captivate the reader's imagination by prompting the mental image of a bubbling pond. This may well be how things stand in this case. Our intuitions arguably diverge as to how often metaphors are used playfully, and how often they are used to communicate content. Be this as it may, for Davidson's argument against metaphorical speaker's meaning to go through, he would have to show that metaphors cannot serve the purpose of communication by their very nature. If the distinctive effect of metaphor were to induce a state relevantly similar to seeing-as, and if this state defies description in propositional terms, then Davidson would have such a modal argument at his command. Yet there is another, possibly more promising line of argument open to him as well. It is to this issue that I turn now.

6. Conclusion

In the previous section, I concluded my systematic reconsideration of Davidson's wrong-way objection to metaphorical speaker's meaning by arguing that it needs to be understood as a modal claim. Something about metaphor would have to violate, by necessity, the following rationality constraint on speaker's meaning, on which Davidson seems to agree with Grice: The speaker

[21] This question refers to the point I raised in footnote 20.

needs to have a reason to expect that she will succeed in conveying a particular propositional content to her audience by means of metaphor, and the audience's recognition of her intention serves as reason for them to retrieve it. The wrong-way objection would have it that the speaker cannot reasonably expect to succeed in conveying a propositional content to her audience, even if she should wish to. While Davidson seemed to hold, as his reply to Scholz shows, that something like seeing-as fits this bill, I argue now that there is another model available, one potentially less fraught with problems. Davidson himself adopts this model in his reply to Quintanilla:

> I agree with Professor Quintanilla that metaphor can play an important role in conceptual invention and change. […] A more drastic form of conceptual change involves explicit introduction of novel concepts, concepts not definable in terms of the original stock. Here, it is natural to say, metaphor can play an overt role. It enters at first by providing an insight by using a familiar word or phrase in a surprising and suggestive way. But in cases where the insight proves to have a general application, the metaphor hardens, surprise dissipates, and suggestion turns to forthright description (Davidson, 1999: 326-27).

Davidson here offers a different perspective on the wrong-way objection. In cases of conceptual change, the speaker could not reasonably expect that she will succeed in expressing and communicating the concept in question because either she or her audience did not yet possess that concept prior to encountering the metaphor. As a type of a-rationality, conceptual change works differently from the combination of the wrong-way and wrong-kind objections we encountered in section 5. While perception suggested itself as the model of a-rational mental causation for the combined argument,[22] either association or manipulation would fit conceptual change. Which of the two works better depends on whether we assume the likeness the metaphor draws attention to is novel only to the audience, or to the speaker as well. Manipulation would seem to suit the first case, whereas association might capture the second.[23] In any case, this model for metaphor requires that the originator of the metaphor believes that at least her audience, and possibly she as well, are not prepared to instantly grasp the metaphor's "meaning".

Of course, it may be held against this conception that it applies at best to a small fraction of the metaphors we use in everyday contexts. True, but the apparent artificiality of Davidson's account dissipates, to some degree at least, if we take into account that most of these are "dead" or idiomatic metaphors such as "He was burned up" (Davidson 1978: 38). Moreover, Davidson would probably want novelty to be construed on the microlevel of the language users' idiolects where conceptual change arguably occurs more frequently than on the macrolevel of a linguistic community. In addition, it seems to be a merit of Davidson's conception that it aims to do justice to metaphor's creative power in

[22] With some reservations. See footnote 20.

[23] Alternatively, we might also consider the recovery of unconscious attitudes as a model for the kind of a-rational mental causality involved in metaphor (see footnote 11). As mentioned earlier, this line of thinking is tempting, given Davidson's "dreamwork" metaphor. To the extent, however, that this model requires an intrapersonal split into two subsystems, one conscious and one unconscious, this raises certain questions regarding Davidson's philosophy of mind, most notably its holism. For further discussion see Gozzano 1999.

a way that keeps sight of the element of surprise that goes along with this process from the subject's point of view. The price to pay for accommodating this creativity in the form of conceptual change may be that the production of novel metaphors, at its very initial stage, turns out to be a-rational. This holds so long as we believe, as Davidson seems to have done, that conceptual change by means of metaphor is a process that cannot be moulded into reason explanations. In this sense, then, metaphor might indeed resemble dreamwork, though formal pragmaticists, cognitive linguists, and psychoanalysts will wake up with a start upon hearing this.[24]

References

Austin, J.L. 1962, *How to Do Things With Words*, Oxford: Oxford University Press.

Bergmann, M. 1982, "Metaphorical Assertions", *The Philosophical Review*, 91, 2, 229-45.

Black, M. 1979, "More about Metaphor", in Ortony, A. (ed.), *Metaphor and Thought*, Cambridge: Cambridge University Press, 19-43.

Camp, E. 2013, "Metaphor and Varieties of Meaning", in Lepore, E. and Ludwig, K. (eds.), *A Companion to Donald Davidson*, Chichester: Wiley-Blackwell, 361-78.

Carroll, L. 1994, *Through the Looking Glass*, London: Penguin.

Carston, R. 2010, "Metaphor: Ad Hoc Concepts, Literal Meaning and Mental Images", *Proceedings of the Aristotelian Society*, 110, 3, 297-323.

Carston, R. 2012, "Metaphor and the Literal/Nonliteral Distinction", in Allan, K. and Jaszczolt, K. (eds.), *The Cambridge Handbook of Pragmatics*, Cambridge: Cambridge University Press, 469-92.

Cook, J. 2009, "Is Davidson a Gricean?", *Dialogue*, 48, 3, 557-75.

Davidson, D. 1978, "What Metaphors Mean", *Critical Inquiry*, 5, 1, 31-47.

Davidson, D. 1986, "A Nice Derangement of Epitaphs", in Lepore, E. (ed.), *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, Oxford: Blackwell, 433-46.

Davidson, D. 1990, "The Structure and Content of Truth", *The Journal of Philosophy*, 87, 6, 279-328.

Davidson, D. 1993a, "Locating Literary Language", in Dasenbrock, R.W. (ed.), *Literary Theory After Davidson*, University Park, PA: The Pennsylvania State University Press, 295-308.

Davidson, D. 1993b, "Reply to Oliver Scholz", in Stoecker, R. (ed.), *Reflecting Davidson. Donald Davidson Responding to an International Forum of Philosophers*, Berlin: de Gruyter, 172-73.

Davidson, D. 1994, "The Social Aspect of Language", in McGuinness, B. and Oliveri, G. (eds.), *The Philosophy of Michael Dummett*, Dordrecht: Kluwer Academic Publishers, 1-16.

Davidson, D. 1999, "Reply to Pablo Quintanilla", in Caorsi, C.E. (ed.), *Ensayos sobre Davidson*, Montevideo: Fundación de cultura universitaria, 326-28.

[24] I wish to thank the following people for feedback on previous versions of this paper: Guy Longworth, Hanna Lukkari, Francesca Ervas, and François Recanati. Moreover, I acknowledge support by the following two grants: ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL*.

Davidson, D. 2001a, "Communication and Convention", in Davidson 2001c, 265-80.

Davidson, D. 2001b, "First Person Authority", in *Subjective, Intersubjective, Objective*, Oxford: Oxford University Press, 3-14.

Davidson, D. 2001c, *Inquiries into Truth and Interpretation*, Oxford: Oxford University Press.

Davidson, D. 2001d, "Moods and Performance", in Davidson 2001c, 109-21.

Davidson, D. 2004, "Paradoxes of Irrationality", in *Problems of Rationality*, Oxford: Oxford University Press, 169-87.

Davidson, D. and Glüer, K. 1995, "Relations and Transitions – An Interview with Donald Davidson", *Dialectica*, 49, 1, 75-86.

Glock, H.-J. 1996, *A Wittgenstein Dictionary*, Oxford: Blackwell.

Gozzano, S. 1999, "Davidson on Rationality and Irrationality", in De Caro, M. (ed.), *Interpretations and Causes: New Perspectives on Donald Davidson's Philosophy*, Dordrecht: Kluwer, 137-49.

Grice, P. 1957, "Meaning", *The Philosophical Review*, 66, 3, 377-88.

Grice, P. 1989, "Logic and Conversation", in *Studies in the Ways of Words*, Cambridge, MA: Harvard University Press, 22-40.

Haack, S. 1987, "Surprising Noises: Rorty and Hesse on Metaphor", *Proceedings of the Aristotelian Society*, 88, 293-301.

Kemp, G.N. 1991, "Metaphor and Aspect-Perception", *Analysis*, 51, 2, 84.

Longworth, G. 2017, "Faith in Kant", in Faulkner, P. and Simpson, T.W. (eds.), *The Philosophy of Trust*, Oxford: Oxford University Press, 251-71.

McGinn, C. 2004, *Mindsight. Image, Dream, Meaning*, Cambridge, MA: Harvard University Press.

McGuire, J.M. 2001, "Sentence Meaning, Speaker Meaning, and Davidson's Denial of Metaphorical Meaning", *Dialogue*, 40, 3, 443-52.

Moran, R. 1989, "Seeing and Believing: Metaphor, Image, and Force", *Critical Inquiry*, 16, 1, 87-112.

Moran, R. 2017, "Metaphor", in Hale, B., Wright, C. and Miller, A. (eds.), *A Companion to the Philosophy of Language*, 2nd edition, Chichester: Wiley-Blackwell, 375-93.

Quintanilla, P. 1999, "La hermenéutica de Davidson: metáfora y creación conceptual", in Caorsi, C.E. (ed.), *Ensayos sobre Davidson*, Montevideo: Fundación de cultura universitaria, 75-98.

Ransmayr, C. 1987, *Die Schrecken des Eises und der Finsternis*, Frankfurt a.M.: Fischer.

Reimer, M. 1996, "The Problem of Dead Metaphors", *Philosophical Studies*, 82, 13-25.

Reimer, M. and Camp, E. 2006, "Metaphor", in Lepore, E. and Smith, B.C. (eds.), *The Oxford Handbook of Philosophy of Language*, Oxford: Oxford University Press, 845-63.

Rorty, R. 1987, "Unfamiliar Noises: Hesse and Davidson on Metaphor", *Proceedings of the Aristotelian Society*, 61, 283-96.

Scholz, O.R. 1993, "'What Metaphors Mean' and How Metaphors Refer", in Stoecker, R. (ed.), *Reflecting Davidson. Donald Davidson Responding to an International Forum of Philosophers*, Berlin: de Gruyter, 161-71.

Stalnaker, R. 1972, "Pragmatics", in Davidson, D. and Harman, G. (eds.), *Semantics of Natural Language*, Dordrecht: Reidel, 380-97.

Demystifying Davidson: Radical Interpretation meets Radical Enactivism

Daniel D. Hutto and Glenda Satne***

* *University of Wollongong*
** *University of Wollongong and Alberto Hurtado University*

Abstract

Davidson's signature ideas on the holism and autonomy of propositional thought have led some exegetes to hold that he advances a kind of transcendentalism that is discordant with a satisfactory naturalism. On the other hand, Davidson's work has strong connections with naturalism, as some Quinean strands of his thinking make apparent. Two strands can thus be identified in Davidson's thought. One emphasizes features of thought that set it apart from the rest of nature. The other seeks to locate thought within nature. Taken to extremes these different strands in Davidson's thinking come into tension. After summarizing both strands, we diagnose the apparent tension between them and propose a way to overcome it by making central appeal to the Radical Enactivist claim that minds can be intentionally directed to the world without contentfully representing it. By expanding our thinking about the character of the mental along radically enactivist lines it becomes possible to defend some of Davidson's most important insights about minds while also promoting a satisfactory and demystifying naturalism.

Keywords: Radical Interpretation, Naturalism, Intentionality, Radical Enactivism

«*I don't think the issue whether animals have beliefs is in itself of any importance—one can use words as one pleases. But if you want to talk about pre-linguistic thought, you need to explain precisely what you have in mind*».
Donald Davidson, 10th November 1991 (personal communication).

1. Introduction

Any naturalist worth his or her salt, even if methodologically non-reductionist, should seek to make the connections between contentful thought and the natural world non-mysterious.

Davidson's signature ideas on the holism and autonomy of propositional thought have led some exegetes to hold that he advances a kind of transcenden-

Argumenta 3, 1 (2017): 127-144
ISSN 2465-2334

DOI 10.14275/2465-2334/20175.HUT

talism that is discordant with a satisfactory naturalism (see for example Maker 1991; Cutrofello 1999; Genova 1999; and Barth 2011).

2. Two Strands in Davidson's Thought about Thought

Many of Davidson's central claims promote a reading of his work that is hard to square with the more naturalistic strands of his thinking. These claims are that:

1. Mastery of natural language is a condition for having objective thoughts—namely, thoughts that can be true or false, correct or incorrect—about anything at all.
2. There are holistic connections between thought and language; between meaning and belief.
3. It is only by mastering natural language that one enters into or breaks into the holistic interpretative circle that holds between belief and meaning.
4. How we come to master contentful language and thought cannot be understood or explained 'from the outside'—for example, by adopting the perspective of and using the resources of the empirical sciences.

By Davidson's lights the domain of propositional thought depends on interpretative practice and this reveals the former to have special constitutive features that distinguishes it from the rest of nature.

Davidson reaches this conclusion by building on and substantively adjusting Quine's thought experiment of the radical translator. Davidson introduces the idea of a field interpreter—an interpreter who has nothing but his observations to go on when interpreting others.[1] Ultimately, this imaginative exercise is meant to bring out why someone in such circumstances has no choice but to rely on constitutive principles of charity if she is to recognize the existence of contentful minds.

The radical interpreter must call on such principles if they are going to break into the holistic circle.[2] As Davidson puts it:

> We do not know what someone means unless we know what he believes; we do not know what someone believes unless we know what he means. In radical interpretation we are able to break into this circle (Davidson 1984: 27).[3]

Radical interpreters must break the holistic circle obtaining between belief and meaning without calling upon either a detailed theory of meaning or a detailed theory of belief for the subject—both of which they are simultaneously trying to develop.

How then can they proceed? As they cannot assign a single propositional attitude to a subject without assigning a host of others it seems that getting radi-

[1] See Malpas 1992, for a discussion of the important differences between Quine's thought experiment and its grounding assumptions and those of Davidson's. As Malpas stresses, in Davidson's hands, "the horizons of translations become much wider […] talk of interpretation rather than translation is a mark of this broadening in conception as much as a of a more semantic emphasis" (1992: 43).

[2] Davidson 1984: 167, 1986a: 315, 1990b: 309.

[3] See also Davidson 1984: 101, 127, 134, 141-42, 144, 146, 153, 156, 186 and Davidson 1986a: 314.

cal interpretation off the ground is a straightforward impossibility. Interpreters need to make some initial assignments prior to having either a developed theory of meaning or belief and yet they cannot make any contentful attributions without such theories. What can be done?

The problem can be resolved if there is "some simple attitude that an interpreter can recognize in an agent" (Davidson 1990b: 322). Of these simple attitudes Davidson writes:

> The assumption that such attitudes can be detected does not beg the question of how we endow the attitudes with content, since a relation, such as holding true, between a speaker and an utterance is an extensional relation which can be known to hold without knowing what the sentence means. I call such attitudes non-individuative, for *although they are psychological in nature, they do not bestow individual propositional contents on the attitudes* (Davidson 1991a: 158, emphasis added).

Elsewhere Davidson tells us that, "certain attitudes toward sentences can be fairly directly inferred [...] From such acts it is possible to infer that the speaker is caused by certain kinds of events to hold a sentence true" (Davidson 1990b: 318). From this humble beginning, anyone in the situation of a radical interpreter would need to carefully observe speech behavior in relation to the environment.

The radical interpreter's method would be to carefully observe a speaker in various situations, over time. Should a consistent and coherent pattern be discerned, the radical interpreter would be able, in principle, to discover any reoccurring structures within the series of utterances. On this basis, it would be possible, in principle, to construct empirical hypotheses about what any given sentence in the other's language means.

Still, locating such patterns—however coherent they are—is not sufficient for assigning contents to another's utterances. As long as it assumed that the constraints imposed by content holism are in play the radical interpreter is still in a predicament. For we must wonder how it is possible to move from finding appropriately robust patterns in another's utterances to assigning content to those utterances.

The way Davidson answers this question reveals what is most important to his vision of mind and language. He maintains that a radical interpreter would have no choice but rely on normative principles of charity if she were to get at the propositional content of another's speech and thought. For a radical interpreter to discover a complex pattern of contentful speech at all she must be making certain important *a priori* assumptions about the other. The basic assumptions a radical interpreter must make are:

> A. The subject is trying to make assertions about certain features of the world.
> B. The subject's assertions are mostly competent and correct (Davidson 1980: 256).

Consequently, the very possibility of contentful interpretation rests on making the charitable assumption "that we can dismiss *a priori* the chance of massive error" (Davidson 1984a: 169). Of course, local errors must be allowed for, since no one is ever perfectly consistent or logical in his or her speech and thought. Showing sensitivity to this fact, one of Davidson's principles of charity bids the radical interpreter to note and forgive occasional errors in trying to make best

overall sense of the other's patterns of responses. Thus, he tells us "it cannot be assumed that speakers never have false beliefs. Error is what gives belief its point".[4]

Going the other way, there are—of course—limits to charity. Should the radical interpreter fail to find a sufficiently robust and coherent pattern in the other's behavior that fits with these charity assumptions she would have to abandon the idea that she is dealing with a thinking, speaking agent at all. Crucially, either the radical interpreter makes charitable assumptions and discerns sufficient consistency in the patterns of behavior of the other or she foregoes the attempt to ascribe propositional content.[5] In the absence of a sufficiently robust pattern of behavior there is no basis for ascribing propositional content or supposing it exists.

These reflections on radical interpretation lead Davidson to endorse the thesis of the autonomy of the mental—a thesis which can be understood in more or less realistic terms. In all versions, the root idea is this: propositional attitudes stand in appropriate kinds of holistically and normative relations. The mental exists if, and only if, the relevant forms of rationality are present. Minds only exist in the space of reasons. This is allegedly why when rationality is missing we must switch to another scheme for understanding an agent's behavior. In such cases a move to non-mental concepts and explanatory schemes becomes necessary precisely because minds, properly understood, are absent.

Mental concepts, for Davidson, are irreducible to the concepts of other discourses, most saliently those of the natural sciences, because of "the normative character of mental concepts" (Davidson 1987: 46). The idea is roughly this: we cannot assign length without a physical framework. Similarly, if Davidson is right about radical interpretation, we cannot ascribe propositional attitudes without a normative, interpretative framework. The mental has special constitutive features. Thus, as long as we conceive of people as rational we cannot operate with a system for ascribing propositional content that can be reduced to a system of descriptions given in, say, the vocabulary of an impersonal scientific discourse. Ascribing propositional content is, for this reason, irredeemably unlike the way in which we understand the behavior of 'mindless' entities (Davidson 1991a: 162-163, see also Davidson 1996).

On the one hand, the thought experiment of radical interpretation is meant to reveal how, for creatures like us, it would be possible in principle to make attributions of content. In this way, contemplating the extreme limit case of the radical interpreter is meant to reveal the essential contours of our actual interpretative practice. Davidson aims to show how the mental can be made accessible by extensional tools and thus made amenable to empirical test. After all, Davidson tells us that "A theory of meaning is […] an empirical theory: its ambition is to account for the workings of a natural language. Like any theory it may

[4] Davidson 1984: 168. Hence, in such cases "The best we can do is cope with error holistically, given his actions, his utterances and his place in the world. About some things we will find him wrong, at the necessary cost of finding him elsewhere right" (Davidson 1986a: 318). Errors must arise against the background of a largely coherent pattern of successful utterances. Hence we can only ascribe error if we assume that the speaker/thinker has largely correct views about the world (cf. Davidson 1980: 221).

[5] Davidson 1984: 152, 159, 197; 1986b: 323, 1986a: 317, 319.

be tested by comparing some of its consequences with the facts" (Davidson 1984: 24). Noting its origins and inspiration, Davidson's project, in this regard, has "naturalistic commitments of a recognizably Quinean kind" (Sinclair 2002: 162).

On the other hand, pulling in a different direction, what the thought experiment of radical interpretation reveals is that there is no possibility of making intelligible the connections between the domain of the mental and the rest of the world using the resources of the natural sciences.

These observations reveal that there are two strands that can be identified in Davidson's thought. One emphasizes features of thought that set it apart from the rest of nature. The other seeks to locate thought within nature. Taken to extremes these different strands in Davidson's thinking come into tension. For example, some exegetes of Davidson hold that our capacity to think contentful thoughts should be regarded as transcendental in a particular sense: namely, that our ability to think such thoughts is a condition on the possibility of having an objective view on the world—a view that cannot be made intelligible within an exclusively scientific image of the world.

Barth (2011) exemplifies. He emphasizes the strand in Davidson's thought that focuses on the autonomy of the mental and sees thought as dependent on language in a way that makes it difficult to square with Davidson's naturalistic agenda. Thus Barth (2011) provides a shining example of an interpreter of Davidson who regards the latter's principle of charity, "as a kind of transcendental principle, and, further, take[s] Davidson's defense of [it] as involving a transcendental proof or argument" (Barth 2011: 174).

Barth (2011) offers a transcendental argument for holding that the general capacity to have thoughts depends on linguistic mastery.[6] He attempts to demonstrate that mastery of language is what makes thought possible. In doing so, he defends a version of what he calls Enabling Ontological Lingualism, advancing an ontological version of a strong dependency claim—namely, that "a subject can only possess thoughts if she also masters a natural language" (Barth 2011: 12). Barth's claim is both universal in scope and conceptual in its modal strength. It is universal in scope because Barth aims to establish that "all thought depends on the mastery of a natural language"; and it is conceptual in character because he holds that "the possession of (propositional and non-propositional) thoughts conceptually depend on a mastery of a natural language".[7] Thus, according to Barth, "we cannot conceive of a subject possessing thoughts without conceiving of her mastering a natural language" (*ibid.*). He dubs the total package of his position Universal Conceptual Lingualism (*ibid.*).

Importantly, even though Barth (2011) advances an a priori conceivability argument that is modally strong, it is not grounded in any of form of conceptual analysis but rather takes the form of a transcendental argument. His argument satisfies two requirements revealing it to be transcendental in character. The first is that it is "an a priori investigation into the conditions of possibility of inten-

[6] In this vein Barth 2011 seeks to improve Davidson's 'belief-argument', described in section 3 below, in order to bring it into "convincing shape" (Barth 2011: 17).

[7] Barth 2011: 13. In saying this Barth 2011 is not endorsing the trivial idea, advanced under the auspices of Local Ontological Lingualism, that having certain kinds of thoughts—say thoughts about atoms or quarks—depends on mastery of sophisticated theoretical discourse, hence mastery language (Barth 2011: 13).

tionality"; the second is that the outcomes of his "investigation are neither synthetic judgements a posteriori gained by empirical research, nor analytic judgements a priori gained by conceptual analysis" (Barth 2011: 15). As a consequence, Barth's reading of Davidson highlights aspects of the latter's thinking that make it seem incompatible with a naturalistic agenda.

Going in the other direction, some authors emphasize a naturalistic strand in Davidson's thinking. Sinclair (2002), for example, argues that the most serious objection to readings of Davidson that exclusively focus on the transcendental elements of his philosophy is that they misconstrue the nature of the divide between the a priori and empirical. According to Sinclair we need to recognize that "Davidson's use of a priori principles maintains a much tighter connection with the empirical by being responsive to empirical facts about us humans" (Sinclair 2002: 175).

Thus Sinclair reminds us that:

> Davidson's interest in what makes interpretation possible can then be captured in this question: what conditions need to be fulfilled so that creatures like us, creatures with a specific evolutionary history, certain inherited, and learned traits, are able to participate in the activity known as interpretation? The principle of charity emerges as an answer to this question, not solely based on a priori considerations but by paying close attention to our nature as biological creatures. 'Necessary' should be read here as necessary for creatures like ourselves, creatures with a certain evolutionary history, and a specific set of sensory modalities and traits that are specific to us (Sinclair 2002: 179).

Underscoring these points, Sinclair (2002) reveals why it is a mistake to forget Davidson's Quinean background. A fundamental Quinean idea is that philosophy and science are continuous. Thus, for Quine, there is no hard and fast distinction—no in-principle barrier between the two. This assumption is strongly linked to Quine's views that no belief is beyond revision and that when deciding which of our beliefs we should revise, we must take our lead from developments in and findings of the natural sciences.[8]

In Davidson's hands we find a remnant of this Quinean naturalistic legacy in that Davidson's use of a priori principles, such as the principle of charity, is "informed by empirical facts about us human creatures" (Sinclair 2002: 171). There is at least this much residual Quinean influence on Davidson's approach.[9]

[8] Sinclair 2002 calls this 'the continuity requirement'. As he describes it, for Quine "there is no independent a priori philosophical perspective that remains insulated from scientific inquiry. To engage in philosophical investigation is to work from within the same understanding of the world provided by science, and to reject the claim that philosophy can justify the results offered by science" (Sinclair 2002: 165).

[9] While we agree with Sinclair that Davidson is a naturalist of sorts, Sinclair 2002 occasionally goes too far and reads too much Quine into Davidson. For example, at one point Sinclair says that, "Davidson's constitutive principles are themselves susceptible to empirical revision, since they are responsive to empirical features of human biological creatures. Empirical discoveries that suggest changes in our understanding of ourselves may then prompt changes to these constitutive principles" (Sinclair 2002: 177). This can make it sound as if there exist principles of charity that we might actively update in the light of empirical findings. Whereas at most what might be said is that things could contingently

Even so, Davidson is a quite different kind of Quinean than those who maintain that all bona fide philosophical questions must be answered, in the end, by testing out empirical hypotheses.[10]

Naturalists such as Fodor insist that we can only use the resources of the special sciences for understanding mental phenomena if those sciences assume the mental operates in a law-like manner and exhibits a nomological dependence on the laws of a more basic science. Davidson famously disagrees. He promotes the view that the mental exists in its own autonomous, anomalous domain and that we neither need nor should expect to find any strict psychophysical laws that will connect that domain to the other sciences. Hence, his "conception of naturalism recognizes a set of rational normative concerns that cannot be addressed within the explanatory interests of natural science" (Sinclair 2002: 180).

Indeed, it is these transcendental aspects of Davidson's thinking that reveal that the questions of interest to him were never wholly empirical. In the end, to make good on Davidson's brand of naturalism, we must see the development of radical interpretation as being "informed by a commitment to a naturalistic view of philosophy, but one that does not look to a unified scientific methodology as the sole model for explanation. This then loosens the constraints on what counts as legitimate explanation, *making room for a kind of inquiry that is not itself part and parcel with natural science*" (Sinclair 2002: 162, emphasis added). Thus, speaking of the role of radical interpretation in Davidson's thinking, Sinclair tells us that "Davidson's insistence that there is an additional question to be pursued here beyond an empirical concern with actual interpretation is not easy to make sense of in naturalist terms" (Sinclair 2002: 171).

To fully explicate the character of this sort of relaxed naturalism would require providing—as Sinclair observes:

> a *characterization of normative phenomena which demonstrates how they can be seen as the product of natural capacities*, capacities that are explained through scientific methods. This is an important aspect of radical interpretation not often emphasized, where our interpretive abilities are depicted as the result of natural capacities, and as being the product of innate and learnt traits. Radical interpretation purports to show how it is possible for us, given such natural capacities, to accomplish our interpretive feats successfully (Sinclair 2002: 178, emphasis added).[11]

change about our interests and practices such that the constitutive principles might alter or cease to apply altogether.

[10] As Davidson says—in reply to Fodor and Lepore's accusations that radical interpretation is empirically refuted—"I do not think I have ever conflated the (empirical) question how we actually go about understanding a speaker with the (philosophical) question what is necessary and sufficient for such understanding. I have focused on the latter question" (Davidson 1994: 3).

[11] Expressing the same point elsewhere Sinclair 2002 emphasizes that, "Davidson uses his model of radical interpretation to highlight these important irreducible features of our intentional vocabulary, features that reflect our interest in viewing others as rational agents. The project is also informed by our empirical conception of ourselves as biological creatures, demonstrating that our view of ourselves as agents cannot be separated from important empirical features concerning the type of creatures we are" (Sinclair 2002: 180).

We believe it is possible to pursue a satisfactory naturalism that is importantly relaxed in these respects and resists the reductive agenda of strict naturalists.[12] We also believe it is possible to address the aforementioned characterization challenge in a way that meets Sinclair's demand. However, we hold that doing so requires adjusting Davidson's thinking in some important respects. Before saying how we propose to pull off the trick (which is the work of section 5) we will first say a bit more about what we find attractive in Davidson's views and in what ways we find his official position to be problematic.

3. What is Right in Davidson's Thought about Thought

Davidson is renowned for holding that being able to think contentful thoughts about an objective world requires mastery of natural language. He defends this position through a mix of philosophical arguments, supported by observations about a range of relevant facts. As he puts it, when asking whether animals are rational creatures capable of propositional thought "the question is not entirely empirical, for there is the philosophical question what evidence is relevant to deciding when a creature has propositional attitudes" (Davidson 1982/2001: 95).

Observations about the holistic nature of propositional thought provide one reason for thinking there is a connection—perhaps a strong one—between thinking such thoughts and the mastery of natural language. Although when discussing the holism of the mental Davidson tends to focus on and privilege beliefs, his observations apply to the content of propositional attitudes quite generally. If content holism is true then what a person thinks about a given topic is constrained, in part, by the content of their other thoughts. Thus the content of my thought that 'Australia is teeming with dangerous flora and fauna' is fixed by other things I think—'Australia is south of Indonesia', 'The box jellyfish is a dangerous animal', 'South is not east', 'Australia is a country', and so on (Davidson 1985: 475; Davidson 1984: 257). The content of any propositional attitude is fixed by such connections and thus exists in a 'logical geography' of contents.

To chase out all the connections of any given propositional content with precision would be to pin down its intensional (with-an-s) content. It is only if we discern the existence of such holistic patterns that we can legitimately ascribe thoughts with propositional content. Yet Davidson argues that the only evidence we find of the existence of such holistic contents is in the fine-grained patterns inherent in speech.[13] For without speech it will not be possible to differentiate between a wealth of possible contents and to justify any particular attribution.

As a matter of fact, we only find the kinds of finely discriminating patterns of behavior—those that would warrant the ascription of contentful thought—in the speech acts of those who have mastered natural language (Quine 1960: 3).

On its own this observation does not, as Barth (2011) recognizes, establish that holistic thought depends on language. At most it establishes that speech acts in natural language provide our best evidence for the existence of contentful

[12] See Hutto and Satne 2015, and Hutto and Satne forthcoming-b.

[13] On the basis of observations about content holism, he claims "it is clear that a very complex pattern of behavior must be observed to justify the attribution of a single thought […] I think there is such a pattern only if the agent has language" (Davidson 1982/2001: 100).

thoughts. As such this observation of Davidson's, even if true, at best only establishes that the 'attribution conditions' for thought depend on natural language (Barth 2011: 55).

The strongest line of argument for the idea that the ability to think contentful thoughts depends on mastery of natural language is found in Davidson's claim that it is necessary to master the concept of belief in order to have a belief, and that it is necessary to master natural language in order to master the concept of belief (Davidson 1982/2001). For Davidson, having the concept of belief is a necessary ingredient for having a contentful perspective on an objective world.

If Davidson is right, the only way of becoming acquainted with the subject-object contrast is to become acquainted with and sensitive to relevant intersubjective standards. There is no other way—no other path—for acquiring the idea that there are other—divergent, contrasting—contentful perspectives on things. Command of the notions of objective truth and error only arise in the context of interpretation: it is in this context that notions of subjective and objective emerge, as it were, simultaneously. In Davidson's words:

> Communication depends, then, on each communicant having, and correctly thinking the other has, the concept of a shared world. But the concept of an intersubjective world is the concept of an objective world, a world about which the communicant can have beliefs (Davidson 1985: 480, 1984: 170, 1990b: 314).

The concept of shared world is a necessary basis for having contentful thoughts and, if Davidson is right, that concept only arises in the context of mastering a language. For him, it is only through learning how to interpret the speech of others that it becomes possible for a creature to adopt a contentful perspective on the world. This is because, he argues, only creatures that are aware of a subjective-objective contrast can ascribe propositional contents to the speech and thought of others, and hence, are able to have contentful attitudes themselves. Putting all of this together we reach the conclusion that mastery of natural language is necessary both for interpreting the contentful utterances of others and having contentful thoughts oneself. This line of thought is repeated in Davidson's remarks about triangulation creating the space needed for error (Davidson 1986a). It is safe to say that it constitutes his basic argument for the dependency of thought on language.

In sum, Davidson thinks mastery of natural language makes propositional thought available for an agent because mastery of natural language requires engaging in special sorts of intersubjective practices—practices that put agents in a position to grasp the notion of having a contentful perspective on an objective world. For all of these reasons, Davidson maintains, to be a believer of propositions requires "the gift of tongues" (Davidson 1985: 473).

Davidson presents his main argument about what is involved in acquiring a sensitivity to the requisite intersubjective standards in terms of mastering a nest of inter-related concepts—of belief; of a shared world; of an intersubjective world; of an objective world and so on. We deem this commitment to be prob-

lematically overly intellectualist. Softer, and more plausible variants of this basic argument are, however, available.[14]

At its simplest, it is possible to reformulate the central idea of Davidson's argument so that it does not require mastery of concepts per se but holds only that creatures capable of contentful thought will have adjusted to the norms of communicating with and interpreting others in language. Natural language is a practice that enables a meeting of minds that generates the right kind of cognitive friction—namely, the kind of cognitive friction that is needed to develop the shared norms that enable speakers to get to grips with the possibility of there being contrasting perspectives on a shared world. A much revised, and less conceptually grounded, version of Davidson's belief argument might be developed to show that having a contentful perspective requires being a creature that is acquainted with the possibility of there being other contentful perspectives on a shared world—perspectives that can be true or false. It is possible to make such adjustments while agreeing with Davidson that mastery of natural language is one way—our way—of coming to be acquainted with the possibility of there being contentful perspectives on a shared world.

Still, even modified in these important respects, on their own, these Davidsonian considerations only succeed in showing that mastery of natural language is sufficient, but not necessary, for understanding and developing a contentful perspective—a perspective that can be right or wrong—on a shared world (Barth 2011: 60). Learning to interpret others by participating in discursive, linguistic practices is at least one way to acquire a contentful view on things: it is one way to acquire the capacity to think thoughts for which the question of truth can arise. Those who master a particular kind of intersubjective practice—one that respects special kinds of norms—can master contentful thinking.

Drawing these threads together, if modified in important respects, there is a version of Davidson's master argument for thinking that contentful thought depends on language that both holds promise and which has the potential to be rendered compatible with the naturalistic strands in his thinking.

4. Challenges to Davidson's Thought about Thought

More work needs to be done if we are to take full advantage of the proposed adjustments to Davidson's dependency claims outlined in the previous section and to show how Davidson's thinking can be rendered fully compatible with a satisfactory naturalism.

There are residual issues to address before the apparent tension can be resolved. The main difficulty is to see how Davidson's conception of the mental as autonomous and holistic can be thought to fit within the natural world. Namely, we need to determine whether—and how—it is possible to make the connections between contentful thought and the rest of the natural world non-mysterious. Davidson was famously skeptical about providing a positive answer to the question of how the mental can emerge in a natural world. He held that

[14] Barth 2011 holds that despite many formidable arguments designed to defeat it, a modified version of Davidson's belief argument can be fashioned that avoids the standard objections and which is promising. See Barth (2011: 53-74) for a detailed discussion of the options and his own reconstruction of Davidson's belief argument.

we lack the descriptive resources needed to give such an account—namely, that we face a characterization problem. He writes:

> The difficulty in describing the emergence of mental phenomena is a conceptual problem: it is the difficulty of describing the early stages in the maturing of reason, the stages that precede the situation in which concepts like intention, belief, and desire have clear application. In both the evolution of thought in the history of mankind and the evolution of thought in an individual, there is a stage at which there is no thought followed by a subsequent stage at which there is thought. To describe the emergence of thought would be to describe the process which leads from the first to the second of these stages. What we lack is a satisfactory vocabulary for describing the intermediate steps (Davidson 1997/2001: 127).

Davidson thinks that we lack the requisite vocabulary because he is committed to the idea—in line with his views on the holism of the mental—that minds can only be discerned and characterized by ascribing propositional contents to them. In his view, "words, like thoughts, have a familiar meaning, a propositional content, only if they occur in a rich context, for such a context is required to give words or thought a location and a meaningful function" (Davidson 1997/2001: 127).

A fortiori, for him, nonverbal thought cannot be characterized because it lacks the necessary links with contentful attitudes—it stands outside of the network of propositional attitudes. For this reason, Davidson doubted that there could be "a sequence of emerging features of the mental […] described in the usual mentalistic vocabulary" (*ibid.*).

As a consequence of this lack of vocabulary—this characterization problem—Davidson thinks we are without the resources for making sense of the connections between contentful attitudes and the rest of nature. In his way of setting things out, the characterization problem leads to a connection problem, which in turn generates an explanatory continuity problem as a special instance.

In the end, Davidson sees no way to draw intelligible connections between our capacity for contentful thought and the cognitive capacities of our younger selves and our immediate evolutionary ancestors. Anyone convinced of Davidson's package of views about holism will see necessary links between thought, talk and interpretation which imply that there is no way of making intelligible or explaining the natural origins of content. Such an approach renders mysterious the ontogenetic and phylogenetic history and development of propositional forms of thought.

These lines of reasoning explain why Davidson was not interested in empirical speculations about, investigations into, or attempts to explain how our capacity to think propositional thoughts actually arose in ontogeny and arises in phylogeny. As he says:

> The approach to the problems of meaning, belief, and desire which I have outlined is not, I am sure it is clear, meant to throw any direct light on how in real life we come to understand each other, nor how we master our first concepts and our first language (Davidson 1990b: 325).

Davidson, on the one hand, recognizes that contentful perspectives arose and arise in the world and yet holds, on the other hand, that we are conceptually debarred from explaining how this could be so. Pointing to this combination of

views, McDowell observes that Davidson's position "smacks of magic" (McDowell 1998: 410). Is it possible to keep what is best in Davidson's work while avoiding this charge? Is it possible to demystify Davidson?

5. A Radically Enactive Answer

Davidson thinks that we lack the requisite vocabulary for describing the stages that precede the emergence of thought. This is because he holds fast to the idea that minds can only be discerned and characterized by ascribing propositional contents to them. For him, the problem boils down to this:

> We have many vocabularies for describing nature when we regard it as mindless, and we have a mentalistic vocabulary for describing thought and intentional action: what we lack is a way of describing what is in between. This is particularly evident when we speak of the 'intentions' and 'desires' of simple animals. We have no better way to explain what they do (Davidson 1997/2001: 128).

We think the characterization problem, as Davidson presents it, can be dissolved. Our diagnosis of how to achieve this is that it requires relaxing the condition on how we discern and characterize minds. This proves pivotal, since once the characterization problem is dealt with—once we clarify why it is not a problem—it becomes clear that there is no conceptual barrier that prevents us from dealing adequately with the connection and continuity problems.

The first step is to dissolve the characterization problem as Davidson sets it up. We think that can be achieved by expanding and enriching our ways of thinking about the mental so as to include recognition of the world-directed, intentional attitudes that lack fine-grained content—indeed, that lack any kind of content whatsoever.

For some the very idea of contentless intentional attitudes is a nonsense—it seems a conceptual impossibility. Such resistance is to be expected from anyone who holds that intentionality—whatever form it may take—necessarily entails content. If we combine Davidsonian observations about the dependency of contentful thought on language with the idea that all forms of intentionality are necessarily contentful, then we reach the strong conclusion that "*all* thought depends on the mastery of a natural language" (Barth 2011: 13, emphasis added); namely, that "the possession of (propositional and non-propositional) thoughts conceptually depend on a mastery of a natural language".[15]

> Propositional as well as non-propositional thoughts are intentional in that they are of or about something. What they are of or about is their intentional object […] *The contents of thoughts have a representational dimension in virtue of being intentional*. They represent objects as being so-and-so in virtue of referring to objects and in virtue of characterizing these objects under some aspects (Barth 2011: 9, emphasis added).

[15] *Ibid.* Or again, as Barth elsewhere puts it, "the possibility of both propositional and non-propositional thoughts depends on language" (Barth 2011: 8). Barth distinguishes propositional and non-propositional thought in the following way: "Propositional thoughts do not only exhibit a representational dimension but also an inferential dimension […] Non-propositional thoughts are not inferentially significant" (Barth 2011: 9-10).

Every form of thought exhibits intentionality and hence is representationally contentful, according to Barth. If we accept Barth's conceptual stricture then the very idea of a contentless intentional attitude is a non-starter and the characterization problem stands.

To escape this trap we need to show that we can make sense of the idea of contentless intentional attitudes and that we can understand their characteristics. Picking up on Davidson's comment about non-verbal animals, it helps to focus attention on the much-discussed example of Malcolm's barking dog (Malcolm 1997: 49-50). How should we characterize the mind of the dog that finds itself in the following circumstances: The dog sees a cat. It gives chase. The cat leaps into a tree. The dog circles around the base of the tree, barking. Yet unbeknownst to the dog, the cat slips away. The dog continues to bark.

We are naturally inclined to say that the dog believes that the cat is up the tree and that it wants to get at the cat. Yet, as Davidson cautions when discussing this very case, there is not enough in the totality of the dog's patterns of behavior to justify ascribing it any contentful attitudes (Davidson 1982/2001: 97-100). The trouble is that "it does not seem possible to distinguish between quite different things that the dog might be said to believe" (Davidson 1982/2001: 97). For example, we lack grounds for ascribing the concepts 'tree' or 'cat' to the dog as opposed to a multitude of other possible concepts.

Maybe the dog is not thinking about the cat as a cat. Maybe it is operating with a more general concept of 'animal'. Or perhaps it is thinking that there is 'something chaseable' in the tree. Or it might be having countless other possible thoughts on the topic of its quarry. What sort of mistake the dog makes, if any, depends on the precise content of its thoughts, but we have no principled way of determining what those putative contents might be, or indeed if there are any such contents in play. The crux is that, "We want to say the dog believes something—but we do not seem able to say what" (Armstrong 1973: 25; see also Stich 1979: 18).

Taking everything into account about the full repertoire of the dog's behavior, Davidson's lesson is that we lack evidence for assigning it any particular set of contentful attitudes and, thus, we lack any justification for supposing that it has any such attitudes.

Our difficulty in assigning any content to the dog's thoughts in this case reveals that we have no reliable way of characterizing its state of mind in mentalistic terms as long as we restrict ourselves to using the machinery of the sort that would be available to a Davidsonian radical interpreter.

One way of going beyond the resources of radical interpretation would be to bet that the notion of content will be vindicated and shown to be part of the theoretical vocabulary of mature sciences of the mind. Should that prove true, then we might rely on such sciences, as opposed to our interpretative practices, to make well-grounded assignments of contents to non-verbal states of mind. In that case, we could join with Carruthers in saying that although we "find ourselves forced, implausibly, to describe animal and infant thoughts using adult human concepts and categories, this is our problem, not theirs" (Carruthers 1998: 220). Those who assume that the notion of content will feature in the mature sciences of the mind thereby have a basis for remaining faithful to the idea that all thought must be contentful. They can hold that if non-verbals have intentional attitudes then these attitudes must be contentful attitudes, even if we have difficulty knowing which contents to ascribe to them using our everyday

resources. The issue is highly contentious. Nevertheless, there are many reasons for doubting that the notion of content will feature in the mature sciences of the mind (see Hutto and Myin 2013 and 2017 for detailed discussion).

Alternatively, we might not restrict ourselves to using only the resources available to a radical interpreter. We might come at the issue from a different angle—saying instead that the dog has attitudes that are directed towards the cat and the tree without assuming that such attitudes are contentful. If we can make sense of the idea of contentless intentional attitudes, then we can avoid the intractable problem of trying to characterize, *per impossibile*, the content of such attitudes. Crucially, to accept that the dog has intentional attitudes that are not propositional attitudes absolves us of trying to specify the content of the dog's attitudes. This is good news because, as we have seen, the dog's behavior does not exhibit a pattern that would warrant the ascription of content.

Nevertheless, in chasing the cat up the tree the dog still exhibits a complex pattern of behavior that exemplifies a world-directed mentality—even if the dog's intentional attitudes are contentless there can be rich connections between what the dog thinks, feels, intends and desires. Our awkward attempts to assign content to the dog's attitudes can be understood as a way of picking out which aspects of the situation that the dog is directed at non-contentfully. We can say of the dog—and other creatures of a similar mindset—that it is directed at the situation—and so qualifies as having intentional attitudes—even though such attitudes are not contentful (see Hutto 2008). The idea that the most basic kinds of mentality are world-directed yet contentless is the driving idea behind radical enactivism (Hutto and Myin 2013, 2017).

Why believe in such non-contentful yet intentional attitudes? As Davidson himself implicitly and uneasily acknowledges when introducing the characterization problem, in certain circumstances we need to make sense of the attitudes of non-verbal animals, infants and adult humans even though we lack any justification for ascribing them contentful attitudes.

On a more positive note, once unshackled from the restricting idea that all intentionality must be contentful, we can find plenty of examples of non-verbal mentality that cry out to be understood and characterized by making comparisons—noting similarities and differences—with our most basic and more sophisticated ways of thinking about the world. Wittgenstein chides those who assume that animals are incapable of thought, merely because they cannot talk, along these lines. Challenging this assumption, he stresses: "[T]hey simply do not talk. Or to put it better: they do not use language—if we except the most primitive forms of language" (Wittgenstein 1953, §25).

It is important to be clear that, in saying this, we are not offering a straight solution to Davidson's characterization problem but rather showing how—by appealing to a richer conception of the mental—the characterization problem *as he poses it* can be defused.[16]

[16] There are other versions of the characterization problem that should also be avoided. In particular, it is important not to construe it as a "missing link" problem, the solution to which is supposed to consist in finding intermediate steps. We agree with Sultanescu (2015) that seeking to solve Davidson's characterization problem is a fool's errand if doing so requires being able to positively characterize each stage of thought, from the inside. For if that were necessary for solving the problem then, a bit like Zeno's paradox,

Bar-On (2013) makes a similar move. She attempts to address Davidson's philosophical challenge of explicating the relevant connections and continuities head on by first addressing the characterization problem. She aims to achieve the latter by making appeal to expressive attitudes in order to characterize non-propositional states of mind. As she explains, "although these are mentalistic descriptions, which do not carve behavior in purely causal terms, they do not presuppose the full battery of concepts that inform our descriptions of each other" (Bar-On 2013: 329). Thus she aims to show that there are commonsense descriptions of the expressive behavior available that "can guide us towards a natural intermediate stage in a diachronic path connecting the completely unminded parts of the animal world with the fully minded, linguistically infused parts that we humans now occupy" (Bar-On 2013: 330).

However, she does not sufficiently disentangle the characterization problem, connection and continuity problems. This failure leads her to misrepresent what needs to be done in order to deal with the latter problems. Hence, Bar-On holds dealing with the latter problems requires doing the conceptual work of fusing "the scientific image and the naive commonsense image" (Bar-On 2013: 329). Although we agree that all these problems are related, we do not think that solving the connection and continuity problems requires the kind of fusion that Bar-On describes.

We should not conceive of solving the connection and continuity problems as requiring the fusing of the two images. Rather such problems and mysteries can be dealt with by making illuminating connections between relevant domains of discourse. In this latter vein, we propose a different way of showing how there can be, as Bar-On (2013) puts it, a "*scientific* account of the emergence of our mental states and the sort of communication they underwrite" by providing a "legitimate *philosophical characterization* of such a progression".[17]

In sum, we can avoid having to solve the characterization problem in Davidson's terms if we recognize the possibility of there being intentional attitudes that lack content. Upon doing so, it becomes possible to see how to overcome the connection problem, and its more specific instantiation of phylogenetic and ontogenetic continuity problems (for further details on how to deal with these latter problems see Hutto and Satne, forthcoming-b).

To understand our preferred way of dealing with the characterization problem it is important to note that contentless intentionality is not supposed to characterize an intermediate evolutionary stage that sits between contentful

we could replay the worry at every micro-step of the process with the result that "the intermediate steps between primitive intentionality and contentful intentionality cannot in fact fully be accounted for" (Sultanescu 2015: 639). Accordingly, however much we might succeed in narrowing the imaginative gap there would be no way to close it completely. Thus even if expressive or intentional attitudes are allowed into the story, if they are used to fill in the "intermediate steps" between contentful and non-contentful attitudes, we can always ask how exactly the gap between such attitudes and "contentful goings-on is supposed to be bridged" (Sultanescu 2015: 646).

[17] Bar-On 2013: 303, emphases added. Accounting for the emergence of the mental requires working under the auspices of Relaxed Naturalism (as we argue in Hutto and Satne 2015, Hutto and Satne, forthcoming-b). For Relaxed Naturalists philosophy provides the structural steps of the story while the human, social and natural sciences are called in to fill in the details.

thinking and non-intentional behavior. On the contrary, contentless intentionality works as a platform that enables the emergence of complex practices—namely discursive practices that provide the special resources for bringing content into being in the natural world.[18]

Distinguishing contentless Ur-intentionality from contentful intentionality enables us to understand how it is possible for interpreter and interpreted to triangulate, to target and be directed at the same focal points, and to do so in similar ways even though they may lack thoughts with propositional or other content.

In addition, broadening our understanding of the varieties of intentionality in this way opens the door to giving an account of the directedness of contentless intentional attitudes in biosemiotic terms. Such an account of basic intentionality is wholly compatible with the possibility that having contentful attitudes depends, as a matter of fact, on mastering special kinds of linguistic practice (Hutto and Myin 2013: ch. 4; 2017: ch. 5; Hutto and Satne 2015). By availing ourselves of a distinction between contentless and contentful forms of intentionality, it becomes possible to connect Davidson's vision of the mental with explanations in the sciences of the mind.

To illustrate, consider the role Davidson suggests that triangulation plays in the primitive learning situation. In a number of places he emphasizes the importance of learners and teachers exhibiting a similarity of response to similar objects or features of the world. In triangulating, he assumes that there must be a commonality to what the learner and teacher target—what they naturally group together (Davidson 1992: 264). At its simplest, our suggestion is that it is no accident that learners and teachers are capable of such acts of joint attention. We can make sense of the attitudes involved in such primitive feats of triangulation by understanding them as contentless but world-involving intentional attitudes—attitudes that can be primarily understood as a gift of our biological heritage. This proposal is perfectly in tune with the naturalistic strand in Davidson's thought—the one that emphasizes that it is "because of the way we are constructed (evolution has something to do with this), that we find these responses natural and easy to class together" (Davidson 1991b: 200).

6. Conclusion

We contend that it is, in the final analysis, possible to show how content could have arisen in the natural world without gaps. This can be achieved without having to attempt the impossible—namely, without having to solve the characterization problem in Davidson's terms. That would require imagining the content of a missing mental link; a strange centaur; an intermediate state of mind that sits somewhere between purely intentional attitudes and properly contentful attitudes. Instead of providing such a contentful characterization, we propose expanding our thinking about varieties of intentionality and thus making it possible to defend some of Davidson's important insights about minds— specifically, modified versions of the four claims set out in section 1—while also promoting a satisfactory and demystifying naturalism.

[18] We have developed this kinky account of cognition more fully in Hutto and Satne (2017) and Hutto and Myin (2017).

References

Armstrong, D.M. 1973, *Belief, truth and knowledge*, Cambridge: Cambridge University Press.

Bar-On, D. 2013. "Expressive communication and continuity scepticism", *Journal of Philosophy*, 110, 6, 293-330.

Barth, C. 2011, *Objectivity and the language-dependence of thought*, London: Routledge.

Carruthers, P., 1998, *Language, thought and consciousness*, Cambridge: Cambridge University Press

Cutrofello, A. 1999, "The transcendental pretensions of the principle of charity", in L. Hahn (ed.), *The philosophy of Donald Davidson*, La Salle: Open Court, 333-41.

Davidson, D. 1980, *Essays on actions and events*, Oxford: Clarendon Press.

Davidson, D. 1982, "Rational animals", in Davidson 2001.

Davidson, D. 1984, *Inquiries into truth and interpretation*, Oxford: Clarendon Press.

Davidson, D. 1985, "Rational animals", in Lepore, E. and MacLaughlin, B. (eds.), *Actions and events: Perspectives on the philosophy of Donald Davidson*, Oxford: Blackwell.

Davidson, D. 1986a. "A coherence theory of truth and knowledge", in Lepore, E. (ed.), *Truth and interpretation: Perspectives on the philosophy of Donald Davidson*, Oxford: Basil Blackwell.

Davidson, D. 1986b, "Empirical content", in Lepore, E. (ed.), *Truth and interpretation: Perspectives on the philosophy of Donald Davidson*, Oxford: Basil Blackwell.

Davidson, D. 1987, "Problems in the explanation of action", in *Metaphysics and morality*. Oxford: Blackwell, 34-49.

Davidson, D. 1990a, "Afterthoughts, 1987", in Lepore, E. (ed.), *Reading Rorty: Critical responses to philosophy and the mirror of nature*, Oxford: Blackwell.

Davidson, D. 1990b, "The structure and content of truth", *Journal of Philosophy*, 88, 6.

Davidson, D. 1991a, "Three varieties of knowledge", in Griffiths, P. (ed.), *A.J. Ayer: Memorial essays*, Cambridge: Cambridge University Press, 153-66.

Davidson, D. 1991b, "Epistemology externalized", *Dialectica*, 45, 2-3, 191-202.

Davidson, D. 1992, "The second person", in *Midwest Studies in Philosophy*, 17, 255-67.

Davidson, D. 1994, "Radical interpretation interpreted", *Philosophical Perspectives*, 8, 121-28.

Davidson, D. 2001, *Subjective, intersubjective, objective*, Oxford: Clarendon Press.

Davidson D. 1996, "Subjective, intersubjective, objective", in Coates, P. and Hutto, D. (eds.), *Current issues in idealism*, Bristol: Thoemmes Press.

Davidson, D. 1997, "The emergence of thought", in Davidson 2001.

Davidson, D. 1999, "The emergence of thought", *Erkenntnis* 51, 1, 7-17.

Genova, A.C. 1999, "The very idea of massive truth", in L. Hahn (ed.), *The philosophy of Donald Davidson*, La Salle: Open Court, 167-91.

Hutto, D.D. 2008, *Folk psychological narratives: The sociocultural basis of understanding reasons*, Cambridge, MA: MIT Press.

Hutto, D.D. and Myin, E. 2013, *Radicalizing enactivism: Basic minds without content*, Cambridge, MA: MIT Press

Hutto, D.D. and Myin, E., 2017, *Evolving enactivism: Basic minds meet content*, Cambridge, MA: MIT Press.

Hutto, D.D. and Satne, G. 2015, "The natural origins of content", *Philosophia*, 43, 3, 521-36.

Hutto, D.D. and Satne, G. 2017, "Continuity scepticism in doubt: A radically enactive take", in Durt, C., Fuchs, T. and Tewes, C. (eds.), *Embodiment, enaction, and culture*, Cambridge, MA: MIT Press. 107-128.

Hutto, D.D. and Satne, G. forthcoming-b, "Naturalism in the Goldilock's zone: Wittgenstein's delicate balancing act", in Raleigh, T. (ed.), *Wittgenstein, philosophy of mind and naturalism*, London: Routledge.

Maker, W. 1991, "Davidson's transcendental arguments", *Philosophy and Phenomenological Research*, 51, 345-60.

Malcolm, N. 1977, "Thoughtless brutes", in his *Thought and knowledge*, New York: Cornell University Press.

Malpas, J.E. 1992, *Donald Davidson and the mirror of meaning*, Cambridge: Cambridge University Press.

McDowell, J. 1998, "Reply to commentators", *Philosophy and Phenomenological Research*. 58, 2, 403-31.

Quine, W.V.O. 1960, *Word and object*, Cambridge, MA: MIT Press.

Sinclair, R. 2002, "What is radical interpretation? Davidson, Fodor, and the naturalization of philosophy", *Inquiry*, 45, 2, 161-84.

Stich, S. 1979, "Do animals have beliefs?", *The Australian Journal of Philosophy*, 57.

Sultanescu, O. 2015, "Bridging the gap: A reply to Hutto and Satne", *Philosophia* 43, 639-49.

Wittgenstein, L. 1953, *Philosophical investigations*, Oxford: Basil Blackwell.

Davidson's Semantic Externalism: From Radical Interpretation to Triangulation

Claudine Verheggen

York University

Abstract

The received interpretation of Donald Davidson's philosophy has it that his thoughts underwent a significant change between his early work and his later work, in particular, between his work on radical interpretation and his work on triangulation. It is maintained that the kind of semantic externalism Davidson advocated in his later work is importantly different from that advocated in the early work. Indeed, it is sometimes even maintained that his semantic externalism emerged only, roughly, in his later work. I argue that Davidson's semantic externalism has always been not only holistic and historical, but also social and non-reductionist. His work on triangulation, by supplementing the early work, reinforces these earlier conclusions and vindicates some of his early assumptions, in particular, his claims that language and thought are essentially public and that their possession requires having the concept of objectivity. I end the paper by articulating what I take to be the most significant differences between Davidson's version of externalism and more orthodox versions.

Keywords: Radical interpretation, triangulation, semantic externalism, meaning.

1. Introduction

In the last forty years of his life, Donald Davidson developed a highly distinctive version of semantic externalism, which has been largely unrecognized as such, and which has important consequences for his philosophy unrecognized by Davidson himself. The main purpose of this paper is to correct these lacunae.

Semantic externalism can be either physical or social, or both. According to physical externalism, the meanings of utterances and the contents of thoughts are determined in part by factors belonging to the physical environment of speakers and thinkers. According to social externalism, they are determined in part by factors belonging to the social environment of speakers and thinkers. Davidson advocates both kinds of semantic externalism (externalism for short in what follows). What makes his view distinctive is that, according to him, the physical side of externalism can be secured only through the social side. This, to

Argumenta 3, 1 (2017): 145-161
ISSN 2465-2334

DOI 10.14275/2465-2334/20175.VER

begin with, makes room for a unique version of physical or, as Davidson calls it, perceptual externalism. But the social side is itself unorthodox in that it is not a version of the community view most frequently propounded by social externalists.

Though Davidson's version of externalism came to be fully developed in his work on triangulation, it does have its seeds in his work on radical interpretation. As he himself writes in 1991, he "has for some thirty years been insisting that the contents of our earliest learned and most basic sentences ('Mama', 'Doggie', 'Red', 'Fire', 'Gavagai') must be determined by what it is in the world that causes us to hold them true".[1] And earlier on: "The causality plays an indispensable role in determining the content of what we say and believe. This is a fact we can be led to recognize by taking up…the interpreter's point of view".[2] And towards the end of his life: "what a speaker means by what he says, and hence the thoughts that can be expressed in language, are not accidentally connected with what a competent interpreter can make of them, and this a powerfully externalist thesis".[3] However, that reflecting on radical interpretation yields externalism has seldom been emphasized by his commentators,[4] and Davidson's writings about triangulation have not received the attention they deserve, in part, I think, because they have been deeply misunderstood—a secondary aim of this paper is to correct this, too.[5] In fact, the argument for externalism can be seen as coming in two steps, one provided by the considerations of radical interpretation and the other by the considerations of triangulation. Thus, the work on radical interpretation establishes the broad externalist claim according to which the causes of speakers' basic utterances, such as "There is a cow", play a crucial role in determining their meaning (and the causes of their basic (propositional) thoughts play a crucial role in determining their content—from now on, for simplicity's sake, I shall focus on language). The work on triangulation answers the further question how the relevant causes are isolated as the determinants of meaning. The answer to this question reveals the indispensable role of the social.

I start by reviewing Davidson's motivations for reflecting on radical interpretation, the assumptions it relies on, and its procedure. Next I present the triangulation argument, focusing on how reflections on triangulation supplement reflections on radical interpretation. I argue that, with the work on triangulation, some of Davidson's earlier conclusions are being reinforced and some of his early assumptions, in particular, his claims that language is essentially public and that its possession requires having the concept of objectivity, are being vindicated. I end by articulating what I take to be the most significant differences between Davidson's version of externalism and more orthodox versions.

[1] Davidson 1991a: 200.
[2] Davidson 1983: 150.
[3] Davidson 2001b: 11.
[4] Peter Pagin, e.g., maintains that it emerges in the early 1980s (Pagin 2013: 235). And there is no recognition of Davidson as an externalist in Burge's 1992 "state of the art" article, nor, more recently, in Haujioka 2017.
[5] Of course many have recognized the externalism yielded by triangulation—see Bridges 2006, Amoretti 2007 and 2013, Bernecker 2013, among others.

2. Radical Interpretation

As Davidson makes clear in his introduction to *Inquiries into Truth and Interpretation*, his goal in engaging in the radical interpretation thought-experiment is to answer the question, "What is it for words to mean what they do?".[6] He thinks that the best way to answer it is by considering what it would take to understand a foreign speaker from scratch, that is, without any knowledge of what her words mean or any detailed knowledge of her propositional attitudes (as detailed knowledge of these would require knowledge of her language as well), and of course without the benefit of bilingual intermediaries or dictionaries. This approach is non-question-begging in so far as it does not involve at the start the notion that needs to be explained. It does, however, make an important assumption about meaning, namely, that it is essentially public, which Davidson always urged: "[t]he semantic features of language are public features".[7] "There can be no more to meaning than an adequately equipped person can learn and observe".[8] It must also be stressed that reflecting on radical interpretation, precisely because it is designed to answer the question what it is for words to mean what they do "in a philosophically instructive way",[9] is not supposed to tell us simply how meanings can be attributed to speakers, but also, and more importantly, how meanings are determined or constituted. The ultimate goal is not the semantic theory—a description of a speaker's meanings—that doing radical interpretation is supposed to yield but the meta-semantic or foundational theory that emerges from reflecting about radical interpretation; it is to illuminate philosophically the nature of meaning by telling us how meanings are determined or constituted.

As is well-known, Davidson argues that we would have an answer to the question what it is for words to mean what they do if we reflected on how to construct a Tarski-style theory of truth for a speaker. For, if properly constructed, such a theory would enable us to understand any utterance of a speaker by giving us, for any such utterance, its truth-conditions. These could be derived from a final set of axioms that would tell us, for every primitive semantic expression and every rule of combination, how they contribute to the truth-conditions, and hence to the meaning, of any utterance in which they occur. Davidson himself came to acknowledge that the prospects of an all-encompassing theory along these lines are dim, wondering "whether, or to what extent, such theories can be made adequate to natural languages".[10] At the same time, however, he never doubted that "they are adequate to powerful parts of natural languages".[11] Note also that he did not think that "speakers and interpreters actually formulate such theories", but that, "if we can describe how they *could* formulate them, we will gain an important insight into the nature of the intentional (including, of course, meaning)".[12] Moreover, an answer to the ques-

[6] Davidson 1984: xv.
[7] Davidson 1979: 235.
[8] Davidson 1990b: 62.
[9] Davidson 1984: xv.
[10] Davidson 1993: 83.
[11] *Ibid.*
[12] Davidson 1993: 84.

tion "how a competent interpreter might come to understand the speaker of an alien tongue" not only should "reveal important features of communication", but also "throw indirect light on what makes possible a first entry into language".[13] Therefore, despite the limits of the theory, there may still be a lot to learn from reflecting on how a theory of truth and interpretation could be constructed.

The basic idea is for the radical interpreter initially to connect utterances of sentences held true by the speaker with observable events in their shared surroundings. To use held true sentences as part of the primary evidence on which to build a theory of interpretation is non-question-begging, since Davidson thinks that held true sentences can be detected independently of knowing either their meaning or the belief they express. The next step is to employ the principle of charity and to assume that speaker and interpreter share many of their beliefs, focusing, in the first instance, on beliefs about current events around them. Thus it is, tentatively at least, to interpret utterances of sentences held true as expressing the beliefs the radical interpreter has herself formed in the circumstances, e.g., the belief that there is a rabbit in their vicinity or the belief that it is snowing around them. Of course, single occasions of utterance will not do to finalize an interpretation. Initially it could be thought that the foreign speaker's utterance of "Gavagai" means there goes something furry, or there goes something four-legged, or there goes something cute, rather than there goes a rabbit. It is only when the expression will have been used in numerous other circumstances, and contrasted with numerous other expressions, that its meaning may become settled. In fact, Davidson thinks that it is only when the interpreter knows all the axioms of the theory for a speaker that she may be able to understand the speaker's language.[14] And knowing an entire theory will require that the interpreter expand the principle of charity and make further assumptions concerning the speaker, such that by and large her beliefs are rational, i.e., true, justified and consistent, and her desires and patterns of preferences reasonable and coherent. As Davidson came to insist, however, "charity prompts the interpreter to maximize the intelligibility of the speaker, not sameness of belief",[15] nor, I should add, rationality.

For the purposes of this paper, knowing the bare bones of the process of radical interpretation should suffice.[16] One thing to emphasize at this stage is that the thought-experiment reveals the thoroughly holistic aspect of meaning.[17] Recall that what we learn about how a radical interpreter attributes meanings to a speaker's utterances also teaches us how meanings are constituted. Thus, since meanings can be attributed only holistically, they are also constituted holistically. Indeed, meaning is holistic in that the meaning of an expression depends on how it is used in many different circumstances in connection with many other expressions. To put it in another way, the meaning of an expression depends on the many beliefs one has about what the expression is about. It is not the case, though, that all the uses or beliefs are relevant to determining its meaning—this

[13] Davidson 1991b: 210.
[14] Davidson 1973: 139.
[15] Davidson 1984: xix.
[16] For details, see especially Davidson 1973, 1974, 1975, 1976, 1990a and 1993.
[17] Davidson also takes it to be a presupposition. See Davidson 1984: xv.

would have the odd result that any change of belief results in a change of meaning. Thus the belief that rabbits are four-legged animals may be central to what 'rabbit' means for a speaker. But the belief that rabbits make for delicious dinners may not be so.[18] There is, for Davidson, no principled way to draw the line between those beliefs that are essential to the meaning of a word and those that are "merely" beliefs about the referent or extension of the word, apart from saying that the latter may come and go whereas a speaker would be more reluctant to relinquish any of the former.

Another noteworthy consequence of reflecting on radical interpretation in order to illuminate philosophically the nature of meaning is that the theory the radical interpreter comes up with cannot be stated in reductionist terms, that is, without saying what it is that the speaker means by her words. Recall that the truth-conditions that can be derived from the axioms of the theory are to be meaning-giving. The goal of the radical interpreter is to match every simple expression of the speaker's repertoire with one of her own. Initially this is accomplished by looking for something in the world the expression refers to or designates. This is why the theory is in the first instance extensional—meanings are given in terms of items in the world that they are about. But the theory cannot be purely extensional, on pain of providing interpretations that cannot explain the speaker's behaviour, linguistic or otherwise. A purely extensional theory would not distinguish between 'Hesperus' and 'Phosphorus' as that which matches a speaker's utterance of 'Hesperus'. The same entity is being picked out by each name. But the failure to distinguish between the names would make it impossible to explain why the speaker is thrilled to discover that Hesperus is identical with Phosphorus. If the names contributed in the same way to the truth-conditions of the speaker's utterances, and hence to their meanings, there would be no room for thrill—the speaker would have known all along that Hesperus and Phosphorus are identical. But this is to say that, in stating the axioms of the theory, the radical interpreter must be careful to pick out the entities referred to or designated in a way that reflects how the speaker thinks about them, which is another way of saying that the interpreter must pay attention to many of the beliefs a speaker expresses with a given expression.

Meaning is also obviously externalist, since what speakers mean by their basic utterances necessarily depends, at least in part, on what in the world around them cause them to produce those utterances. This is clearly brought out by the radical interpretation thought-experiment, the lessons of which, again, concern not just the attribution of meaning but also its constitution. The radical interpreter has no choice but to take the objects and events around them that cause the speaker's utterances to be determining, at least in part, their meaning as well as the content of the beliefs the interpreter herself takes the utterances to be expressing. This becomes all the more obvious when an interpreter's initial assignment of meaning turns out to be untenable, forcing the interpreter to probe the speaker and to scrutinize their common surroundings to arrive at the proper interpretation. I take this externalism to be the most fundamental significance of the claim that meaning is truth-conditional, as Davidson understands

[18] Davidson writes: "we have no [firm] way of distinguishing between the relations [among beliefs] that define the state of mind (or the meaning of an utterance) and those that are 'merely' contingent, and so do not touch content" (Davidson 1995: 15).

truth-conditions. What partly determines the meaning of basic utterances is what in the world makes them true. That is, it is the obtaining of the truth-conditions of basic utterances that partly determines their meaning. This remains the case even if the theory of truth cannot apply to every part of language, even if there are expressions, or grammatical structures, that cannot fit the truth-conditional mould. The core of the theory that provides us with the meanings of a speaker's utterances is truth-conditional. Indeed this is what makes it possible for the radical interpreter to get started.

The answer so far to the question what it is for words to mean what they do is for them to be used, in the first instance, in such ways that they can be understood by a radical interpreter; thus it is for at least some words to be used to refer to features of the environment shared by speaker and interpreter and which have caused them both to respond in certain ways. Therefore reflecting on radical interpretation establishes perceptual externalism in its broad terms: as Davidson puts it before claiming he had advocated the view for three decades, "the contents of our thoughts and sayings are partly determined by the history of causal interactions with the environment".[19] What Davidson did not do, however, when reflecting on radical interpretation, is spell out how the causes of speakers' basic utterances are singled out. This is done in his writings on triangulation.

3. Triangulation

The conclusion of the triangulation argument is that only someone who has interacted linguistically with another person and the world they share, that is, only someone who has triangulated linguistically, could have a language and thoughts.[20] As Davidson makes clear, the argument is initially premised on perceptual externalism, which he expresses in this context as "what determines the content of …basic thoughts (and what we mean by the words we use to express them) is what has typically caused similar thoughts".[21] This is a more precise formulation of the broad externalist thesis, for what is introduced here is the idea of the meaning-determining causes being *typical*. This is as it should be, for we would not want to say that, even for every basic utterance a speaker produces, something in her environment is currently causing it, or that, for every basic utterance which is currently being caused by something in the speaker's environment, it has to be true. (Also, no externalist claims that, for every basic word a speaker uses, she must have at some point been responding to what in the world caused her to use the word in the way she does.) But this brings into focus a problem that was not addressed before: what are the typical causes of speakers' basic utterances?

At first blush, it might seem that this is an easy question to answer. The typical causes are those that a speaker is responding to when her utterances are sincere and the conditions of perception are good, such that, say, a table, or a rabbit, is clearly visible in the vicinity of the speaker and the speaker herself has

[19] Davidson 1991a: 200. Note that this also indicates that Davidson took his early version of externalism to be historical, contra what some commentators have thought, e.g., Lepore and Ludwig 2005: 337. For discussion see Myers and Verheggen 2016: 68-71.
[20] The circularity of the claim will be addressed in due course.
[21] Davidson 1991a: 201.

good eyesight and is not under the influence of, say, drugs or alcohol. After all, it might be added, why have other externalist theorists not asked that question? Thus Tyler Burge, for instance, simply declares that "[t]he natures of such states [thoughts about water] are determined partly by normal relations between the person…and the environment".[22] And Hilary Putnam simply assumes that the meaning of 'water' is determined by the liquid around us that, in normal circumstances, causes us to use the word.[23] But note here that I have just introduced another word—'liquid'—in order to pick out the typical cause of uttering 'water' and what in the world determines, at least in part, the meaning of 'water'. This points us towards the problem adumbrated by Davidson: how are we to decide that it is tables or rabbits rather than colours or materials or surfaces or shapes or chunks of the world surrounding tables or rabbits that are the typical causes of a speaker uttering 'table' or 'rabbit'? Of course, we will be helped to do so if we can ascertain that the speaker who utters 'table' or rabbit' is talking about a piece of furniture or an animal, but how do we ascertain that? What are the typical causes of a speaker uttering 'piece of furniture' or 'animal'?[24] To be sure, if the typical causes of some utterances are already fixed, we may be able to use these utterances to fix the typical causes of other utterances. But how do any typical causes get fixed to begin with?

Davidson maintains that the question what the typical causes of her basic utterances are is not a question a person who has never interacted with others, a lifelong solitary person, could answer. Consequently, it is not just that a solitary person could not know what she means, but that there is nothing that she could mean by her utterances.[25] As Davidson stresses in his seminal article, the problem is not "one of verifying what objects or events a creature is responding to; the point is that without a second creature responding to the first, there can be no answer to the question".[26] "[T]here would be no saying what a speaker was talking or thinking about, no basis for claiming he could locate objects in an objective space and time, without interaction with a second person".[27] This indicates that the question what the typical causes of a speaker's basic utterances are is not a question we could answer if we were just observing that speaker. A fortiori, neither could the radical interpreter who is simply observing the foreign speaker.

As the latter quote might suggest, the problem, for Davidson, is in fact twofold. For it is not just that, for a person who has not interacted with others, the distal causes of her responses are ambiguous or indeterminate. According to Davidson, such a person could not even distinguish between distal causes and proximal causes, such as stimulations at the surface of her skin, or between typical causes and other causes in the causal chain that led to her utterance, all the

[22] Burge 1986: 125.
[23] Putnam 1975: 225.
[24] See Wittgenstein's remarks on ostensive definition (Wittgenstein 1953: §§ 28-30).
[25] Thus the problem is not merely epistemological but metaphysical, which is of course reminiscent of how Kripke understands the sceptical problem about meaning and rule-following he finds in Wittgenstein's paradox. (See Kripke 1982. I compare Wittgenstein's and Davidson's treatment of the sceptical problem in Verheggen 2017.)
[26] Davidson 1992: 119.
[27] Davidson 1992: 121.

way to the original big bang.[28] As he writes: the causes of a solitary person's responses are "doubly indeterminate: with respect to width and with respect to distance. The first ambiguity concerns how much [what 'part or aspect'] of the total cause of [an utterance]…is relevant to [meaning]…The second problem has to do with the ambiguity of the relevant stimulus, whether it is proximal (at the skin, say) or distal".[29] Call these problems respectively the "aspect problem" and the "distance problem".

Now Davidson distinguishes between two kinds of triangulation. Primitive triangulation is the kind of triangulation even non-linguistic creatures can engage in. It occurs when creatures are reacting simultaneously to each other and to common stimuli in their surroundings, as in the example of two lionesses trying to catch a gazelle and coordinating their behaviour by watching each other and the gazelle and reacting to each other's reactions.[30] According to Davidson, the distance problem can be solved for creatures who engaged merely in primitive triangulation. Thus the interacting lionesses can be said to be reacting to a distal cause, say, the gazelle, rather than to their sensory stimulations. Given the interaction, the cause of their reactions can be isolated as the common cause of their reactions, what is situated at the intersection of the lines that can be drawn between the lionesses and the object of their reactions. It must be stressed, however, that these creatures have no concept of what it is they are reacting to. Concepts, understood as elements of propositional thoughts, come only with language and thus with the second kind of triangulation. This kind of triangulation is fully linguistic. It occurs when creatures are reacting linguistically to each other and to common stimuli in their surroundings, as when a child who has become "aware of the possibility of error" triangulates with her teacher on objects or events in their surroundings; or as in a "situation in which two participants are equipped with thought and a language, but lack a common language [and the] problem is for each to understand the other: the problem of radical interpretation";[31] or when participants do by and large understand each other but ostension is needed to determine the cause, and hence the meaning, of one of the speakers' particular utterances. In short, linguistic triangulation is just a subset of interpersonal linguistic communication.

I have argued elsewhere that solving the distance problem does not actually require even primitive triangulation.[32] Non-triangulating creatures can rightly be described as reacting to features of their environment. I think, in fact, that the distance problem is not a problem. But I must emphasize that, even for Davidson, solving the distance problem is of no help in solving the aspect problem and thus of little help in answering the question what the typical causes of speakers' basic utterances are. For Davidson, to say that the distance problem can be solved by triangulating non-linguistic creatures is only to say that we are allowed to describe these creatures as reacting to features of their environment, that is, to distal causes. Solving the problem does not yield an answer to the question what specifically these features, i.e., what the relevant distal causes,

[28] Davidson 2001b: 4.
[29] Davidson 1999: 129-30.
[30] Davidson 2001b: 7.
[31] Davidson 2001c: 294.
[32] See Myers and Verheggen 2016: Ch. 1.

are; and so to say that some responses can be regarded as responses to distal causes is not tantamount to saying that the meanings of these responses are now determined. They are not because the crucial problem, the aspect problem, still needs to be solved.

Why is this a problem a solitary person cannot solve? First it should be made clear why it is speakers who have to determine what the causes of their basic utterances are. The reason, as I already suggested, is that, considered by themselves, the features of the world causing us to respond in certain ways are multifarious. For any given cause, it is similar to others in many respects. This is so no matter how regular a speaker's responses may appear to be. Indeed, this is so no matter how many people may together be giving what appear to be the same responses.[33] As long as they produce these responses passively, so to speak, these can be described in any number of ways. Davidson expresses the problem this way:

> Since any set of causes whatsoever will have endless properties in common, we must look to some recurrent feature of the gatherer, some mark that he or she has classified cases as similar. This can only be some feature or aspect of the gatherer's reactions […], in which case we must once again ask: what makes these reactions relevantly similar to each other? (Davidson 2001b: 4-5).

This is a question speakers need to answer if the causes, and hence the meanings, of their reactions are to be fixed. But this is not a question a solitary person, or a group of non-triangulating people, could answer. Again, why not?

The reason, as Davidson has often acknowledged,[34] was first brought to the fore by Wittgenstein. For a solitary person to determine what the causes of her responses are is for her to determine which causes are the same as which or, to put it differently, it is to determine which responses are correct and which incorrect. After all, to determine the meaning of an expression is thereby to determine what conditions of correctness govern its applications. Thus a solitary person needs to be in a position to distinguish between what seems to her to be the same cause and what is in fact the same cause, or between what seems to her to be the correct response and what is in fact the correct response. And she needs to do this in an objective way. If she can draw the distinction in any way she pleases, the distinction is not in fact genuine. To put it differently, a solitary person needs the idea of independently existing things affecting her in certain ways and of her responses to those things being correct or not depending on the things themselves and not on her saying so. But a solitary person is not in a position to have this idea and thus to make the relevant distinction. How indeed could she get the idea of objectivity when all she has at her disposal is her own subjective point of view?

It is easier, I think, to understand the predicament the solitary person is in if we first look at what makes it possible for people triangulating linguistically to distinguish in an objective way between what seem to them to be the same causes and what are the same causes. They are in a position to do this precisely because, by interacting with each other and the world they share, they are in a po-

[33] As made clear by Davidson 2001b: 8.
[34] See, e.g., Davidson 1994.

sition to recognize the possibility of different perspectives on their environment and the possibility of being sometimes mistaken about it. Davidson writes:

> Once these correlations [between interlocutors' reactions and external phenomena] are set up, each creature is in a position to expect the external phenomenon when it perceives the associated reaction of the other. What introduces the possibility of error is the occasional failure of the expectation; the reactions do not correlate (Davidson 1999: 129).

By triangulating linguistically with each other and the world they share, interlocutors are in a position to disagree on what is currently happening around them. More importantly, what people who triangulate linguistically can do, and the solitary person cannot, is settle their disagreement in a way that is not simply up to one or the other interlocutor. They can do this, not only because there are two of them, but also because their dispute is linguistic. A solitary person who "settled a dispute" with a non-linguistic creature would again settle it in a way that is entirely dependent on her. Note further that settling their disagreement can occur only if there are also things they agree on and, to begin with, if they agree on what their basic utterances mean and so on what in the world has contributed to determining their meaning. As Davidson has urged repeatedly, they do not have to agree to mean the same thing by the same words—though given how meanings are determined, in part by shared features of their environment, they are bound to agree on many. But they have to agree on what the speaker means by her words. Now, this agreement is also something that they have worked out together. They have worked together at narrowing and nailing down the causes of their basic utterances so that their meanings are eventually determined. But then it might be said that, since the meanings of their words are the product of a decision, since they are partly determined by the causes they took to be the same, triangulating people, too, are not really distinguishing between what are the same causes and what seem to them to be the same causes. This, however, would be to forget that their agreeing on the meanings of their utterances is the result of negotiations that did require distinguishing between what is the same and what seems to be the same. Moreover, once the meanings of expressions are fixed, that is, once the conditions of correctness governing their applications are fixed, whether these conditions are met or not is an objective matter, which cannot depend on the mere say of the speaker as it does with the solitary speaker. In other words, triangulating people eventually settle together what conditions of correctness govern the applications of their words, but they do not decide when these conditions are met.

What is the answer now, in view of the triangulation argument, to the question what it is for words to mean what they do? It is, for some of them at least, to have been used in linguistic triangulating situations. The meanings of basic words are partly determined through regular connections between triangulating interlocutors' utterances and features of the world around them that caused them to produce the utterances and that have been understood as the particular features they are by both interlocutors, that is, features which both interlocutors have agreed are the features relevant to determining the meanings of the speaker's words. Thus, the social view Davidson advocates is not of the communitar-

ian variety—as I noted earlier, speakers do not have to mean the same thing by any particular expression.[35] Of course, not all words have their meaning determined in a triangular way. The meaning of many words can be explained by means of other words, and someone could in that way introduce a new word that would never be used. As I suggested earlier, it is also not true that, for every word that refers to some feature of the world around us, one must have triangulated on that feature in order to mean something by it. But "there must be a causal history of that person that traces back, directly or indirectly, to triangular experiences".[36]

4. From Radical Interpretation to Triangulation

I take Davidson's views from radical interpretation to triangulation to be continuous. This has been widely contested. Kathrin Glüer, for instance, has suggested that the radical interpreter is a mere "dramatic device", which can be dismissed once the thought-experiment is over.[37] This, if right, would be significant, for it would indicate that Davidson shifted from the claim that one needs only to be interpretable to the claim that one must actually have been interpreted, in order to be a speaker.[38] For one thing, however, there never was, on Davidson's part, any acknowledgement of such a significant change of mind. On the contrary, many of the quotations I have provided are evidence that he meant the triangulation argument further to develop and refine views he had introduced earlier on. As already mentioned, he even explicitly says that radical interpretation is an instance of triangulation.[39] For another, philosophically more important, thing, the question to which triangulation is supposed to be an answer, viz., the question what the typical causes of speakers' utterances are, is a question that the radical interpreter would need to answer as well. If she did not, she would simply not be in a position to understand the speaker since she could not know what specific features in their surroundings contribute to determining the meanings of the speaker's basic utterances. And of course, if the radical interpreter needs to triangulate with a speaker in order to understand her, then the speaker herself must have triangulated with others in order to have a language. (Recall once more that reflecting on radical interpretation is meant to tell us not only how meanings can be attributed but also how they are determined or constituted.)

Some of the conclusions established by the triangulation argument reinforce those that follow from the considerations of radical interpretation. Pretty evidently,[40] the triangulation argument further demonstrates that meaning is thoroughly holistic since meanings cannot be determined piecemeal but as words are used in multiple triangular situations, which are needed for speakers

[35] Thus Davidson never gave up the view that meaning is not essentially conventional. See, e.g., Davidson 1986.
[36] Davidson 2001c: 293.
[37] Glüer 2011: 136. See also Pagin 2013: 230-31.
[38] This interpretation is also advanced by Lepore and Ludwig 2005: Ch. 19.
[39] Davidson 2001c: 294.
[40] Though this, too, has been contested—see Bernecker 2013: 447.

to single out the specific causes of their basic utterances.[41] What the triangulation argument also further establishes is the thoroughly non-reductionist character of Davidson's account of meaning.

Recall that the radical interpreter cannot in the end give a description of the meanings of a speaker's expressions—a semantic theory—without saying what the expressions mean. Likewise, as we have seen, merely describing regular connections between speakers' utterances and items in the world around them would not amount to specifying their meanings. But neither can we say that merely describing regular connections between triangulating speakers' utterances and items in the world around them is sufficient to capture their meanings. We have to think of the speakers themselves as taking those connections in specific ways, that is, we have to think of them as speaking meaningfully. What speakers must do in order to have a language, i.e., fix the specific causes of their basic utterances, is something we can think of them as having done only if we think of them as having a language. Thus in the end we cannot give an account of the nature of meaning without thinking of people who already use expressions meaningfully. This has of course struck many commentators as blatantly circular. Davidson's account is indeed circular, as is to be expected from a non-reductionist account. But it is not viciously circular—it is still a constructive account even though it is non-reductionist. As might also be expected from a non-reductionist account, it provides only necessary conditions for words meaning what they do and not sufficient ones. But the account, I should think, is far from uninteresting, indeed, far from uncontroversial.

The triangulation argument does not just reinforce claims Davidson had made while reflecting on radical interpretation; it also vindicates claims he endorsed as soon as he started reflecting on language and thought.

To begin with, the argument, as I have construed it, vindicates Davidson's claim that only a person who has the concept of objectivity can have a language and thoughts. This is a claim Davidson made shortly after he introduced the radical interpretation thought-experiment, together with the claim that being a "member of a speech community" is needed for possession of the concept. The contrast between truth and error, he maintained, "can emerge only in the context of interpretation, which alone forces us to the idea of an objective, public truth".[42] But he himself never made explicit the connection between, on the one hand, the claim that possession of the concept of objectivity and possession of a language necessarily go hand in hand and, on the other hand, what is required for meaning to be determined.[43] As a result, most commentators have understood his triangulation argument as an attempt to establish two independent claims, one about meaning determination and one about the possession of the

[41] Needless to say as well, only a significant history of using words in triangulating situations could make it possible for their meanings to be determined.

[42] Davidson 1975: 70.

[43] Though his insistence that our being able to say that creatures are reacting to distal causes is not sufficient for their having a language but that they must also be reacting to the interaction strongly suggests the connection between the two claims. See Davidson 1992: 120 and 2001c: 13.

concept of objectivity.[44] And it has been commonly argued that Davidson contended that primitive triangulation is needed to solve the problem of meaning determination, and that he was wrong about this. To repeat, I do not think that primitive triangulation needs to be present in order to describe creatures as reacting to distal causes, but doing this is not sufficient for saying that meanings are determined, for the aspect problem still needs to be solved, and linguistic triangulation is needed to solve it. Furthermore, though it may be conceded that linguistic triangulation is needed to have the concept of objectivity, it has been argued that people do not need to have the concept in order for the meanings of their words to be determined and thus in order for them to have a language.[45] What I have argued is that the two tasks cannot be separated. Linguistic triangulation is needed to determine meanings because it is needed to have the concept of objectivity which is required to determine meanings. Note that the tasks are accomplished concurrently. Given the non-reductionism—since there is no explaining in detail how a language is first acquired[46]—the best we can say is that it is only of people who have triangulated linguistically that we can make sense of their having the concept of objectivity and of their using expressions meaningfully. This is the result of acknowledging that speakers themselves have an important role to play in determining meanings. And this role is essential to the solution of the aspect problem.

The triangulation argument also vindicates an assumption Davidson made at the very start of his enquiry into meaning, viz., that it is essentially public. In effect, the argument does this by vindicating perceptual externalism. Here is how.

Recall that the assumption that meaning is essentially public prompted Davidson to reflect on radical interpretation. As we saw, externalism easily follows from these reflections. But suppose that the assumption is not being made; and suppose that externalism is being denied and that it is maintained that the determinants of meaning are to be found instead within speakers, in the form of mental pictures or representations, or of abstract entities grasped by the mind, or of dispositions. The problem, then, for the internalist is similar to that facing the externalist asking the question what the typical causes of speakers' basic utterances are. What the internalist needs to specify is how the allegedly meaning-determining picture, or abstract entity, or disposition is to be taken. For, in and of itself, a picture can be understood as representing many different things, an abstract entity can be taken as the symbol of many different things, and a disposition to utter words in certain ways can be understood as a disposition to use words with many different meanings. If Davidson's non-reductionist lesson holds, then anything, internal or external, that is not yet seen as meaningful, could not, by itself, determine the meaning of an expression.[47] But, if the triangulation argument holds, then it takes two to fix whatever specific features contribute to determining meaning. These features have to be triangulated upon. But then the items that, in the first instance, determine meaning can only be ex-

[44] See, among others, Pagin 2001: 201, Lepore and Ludwig 2005: 408, Bridges 2006: 295, Glüer 2011: 235, Bernecker 2013: 450.
[45] See, e.g., Andrews and Radenovic 2006.
[46] Davidson 2001c: 293.
[47] Even meanings will not do, as it must be specified which words are associated with which meanings.

ternal, for no internal items could be triangulated upon. Thus perceptual externalism is vindicated and, with it, the claim that meaning is essentially public.[48]

I end with a brief comparison of Davidson's externalism and orthodox versions.

5. Davidson's Externalism vs. Orthodox Versions of Externalism

I shall focus on an orthodox version of perceptual externalism, in part because Davidson himself did, on multiple occasions, address orthodox versions of social externalism.[49] However, though he did discuss orthodox versions of perceptual externalism as well, he never articulated the deep metaphysical difference that, I think, is the origin of the different versions of perceptual externalism, and which the triangulation argument brings into relief.

The version of perceptual externalism I have in mind is the one Putnam introduced over four decades ago with his Twin Earth thought-experiment. Suppose there is a planet, Twin-Earth, which is identical with Earth in all but one respect: the liquid called water on Twin Earth, which tastes and quenches thirst like what is called water on Earth, is not composed of H_2O molecules but of XYZ. As a result, according to Putnam, the very meaning of 'water' is different on Earth and Twin Earth. When an inhabitant of Earth and her doppelgänger on Twin Earth utter the word 'water', they are talking about different liquids. They may do so even unbeknownst to them, as they definitely would if they were living in 1750, before the molecular structure of what is called water was discovered. This is to say, according to Putnam, that the extension of words, when this is understood, at least in the case of most natural kind words, as the fundamental nature of the referent of a word, plays a crucial role in determining their meaning. Indeed, the extension plays the crucial meaning-determining role, as it is the only feature a change in which could cause a change in meaning. Stereotypical properties associated with the referent of the word, such as, in the case of 'water', colorlessness, transparency, tastelessness, etc., though they may initially help us to identify the referent of the word, are neither necessary nor sufficient for something to fall under a given kind, and so for the word to have the meaning that it has. In the end the only external feature that determines the meaning of a word like 'water' is its extension.[50]

This version of perceptual externalism is strikingly different from that defended by Davidson, for whom the specific features that cause us to respond in certain ways and determine in part the meanings of these responses are to be fixed through the multiple linguistic uses and beliefs of interlocutors who triangulate on those features. Putnam's idea, as Davidson expresses it, "is that if I learn the word 'water' while experiencing H2O, the word must henceforth refer only to substances with the same microstructure".[51] However, Davidson continues, "I do not see why sameness of microstructure is necessarily the relevant

[48] Of course, if the internal items we are thinking of do have a definite meaning, then they no longer have to be taken in specific ways. But the question then becomes, what determined their meaning? And, if Davidson is right, the answer can only be externalist.
[49] See, e.g., Davidson 1994 and 2001b. See also Myers and Verheggen 2016: Ch. 3.
[50] Putnam 1975: 234.
[51] Davidson 1991a: 198.

similarity that determines the reference of my word 'water'".[52] Indeed, as we saw, according to Davidson, what makes a given cause similar to another and thus the specific feature that determines in part the meaning of a word is itself fixed not simply or necessarily by the beliefs speakers may have about the microstructure of its referent, but by beliefs concerning other properties such as, e.g., in the case of water, its being odorless, potable, etc.[53] Now, what accounts for this striking difference between the two versions of perceptual externalism?

The key, I believe, to understanding this lies in the fact, alluded to earlier, that orthodox externalists like Putnam never pause to ask the question what the typical causes of speakers' basic utterances are. Why do they not? The reason may be two-fold. On the one hand, it is because, I think, they take the world around us to be ready-made, so to speak, that is, to be structured in such a way that language latches on its components without our having to contribute to this. The features of the world that cause us to use words in certain ways are already determined. All we need to do is to try to discover what they are. On the other hand, someone like Putnam does not seem to be in the business of giving an overall account of meaning. As I alluded to earlier, he uses the word 'liquid' in order initially to identify the referent of water. But he does not tell us in turn how the meaning of 'liquid' was determined to begin with. Even if this is indeed not Putnam's aim, it remains puzzling why, according to him, some beliefs rather than others are to play a privileged role in determining the relevant features of the world some words refer to and in turn in fixing their meaning, how they could indeed do this before anyone was in a position to have the relevant beliefs, and how they could do this even for speakers who lack the relevant beliefs. Only a certain kind of metaphysical picture, of the sort suggested above, can motivate this, leaving the relation between language and world rather mysterious. But if Davidson is right that the connections between speakers' basic utterances and their typical causes are not ready-made, the relation between language and world as conceived by orthodox externalists is not just mysterious; it is incoherent. There is no way we can think of the world as causing us to react in certain ways unless we already have reacted and thought of these reactions in certain ways. And this we have been able to do by triangulating linguistically on features of the world we share with our interlocutors.[54]

6. Conclusion

To sum up the main train of thought of this paper, I see no reason to deny that Davidson's triangulation argument builds on his earlier work on radical interpretation in a way that involves no significant changes. The kind of semantic externalism Davidson himself claims to have always endorsed is thereby better supported, as some of its fundamental assumptions have been vindicated.[55]

[52] *Ibid.*
[53] Davidson 1987: 29.
[54] For further discussion of Putnam's views, see Myers and Verheggen 2016: 77-83. See also Amoretti 2007.
[55] What about, it might be asked, Davidson's famous (or rather infamous) claim that meaning is indeterminate, such that, for any theory of interpretation an interpreter may come up with for a speaker, there might be another that fits the evidence equally well,

References

Amoretti, M.C. 2007, "Triangulation and Rationality", *Epistemologia,* 30, 307-26.

Amoretti, M.C. 2013, "Concepts Within the Model of Triangulation", *Protosociology,* 30, 49-62.

Andrews, K., and Radenovic, L. 2006, "Speaking without Interpreting: a Reply to Bouma on Autism and Davidsonian Interpretation", *Philosophical Psychology*, 19, 5, 663-78.

Bernecker, S. 2013, "Triangular Externalism", in Lepore and Ludwig 2013, 444-55.

Bridges, J. 2006, "Davidson's Transcendental Externalism", *Philosophy and Phenomenological Research*, 73-82, 290-315.

Burge, T. 1986, "Cartesian Error and The Objectivity of Perception", in Pettit, P. and McDowell, J. (eds.), *Subject, Thought, and Context*, Oxford: Oxford University Press, 117-36.

Burge, T. 1992, "Philosophy of Language and Mind, 1950-1990", *The Philosophical Review*, 101, 3-51.

Davidson, D. 1973, "Radical Interpretation", in Davidson 1984, 125-40.

Davidson, D. 1974, "Belief and the Basis of Meaning", in Davidson 1984, 141-54.

Davidson, D. 1975, "Thought and Talk", in Davidson 1984, 155-70.

Davidson, D. 1976, "Reply to Foster", in Davidson 1984, 171-79.

Davidson, D. 1979, "The Inscrutability of Reference", in Davidson 1984, 227-41.

Davidson, D. 1983, "A Coherence Theory of Truth and Knowledge", in Davidson 2001a, 137-53.

Davidson, D. 1984, *Inquiries into Truth and Interpretation*, Oxford: Clarendon Press.

Davidson, D. 1986, "A Nice Derangement of Epitaphs", in Davidson 2005, 89-108.

Davidson, D. 1987, "Knowing One's Own Mind", in Davidson 2001a, 15-38.

Davidson, D. 1990a, "The Structure and Content of Truth", *Journal of Philosophy*, 87-6, 279-329.

Davidson, D. 1990b, "Meaning, Truth, and Evidence", in Davidson 2005, 47-62.

Davidson, D. 1991a, "Epistemology Externalized", in Davidson 2001a, 193-20.

Davidson, D. 1991b, "Three Varieties of Knowledge", in Davidson 2001a, 205-20.

Davidson, D. 1992, "The Second Person", in Davidson 2001a, 107-22.

Davidson, D. 1993, "Reply to Jerry Fodor and Ernest Lepore", in Stoecker, R. (ed.), *Reflecting Davidson*, Berlin-New York: Walter de Gruyter, 77-84.

Davidson, D. 1994, "The Social Aspect of Language", in Davidson 2005, 109-26.

and his related claim that reference is inscrutable, such that there may be many equally good ways that we could keep track of it? (See, e.g., Davidson 1979.) Are these claims compatible with those made about triangulation? It might be thought that they are not since, after all, the triangulation argument tells us exactly how the typical causes of speakers' basic utterances are fixed, and hence how their meanings are determined. But Davidson himself never gave up the indeterminacy and inscrutability theses. I think the key to understanding how these theses are compatible with the claims made about triangulation lies in the holistic character of meaning. But establishing the compatibility is a difficult task which, partly for want of space, will have to await another occasion.

Davidson, D. 1995, "The Problem of Objectivity", in Davidson 2004, 3-18.

Davidson, D. 1999, "The Emergence of Thought", in Davidson 2001a, 123-34.

Davidson, D. 2001a, *Subjective, Intersubjective, Objective*, Oxford: Clarendon Press.

Davidson, D. 2001b, "Externalisms", in Kotatko *et al*. 2001, 1-16.

Davidson, D. 2001c, "Comments on Karlovy Vary Papers", in Kotatko *et al*. 2001, 285-308.

Davidson, D. 2004, *Problems of Rationality*, Oxford: Clarendon Press.

Davidson, D. 2005, *Truth, Language, and History*, Oxford: Clarendon Press.

Glüer, K. 2011, *Donald Davidson: A Short Introduction*, Oxford: Oxford University Press.

Haukioja, J. 2017, "Internalism and Externalism", in Hale, B., Wright, C. and Miller, A. (eds.), *A Companion to the Philosophy of Language*, Volume 2, Second Edition, West Sussex: Wiley & Sons, 865-880.

Kotatko, P., Pagin, P., and Segal, G. (eds.) 2001, *Interpreting Davidson*, Stanford: CSLI.

Kripke, S. 1982, *Wittgenstein on Rules and Private Language*, Cambridge, MA: Harvard University Press.

Lepore, E, and Ludwig, K. 2005, *Donald Davidson: Meaning, Truth, Language, and Reality*, Oxford: Clarendon Press.

Lepore, E. and Ludwig, K. (eds.) 2013, *A Companion to Donald Davidson*, London: Wiley Blackwell.

Myers, R.H. and Verheggen, C. 2016, *Donald's Davidson Triangulation Argument: A Philosophical Inquiry*, New York and London: Routledge.

Pagin, P. 2001, "Semantic Triangulation", in Kotatko *et al*. 2001, 199-212.

Pagin, P. 2013, "Radical Interpretation and the Principle of Charity", in Lepore and Ludwig 2013, 226-46.

Putnam, H. 1975, "The Meaning of 'Meaning'", in his *Language, Mind and Reality*, Cambridge: Cambridge University Press, 215-71.

Verheggen, C. 2017, "Davidson's Treatment of Wittgenstein's Rule-Following Paradox", in Verheggen, C. (ed.), *Wittgenstein and Davidson on Language, Thought, and Action*, Cambridge: Cambridge University Press, 69-96.

Wittgenstein, L. 1958, *Philosophical Investigations*, translated by G.E.M. Anscombe, New York: Macmillan.

Advisory Board

On Searle on Austin on Truth

Odai Al Zoubi

University of East Anglia

Abstract

John Searle gives two different interpretations to Austin's view on truth: 'the propositional interpretation' and 'the stating interpretation'. The former identifies what is true or false with the locutionary meaning, and the latter with the illocutionary act of stating. In this article, I argue that both interpretations are inaccurate, and I introduce a fresh interpretation that identifies what is true or false with the whole speech act.

Keywords: Truth, Speech Act Theory, Propositions, J.L. Austin, John Searle

According to J.L. Austin, in analysing utterances we need to distinguish between a *locutionary* act and an *illocutionary* act.

The locutionary act is "the utterance of certain noises, the utterance of certain words in a certain construction, and the utterance of them with a certain 'meaning'" (Austin 1975: 94). This contrasts with the illocutionary act. As Austin puts it:

> To determine what illocutionary act is so performed we must determine in what way we are using the locution:
> asking or answering a question,
> giving some information or an assurance or a warning,
> announcing a verdict or an intention,
> pronouncing a sentence,
> making an appointment or an appeal or a criticism,
> making an identification or giving a description,
> and the numerous like (Austin 1975: 98-99).

Every utterance[1] possesses a locution and an illocution, or what Austin sometimes calls 'meaning' and 'force' respectively.[2] Thus, for example, we distinguish between the meaning of the utterance: "Shoot her!", and the force of that utterance,

[1] Actually, *almost* every utterance: "whenever I 'say' anything (except perhaps a mere exclamation like 'damn' or 'ouch') I shall be performing both locutionary and illocutionary acts" (Austin 1950: 133).

[2] Note that Austin gives a technical sense to both 'meaning' and 'force'. "Admittedly we can use 'meaning' also with reference to illocutionary force—'He meant it as an order', &c. But I want to distinguish *force* and meaning" (Austin 1975: 100).

Argumenta 3, 1 (2017): 165-180
ISSN 2465-2334

DOI 10.14275/2465-2334/20175.ALZ

which depends on the circumstances but could consist in urging, or advising, or ordering me to shoot her.[3]

Precisely how to interpret this distinction, and how it relates to truth is disputed. In this article, I discuss John Searle's interpretations to Austin's view on truth. He gives two different interpretations, and the reason is that he does not find in Austin's writings a clear and conclusive explanation. The first interpretation, which I call *the propositional interpretation*, says that the locutionary meaning is the part of the speech act which corresponds to the facts, and which is true or false. The second interpretation, which I call *the stating interpretation*, says that the illocutionary act of stating is to be judged as true or false. I will argue that both interpretations are inaccurate, and that it is better to read Austin in a third way: what is true or false is the whole speech act.

I will start in 1 with Searle's first interpretation. This interpretation depends on a specific way of reading Austin's distinction between performatives and constatives, and his reasons for abandoning this distinction and to introduce instead the theory of speech acts. We therefore need to examine the first distinction. I will do this in 2, where I introduce my reading of Austin's text. In 3, I go back to the propositional interpretation and show why I find it incompatible with Austin's text. In 4, I discuss and criticise the stating interpretation. Next, in 5, I address a problem in my interpretation regarding the assessment of illocutionary forces and speech acts. Finally, in 6, I draw some conclusions.

1. Searle's First Interpretation

Searle's first interpretation, the propositional interpretation, suggests that for Austin what is true or false is the locutionary meaning.

To explain his interpretation, Searle starts with his understanding of Austin's distinction between locutionary meaning and illocutionary force:

> Austin may have had in mind the distinction between the content or, as some philosophers call it, the proposition […] and the force or illocutionary type of the act. Thus, for example, the proposition that I will leave may be a common content of different utterances with different illocutionary forces, for I can threaten, warn, state, predict, or promise that I will leave. […] the same propositional act can occur in all sorts of different illocutionary acts (Searle 1973: 155).

It seems that Austin would agree with the main idea here. As we have seen above, "shoot her" might be taken as advising, ordering, urging, etc., and these are different illocutionary forces, but the 'content', the 'locutionary meaning', is the same in all of them.[4] However, I will argue that Austin would not agree with what follows.

Searle continues, "it is the proposition which involves 'correspondence with the facts' […] Propositions […] can be true or false" (Searle 1973: 158-59). Searle then takes the content, the locutionary meaning, to be the part which is either true or false.

[3] The example is from Austin 1975: 101-102.

[4] We need to add, on behalf of Austin and Searle, that only after we fix reference and sense of the sentence used on different occasions, we say that, in these different occasions, the different uses of the same utterance share the same locutionary meaning, but have different illocutionary forces.

Note that this is the pervasive interpretation, in two ways: it is attributed to Austin, but it is also the reading which Searle himself and many philosophers adopt as the right way to relate truth to speech acts.[5] For example, P.F. Strawson has a similar view:

> Propositions […] are supposed to be bearers of truth-value […]. On any view, propositions may be expressed by parts of utterances […] parts which are not themselves advanced with the force which belongs to the utterance as a whole; and it may be expedient to [replace] the term 'propositions' […] with one less general. For the purpose Austin's own term 'constative' offers itself as a convenient candidate (Strawson 1973: 59-61).

Strawson suggests that we can abstract from the whole utterance the locutionary meaning, and separate it from the force. The locutionary meaning is the proposition, or the constative, and is to be assessed as true or false.[6]

It is important to note that both Searle and Strawson do not claim that they are just giving an interpretation to Austin's distinction: according to them, Austin himself is not completely clear about the distinction. However, they find indications in Austin's account of locutionary meaning which encourage them to adopt the propositional interpretation. These largely consist of what they take to be continuity between, on the one hand, the distinction between performatives and constatives, and, on the other, that between the locutionary and the illocutionary. Since this is the case, we need to regress for a while to introduce the distinction between performatives and constatives, in some length. It is only on the background of the relation between the two distinctions that we will be able to understand the propositional interpretation and what I claim to be its flaws.

2. The Performative/Constative Distinction and its Collapse

In his early writings, such as "Other Minds" and "Truth", Austin proposes that we can distinguish between 'performatives' and 'statements'. In *How to Do Things with Words*, "Performative Utterances" and "Performatives-Constatives", however, Austin later finds that the distinction is unstable, and he comes to realize that a new theory of speech acts is needed as a result. In what follows, I trace his thought through this development, starting with the performative/constative distinction in 2.1, and its collapse in 2.2. After that, in 2.3, I will explain what he means by a *dimension-word*, the key to understand Austin's position.[7]

Let me be clear that the following is my reading to Austin's text. My claim is that Searle ignores the importance of Austin's view that 'true' is a dimension-word, and that this is the reason he fails to give the right interpretation to Austin's theory of truth.

[5] Marina Sbisà, against the pervasive interpretation, argues that the propositional interpretation is incompatible with different aspects of Austin's philosophy (Sbisà 2006). She identifies his views on truth as one of them. I agree with most of her remarks. I focus here only on one aspect: Austin's claim that 'true' is a dimension-word, a claim that she touches on only briefly.

[6] Nat Hansen, in a recent paper, gives a similar interpretation, see Hansen 2012.

[7] In Al Zoubi 2016, I used the same argument introduced here in section 2 to discuss a recent debate between Alice Crary 2002 and Nat Hansen 2012 regarding Austin's views on literal meaning.

2.1 The Performative/Constative Distinction

According to Austin, philosophers used to take every utterance of the declarative grammatical form (an utterance which is a not question, command, etc.) to describe state of affairs, or report or state facts. As a result, they thought that they must be either true or false.[8] Other utterances, which do not take the declarative form, such as questions or commands, are not true or false. Let us call utterances which are either true or false *statements*.[9] However, says Austin:

> it has come to be realized that many utterances which have been taken to be statements (merely because they are not, on grounds of grammatical form, to be classed as commands, questions, &c.) are not in fact descriptive, nor susceptible of being true or false (Austin 1950: 131).

Austin observes that an utterance which takes the declarative form is not a statement "when it is a formula in a calculus: when it is a performatory utterance: when it is a value-judgement: when it is a definition: when it is part of a work of fiction" (Austin 1950: 131). These are different kinds of utterances: they take the declarative form, but are not descriptive. One important kind of such an utterance is the 'performative'.

According to Austin, the distinction between performatives and constatives is as follows.[10] Constatives are utterances which are either true or false. For example, when you state something, or describe something, or report something, your utterance is either true or false. Take for example "the cat is on the mat". This is a declarative sentence, which is descriptive. It describes how things are, and it is true or false, if the state of affairs is, or is not, as it states.

In uttering a performative, on the other hand, I do not describe a state of affairs, or report something, and my utterance cannot be taken to be true or false. Instead, I *do* something. For example, in a marriage ceremony, when I say "I do", "I am not reporting on a marriage: I am indulging in it" (Austin 1975: 6); or when in some official ceremony I am supposed to name a ship, I say, "I name this ship the Queen Elizabeth"; or when I say "I bet you sixpence it will rain tomorrow". Other examples include: "I promise that ..." and "I apologize". Thus, in uttering a performative we get married, or name something, or promise, or apologize.

[8] Austin discusses this *descriptive fallacy* in a number of different places: see Austin 1946: 97-103; 1950: 130-32; 1956: 233-34; Austin 1975: 1-4 and 100.

[9] Austin was suspicious of the two terms, 'descriptive fallacy' and 'statements', which he himself employed in his early writings: "perhaps this is not a good name, as 'descriptive' itself is special. Not all true or false statements are descriptions, and for this reason I prefer to use the word 'Constative'" (Austin 1975: 3). The point is this: the fallacy takes all utterances of the affirmative grammatical form as either true or false. Austin was led to see that there is a problem in lumping all these terms, such as stating, describing, reporting, etc. under the heading 'descriptive' or 'statement'. See "How to Talk", where Austin tries to give an account of the differences between these different terms. We need a term to describe what seems to be either true or false, and 'constative' is the one Austin used in his major work, Austin 1975. This will be an important observation when we discuss Searle's second interpretation.

[10] In the three later works mentioned above, Austin examines the distinction before declaring that it does not work. Most of what follows depends on the characterization of the distinction as it appears in the major work, Austin 1975.

What we say is not true or false, and we do not state, or describe, or report anything. We do something else.

However, simply uttering a performative is not sufficient to constitute the specific act. Saying a few words is not marrying: "The words have to be said in the appropriate circumstances" (Austin 1956: 236). One way to highlight this dependence on appropriate circumstances is to consider how we might *fail* in doing the act. For example, if I am married already, then saying "I do" in the ceremony, will not make me married. If I am not the person who was chosen to name the ship, then saying "I name this ship…" fails: the ship was not named, even though I uttered the words; and if no one wants to bet with me, then I have not bet anyone. In each of these situations something goes wrong because some factor in the context is inappropriate. In such circumstances, according to Austin, the act is "to some extent a failure: the utterance is then, we may say, not indeed false but in general *unhappy*" (Austin 1975: 133).

However, in the next section we will see that Austin comes to realise that the constative/performative distinction is unstable, and in accordance with its collapse, he offers his theory of speech acts.

2.2 The Collapse of the Distinction

In "Truth" Austin says that "many utterances which have been taken to be statements […] are not in fact descriptive, nor susceptible of being true or false" (Austin 1950: 131). He gives some examples, performatives being one.[11] However, in the same paper, he states that it is common for statements to have a performatory aspect. He explains that "[I]t is common for quite ordinary statements to have a performatory 'aspect': to say that you are a cuckold may be to insult you, but it is also and at the same time to make a statement which is true or false" (Austin 1950: 133).

The utterance "you are a cuckold" is both: it is performative, to insult you, and it is descriptive, it is a statement, which is either true or false.

The difficulty is that this position seems inconsistent: on the one hand Austin seems to be denying performatives the capability to indicate truth or falsehood, but, on the other, he seems to grant them this ability. As a result, the fundamental distinction between performatives and constatives seems to be threatened, and Austin himself quickly realises this.

In particular, he recognises that for both kinds of utterances we often appraise the relation between the words and the world in the same way, using the same family of terms which belong to the dimensions of truth. Any utterance is appraised in relation to both the appropriate circumstances under which it is uttered, and the facts which the utterance somehow 'correspond to'. Thus, constatives are assessed (being true or false) in relation to facts, as is the 'happiness' of performatives: we estimate rightly or wrongly; we find correctly or incorrectly; we argue soundly; we advise well; we judge fairly; we blame justifiably. In all these cases, our assessment relies on the facts: "the question always arises whether the praise, blame, or congratulation was merited or unmerited" (Austin 1975: 141).

Equally, "such adverbs as 'rightly', 'wrongly', 'correctly', and 'incorrectly' are used with statements too" (Austin 1975: 141). All this makes us question the original distinction between two kinds of utterances, constatives which are merely

[11] See Austin 1950: 133.

true or false and correspond to facts, and performatives, which were not thought to be either true or false because they neither describe nor state how things are, and therefore do not correspond to facts. As a result, Austin asks:

> Can we be sure that stating truly is a different class of assessment from arguing soundly, advising well, judging fairly, and blaming justifiably? Do these [performatives] not have something to do in complicated ways with facts? (Austin 1975: 142).

In assessing a performative to be happy or unhappy, using the adjectives above, "[F]acts come in as our knowledge or opinion about facts" (Austin 1975: 142). In other words, the happiness or unhappiness of performatives, which originally were thought to be independent of correspondence to facts, turns out to be related to facts, as are constatives.

A similar difficulty arises when we consider constatives, whose truth value were originally thought to be independent of the circumstances under which of uttering the words. Austin gives the following example:

> Suppose that we confront 'France is hexagonal' with the facts, in this case, I suppose, with France, is it true or false? Well, if you like, up to a point; of course I can see what you mean by saying that it is true for certain intents and purposes (Austin 1975: 143).

According to Austin, it is a "rough description". But we cannot simply assess if it is true or false. "It is good enough for a general top-ranking general, but not for a geographer" (Austin 1975: 143). It is difficult to see how we can say it is true or false, without taking the circumstances of uttering it into account. Take another example: "Lord Raglan won the battle of Alma". This is good enough for a school book, but not for historical research (Austin 1975: 143-44). More examples from "Truth" include: "Belfast is north of London", and "the galaxy is the shape of fried egg". In all these cases, it seems that we cannot tell if the statement is true or false without taking into account the circumstances under which it was uttered.

The upshot of this is that the distinction between performatives and constatives collapses.[12] The distinction was supposed to show us that we have, on one hand, utterances which are true or false, which correspond to the facts, and, on the other hand, utterances which are not true or false, and are assessed according to the circumstances under which they are uttered. The above examination shows us that both kinds of utterances are often related both to the facts and to the circumstances under which they are uttered, and that they are both assessed in similar ways. The key reason for this, according to Austin, is his view of 'true' as a dimension-word: The terms which we use in assessing the performatives overlap with the terms we use in assessing constatives: we use the same family of words to describe and assess both performatives and constatives. Austin concludes "[W]hen a constative is confronted with the facts, we in fact appraise it in ways involving the employment of a vast array of terms which overlap with those that we use in the appraisal of performatives" (Austin 1975: 142-43).

Next, we will examine what exactly Austin means by a dimension-word.

[12] Austin abandons the distinction for other reasons as well, reasons that both Searle and Strawson rightly highlight, but they ignore the reason I highlight here. We will come back to this in 4.

2.3 'True' as a Dimension-Word

Austin distinguishes between two kinds of words: words that have one meaning, and words that have multiple, unrelated meanings. In his examination of the different uses of 'real' Austin points out that this word "does not have one single, specifiable, always-the-same *meaning* [...] *Nor* does it have a large number of different meanings—it is not *ambiguous*" (Austin 1962: 64). According to Austin, there are words that have always-the-same-meaning, like 'yellow' or 'horse', and, on the other hand, there are ambiguous words like 'bank', which can mean either a financial institution or the edge of a river. These are completely different meanings.[13] There is, nevertheless, a middle ground between these two kinds of word. Many philosophers neglect this middle ground, he thinks, and, as a result, they fall into a false dichotomy: '*one meaning vs. ambiguity*', which often causes them erroneously to look for one meaning for each word.

In particular, Austin argues that with certain types of word that have multiple meanings there might be something in common between all the uses of the word, but that this commonality exists at an 'abstract' level, and that focusing on this common factor obscures the many differences that exist at the 'concrete', contextual level. In other words, the meaning of these words involves two levels: what we might term 'abstract meaning'/'semantic function' and 'specific meaning'. The former, in virtue of being abstract, might well be consistent across uses of the word in different contexts and cases, whereas the latter is likely to vary depending on the circumstances and contexts in which the word is used.

One type of such words, which Austin studies in depth, is dimension-words.[14] The dimension-word "is the most general and comprehensive term in a whole group of terms of the same kind, terms that fulfil the same [semantic] function" (Austin 1962: 71). Dimension-words define a semantic dimension and the range of terms appropriate to the particular abstract meaning or semantic function of the particular dimension-word. The dimension-word could, in fact, substitute for any of the members of the family of words within its dimension because all members possess this abstract meaning along with their own context-related specific meaning. However, the necessarily abstract nature of the meaning of the dimension-word means that its usage in particular situations would be unlikely to convey the required specificity of specific meaning. Thus, although the abstract meaning/semantic function of all the terms in one family is the same and is constant in all the uses of a dimension-word, Austin wants to show that identifying this common thing and focusing on it will not provide a sufficiently robust or accurate basis on which to determine meaning. The semantic function is too thin; it needs to be supplemented by the specific meaning, which changes according to the context.

It is the combination of the shared abstract meaning and the context-related specific meaning which suggests that dimension-words do not have one meaning in all of their uses, and yet are not ambiguous. Rather, they have a number of different-but-related specific meanings which are unified by their common possession of the 'abstract meaning' of the term.

[13] 'Yellow' and 'horse' are Austin's examples; 'bank' is mine.

[14] The other types are trouser-words and adjuster-words. I discuss all three types of word and Austin's views on the meaning of words in my forthcoming "Austin on the Unity of Meaning".

Austin thinks that 'true' is a dimension-word, in virtue of which it has something in common in all of its uses, what we called the 'abstract meaning'/'semantic function', but no one *specific* meaning in all of its contexts or circumstances of use.[15] The semantic function associated with 'true' fulfils the following purpose: "true and false are just general labels for a whole dimension of different appraisals which have something or other to do with the relation between what we say and the facts" (Austin 1956: 250-51). In addition, he notes that the different terms which belong to the family, and share this semantic function, are quite diverse. Thus, we find within its ambit terms such as 'exaggerated', 'vague', 'bald', 'rough', 'misleading', 'not very good', 'general', 'too concise', 'fair', etc. These are the terms which we, in ordinary language, use for the appraisals of utterances. All members of the family share the same semantic function, but differ from each other in other aspects and characteristics.

According to Austin, it is rare that we use 'true' or 'false' in ordinary language. Austin, as we shall see, thinks that ordinary users employ these abstract terms only in logic and mathematics. Instead, we tend to pick a member of the family (such as 'exaggerated' or 'vague') that better represents the particular aspect of truth or falsity appropriate to the situation.[16]

Now we can understand Austin's position on truth: The different terms we use to assess the relation between our utterances and the world are diverse and rich. All these terms, in different ways, share the same semantic function of that assessment, but they differ according to the context, which specifies one of the different specific meanings. This is one of the reasons that the distinction between performatives and constatives collapses, as we have seen above, and Austin proposes a new theory of speech acts, where he distinguishes between locutionary and illocutionary acts.

The relation between the two distinctions is crucial if we want to understand Austin's view on truth. As we have seen in section 1 Searle suggests that there is a continuity between the two distinctions, in the sense that the locutionary meaning inherits the feature of the constative of being true or false. At the same time, the illocutionary force inherits the feature of the performative of being an act of a specific type, such as stating, or promising, or asking, etc., as we shall see in section 4.

I argue that the first of these two points is inaccurate, while I agree with the second. In arguing against Searle, I will explore in more depth Austin's views on truth and give a fuller account of it. My reading is that being true or false is to be assessed in relation to the whole speech act, and not any part of it. This is explained next in 3, and then utilised in criticising Searle's second interpretation later in 4.

3. The Problem with Searle's First Interpretation

Austin comments on the relation between the two distinctions as follows.

[15] Austin gives other examples in his writings, such as 'real', 'good', and 'freedom'.

[16] Austin discusses the different terms of the family of words we use to assess an utterance in ordinary language in Austin; 1950: 129-30; 1956: 250 and 1975: 122-74.

> With the constative utterance, we abstract from the illocutionary […] aspects of the speech act, and we concentrate on the locutionary […] With the performative utterance, we attend as much as possible to the illocutionary force of the utterance, and abstract from the dimension of correspondence with facts (Austin 1975: 145-46).

Both Searle and Strawson cite this remark to motivate the propositional interpretation,[17] and, taken in isolation, it perhaps seems reasonable to infer that Austin's view is that the locutionary meaning is the heir of the constative, what is true or false, and the illocutionary force is the heir of the performative, doing something like arguing, stating, warning, etc.

However, on the same page Austin also writes:

> Perhaps neither of these abstractions [constative as focusing on the locutionary, and performative as focusing on the illocutionary] is so very expedient: perhaps we have here not really two poles, but rather an historical development. Now in certain cases, perhaps with mathematical formulas in physics books as examples of constatives, or with the issuing of simple executive orders or the giving of simple names, say, as examples of performatives, we approximate in real life to finding such things. It was examples of this kind, like "I apologize", and "The cat is on the mat", said for no conceivable reason, extreme marginal cases, that gave rise to the idea of two distinct utterances (Austin 1975: 146).

Austin here makes explicit the instability of the distinction between constatives and performatives that I identified above. Whilst there are extreme cases where the distinction is clear, the vast majority of constatives and performatives fail to conform to this strict interpretation. Austin seems to argue precisely for making a break with the very notion of such a distinction in practice.[18] It is on the basis of this realisation that Austin wants to introduce his new theory of speech acts, the collapse of the old distinction having been driven, as we saw earlier, by the recognition that 'true' is a dimension-word, and that we use the same family of words to appraise both kinds of utterances.

The strange thing about Searle's and Strawson's reading (apart from failing to place in context the quotation on which they rely) is that it seems to fall back into the same problem that led Austin to move away from the constative/performative distinction and propose the new speech act theory. In particular, it seems that Austin recognises that at the heart of the collapse of the distinction is the realisation that we cannot separate the two categories by appealing to two distinct notions of appraisal: true/false and happy/unhappy. However, if the locutionary meaning is not directly heir to the constative, what is it in an utterance that can be validly appraised as being true or false?

I argue that the most plausible reading of Austin's position is that we should assess an utterance as true or false in relation to the whole speech act, and not just one part of it. This position is consistent with the moral of the collapse of the first distinction, where we had to take into account that the terms of assessment of true and false merge and overlap with terms of assessments of happy and unhappy.

[17] See Strawson 1973: 53 and Searle 1973: 155.

[18] I do not have space here to discuss the cases Austin mentions where the truth of what is said is not related to the circumstances under which the utterance is issued.

The lesson there was that both types of assessment generally depend on, and are determined by both facts *and* circumstances of utterance.[19]

Finally, let me clarify one aspect of my objection to the propositional reading. The problem with identifying the locutionary meaning as that which is true or false is that it treats it as a 'proposition' which is to be appraised as true or false *regardless of the circumstances under which it is uttered*. Whilst I take Austin to agree with Searle that the locutionary meaning might be shared by different speech acts[20] and that it is something which we abstract from those different speech acts, Searle identifies the locutionary meaning with what is true or false, whereas I argue that Austin does not. If Searle is right, then the locutionary meaning, which we abstract from different speech acts, can by itself—and independently of being uttered under specific circumstances, since it is abstracted from the actual circumstances under which it is uttered—be true or false. This is precisely the opposite of what I have tried to show for Austin: that the circumstances under which we utter the words is vital for applying the terms of the truth family.

This account is symmetrical with Austin's account of dimension-words, and truth in particular. Whilst, in extreme cases, the abstract component of the dimension word can be used on its own without reference to the circumstances of use, in almost every normal case the abstract element is too weak to be used and, instead, other words in the same family are employed, words which better matches the context.

In summary, I argue that the propositional interpretation is not compatible with Austin's text, and that, in the general case, it is the whole speech act which is to be judged as true or false. I do not deny that there is a relation between the constative/performative distinction and the locutionary/illocutionary distinction. Indeed, Austin himself thinks that there is such a relation: "[T]he doctrine of the performative/constative distinction stands to the doctrine of locutionary and illocutionary acts in the total speech act as the special theory to the general theory" (Austin 1975: 148). However, as I have shown, Austin does not think that the locutionary meaning is the heir of the constative in the crucial sense that it is not to be assessed as true or false.

4. Searle's Second Interpretation

Now we move to Searle's second interpretation. Here, Searle attributes to Austin the view that what is true or false is the illocutionary act of stating. I will argue that Austin's text does not support such a reading.

Searle says that he wants to examine "one of Austin's most important discoveries, the discovery that constatives are illocutionary acts as well as performatives, or, in short, the discovery that statements are speech acts" (Searle 1973: 157). It is true, as Searle explains, that Austin in the new theory regards stating, describing, arguing, warning, etc., as illocutionary forces. "Stating, describing, &c., are just two names among a very great many others for illocutionary acts" (Austin 1975: 148-49).

[19] As mentioned above, Sbisà takes Austin's views on truth to be incompatible with the propositional interpretation, and her sketchy remarks are in line with my own interpretation. For example, she writes that Austin wants "to interpose the illocutionary dimension between meaning and truth/falsity assessment, thus contextualizing truth/falsity assessments to the 'speech act in the speech situation'" (Sbisà 2006: 167).

[20] As we said above, after we fix the sense and reference.

Before going on to Searle's second interpretation, let me regress to the collapse of the distinction between constatives and performatives: In secondary literature, the collapse is seen, mainly, as a result of this discovery. The constatives/performative distinction is thus flawed because it is based on a distinction between saying something and doing something, and Austin finds that constatives and performatives are both 'doing' and 'saying'. The new theory developed in Austin 1975 gives him a sophisticated tool to clarify how each utterance consists of different doings (acts), in different ways.[21]

My discussion in 2.2 above on the collapse of the distinction does not exclude the importance of this 'discovery', but it focuses on a neglected aspect of Austin's work: his claim that 'true' is a dimension-word.[22]

Now let us go back to Searle's second Interpretation: It is this discovery with which Searle in fact agrees that he identifies as the source of the mistakes in Austin's theory of truth.

Searle starts by explaining that 'statement' "is structurally ambiguous" (Searle 1973: 157). It has two meanings: "'Statement' can mean either the act of stating or what is stated" (Searle 1973: 157). He calls the former *statement-acts*, which are illocutionary acts, and the latter *statement-objects*, which are the propositions/locutionary meanings stated. According to Searle, the distinction helps us to identify clearly what is true or false: "Propositions but not acts can be true or false; thus statement-objects but not statement-acts can be true or false" (Searle 1973: 159). It is the statement-object, the proposition, and not the illocutionary act of stating, Searle claims, which is to be identified as true or false.

Austin, Searle thinks, has confused the two:

> [T]he failure to take into account the structural ambiguity of 'statement' […] had very important consequences [...] For since statements are [illocutionary] speech acts, and since statements [the statement-objects, the propositions] can be true or false, it appears that that which is true or false is a [illocutionary] speech act. But this inference is fallacious, as it involves a fallacy of ambiguity […] And the view that it is the act of stating which is true or false is one of the most serious weaknesses of Austin's theory of truth (Searle 1973: 157).

Searle concludes,

> Statement-acts are illocutionary acts of stating. Statement-objects are propositions […] The latter but not the former can be true or false. And it is the confusion between these which prevented Austin from seeing […] [that illocutionary] acts cannot have truth values (Searle 1973: 159).

[21] We do not need here to go into details with how to distinguish exactly between the different acts.

[22] Similar readings are offered by other interpreters, Max Black 2011 and Jennifer Hornsby 1988 and 1994, for example. All these readings, correctly, point out that the saying/doing criterion is flawed, and they point out that the true/false and happy/unhappy criterion is problematic, in different ways. However, none of them brings out the importance of 'true' as a dimension-word.

For Searle, it is the locutionary meaning / the proposition/ the statement-object which is true or false. Austin was mistaken in taking the illocutionary act of stating to be true or false because Austin confused the act of stating with what is stated.

I find this reading problematic for two reasons. Firstly, as we have seen in section 2, Austin uses the term 'constatives' rather than 'statements' in his later writings, such as his major work on speech acts, Austin 1975[23]. It is therefore difficult to understand the suggestion that he equivocates on the term 'statements' in his argument. Indeed, it seems from Austin's reservations about the terms used to designate what is true or false in the initial distinction he makes that he was at pains to avoid using terminology that carries any specific traditional philosophical charge, precisely to avoid misleading himself or the reader.

Secondly, Searle's reading does not engage with the idea that for Austin 'true' is a dimension-word. This means, as we have seen, that Austin thinks that we apply a family of different terms to appraise the relation between utterances and the world, and that it is therefore the full speech act which is liable to be true or false. In particular, it seems clear that Austin believes that the whole speech act is assessed for truth or falsehood *whatever the illocutionary force*. In addition, Searle's position here is weak because of the lack of pertinent textual evidence in support of his claim. Although Searle is perfectly right in saying that, for Austin, stating is an illocutionary force, there is no textual evidence to suggest that Austin might have thought that what is true or false is the illocutionary act of stating. In fact, there is a paragraph where Austin seems explicitly to reject Searle's reading. Here is what Austin writes in the last lecture of Austin 1975, on the very page where he also writes that stating is an illocutionary force:

> Stating, describing, &c., are just two names among a very great many others for illocutionary acts; they have no unique position […] In particular, they have no unique position over the matter of being related to facts in a unique way called being true or false, because truth and falsity are (except by an artificial abstraction which is always possible and legitimate for certain purposes) not names for relations, qualities, or what not, but for a dimension of assessment-how the words stand in respect of satisfactoriness to the facts, events, situations, &c., to which they refer (Austin 1975: 148-49).

Here, then, Austin re-states the position that we examined earlier: except in extreme cases or artificial circumstances of abstraction, truth and falsity represent a family or dimension of terms the use of which depends upon the circumstances (facts, situations, etc.), and illocutionary acts of any type, whether or not they consist in stating or describing, are insufficient on their own to determine truth or falsity. Instead, consideration of the speech act in the round is necessary for such an assessment.

It therefore does not seem that Austin was misled by the two meanings of 'statement', as Searle claims. First this assertion is not backed up by the text, and second it is not compatible with Austin's explicit perspective on 'true' as a dimension-word, a factor which Searle ignores.

[23] See footnote 9 above.

5. The Fate of 'Happiness'

In this final section, I will point out a tension in Austin's conclusions on the fate of the happy/unhappy assessment: on one hand, he suggests that it survives the collapse of the performative/constative distinction, and that it is inherited from the performative by the illocutionary force; on the other hand we might read him as saying that it emerges with the true/false assessment, as I argued above. I will examine this tension and suggest that the second option seems more plausible.

Let us start with the first. Austin seems to suggest that the distinction between two kinds of assessment, happy/unhappy and true/false, survives the collapse. He ends lecture XII of Austin 1975, where he examines the performative/constative distinction and draws his final conclusions in what he calls the "real conclusion", stating what needs to be done after the collapse:

> [we need] critically to establish with respect to each kind of illocutionary act […] what if any is the specific way in which they are intended, first to be in order or not, and second, to be 'right' or 'wrong'; what terms of appraisal and disappraisal are used for each and what they mean (Austin 1975: 146-47).

It seems here as though Austin has in mind first the happiness of the illocutionary ("in order or not") and second its truth or falsity ("to be 'right' or 'wrong'"). It might be, then, that Austin thinks that being true or false is not inherited by the locution, but being happy or unhappy is inherited by the illocution.

However, there is a problem with this line of thinking, and we must investigate it. Austin says at the beginning of this very lecture that he will examine the performative/constative distinction in two steps: first, looking at the doing/saying distinction, and second, looking at the two kinds of assessment (the true/false and happy/unhappy). I will repeat quickly the main arguments discussed in detail above in order to explain our problem. In the first step, he rejects the saying/doing distinction, suggesting that all utterances are both 'sayings' and 'doings'. Next, he examines in detail the two kinds of assessment, or the "alleged contrast" between them, as he calls it (Austin 1975: 136). It is this second step which is relevant to us here. He starts "from the side of the supposed constative utterances", where he argues at length that "we find that statements are liable to every kind of infelicity of which performatives are liable" (*ibid.*). After that, he moves on to "looking at the matter from the side of performatives" (Austin 1975: 140), where he argues, again at length, that we use the same terms in assessing constatives and performatives. The core argument of his second step is that the true/false assessment was supposed to show whether the constative corresponds to the facts, and the happy/unhappy assessment was supposed to show whether the performative is uttered under the appropriate circumstances. But both performatives and constatives are, in real life, assessed according to both considerations, and we use the same family of terms in these assessments. Consequently, it seems plausible to conclude that there are not two kinds of assessment, but only one.

What I find problematic in Austin's first position, where he keeps the distinction between true/false and happy/unhappy assessments, is the following: being uttered under the appropriate circumstances, the consideration taken into account for the happy/unhappy assessment, is part of our consideration when assessing the truth or falsity of the whole speech act, and, at the same time, when assessing the happiness of one part of it, the illocutionary force. This does not seem very

convincing. In addition, he argues that 'corresponding with the facts', the consideration taken into account for the true/false assessment, is to be considered when we assess the happiness of performatives, and this is one of the reasons why the first distinction collapses. Why, then, does he think that the happy/unhappy assessment will survive? Again, this does not seem convincing. Finally, he argues that the terms we use to assess constatives and performatives overlap. If this is so, is it not more plausible to think that the two kinds of assessment merge into one big family rather than stay separate? The first position, then, seems to run against what he tried to establish throughout the lecture. The second option, where you look at the whole speech act and assess its success with one family of terms, is more plausible.

However, the text is inconclusive. We probably have to follow Austin's advice and think about the different kinds of illocutionary forces: what terms do we use to assess them? Do they overlap? What are the considerations we take into account when we assess the success of speech acts with different kinds of illocutionary force? Such an examination is beyond the scope of this paper, though. I will leave it here, suggesting that, initially at least, the merging of the two kinds of assessment into one is more plausible, but admitting that only the examination of how to assess speech acts and different kinds of illocutionary forces would settle the issue.

6. Conclusions

I have examined Searle's two interpretations to Austin's theory of truth, and argued that neither is accurate. Searle's first interpretation identifies what is true or false with the locutionary meaning, while his second interpretation identifies it with the illocutionary act of stating. I suggest that there is a third reading of Austin, to which I subscribe. It identifies what is true or false with the whole speech act.

The key to understand Austin's position, which Searle ignores, is that 'true' is a dimension-word: It has one and the same semantic function in all its uses, but different specific meanings according to the context. It is the most abstract word in a family of words that are used to assess the relation between words and the world. It is only by appreciating the importance of Austin's view that 'true' is a dimension-word that we will be able to properly interpret his theory of truth, and the relation between the locution and illocution. We saw that the distinction between constatives and performatives fails *partly* because 'true' is a dimension-word and the terms we use to assess both kinds of utterance overlap. These terms belong to the family of 'true' where all terms are used to assess the relation between utterances and the world in different dimensions and degrees. The propositional interpretation fails to appreciate the importance of, and the reasons for, the collapse of the constative/performative distinction; it sees a continuity between the constative and the locutionary meaning in the sense that the latter inherited the feature of being true or false. This, as we have seen, is incompatible with Austin's text and arguments. Again, the stating interpretation, which identifies what is true or false with the illocutionary act of stating fails for the same reason. I have argued that, for Austin, contrary to both interpretations, the whole

speech act is to be assessed as true or false.[24] I ended up with a question regarding how to assess illocutionary forces, where I pointed out a tension between Austin's views on two positions. To solve this tension, we need to look at how we assess different kinds of speech acts with different illocutionary forces. This is a project that will have to be undertaken in the future.

References

Al Zoubi, O. 2016, "Austin and Literal Meaning", *Kriterion*, 30, 3, 41-64.

Al Zoubi, O., "Austin and the Unity of Meaning", forthcoming.

Austin, J.L. 1946, "Other Minds", in Austin 1979, 76-116.

Austin, J.L. 1950, "Truth", in Austin 1979, 117-33.

Austin, J.L. 1953, "How to Talk", in Austin 1979, 134-53.

Austin, J.L. 1956, "Performative Utterances", in Austin 1979, 233-52.

Austin, J.L. 1962, *Sense and Sensibilia*, Oxford: Oxford University Press.

Austin, J.L. 1963, "Performative-Constative", in Caton, C.E. (ed.), *Philosophy and Ordinary Language*, Urbana: University of Illinois Press, 22-54.

Austin, J.L. 1975, *How to Do Things with Words*, Oxford: Oxford University Press.

Austin, J.L. 1979, *Philosophical Papers*, Oxford: Oxford University Press.

Berlin, I. (ed.) 1973, *Essays on J.L. Austin*, Oxford: Clarendon Press.

Black, M. 2011, "Austin on Performatives", in Fann, K.T. (ed.), *Symposium on J.L. Austin*, London: Routledge & Kegan Paul, 401-11.

Crary, A. 2002, "The Happy Truth", *Inquiry*, 45, 1, 59-80.

Hansen, N. 2012, "J.L. Austin and Literal Meaning", *European Journal of Philosophy*, Early view 19 Jan.

Hornsby, J. 1988, "Things Done with Words", in Dancy, J., Moravcsik, J. and Taylor, C.C.W. (eds.), *Human Agency: Language, Duty and Value*, Stanford: Stanford University Press, 27-46.

Hornsby, J. 1994, "Illocution and its Significance", in Tsohatzidis 1994, 187-207.

Recanati, F. 1994, "Contextualism and Anti-Contextualism in the Philosophy of Language", in Tsohatzidis 1994, 156-66.

Recanati, F. 2001, "What is Said", *Synthese*, 128, 75-91.

Sbisà, M. 2006, "Speech Acts without Propositions?", *Grazer Philosophische Studien*, 72, 155-78.

Searle, J.R. 1969, *Speech Acts*, Cambridge: Cambridge University Press.

Searle, J.R. 1973, "Austin on Locutionary and Illocutionary Acts", in Berlin 1973, 141-59.

Strawson, P.F. 1973, "Austin and 'Locutionary Meaning'", in Berlin 1973, 46-68.

[24] My interpretation is compatible, in its main lines, with Charles Travis's and Francois Recanati's influential views on truth and meaning. See Travis 2008 and Recanati 1994 and 2001. However, 'true' as a dimension-word does not play the role in their work I claim it plays in Austin's work. Furthermore, if my claim is correct, there will be some considerable differences between Austin and their views on truth. It is, however, beyond the scope of this paper to pursue these differences further.

Tsohatzidis, S. (ed.) 1994, *Foundations of Speech Act Theory: Philosophical and Linguistic Perspectives*, London: Routledge.

Travis, C. 2008, *Occasion-Sensitivity: Selected Essays*, Oxford: Oxford University Press.

Di Francesco, Michele, Marraffa, Massimo and Paternoster, Alfredo, *The Self and Its Defences. From Psychodynamic to Cognitive Sciences.*
London: Palgrave MacMillan, 2016, pp. vii + 219.

Today more than ever before, the philosophical debate on the self is situated at the crossroads of diverse disciplines.[1] The multifaceted and integrative nature of the contemporary researches on subjectivity is the first noticeable aspect widely rendered in *The Self and Its Defences. From Psychodynamic to Cognitive Sciences,* where the Authors M. Di Francesco, M. Marraffa and A. Paternoster outline their theory of the self and self-consciousness. Framed in a naturalistic as well as theoretically informed scenario, the theory they propose stems from the puzzling attempt to define what is a self, in the light of a conceptual and empirical inquiry, along with an historical and analytical reconnaissance of some relevant issues in philosophy of psychology and philosophy of mind.

From a philosophical point of view, the reflection on the self and self-consciousness is currently located between apparently opposed and yet complementary and sound conceptual options, whose differing accounts mirrors to a certain extent the common sense twofold intuition. Indeed, on the one hand, the self appears undoubtedly a product of the cultural realm, something constructed through social interaction, communication and narratives, i.e. the elemental, psychological unit of the broader, juridical concept of *person*. On the other hand, a self seems to be the starting condition we need in order to consider an organism a subject of experience, a psychobiological system bounded in his first-person perspective, somehow emerging from neural and biological mechanisms. It can easily be expected though, that even common sense is reluctant to define the self as merely an abstract entity constructed through inferences and interpretations; or conversely to think that it merges entirely with the primordial, bodily and perceptual domain of the first-person perspective; or farther that it overlaps with the mechanisms underlying the personal domain. *The Self and Its Defences* aims at combining these constitutive aspects emerging from the philosophical and common sense reflection, and at maintaining both a *top-down* and a *bottom-up* explanation, in the wake of the challenge promoted by cognitive science, which consists in "widening the naturalistic realm […], denying that our choices and actions can be exhaustively attributed to historical and interpretative factors, and thus trying to overcome the dichotomy, dear to the hermeneutic tradition, between *Naturwissenschaft* and *Geisteswissenschaft*" (11).

In the domain of the contemporary debate on subjectivity, the project of making this connection possible is relevant and worthy, as in the last decade a certain amount of philosophical works has dealt with the abovementioned alternative.[2]

Non reductive, explanatory pluralism is the compass of the project endorsed in the book, where philosophy of psychology constantly dialogues with

[1] See Gallagher, S. (ed.) 2011, *The Oxford Handbook of The Self,* Oxford: Oxford University Press.

[2] See Gallagher, S. 2000, "Philosophical Conceptions of The Self: Implications for Cognitive Science", *Trends in Cognitive Science*, 4, 1, 14-21; Zahavi, D. 2007, *Self and Other: The Limits of a Narrative Understanding*, in D.D. Hutto (ed.), *Narrative and Understanding Person*, Royal Institute of Philosophy Supplement, 60, Cambridge: Cambridge University Press.

Argumenta 3, 1 (2017): 181-195
ISSN 2465-2334

DOI 10.14275/2465-2334/20175.BOO

several domains that ranges from philosophy and metaphysics of mind, to phenomenology, anthropology, developmental and social psychology.

The derivative approach to the self is usually connected to the philosophical thesis of anti-realism, that is, whereas the self is not considered as something primitive and logically prior, or as just a bodily, bounded entity that perceives the world from its first-person perspective, but rather as a product, it ends up being an illusion or a useful fiction, a work of art, namely an entity which is not part of the equipment of the natural, physical world. But a consideration of the self in terms of solely neurocognitive, bodily processes underlying the personal domain determines a reductive approach as well. More specifically, the effort to outline a naturalistic scenario where elusive notions like that of self can be included has to face the well-known trouble of the overlapping between the personal and the automatic, blind, sub-personal domain. Historically, the attempt to combine naturalism and realism when treating the abstract entities posited by psychology lies at the very ground of almost every debate in philosophy of mind. As of the more specific debate on the self, this effort arises from the overcoming of an old, idealistic approach, back to Descartes' *Cogito* to the Kantian *I Think* and the Husserlian transcendental phenomenology. *The Self and Its Defences* continues and renews the philosophical tradition that has questioned the given, innate nature of the self, and combines this tradition with the conceptual framework of the contemporary, empirical researches on subjectivity. The sound intuitions and the large empirical data that support a naturalistic account of the self are explored and evaluated without accomplishing a reduction to the subpersonal domain.

The Authors state at the very beginning that the theory of the self they outline is at the same time a theory of self-consciousness, declaring a pivotal idea developed in the book. Indeed, the connection between the two notions (i.e., self and self-consciousness) reveals not only that the self is considered in its reflexive and "objectifying" component, i.e. in its being both a subject and an object of consciousness. It also establishes that in order for an organism to be considered a self, it has to be involved in a process of knowledge, namely of transitive consciousness, where the object of consciousness is the self. The two core ideas are that the self is "a psychobiological system activity of self-representing", and that this representational process is aimed at "defending the self-conscious subject against the threat of its metaphysical inconsistence" (1). Following William James, the theory here proposed is focused on the mechanisms that allow the I to make the Me, i.e. on both the processual, representational activity of the mind, and the representation of the self as an object of consciousness. Chapter 2 (*The Unconscious Mind*) and Ch. 3 (*Making the Self, I: Bodily Self-Consciousness*) are mainly dedicated to the elaboration of the first thesis, i.e. to the description of the self's processual, representational activity, starting from the unconscious mechanisms to the gradual emergence of object consciousness. In Chapter 4 (*Making the Self, II: Psychological Self-Consciousness)* and Ch. 5 (*The Self as a Causal Centre of Gravity*), it is presented the second core idea of the theory, i.e. the defensive nature of the self-*ing* process and the more general theme of the fragility of the subject.

The structure of the book mirrors the bottom-up route aimed at individuating the several stages towards the construction of personal identity. After having introduced the methodological and conceptual frame in which their research is conducted (Ch. 1), the Authors present an analytical and historical examination

of the notion of the unconscious, which lies at the very foundation of the subjectivity's development (Ch. 2). The core of this section is that the abilities manifested in behaviour have to be explained by the unconscious mechanisms underlying them. This thesis calls into question the puzzling mark of the mental problem (i.e., how are we to define the boundaries between what is mental and what is not?), the connected interface problem (i.e., how are we to combine and make the personal level of explanation interacting with the non-personal one?), as well as an inquiry on the difference between the Freudian and the psychodynamic unconscious. As of the first issue, while trying to individuate the right criterion for demarcating the mental, it is auspicated a strategy for the indirect incorporation of the sub-personal as well as its dialectical relationship with the personal level. As of the second issue, in overcoming both Freud's positivistic naturalism and the hermeneutical, post-modern reading of psychoanalysis, cognitive science is situated halfway between the personal sphere of the first-person phenomenology and the non-personal domain of neurobiological events. The historical inquiry of this section brings out a fundamental issue: if the cognitive unconscious, far from reflecting the conscious level, is not mental, how to interface the computational level of explanation with the ordinary psychological explanations? What it is auspicated is a coexistence and compatibility between the ordinary image of the subject as a rational, free agent, and the scientific conception of subjects as computational machines, in a strategy where both dependence and autonomy are worth to be maintained. In this perspective, the concept of attachment analysed in a cognitive and evolutionary framework is taken to be a pivotal research area in which the relationship is possible. The rehabilitation of psychoanalysis in a cognitive-evolutionary framework, where dynamic psychology and the cognitive sciences are connected through the notions of motivation and attachment, determines a systemic–relational framework where neither radical social constructivism nor strong individualism are embraced. In this developmental and evolutionary framework, minds are shaped in early interactions with others, and personal identity's assembling is followed starting from the analysis of the young child's affective and cognitive relationality.

In Ch. 3 a non-idealist, non-eliminative view of the self is outlined. The section contains an analytical discussion of Humean eliminativism on the self, of analytic Kantianism, and of the contemporary phenomenological accounts of self-consciousness. This examination aims at combining naturalism and realism, thus at questioning the thesis according to which the self has no room in the natural world. The worth-mentioning insight of this section is that the reality of the self does not rest on a phenomenological, bodily and innate sense of *me-ness* or *for-me-ness*, which turns out to merge with a weak form of first-person perspective, or with a transcendental precondition evocating the Kantian *I Think.*[3] In the bottom-up construction of the self here endorsed, the early intelligent behaviour as well as the organism's experiential interaction with the environment and the boundedness of its body are not considered as developmental stages where a form of minimal, pre-reflective self-consciousness makes the experience possible. In particular, it is argued that in order to consider an organism self-conscious, it has to be conscious of its body taken as a whole, namely it has to be able to represent himself as the object of its consciousness. Therefore, it is

[3] See also Marraffa, M. and Paternoster, A. 2014, "A Third-person Approach to Self-Consciousness", *A&P, International Multidisciplinary Journal*, 11, 107-19.

questioned the very popular hypothesis that a form of pre-reflective self-consciousness grounds the reflective one as well as consciousness in general.[4] The pivotal thesis that emerges in this section, i.e. that self-consciousness does not ground conscious experience but is rather acquired during development, is then sustained by a large review of developmental psychology and cognitive ethology, and through a critical reading of some philosophical attempts to connect the a-priori methodological approaches on the study of subjectivity with the empirically informed ones. Besides the fact that the possession of a first-person perspective—with which the "empirically void" notion of pre-reflective self-consciousness seems finally to merge with—is not a sufficient condition for self-consciousness, the Authors finally argue that bodily self-consciousness is far from being consciousness of oneself as a continue entity through time, namely it is far from being a psychological self.

The emergence of an "extended"[5] self (Ch. 4) is due to an interplay of mentalization, memory and interpersonal skills regulated by cultural and environmental cues. The construction of the virtual inner space of the mind is followed within the framework of attachment theory, along with a systematic, remarkably large review of the literature ranging from developmental, to social and personality psychology. It is retraced the process towards the subject's awareness of the mind as an inner dimension, as well as its evolution into the cognitive more complex form of the narrative/autobiographical self.

The two core ideas underlying the theory outlined in *The Self and Its Defences*—i.e. the processual and defensive nature of the self-*ing* activity—are elaborated in a framework aimed at avoiding both idealism, according to which the self is somehow logically prior, and eliminativism, that conversely stems from a derivative account of the self. The naturalistic scenario that frames the book, far from embracing a consideration of the self as an illusion or an epiphenomenon, is then aimed at drawing a realistic account of the self. Through a strategy that contemplates the constant examination of bottom-up and top-down explanatory paths, the Authors' intent is to avoid the excesses of these approaches, that is, as of the latter, the reductive approach that stems from an analysis in terms of solely neurocognitive mechanisms, and as of the former, the inflationary and radical top-down approaches. Naturalism is endorsed through the systematic attention to the scientific, empirical researches of cognitive sciences, and to the contribution given by the attachment theory and the psychodynamic tradition.

Given the conciliatory project of this work, the demanding, fond-of-metaphysics reader could finally feel a bit unconvinced. At the very beginning, the Authors state that their project is more an exercise in the philosophy of psychology rather than in the metaphysics of mind. Nevertheless, the book is necessarily involved with fundamental issues such as the already mentioned mark of the mental problem, the causal closure of the physical domain, the overdetermination thesis, just to mention few controversial themes to which the Authors themselves allude. As of the mark of the mental problem, the metaphysical grounding is far from being insignificant. The sub-personal level is taken to be

[4] See Gallagher, S., Zahavi, D. 2005, "Phenomenological Approaches to Self-Consciousness", *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition), E.N. Zalta (ed.), http://plato.stanford.edu/archives/spr2015/entries/self-consciousness-phenomenological

[5] The term is drawn from Damasio, A. 1999, *The Feeling of What Happens. Body, Emotion and the Making of Consciousness,* London: Vintage.

connected to the personal one through the undeniable fact that the subject has access to the products of the former, but not to the sub-personal processes themselves (§ 1.3.2). What is mental, according to this proposal, is the integrated sub-personal realm, whose inclusion is due to the transparency of its products. This proposal rests on a metaphysical thesis, because the desired connection between the personal and the sub-personal level of explanation can be established only if sub-personal computations are considered as genuine pieces of mind: that is, just in case it is endorsed a metaphysical commitment to the extended mind hypothesis. Thus, the mark of the mental problem is an ontological, metaphysical problem, and unless the products of the sub-personal are considered as genuine pieces of mind, the purpose of connecting the sub-personal and the personal as much as it is possible hardly establishes what is mental and what is not from an ontological point of view.

As of the difficulty with the abovementioned conciliatory strategy, it lies in a crucial point, on which the Authors themselves take stock at the end of the book, and that could be a worth pursuing extension of their theory of the self. How does the narrativism they endorse differ from the hermeneutical one? And how does it avoid the illusionism *à la* Dennett?[6] As of the former, likely question, the framework in which their narrativism is maintained, they argue, is naturalistic because the theory-driven self-interpretation is taken to be a re-appropriation of the products of the neurocognitive unconscious. Second, it is argued that narrative selfhood would not arise without an affective and bodily self-description. Furthermore, the very reason why the Authors consider their theory a non eliminativist one lies in a sort of "final", double overcoming of the Cartesian account of the self. They observe how this account still echoes in contemporary eliminativism, for it seems to presuppose that if the self exists, it must be a persisting individual substance. In other words, according to the Authors, contemporary illusionism infers the non reality of the self from the lack of a neuronal counterpart of the Cartesian ego.

Another clarification aimed at pointing out the specificity of their theory is worth questioning. The Authors state that their disagreement with the eliminativists is not merely verbal, because they are not arguing that the illusion of the self can have causal powers as a false belief may (66). This is taken to be an "obvious" and "irrelevant" conceptual move (*Ibid.*), useless if considered in order to distinguish their account from the eliminativist one. Nevertheless, given the confabulatory, deceptive nature of the self here widely documented, the reader could ask why the definition of the self as a "causal center of gravity" (Ch. 5) should not be considered also in the sense in which a false belief has causal powers. After all, a crucial assumption is that the selfing process deals with a teleology of self-defense: far from being an epiphenomenon, subjective identity is a layer of personality, a causal center of gravity whose protection is necessary for mental health. As of this issue, a worth pursuing extension of the book could be a clarification of the criteria affording the distinction between self-knowledge and self-deception, in particular regarding their different causal role.

This distinction seems to depend on a fascinating criterion derived from developmental and evolutionary data: if, after all, the problem with antirealism is that it disregards the *defensive* nature of identity self-construction (147), does real-

[6] Dennett, D.C. 1992, "The Self as a Center of Narrative Gravity", in Kessel, F., Cole, P., Johnson, D. (eds.), *Self and Consciousness: Multiple Perspectives,* Hillsdale: Erlbaum.

ism finally rest on the adaptive benefit of the narrative self? If metaphysics is taken to be driven by our current best scientific theories, then the self is real as such theories seem to suggest that the very foundation of human beings' health is the protection from psychological disintegration, which is made possible by the production of causal efficacious' life stories.

University Roma Tre MARIAFLAVIA CASCELLI

Williamson, Timothy, *Modal Logic as Metaphysics*.
Oxford: Oxford University Press, 2013, pp. xvi + 464.

At the end of the preface of *Modal Logic as Metaphysics* (MLAM, from now on), T. Williamson apologises (to his sons) for writing a book "with such a dull title" (p. xvi). Indeed, the title of this book does not even try to be engaging, or cool, but it has surely the merit of highlighting the fundamental methodological thesis on which the entire project is based. If modal logic *is* a metaphysics, then the role of modal logic within modal metaphysics—and that of logic within philosophy in general—is not (anymore) neither that of being a *neutral arbiter* nor—to use David Kaplan's words—that of "striving to serve ideologies".[1] This methodological attitude signs the end of *exceptionality* about logic and the beginning of an *anti-exceptional* attitude about it. Logic is not anymore exceptional, because, even in its meta-logical manifestations, it should not be taken to be the realm of the insubstantial or the uncontentious; logical theories (and their metalogics) are substantial and contentious as any other scientific enterprise: therefore, they should be evaluated by means of an abductive methodology and judged "partly on their strength, simplicity, and elegance, partly on the fit between their consequences and what is independently known. […] Logic […] is no mere background framework but the very thing at issue" (423-24).[2]

Given this methodological framework, MLAM's main aim is that of convincing us that *necessitism*, the view that necessarily everything is necessarily something is the best metaphysics of modality. The endorsement of this *metaphysics*, however, is closely connected to (and probably is) an abductive endorsement of a *modal logic*: *higher-order necessitist S5*. Williamson shows that this logic, or better a specific intended model structure of it, is strong enough to give us the general structure of metaphysical modality, to give us substantial answers to a great numbers of modal metaphysical questions and it is better positioned than its opponents as far as strength, simplicity and fit are concerned. In showing this, Williamson achieves the impressive result of reconfiguring the entire debate on the philosophical foundations of modal logic and modal metaphysics. One way in which this reconfiguration happens is by abandoning (what Williamson believes is) the "badly confused" dichotomy possibilism/actualism and substitute it with the much clearer (at least for him) dichotomy contingentism/necessitism. Williamson's dissatisfaction with the possibilism/actualism dichotomy depends on the view that the only two plausible definitions of actual-

[1] Kaplan, D. 1994, "A Problem in Possible-World Semantics", in Sinnott-Armstrong, W., Raffmann, D. and Asher, N. *Modality, Morality and Belief: Essays in Honor of Ruth Barcan Marcus*, Cambridge: Cambridge University Press.
[2] See also Williamson, T. 2014 "Logic, Metalogic and Neutrality", *Erkenntnis*, 79, 211-31.

ism make this notion either viciously circular or trivially true (22-23). In the first case, the circularity depends on defining what is actual as what is true in the *actual* world, in the second case, the triviality depends on the logical behaviour of the actuality operator which could be inserted, *salva veritate*, in any non-modal context (at least in its standard interpretation); from this it follows that "actualism", the view that what is is what *actually* is, turns out to be a logical truth and possibilism, the negation of actualism, a logically falsehood. The entire possibilism/actualism debate is thus trivial and it should be substituted by the contingentism/necessitism debate which does not rely, at least explicitly, on any notion of actuality.

In this review, I will mainly be concerned with chapters of the book which make clear the sense in which a logic could be a metaphysics and justify, present and defend higher-order necessitist S5. These chapters could be taken as the "core" part of the book (at least in my opinion): action begins at Chapter 3 (from section 3.3) and continues through Chapters 5 (really from 5.5) and 6. Chapters 1-2 and 7-8 could be seen, respectively, as a preparation and an application of the core part.

Chapter 1 is devoted to philosophical preliminaries such as the definition of necessitism and of contingentism (as the negation of necessitism), to the temporal analogues of necessitism/contingentism, and to the critique of the actualism/possibilism distinction (which will be further elaborated in Ch. 7). Chapter 2 is devoted to the emergence of the Barcan Formula and its converse in the twentieth-century logical philosophy in authors such as Barcan Marcus, Carnap, and Prior: the chapter represents an extraordinary excursus on the pre-Kripkean developments of the semantic of modal logic. Chapter 7 is devoted to the possibility of "inter-theoretical communications" between necessitists and contingentists through the method that Williamson calls "of neutral equivalents". The communication between contingentists and necessitists is possible because contingentists and necessitists speak the very same language characterised by an unrestricted interpretation of the quantifiers and a metaphysical interpretation of modal operators. This proves, for Williamson, that, unlike the possibilism/actualism debate, the debate between contingentists and necessitists is non-verbal and substantial. Within the common language, there will be lots of sentences that are *neutral*, namely sentences over which the necessitist and the contingentist do not disagree, and there should be, at least in principle, a systematic way of "mapping" (where a "mapping" is not a translation) the non-neutral sentences into this neutral territory. Assume that there is a sentence D such that the necessitist accepts D and the contingentist rejects it; $(D)^{nec}$ is a mapping of D into this neutral territory (if there is such a mapping at all). Given that the necessitist accepts the equivalence between D and $(D)^{nec}$, the necessitist accepts $(D)^{nec}$, but given that $(D)^{nec}$ is, by hypothesis, neutral, also the contingentist can accepts $(D)^{nec}$. By accepting $(D)^{nec}$, the contingentist can thus extract some "cash value" from a sentence that she does not accept in its non-neutral formulation. It turns out that necessitism has more expressive power than contingentism, because, while everything that the contingentist can express can also be mapped into the neutral part of the language (i.e., where A is a sentence that the contingentist accepts, there is always a mapping $(A)^{con}$ that the necessitist can accept), the necessitist can express some theses that cannot be mapped into the neutral part of the language. Williamson thus concludes that "the necessitist can draw more distinctions than the contingentist can. Every distinction the contingentist can draw

can be drawn in neutral terms, so the necessitist can draw it too. The converse fail. The necessitist can draw distinctions the contingentist cannot, because they cannot be drawn in neutral terms." (364). Finally, Chapter 8 is a very rich and stimulating chapter devoted to the consequences of necessitism: among those worth noticing there is the view that the distinction between modal and non-modal dimension is elusive (and this saves the necessitist from the need to explain how the modal supervenes on the non-modal), the view that the incompatibility of necessitism with the truth-maker principle is a benefit of the view, rather than a cost and the view that necessitism is able (albeit with some difficulties) to account for radical contingency.

Let us now turn to the "core" part. Section 3.3 is a fundamental section of the book, because it makes clear in what sense a modal logic can be treated as a metaphysics of modality. What MLAM tries to establish is that the set of "sufficiently general" modal metaphysical truths has "the formal characteristics of a modal logic" (92). Of course the connection between a modal logic L and metaphysical modality cannot be direct. At least in the tradition of Kripkean model theory, the connection between a logic and a philosophical theory should be established by means of an *intended interpretation*. But an intended interpretation of a modal logic (one where, for example, to a certain predicate P is associated a specific predicate such as "being a tiger") is not sufficiently general. What Williamson is looking for are the general, structural principles of metaphysical modality and, at least in the propositional case, such principles are those that remain true under the permutation of their non-modal parts by holding fixed logical constants and modal operators in their metaphysical interpretation. This kind of permutation could be expressed, for Williamson, by means of quantification into sentence position. A formula like $\Diamond p \rightarrow \Diamond\Diamond p$ is thus a general structural principle of metaphysical modality (is "*metaphysically universal*", 93) if and only if the formula $\forall X\,(\Diamond X \rightarrow \Diamond\Diamond X)$ is true. MLAM tries to show that there is a logic, *MU*, whose theorems are all and only the metaphysically universal truths and thus *MU* is "the uniquely correct logic of metaphysical modality". In particular, the formulas valid in an intended modal structure of *MU* are exactly the metaphysical universal ones and validity in such a model structure *is* metaphysical universality. The next step is to characterise the elements of such a model structure by means of a Boolean algebra of propositions; this move allows Williamson to show that *MU* is none else that S5, namely to show that all and only the theorems of S5 are metaphysically universal. An extension of the same techniques to the predicative case allows Williamson to define metaphysical universality for the first-order case (124). Here the fundamental argument is one where Williamson shows that a formula like $\exists y\,(x = y)$ is valid in an inhabited, pointed, model structure (a model structure with a domain and with the specification of an actual world) only if the Barcan Formula is valid in that model structure. The validity of the Barcan Formula (and, in the end, its metaphysical universality) is thus shown by Williamson to be simply a consequence of the validity and metaphysical universality of $\exists y\,(x = y)$. This is bad news for the contingentists, because this result shows that, if there is an intended inhabited structure for *MU*, then contingentism is false. The only option left for contingentists seems to be that of denying that there is an intended inhabited modal structure or having a totally instrumentalist approach to the model-theoretic apparatus.

Chapter 5 and 6 represents the higher-order extension of what Williamson has done in Chapter 3. According to him, a deeper comparison between contingentism and necessitism could be pursued in the contest of a higher-order modal logic where one can quantify both in predicate and sentence position. Within the most sensible system of higher-order modal logic, ML_p, a comprehension principle of the form $\exists X\,(\Phi\forall x\,(Xx \leftrightarrow A))$ (whose role is that of guaranteeing a suitable array of properties) turns out to be metaphysically universal and thus valid in the intended modal structure (section 5.5, 230). Very briefly stated, the argumentative line of Chapter 6 is the following: necessitists are better positioned than contingentists to accept a robust interpretation of such a comprehension principle and to accept its consequences in a non-instrumental way. One of these consequences is *haecceitism*, the view that for every object, it is necessary that there is an haecceity of that object (where the haecceity of an object *o* is the property of being *o*). Another consequence is a second-order analogue of the necessary existence of individuals, namely the view that necessarily, every property is necessarily something (263). This conclusion is based on the assumption that necessary co-extensiveness of properties is the second-order analogue of first-order identity for objects. Actually, the conclusion is even stronger, because the comprehension principle implies necessary existence at *every* order except the first. The contingentists cannot accept neither haecceitism nor "any-order" necessary existence so they are forced to reformulate the comprehension principle. The problem is that weaker or non-modal formulations of the comprehension principle are, according to Williamson, unsatisfying: they would prevent "second-order logic from adequately serving the logical and mathematical purposes for which we need it" (288). On these abductive grounds, Williamson concludes that, if we want to use higher-order modal logic as a framework for our modal metaphysical inquiries, necessitist interpretations of ML_p should be preferred to its contingentist interpretations and thus necessitist higher-order S5 is our modal metaphysics. It has to be noticed, however, that the abductive grounds presented by Williamson do not seem to be so vast: the "mathematical purposes" served by a necessitist interpretation of ML_p are limited to some inferences within modal set theory, admittedly a quite remote and exotic field of mathematical practice (287).

The achievements of this book are impressive and they will surely change the way in which modal metaphysics will be done in the next decade. Before concluding my review, however, I would like to point out two problems. The first is about the elimination of the possibilism/actualism debate in favour of the contingentism/necessitism debate. Williamson claims that, on the operator account, "being actual had better be actually doing something harder than just being, otherwise the supposed dispute is silly" (23). I think that the critique is not entirely fair. As C. Menzel notices[3], the actualist should not be represented as someone looking for a robust conception of actuality, for a conception of actuality where actuality should be viewed as some "hard"-core feature of the world: rather, an actualist is *denying* the possibilist bifurcation of reality between *possibilia* and *actualia* (or between an heavyweight and a lightweight notion of existence) and she is simply denying the existence of *possibilia* (or the distinction be-

[3] Menzel, C., "In Defence of the Actualism/Possibilism Distinction", unpublished, https://www.academia.edu/20200724/In_Defense_of_the_Possibilism-Actualism_Distinction

tween two senses of existence). Assume that there is a philosopher who defends an analogous bifurcation between physical and "spiritual" things. A physicalist philosopher would be one who is simply denying the bifurcation by denying the existence of spiritual things. Sincerely, I do not see how this dispute (and by analogy the possibilism/actualism dispute) should be seen as "silly". So, I think that, after all, the dichotomy between possibilists and actualists will survive the publication of MLAM. The second problem is this. It is a recurrent theme of the book (specially section 3.6) that contingentists can have at most an instrumentalist attitude towards the model theory for modal logic. The problem for contingentists is that, if there is an intended model structure whose logic is sound and complete for metaphysical universality where $\exists y\,(x = y)$ is true, then, it is quite easy to prove that the Barcan Formula is true in that intended modal structure. Contingentists should therefore either deny that there is an intended model structure or hope that the model structure is only indirectly related to metaphysical universality. In Chapter 8, however, while defending the compatibility of necessitism with radical contingency, it turns out that necessitists too should have an instrumentalist attitude towards the intended modal structure: in particular, they should deny that the intended model structure is explanatorily prior to the contingent propositions from which it was constructed (410). So even necessitists should retain an instrumentalist attitude towards the model theory. Maybe, the instrumentalist attitude of the contingentist needs to be more radical than the one required to the necessitist, but given that the main aim of the book was to show how a logic could be literally taken to be a metaphysics, this instrumentalist residue towards the logic could be problematic for the necessitist as well.

University of Padua VITTORIO MORATO

Vetter, Barbara, *Potentiality*.
Oxford: Oxford University Press, 2015, pp. ix+335.

Potentiality aims to provide a disposition-based account of modality, specifically metaphysical modality. Some interesting twists, however, separate it from other efforts in this direction. First, (Ch. 2 to 4) Vetter develops her preferred view about dispositions, then applies it to the treatment of modality (Ch. 5 to 7). These two parts are not at all disjointed. Firstly, her account of dispositions, and specifically their individuation, dictates the starting point in the treatment of modality: *possibility* rather than *counterfactuals*. Vetter does not believe dispositions to be primarily linked to stimulus/manifestation counterfactuals (Ch. 2). Refreshingly, she does not argue against a conditional analysis of dispositions through the usual interferences counterexamples (one may share the feeling that little to no progress can be achieved, on neither side, by pursuing this weary debate). Rather, as in Manley and Wasserman (2007, 2008),[1] Vetter discusses

[1] Manley, D. and Wasserman, R. 2007, "A Gradable Approach to Dispositions", *Philosophical Quarterly*, 57, 226, 68-75. And from the same authors, 2008, "On Linking Dispositions and Conditionals", *Mind*, 117, 465, 59-84. One can also consult the exchange between Manley, D. and Wasserman, R. 2011, "Dispositions, Conditionals, and

deeper, structural problems within the conditional analysis. In Ch. 3, using linguistic and lexicographic considerations as a jumping point, she develops a new account of dispositions, individuated by manifestations alone; their modal character is (to some approximation) that of possibility.

However, the *trait d'union* between the two parts of the book is the introduction of *potentialities*. Potentialities are a generalization and extension of our ordinary dispositional concepts, ranging from dispositions to powers, capacities, abilities and so forth. There is a number of reasons for their introduction. First of all (Ch. 3), potentialities serve as a vagueness- and context-insensitive metaphysical background for dispositional ascriptions; as such, they come in degrees of modal strength. Potentialities are also introduced to obtain suitable items for the definition of modal operators. The key-word here is "suitable", as Vetter introduces three constraints for a potentiality-based theory of possibility: *extensional correctness* (potentialities must deliver enough possibilities), *formal adequacy* (potentiality must have the right logical form and structure), and *semantic utility* (potentiality must reconstruct a good portion of modal semantics for natural languages as well). Ordinary dispositions, such as my irascibility, or the fragility of a glass, are not up to this task; thus, potentialities (in a number of varieties, as introduced in Chapter 4) function as suitable extensions thereof. Chapters 5 and 6 mostly deal with the second and third constraint; Chapter 7, discussing counterexamples, deals with extensional correctness. An appendix, finally, gives some formal details.

Vetter's goal is, primarily, a definition of the possibility operator in terms of potentialities. The definition has to capture the idea that for each possibility, there is a potentiality possessed by some (usually, but not always, concrete) object. Each possibility is thus a potentiality considered in abstraction from its bearer; the former, unlike the latter, is what Vetter calls *non-localized modality*. The definition, so to speak, tells us where to look to find modality in the world: in potentialities possessed by objects. Vetter urges us to think of the localization of possibility operated by potentialities as analogous to the localization of necessity operated by Finean essences.[2]

Potentiality presents a great deal of interesting issues that cannot possibly be covered in this review. There is, however, a big problem for any account of possibility in terms of dispositions, that *Potentiality* brings to the forefront, perhaps for the first time; its resolution occupies a significant portion of the book, and the strategy adopted is, I believe, especially worth discussing. This problem has both formal and informal components, and comes in the form of a gap to be bridged between dispositions and possibilities. The driving idea behind a treatment of possibility through dispositions is a biconditional schema such as:

(D) It is possible that x is P iff x has a disposition to be P.

If the goal is a definition of possibility, **(D)** can only go so far. Not only it is open to many counterexamples (what happens if x does not exist?); a more general problem is that possibility is standardly formalized as a sentential operator,

Counterexamples", *Mind*, 120, 480, 1191-1227, and Vetter, B. 2011a, "On Linking Dispositions and Which Conditionals?", *Mind*, 120, 480, 1173-1189.

[2] However, she discusses (164 ff) the relation between potentialities and Finean essences to the conclusion, that, unlike possibility and necessity, they are not duals. What the precise relation between them is, is left unsaid.

whereas dispositions are naturally linked to predicate modifiers—as in **(D)**. In short, it is not straightforward how to generalize **(D)** into an explicit definition of the diamond operator.

This is where Vetter's strategy of introducing potentialities kicks in; it consists, as she repeatedly points out (e.g., 141) in taking "the path of least formal resistance"; viz., she will help herself with the simplest formal devices to close the logical gap in **(D)**, and only in a second moment she will worry about the metaphysics backing such a formalization. The formal work is indeed quite simple: Vetter allows sentences of arbitrary logical complexity into the scope of a potentiality "POT" predicate modifier, then uses a lambda abstractor to obtain predicates to put in it. This is why most of the potentialities described display the peculiar form "the potentiality to be such that ϕ", or "the potentiality for ϕ". The quantity of potentialities allowed by this strategy is astounding, but Vetter spends considerable energy (throughout Chapters 4 and 5, primarily 148ff) to assure the reader that, from a metaphysical standpoint, they "make sense". E.g., how can an arbitrary *a* have a potentiality to be such that (distinct) *b* is F? That this is a bogus potentiality is suggested by the fact that in the suggested formalization POT[λx.Fb](a) there is no free occurrence of x in the scope of the lambda abstractor. But maybe there is some sense to be made of some *a* having a potentiality for *b* to F; such a potentiality will more likely be extrinsic (for its possession by *a* entirely depends on how *b* is)—and will presumably depend on some further potentiality that *a* and *b* jointly possess. Additionally, in virtue of *a* having a potentiality for *b* to be F, *a* also has a potentiality for something to be F (viz., POT[λx.∃xFx](a))—thus, quantifiers inside the scope of POT are fine as well. Vetter offers some examples, although—as we move farther and farther away from ordinary dispositional ascriptions—they are not as intuitive as perhaps the reader would have hoped. Through potentialities the expressive reach of ordinary dispositions is considerably extended, helping dispositions meet both the *formal adequacy* and the *extensional correctness* constraint. The final step is the admission of iterated potentialities, viz., potentialities whose manifestations are, or involve, the possession of other potentialities. There is no general pattern in this "potentiality-finding" operation (Vetter clearly favors piecemeal approaches to catch-all solutions, e.g., in dealing to counterexamples in Chapter 7). However, she argues negatively (155-157) that any restriction of acceptable potentialities based on their logical form would be highly arbitrary. Eventually, a schematic definition of possibility (197) is formulated:

(D_1) $\Diamond \phi =_{df} \exists x POT^*[\phi](x)$,

where POT* is the sentential operator for iterated potentiality; the *definiens* of **(D_1)** must be read as "something has an iterated potentiality for it to be the case that ϕ". The existential quantification expresses the localization of possibility: modality is tracked down to at least one specific object, possessing a specific potentiality.

The ontology of potentiality is thus generous. Vetter rejects non-trivial necessary conditions for object(s) having joint potentialities (117ff). Spatiotemporal proximity or causal interaction are no matter: one can have a joint potentiality to sing a duet with their best friend, or with Queen Elizabeth (however, *which* potentialities an arbitrary collection of objects can have is presumably subject to some restriction; 175). Furthermore, Vetter frequently suggests that many object(s) that we would say do not possess a given potentiality, actually *do* possess it to a very low degree—perhaps too low to warrant a dispositional ascription.

One might believe this be a drawback. Not only for parsimony considerations, but also because it seems nothing is left of the reassuringly "naturalistic" ontology of dispositions (23). To be sure, this motley crew of potentialities is not entirely left unruly. We cannot stipulate *a priori* which potentialities are instantiated. First (202), because in most cases, this is the job of empirical enquiry.[3] Secondly, aside from tautological potentialities—which are trivially possessed across the board—Vetter never formulates sufficient conditions for some object(s) possessing potentialities; however (114), some restriction is clearly in place, since it is not the case that everything has every potentiality with everything else.

Unsurprisingly, many, perhaps most, of the potentialities needed to treat modality will reveal themselves as non-fundamental potentialities. Vetter embraces this fact: "we are not formalizing the potentialities that are metaphysically basic. We are formalizing any potentiality, no matter how derivative." (142). This is, to my eyes, one of the greatest strength of her account, but it also draws attention to what one can and cannot accomplish through it. One can see its strength: as she very briefly points out (142), predicate logic does not care about the metaphysical status of properties expressed in it. Something similar, I would say, holds in modal logic as well; e.g. **(D_1)** is a schema in which ϕ can be replaced by any well-formed formula, so one can reasonably expect all kinds of gerrymandered possibilities. The moral of the story appears to be that if the goal is to close the formal and informal gap in **(D)**, and obtain an extensionally correct definitional biconditional, one needs all kinds of derivative potentialities as well (195 offers the what could be a slogan: "every potentiality counts"). It is not something that Vetter lingers upon, but *Potentiality* appears to be making the interesting point that if the *desideratum* is a definition of possibility, one should not have a fixation with fundamentality (e.g., 195-98).

Relatedly, one way to appreciate strengths and limitations of the approach is by considering other non-definitional strategies for a disposition-based account of modality. Someone interested in what is fundamental about modality, might decide to opt for a grounding or truthmaking claim; for instance, the claim that modal statements have dispositional truthmakers. But here is the catch: if the quest of truthmakers is to unveil some dispositional "deep story" about modality, the sub-propositional logical form of the truthbearer is not going to provide any insight on the matter. For one may suspect that, say, a disposition for P might be ultimately responsible for the truth that ◊ϕ even without any logical correlation between P and ϕ. Unfortunately, there is not much else that a mere examination of "ϕ" could reveal about its truthmakers (some would claim that there is *nothing* the examination of "ϕ" could reveal about its truthmakers). E.g., what ultimately makes it possible that the Sun rises in the West? It might be all manners of dispositions. The truthmaking claim does not offer any clue; one needs to develop a (presumably *a posteriori*) epistemology of grounding/truthmaking to find out. This holds for every account that abandons a one-to-one correspondence between possibilities and potentialities (and thus, any hope of a *definition)*, in order to focus on the fundamental potentialities. Surely such an account would not be as bothered by the logical form of potentialities (212) if it were just to claim that for each possibility, there is *some* potentiality or another that accounts for it. But then again, *which* potentiality?

[3] The author's epistemology of potentialities/possibilities is a largely *a posteriori* matter—something that Vetter considers "an advantage rather than a drawback" (268).

This puts Vetter's theory at an advantage when it comes to *informativeness*. For Vetter's definition can be applied, and thus affords somewhat of an "algorithm" that pairs every possibility with a potentiality. Mind you: through **(D_1)** one only knows that something has an iterated potentiality to ϕ conditionally on the knowledge that ϕ is possible (and vice versa: as Vetter discusses in relation with the *extensional correctness* constraint, it is a tricky matter to know whether ϕ is possible before applying the definition). This algorithm, and thus, Vetter's definition, is not a device to find potentialities *a priori*, nor to discover a dispositional "deep story" about possibilities; rather, it is a useful biconditional link between *specific* possibilities and *specific* potentialities.

Assuredly, given this "path of least formal resistance", the switch from possibilities to potentialities may appear to be a pointless manoeuvre: they are, after all, two logically isomorphic operators. According to Vetter, however "this general notion of potentiality has not been formulated directly, or simply stipulated out of nowhere; it has taken me four chapters to reach the generalized notion" (194-95). The reasoning seems to be that the shift from possibility to potentiality, admittedly pointless if merely a logical stipulation, acquires significance insofar as potentiality is a metaphysical generalization of more ordinary dispositions and capacities. Whether or not this is sufficient to blunt the charge, is left to be seen. Another limitation of Vetter's theory appears to be that—exactly because of the overabundance of derivative potentialities—it may be *per se* incapable of completing the desired localization of possibility. Consider the possibility that Lorenzo is a musician; it appears to be located in Lorenzo's potentiality to be a musician (viz., for Lorenzo to be a musician). Yet there is a staggering number of other entities having a potentiality for Lorenzo to be a musician, e.g., all my possible teachers, or all the devices I can use to learn. There is nothing wrong with multiple "witnesses" to this possibility (as there can be more than one variable value turning an existential quantification true, 201). This however means that the modality in "possibly, Lorenzo is a musician" has been equally localized in entities all across the world. I doubt that this is what the author intended. One may have the idea that ultimately, the possibility of Lorenzo being a musician is localized in Lorenzo's potentiality to be a musician; perhaps because only in Lorenzo the potentiality for Lorenzo to be a musician is intrinsic (some characterizations of *localized modality* appear to have something to do with intrinsicness, e.g. 103). For Vetter, Lorenzo's intrinsic potentiality is preferable to the extent that is needed to ground all the others. But, in general, the selection of basic potentialities is not something that transpires from **(D_1)**, nor from Vetter's logical framework for the treatment of possibilities.

Overall, what might be a weakness of this "path of least formal resistance" is that metaphysics is clearly taking the backseat, and is left to deal with the results of formalization, whichever those might be. This is a bit odd, considering that the reasons for this whole enterprise come mainly from metaphysics: one develops a disposition-based account of modality, even if possible world semantics is an extremely powerful tool, in order to have a more desirable metaphysics of modality (as the author maintains in Ch. 1).[4] Yet many potentialities (especially the iterated ones, or those with logically complex manifestations) are metaphysically suspicious. This complaint could take many forms: e.g., is it harm-

[4] And earlier, in Vetter, B. 2011b, "Recent Work: Modality without Possible Worlds", *Analysis*, 71, 4, 742-54.

less to quantify over and refer to non-existents in the scope of POT, as Vetter does throughout the book? POT, like ◊, is not factive in the intended interpretation—so one is not straightforwardly committed to such shadowy beings. However, one may think that they should excluded from having any role in the individuation of potentialities, even if manifestations are described by quantifying over, or referring to them. Vetter does not spell out what the metaphysical nature of manifestations is; nor in what sense they "individuate" potentialities. So it is not easy to formulate an argument; but worries still loom.

These doubts do not detract from the value of *Potentiality*. As innovative and detailed as it is, it is bound to raise new questions and breathe new life into old ones. For a crucial thing to notice about this book, is that it is a book. As of now, no other dispositional account of modality is available, as thorough and extensive as this one. Thus, in a way, *Potentiality* wears one of its merits up its sleeve. Vetter does not claim hers to be the only one viable dispositional account of modality, nor the best; she employs little to no time discussing alternatives, nor comparing her account to them: she is almost entirely interested in developing her option. The book shines so much in that department, that it is easy to see it setting somewhat of a bar in the field, that competing accounts will have to measure up to.[5]

Scuola Normale Superiore of Pisa LORENZO AZZANO

[5] I would like to thank Massimiliano Carrara, Giorgio Lando and Barbara Vetter for reading the first draft of this review. In my work about this book, I was also helped by two events. Massimiliano Carrara, Giorgio Lando and Vittorio Morato organized the *Potentiality* workshop at the University of Padua in the early July 2017. Shortly after, in September, there was the *Potentiality & Possibility* workshop at the University of L'Aquila, organized by Giorgio Lando and Simone Gozzano. I would like to thank everyone who attended those workshops for their contributions.

Advisory Board