Descartes discarded? Introspective self-awareness and the problems of transparency and compositionality ☆

Markus Werning

Ruhr University Bochum, Department of Philosophy, Universitätsstr. 150, 44801 Bochum, Germany

A B S T R A C T

What has the self to be like such that introspective awareness of it is possible? The paper asks if Descartes's idea of an inner self can be upheld and discusses this issue by invoking two principles: the phenomenal transparency of experience and the semantic compositionality of conceptual content. It is assumed that self-awareness is a second-order state either in the domain of experience or in the domain of thought. In the former case self-awareness turns out empty if experience is transparent. In the latter, it can best be conceived of as a form of mental quotation. Various proposed analyses of direct and indirect quotation are discussed and tested regarding their applicability to thought. It is concluded that, on the assumption of compositionality, the inner self is only insofar accessible to awareness as it has an accessible phonological (or otherwise subsymbolic) structure, as apparently only inner speech does.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

"I am aware that I exist; I ask who is that I of which I am aware".[1] When René Descartes poses this question right after his seminal *cogito* argument of the Second Meditation, he raises the issue of self-awareness. He does not question the existence of such a self. The existence of the self, rather, is asserted in advance. His question is also not one about the epistemic status of the process of self-awareness: Is it knowledge that we have or not? The Latin original indicates that Descartes, indeed, takes it to be knowledge.[2] No, what Descartes asks is: What has the self to be like such that reflexive awareness of it is possible? It is thus presupposed that the essential property of the self is an epistemic property, viz. being the object of reflexive awareness.

Descartes's own answer is twofold. The negative part consists in the claim that the object of self-awareness in the intended sense is not the body. The positive part is rather indirect: "that which is doubting, understanding, affirming, denying,

---

[1] Latin original: "Novi me existere; quaero, quis sim ego ille, quem novi" (Descartes, 1641/1904, p. 27) – A frequently noted worry roots in a grammatical peculiarity that concerns the unusual application of a demonstrative – *ille* 'that' – to a personal pronoun – *ego* 'I'. Some philosophers have blamed Descartes for using ungrammatical language and even accused him of "a linguistic derailment with terrible consequences" (Beckermann, 2009, p.7). Demonstratives are akin to definite articles in that they are typically applied to generic terms as in *that/the man*. The Latin *ego* as the English *I* is an indexical term – whose referent varies with the speaker and the situational context – and normally neither allows for a demonstrative nor for a definite or indefinite article. The application of a demonstrative to the first person singular pronoun coerces a grammatical type shift and treats the pronoun as a generic term. It so, in a question-begging way, hypostatizes a class of entities: the Is – and consequently my I, your I, her I, etc. To excuse Descartes, one might point out that natural language is very flexible and provides many examples for forced shifts of grammatical type. There, e.g., do exist accepted phrases like *the here and now* where indexicals occur with a definite article. Grammar alone does not suffice to discard Descartes.

[2] *Novi* is the first person present perfect of *noscere* which means 'get to know'.

willing, refusing, imagining, and experiencing."[3] Descartes, so we may justly read, identifies the object of self-awareness with the subject of mental states and he thereby individuates the self as the unity that carries mental contents. This would make self-awareness a case of *introspective* self-awareness or awareness of an *inner* self.

The idea of an inner self contrasts with the idea of a *bodily* self. The latter can be regarded as the object of perceptual or doxastic attitudes turned towards one's own body. These make up the so-called *body image*. The bodily self can also be regarded as given by a *body schema*: a collection of motor programs or habits that enable and constrain movement and the maintenance of posture (Gallagher, 2005). It is yet an open question whether and to which extent body image and body schema are intertwined. Experiments known as the rubber hand illusion and the out-of-body illusion induce an illusionary body schema – either a mislocation or a partial or global misidentification of one's own body – on the basis of visual and tactile stimulation alone (Botvinick & Cohen, 1998; Lenggenhager, Tadi, Metzinger, & Blanke, 2007). This indicates that body image and body schema may in fact be more closely related to each other than the conceptual differentiation suggests.

An essentially diachronic view to be distinguished from the notions of the inner and the bodily self is the proposal of a *narrative self*. It has been suggested to conceive of it is as an abstraction from an autobiographical narrative process that brings together hitherto only minimally coherent pieces of behavior "enhanced by an illusion of greater unity" (Dennett, 1992, see also Ricoeur, 1992).

The present paper takes up Descartes's original question regarding the nature of the object of introspective self-awareness and asks whether Descartes's idea of an inner self as a unitary carrier of mental contents can be upheld or whether it has to be discarded. In the latter case the self would very likely reduce to nothing but the bodily self or would have to be replaced by the quasi fictional idea of a narrative self. I would like to stress that metaphysical questions regarding the legacy of a *res cogitans* as a substance ontologically independent from the physical world are important in their own right, but will not be of concern to us here. The idea that a carrier of mental contents be the object of self-awareness does *per se* not contradict physicalism, as the various contemporary naturalistic accounts of mental content seem to show (see Greenberg, 2005, for review).

Instead, we will discuss Descartes's legacy by invoking two principles that have been moving ever closer into the center of the debate in the philosophy of mind over the past two decades: the principle of the phenomenal transparency of experience and the principle of the semantic compositionality of conceptual content. Experience is said to be phenomenally transparent just in case having an experience of certain objects, events or situations is for the subject just as if the objects, events, or situations were present (see the following section for elaboration). By saying that a conceptual representation is semantically compositional we mean that the content of a complex concept is a structure-dependent function solely of the contents of the parts of the concept (see below).

The two principles cover two aspects – the conceptual and the phenomenal – that are somewhat heterogeneously distributed over the class of mental states. There are mental states like beliefs, desires, recollections, or expectations that are often regarded as conceptual – these will be called thoughts in the course of the paper. There are mental states like perceptions, hallucinations, and proprioceptions that are widely regarded as phenomenal – they will subsequently be discussed under the label of experience. There may also be mental states that have both a conceptual and a phenomenal aspect: many emotions are eventual candidates here. One may even hold that seemingly phenomenal states after some scrutiny reduce to conceptual states.

Even though the transparency of experience and the compositionality of conceptual content regard two different aspects in the class of mental states, they, from a more abstract point of view, are somewhat akin because they both deny any particular role that access to the intrinsic properties of the vehicles of mental content might have in the determination of content. As we will see later on, this is the main reason why the idea of introspective self-awareness is so difficult to accommodate for if the two principles hold.

Given the main mutually non-exclusive, but probably exhaustive dichotomy in the class of mental states – experiences with phenomenal qualities on the one hand and thoughts with a conceptual structure on the other hand – there are altogether three ways to account for introspective self-awareness: (i) It can be construed as a phenomenon completely in the domain of experience, such that the introspective and the introspected states are experiences. For that case will we show that introspective self-awareness is empty if experience is phenomenally transparent. (ii) Introspective self-awareness can be analyzed fully in the domain of thought. In that case we will argue that introspective self-awareness is best conceived of as a mental analogue of direct quotation. Provided that thought is compositional, we will then show that the inner self must have a cognitively accessible subsymbolic structure which is very likely akin to the phonological structure of natural language. Thus mental states would have to be linguistically structured thoughts in order to be introspected. The inner self would reduce to a stream of inner speech. (iii) Introspective self-awareness is a hybrid phenomenon crossing the domains of thought and experience: it would thus either be a non-conceptual experience directed towards a non-phenomenal thought or a non-phenomenal thought representing a non-conceptual experience. Since this is a rather hypothetical option and to my knowledge nobody has ever tried to analyze introspective self-awareness this way, I will spare a discussion of this option due to space limits. The paper concludes with the conditional claim that if experience is phenomenally transparent and thought is semantically compositional, then introspective self-awareness is either empty or reduces to a representation of inner speech.

---

[3] Latin original: "Nempe dubitans, intelligens, affirmans, negans, volens, nolens, imaginans quoque, et sentiens" (Descartes, 1641/1904, p. 28).

Although I will focus more on an explication than on a detailed justification of the conjunctive antecedent of the conditional in this paper, I expect that only few philosophers will be ready to reject the if-clause of the conditional *tout court*. Whereas compositionality is almost unanimously accepted as a constraint on thought, the situation with regard to the phenomenal transparency of experience, I concede, is more complex. Here the main question is whether perception and eventually other experiential states are regarded as having conceptual or non-conceptual content. If experiences were construed as conceptual (see, e.g., McDowell, 1996), they would count as instances of thought in the sense of this paper. The compositionality conjunct of the antecedent would apply and the truth of the conditional would make its consequent unavoidable even if one were to deny the phenomenal transparency of experience. Only for non-conceptualist views of experience the transparency conjunct is essential to establish the consequent. Note, however, that most non-conceptualists, prominently Dretske (1981/2003b,2003a), do endorse the transparency thesis or related claims (e.g., direct realism). So even if the main claim of this paper is limited to a conditional, the foregoing remarks illustrate that logical space leaves remarkably little room to avoid the antecedent of the conditional and hence its consequent.

2. Phenomenal transparency

This issue of phenomenal transparency was first advanced by G. E. Moore. In his *Refutation of Idealism* he writes:

> [...] that which makes the sensation of blue a mental fact seems to escape us; it seems, if I may use a metaphor, to be transparent – we look through it and see nothing but the blue; we may be convinced that there is something, but what it is no philosopher, I think, has yet clearly recognized. (Moore 1903, p. 446)

In a slightly more poetic way Martin Heidegger formulated the same idea independently and made it a cornerstone of his philosophy:

> We never originally [...] perceive a throng of sensations, e.g., tones and noises, in the appearance of things [...]; rather we hear the storm whistling in the chimney, we hear the three-engine aeroplane, we hear the Mercedes in immediate distinction from the Volkswagen. Much closer to us than any sensations are the things in themselves. (Heidegger, 1935/1977, p. 156)

Gilbert Harman re-introduced the notion of phenomenal transparency into modern analytic philosophy:

> When you see a tree, you do not experience any features as intrinsic features of your experience. Look at a tree and try to turn your attention to intrinsic features of your visual experience. I predict you will find that the only features there to turn your attention to will be features of the presented tree, [...] (Harman, 1990, p. 39).

What the three authors point to is the fact that, when we experience something, we do not experience the intrinsic properties of the vehicles of the experience – what Moore and Heidegger call "sensations".[4] To give a positive account of phenomenal transparency – what it is for the subject to have such an experience rather than what it is not – and to avoid metaphorical language, I propose a definition that uses a counterfactual as-if construction:

Definition *phenomenal transparency*

> A subject's *S* experience of a scenario *X* is called phenomenally transparent just in case *S*'s having an experience of *X* is for *S* just as if *X* were present.

As the *scenario* somebody experiences we understand the totality of the objects and events with their properties and relations that make up the content of the experience at a time. The as-if analysis of transparency exhibits a number of features that allude to aspects prominent in traditional philosophical characterizations of experience and are worth closer investigation. First, the definition presupposes a relation of ownership between the experience of the scenario and the experiencing subject. Moreover, it also covers the aspect of subjectivity in that it claims that, *for the owner* of the experience, having the experience is as if the scenario were present. For a third person, who studies the experience, say, neurophysiologically, the experience may well be different.

Second, our definition contrasts with the position of direct realism which would license the inference from *S experiences X* to *X is present*. This inference is blocked because the as-if clause is counterfactual. So even if the scenario is not present, having the experience might well be for the subject as if the scenario *were* present. "Experience" in our context is not a success word and the definition is not restricted to veridical experience but may as well hold for hallucination. However, what is sometimes called the "feeling of realness" (Brown, 2004) is intended to be entailed. This feeling of realness is typical for perceptions and proprioceptions, but also comes with many hallucinations and illusions. The notion of presence is even stronger

---

[4] Metzinger formulates what he calls the "standard definition" of phenomenal transparency as follows: "Phenomenal states are transparent in that only their content-properties are introspectively accessible to the subject of experience" (Metzinger, 2003, p. 354). In my eyes it is misleading to presuppose that the content-properties are accessible introspectively. For, it is not introspection, but rather perception or – as in the case of bodily states of one's own – proprioception that makes the content-properties available to the subject. However, since the idea of phenomenal transparency denies that there are accessible intrinsic properties of the vehicles of experiences, the whole idea of introspection becomes at least questionable and we should not define phenomenal transparency in terms of introspective accessibility in the first place.

than that of realness since it does not only entail that the scenario is real for the subject, but also that there is some spatio-temporal immediacy of the scenario for the subject.

Finally, the definition contains an important minimality constraint. The use of the word *just* implies that in the counter-factual condition nothing is present for the subject, but the objects and events with their properties and relations of the scenario that make up the content of the experience. As we will see, the minimality constraint is crucial for the problem phenomenal transparency creates for purely experiential accounts of introspective self-awareness.

3. The experiential account of introspective self-awareness

One way to think of introspection is to view it as a phenomenon fully in the domain of experience. Introspection might be regarded as an experience of the experiences one has at a certain time. The idea is that a person might, e.g., see a tree and then in addition experience that she sees a tree. She might have a toothache or feel a sore tooth and then experience that she has a toothache or feels a sore tooth. Introspection according to this view would be a second-order experience.

However, would introspection construed as a second-order experience lead to self-awareness? Would it license the inference to a self? The answer very likely is "no" if experience is phenomenally transparent. To see this, let's first turn to proprioceptions like pain. There are two ways of thinking about pain (and most other proprioceptions). According to the first conception, pain does not necessarily involve consciousness, but simply is a state of the body that one occasionally experiences when one *feels* pain. Under this view feeling pain is a first-order experience and does not at all qualify as an introspective state. According to the alternative conception, pain is conceived of as a state that necessarily involves consciousness. Here, pain is itself a state of experience rather than a content of experience. Having a toothache, according to this view, is the experience of a sore tooth (of one's own). Turning to the second-order experience we now face a dilemma: Experiencing the presence of a sore tooth and experiencing a toothache – i.e., having the second-order experience – are for the subject either the same or not the same. If they are the same for the subject, there is no reason for the subject to infer the existence of anything over and above the tooth and its soreness on the basis of her second-order experience just as experiencing the presence of a sore tooth would not license any such inferences. Any inference to an internal vehicle of the proprioception as part of an inner self would be unmotivated by the second-order experience.

If experiencing the presence of a sore tooth and experiencing a toothache, on the other hand, are not the same for the subject, then phenomenal transparency is violated. For, if the experience of a toothache were phenomenally transparent it should, by definition, be for the subject just as if a toothache were present. Now again, the presence of a toothache for the subject under the second conception of pain should be just the same as experiencing the presence of a sore tooth.

Unfortunately, the states of proprioception do not by introspection contribute to an awareness of the self if phenomenal transparency holds. The same is true for perception: Assume that the perception of a tree be itself an object of introspective experience. The phenomenal transparency of experience would entail that being in the perceptive state and the introspective state at the same time would be for the subject just as if a tree and the perception of a tree were present. After another application of the transparency property we see that the latter – the perception of a tree – is for a subject just as if a tree were present. It follows that being in the combined – perceptive plus introspective – state is for the subject just as if a tree and the same tree were present. It hence is just like being in the state of perception without being in the state of introspection. Introspection adds nothing to perception and gives no cues to an inner self if experience in general is phenomenally transparent.[5]

4. Second-order thoughts

Introspective self-awareness might also be analyzed in terms of conceptually structured thoughts rather than as a form of phenomenal experience. The property of phenomenal transparency would not apply. In those terms an introspective state would be a second-order thought of sorts that reports another thought of one's own. What we are looking for seems closely analogous to – yes, indeed, structurally the same as – quotation in natural language. A quotation in natural language is an utterance that reports another utterance. One generally distinguishes two kinds of quotation: direct quotation and indirect quotation. The reason is that an utterance can either be reported by reference to its vehicle – the actual expression uttered – or by reference to its content or meaning – what was said by the utterance. Compare the following:

> *Donald said, 'In Mark's garden furze is more widespread than ivy'.* (1)

> *Donald said that in Mark's garden furze is more widespread than ivy.* (2)

The truth-conditions of the two sentences differ. The words that were actually used in the reported utterance make a difference regarding direct quotation whereas they do not have an effect on the truth of the indirect quotation if content is left unchanged. This holds for all kinds of verbal variations that would leave content unchanged: (a) Choice of language: Had Donald said what he said in German, (1) would no longer be an accurate report of the utterance, while (2) would remain true. (b) Substitution of synonyms: For the truth of (2) it does not matter whether Donald used the word *gorse* or *furze* whereas it does matter for (1). (c) Substitution of co-referential rigid designators: Assume that *Mark* and *Samuel* be proper

---

[5] The argument of this section largely builds on Dretske (2003). For a closer discussion see Werning (2004a).

names that refer to the same person – as would be the case for *Mark Twain* and *Samuel Clemens*. Then the truth of (2) would remain unaffected even if Donald actually used the word *Samuel* to refer to the garden in question.[6] (d) Analytical or logical equivalence: The utterance *In Mark's garden, ivy is less widespread than furze* would be truly reported by (2), but not by (1). Furthermore, in specific contexts – where e.g., conversational focus, metrics, dialect, or speech impediments are at issue – the accuracy standards for direct quotation may be raised so that even word order, prosody, phonology, or phonetics play a role. All those aspects are irrelevant for indirect quotation because here only the content, not the vehicle of the utterance is at issue. There are numerous other aspects in which direct and indirect quotation differ – the role of indexicals, anaphoric reference, tense, etc. – but these issues will not be dealt with here.

When we now turn back to self-reporting second-order thoughts, we have a choice between conceiving of them as reports of the internal vehicles of other thoughts of the subject – i.e., as direct mental quotations – or as reports only of the contents of those thoughts – as indirect mental quotations. What would those choices imply for the possibility of self-awareness?

For second-order thoughts of the latter kind Fernández (2003, 2005) has coined the term *extrospection*. He argues that any piece of evidence – perception, memory, testimony – that justifies a subject's belief that p also justifies the subject's belief that she believes that p if the subject has formed the belief that p and the belief that she believes that p on the basis of this evidence. His argument presupposes a reliabilist notion of justification and identifies the justifiedness of a belief with being the product of a belief formation process that is grounded in some reliable regularity. The argument can roughly be summarized as follows: If a properly functioning epistemic subject has good evidence for the truth of the proposition p, she normally comes to believe that p. Due to this mind-to-mind regularity, a subject that forms the belief that p on the basis of good evidence for the proposition p is also justified, on the basis of this evidence, to believe that she believes that p. Access to the internal vehicle of the first-order thought is thus not required to justify the second-order thought. One only has to have access – this is the role of good evidence – to the content of the first-order thought, the proposition p, to justify the second-order thought. Thus self-knowledge would be sufficiently grounded in extrospection.

I am in general very sympathetic with this account of self-knowledge. The question in our context however is whether extrospective self-knowledge qualifies as knowledge of an *inner* self. Only thus Descartes's original idea could be defended. I take it that this was not even in the intentions of Fernández's. Just as indirect linguistic quotation gives us only few hints as to which linguistic expressions – which language, which words, and which syntactic structures – were actually used by the reported speaker, extrospection, viewed as indirect mental quotation, provides us with almost no clue with respect to the nature, the constituents, and the structure of the reported thought. All it provides us with is the external content of the latter. The inner self remains in the dark.

Self-reporting second-order thoughts might also be construed in analogy to direct quotation. Here we may justly speak of *introspective* thoughts because thoughts of one's own are reported by reference to their *internal* vehicles. The challenge here is to find a proper way to analyze direct quotation and apply this analysis to thought. In this regard it is of particular significance that thought – as is widely assumed – is compositional. The principle of semantic compositionality is defined in most general terms for representational structures like language or thought that consist of (i) a set of representational vehicles and (ii) a set of combinational operations mapping arrays of constituent vehicles onto a complex vehicle, and for which (iii) a function of semantic evaluation is defined on the set of vehicles:

Definition *semantic compositionality*

A representational structure (i.e., language or thought) is called semantically compositional just in case the semantic value (meaning, or respectively, content) of any complex vehicle (grammatical term or, respectively, mental concept) is a function of the semantic values of its constituent vehicles where the choice of the function depends solely on the operation by which the constituents combine to form the complex.[7]

As Fodor (2001) has pointed out, compositionality is much less controversial for concepts (including truth-evaluable thoughts) with respect to their contents than it is for linguistic expressions with regard to their meanings. Even though many *a priori* justifications for the compositionality of meaning in natural language have been put forward more or less convincingly – learnability, productivity, systematicity, communicative efficiency (for a critical discussion of those arguments see Werning, 2005; Pagin & Westerståhl, in press) – it is not keen to say that a final verdict has not yet been spoken. In particular many tenacious linguistic counterexamples have yet to be defused.

In the case of thought the case seems clearer: What could matter for the contents of the constituent concepts to determine the content of the complex concept other than their combinational structure? In natural language extralinguistic factors that lie neither in syntactic structure nor in the meanings of the constituent expressions may eventually matter to determine the meaning of a complex expression. Think of ostensive gestures that help to determine the reference of demonstratives (see Kaplan, 1989). Or imagine background situations that play a role for the determination of truth-conditions. In sentences like *All guests have gone* uttered at a party or *Some passengers have died* said after a train accident, the situation obviously restricts the domain of the quantifier – if quantifiers ranged over an unrestricted domain, sentences of this kind would either be trivially false or trivially true. In those cases one is typically required to allow for explicit or implicit

---

[6] It is widely, though not unanimously accepted that co-referential rigid designators may be replaced in indirect quotations without changing their truth-conditions. This view grounds in the possible-worlds semantics of propositional attitudes.

[7] For formalizations see Hodges (2001) and Werning (2004b).

contextual arguments in order to force the semantic analysis of a sentence into a compositional framework. This strategy only works because one may assume that both the speaker and the hearer of such a sentence do represent the relevant factors mentally – typically on the basis of perception or memory – even if the factors are not contained in the learned meanings of the constituent expressions themselves. In the domain of thought this strategy is not viable because contextual factors may matter to determine the content of a complex concept only if they are mentally represented. They are thus not extra-representational, but belong to a larger conceptual structure internal to the mind of the subject. To thought, compositionality applies without exception.

If we are hence to construe introspective thoughts in analogy to direct quotations, we have to make sure that the analysis of direct quotation used abides by the principle of compositionality. This is an important point because many accounts of direct quotation in natural language take it virtually as an axiom that quotation be non-compositional. Cappelen and Lepore (2009), e.g., identify the first "basic quotational [feature] of particular importance" with the principle that "in quotation you cannot substitute co-referential or synonymous terms *salva veritate*." If you cannot substitute synonymous terms without a change in the truth-conditions of a sentence, you cannot substitute synonymous terms in a sentence without a change in meaning. Now, the latter, the substitutability of synonyms *salva significatione* is logically equivalent to compositionality (as proven by Hodges, 2001). Cappelen and Lepore's view of quotation hence logically entails that direct quotation be non-compositional. The widespread conviction that direct quotation should be treated in a non-compositional way – quasi as an exception to the principle of compositionality – is probably the reason why so few explicitly compositional treatments of direct quotation can be found in the literature. To my knowledge only three papers, Potts (1997), Pagin and Westerståhl (in press), and Werning (2005), have ever tried so. Due to space and time limits, I will not discuss the first two. A closer investigation, however, would reveal that the analyses Potts and Pagin and Westerståhl provide for direct quotation in natural language cannot be transferred to the domain of thought because they are either not properly compositional or they essentially make use of extralinguistic contextual arguments.

5. Direct quotation and compositionality

To illustrate what kind of questions the compositionality requirement raises for the treatment of direct quotation, I will first turn to the most natural and still very widespread, so-called disquotational analysis of quotation. To make this analysis explicit, we first need a rudimentary formal framework. According to the standard view, syntactic operations are not operative on the set E of expressions of the language directly, but on a set T of underlying grammatical terms. The distinction between expressions and terms is useful to deal with lexically or syntactically ambiguous expressions such as *bank* or *The boy watches the girl with the telescope*. By convention, expressions are set in italics. An expressing function ε with

$$\varepsilon : T \rightarrow E \tag{3}$$

maps the set of terms onto the set of expressions and a meaning function μ with

$$\mu : T \rightarrow M \tag{4}$$

maps the set of terms onto the set of meanings. The grammar of a language is a pair $\langle T, S \rangle$ where S is the set of syntactic operations. Each syntactic operation is a partial function on some Cartesian product of T into T. As a useful convention that allows us to reduce notational complexity we set primitive or complex expressions in indexed angel brackets in order to disambiguate them and to denote the underlying grammatical terms, e.g., $\langle bank \rangle_1$ and $\langle bank \rangle_2$. If an expression is unambiguous the index, by convention, is left away.

In the disquotational analysis a syntactic operation of quotation q is adopted into the grammar. It is a total function defined as follows:

$$q : T \rightarrow T \text{ such that } \varepsilon(q(s)) = \text{`}\varepsilon(s)\text{' and } \mu(q(s)) = \varepsilon(s). \tag{5}$$

The grammatical term s that underlies the quoted expression $\varepsilon(s)$ is the argument of the syntactic operation of quotation q. The expression of a quotation, i.e., $\varepsilon(q(s))$, is the concatenation of the onset quotation mark, the quoted expression, and the offset quotation mark: '$\varepsilon(s)$' – as is common, the operation of concatenation is left implicit to simplify notation. The meaning of a quotation is the quoted expression.

This account of quotation might be called holophrastic because it takes the quoted expression as a whole and sets it in quotation marks. It can be shown that the disquotational analysis violates compositionality, provided the language to be considered contains synonymous expressions, e.g., the pair *gorse* and *furze*:

$$\mu(\langle gorse \rangle) = \mu(\langle furze \rangle) \tag{6}$$

Although the two expressions are synonymous, they are not identical:

$$gorse \neq furze. \tag{7}$$

From our definition of q we derive the following:

$$\mu(q(\langle gorse \rangle)) = gorse. \tag{8}$$

$$\mu(q(\langle furze \rangle)) = furze. \tag{9}$$

From (7), (8), and (9) we may infer:

$$\mu(q(\langle gorse \rangle)) \neq \mu(q(\langle furze \rangle)). \tag{10}$$

If we, for the sake of the argument, assume compositionality, the meaning of a syntactically complex term $q(s)$ should be some function f_q of the meanings of its syntactic part s:

$$\mu(q(\langle gorse \rangle)) = f_q(\mu(\langle gorse \rangle)). \tag{11}$$

Substitution of identicals according to (6) yields:

$$\mu(q(\langle gorse \rangle)) = f_q(\mu(\langle furze \rangle)). \tag{12}$$

Due to compositionality it also holds that:

$$\mu(q(\langle furze \rangle)) = f_q(\mu(\langle furze \rangle)) \tag{13}$$

From (12) and (13), we hence get:

$$\mu(q(\langle gorse \rangle)) = \mu(q(\langle furze \rangle)). \tag{14}$$

We have derived a contradiction with (10). The hypothetical assumption that the language be compositional must consequently be rejected.

Similar arguments can be made for all holophrastic analyses of quotation provided they treat the contents of expressions in a context-invariant unambiguous way. To see this, one may adapt the argument to Davidson's (1984) demonstrative analysis of direct quotation, which is another case of holophrastic quotation. As in any such analysis, quotation is taken as a syntactic operation $q: T \rightarrow T$ such that $\varepsilon(q(s)) = \text{'}\varepsilon(s)\text{'}$. In the demonstrative analysis, for any sentential term $\varphi(\xi)$ in which a quotation may occur at position ξ, the meaning is given as the meaning of a discourse with backward demonstrative reference:[8]

$$\mu(\varphi(q(s)|\xi)) = \mu(s \cdot \varphi(\langle that \rangle|\xi)). \tag{15}$$

Example: The meaning of

John said, 'Furze is beautiful'   (16)

is given as the meaning of the two-sentence discourse

Furze is beautiful. John said that.   (17)

If we were licensed to assume compositionality, we should be allowed to replace synonyms in the discourse without changing its meaning. (17) should hence be synonymous to

Gorse is beautiful. John said that.   (18)

However, this discourse according to (15) would have the same meaning as

John said, 'Gorse is beautiful'.   (19)

We have again derived a contradiction since (16) and (19) obviously have different truth-conditions. Compositionality must not be presupposed in the demonstrative analysis either.

6. Phonological quotation

In two representative examples we have shown that any holophrastic treatment of direct quotation leads to a violation of the principle of compositionality unless one is ready to allow for extralinguistic context arguments.[9] Permitting the latter would, however, block any transfer of the analysis from language to thought. Any analysis of direct quotation applicable to the domain of thought and thus possibly accounting for introspective self-awareness, must hence be non-holophrastic. It must refer to features of the quoted expression that are below the level of the whole expression. Such an analysis is justly called subsymbolic.

To show that quotation in natural language can be analyzed in a compositional way without extralinguistic context arguments, I will here introduce the method of phonological quotation as an example of a subsymbolic analysis of direct quotation. The idea partly builds on Geach's (1970) descriptive analysis of quotation. What's new in the phonological analysis is that the means of reference to expressions is description by imitation. An expression is described by imitating its phonological structure. The phonological analysis is on par with alternative subsymbolic analyses that refer to expressions by imitating and describing graphical features of inscription or gestures of sign languages.

---

[8] The substitution operator | in a term $\varphi(s|\xi)$ reads: replace the variable ξ – i.e. a placeholder that indicates a certain position – in the term $\varphi(\xi)$ with the term s. The syntactic full-stop operator . builds discourses from its left and right argument terms.

[9] The formal proof is immediate and left to the reader.

Let us assume that each expression of our language be a temporal sequence of primitive phonological parts, e.g., phonemes, from a finite set. The expression *dog*, e.g., is a sequence of the plosive voiced dental consonant *d*, the closed-mid back vowel *o*, and the plosive voiced velar consonant *g*. To enable reference to the expressions of the language within the language, the set of terms *T* of the language has to include as a subset a set of atomic terms *P* that denote the elementary phonological parts necessary to generate each and every expression of the language. Neglecting the details of phonology, a crude approximation might be the following (capital letters – except the labels of sets – are used to refer to object language terms that are names of phonemes, whereas lower-case characters in our meta-language are used to refer to phonemes when they are not used as variables. The space bar _ denotes a distinctive length of silence):

$$P = \{A, B, \ldots, Z, \_\} \tag{20}$$

$$\mu(A) = a, \mu(B) = b, \ldots, \mu(Z) = z, \mu(\_) = \text{x milliseconds of silence}. \tag{21}$$

The phonological analysis of quotation to be developed here refers to phonemes and sequences of phonemes by imitation. Formally, the idea of imitation is captured by the fact that the phonemic terms are themselves expressed by the phonemes they denote:

$$\varepsilon(A) = a, \varepsilon(B) = b, \ldots, \varepsilon(Z) = z, \varepsilon(\_) = \text{x milliseconds of silence}. \tag{22}$$

To achieve reference to complex phonological sequences, a binary syntactic operation of concatenation $\frown$ is adopted into the grammar. The set of phonological descriptions is the closure P^* of the set P with respect to concatenation.[10] The syntactic operation of concatenation has a unique semantic counterpart operation: the function f_s that maps pairs of phonological sequences to the sequence of them. Compositionality is warranted by the following homomorphism, with the variables s and t in P^*:

$$\mu(s \frown t) = f_s(\mu(s), \mu(t)). \tag{23}$$

To enable description by imitation also for complex expressions, the expressing function is to satisfy a similar constraint:

$$\varepsilon(s \frown t) = f_s(\varepsilon(s), \varepsilon(t)). \tag{24}$$

Notice that all expressions of the language can be described by phonological descriptions, but usually not all concatenations do actually describe some word or some sequence of words of the language: The concatenation $O \frown G$ denotes the phonological sequences *og*, which is not a meaningful English word, while $D \frown (O \frown G)$ denotes the English word *dog*. The meaning of the quotation is derived compositionally:

$$\mu(D \frown (O \frown G)) = f_s(\mu(D), \mu(O \frown G)) = f_s(d, f_s(\mu(O), \mu(G))) = dog. \tag{25}$$

Likewise the expression of the quotation is generated stepwise:

$$\varepsilon(D \frown (O \frown G)) = f_s(\varepsilon(D), \varepsilon(O \frown G)) = f_s(d, f_s(\varepsilon(O), \varepsilon(G))) = dog. \tag{26}$$

Quotation is hence expressed by an imitation of the phonological structure of the quoted expression. Be aware, however, that the expression of the quotation and the quoted expression are given different syntactic analyses, even though they exhibit the same phonological structure. In (25) the quoted expression *dog* is regarded as syntactically unstructured. It is the expression of the primitive, lexical term $\langle dog \rangle_1$:

$$dog = \varepsilon(\langle dog \rangle_1). \tag{27}$$

In (26) *dog* is regarded as the expression of the quotation $D \frown (O \frown G)$:

$$dog = \varepsilon(D \frown (O \frown G)). \tag{28}$$

Keep in mind that in spoken language there is no phonological expression of the quotation marks we use in writing. The written expressions '*dog*' and *dog* correspond to one and the same phonological sequence, and hence to one and the same expression in spoken language (however, see option three in the passage on higher-order quotation below). Notice that phonology is all there is to the identity of expressions in spoken language. The phonological analysis of quotation is here developed for spoken language and, thus, systematically introduces ambiguities on the level of expressions. These ambiguities, however, are syntactically resolved and do not pose any threat to a strict application of the principle of compositionality. For, compositionality regards the relation between terms and meanings, rather than the one between expressions and meanings. We may sum up that unlike holophrastic analyses, the phonological treatment provides a compositional analysis of direct quotation and avoids context arguments and syntactically unresolved ambiguities.

I am aware that the phonological analysis is not a silver bullet against all sorts of difficulties quotation brings about in natural language. As with other analyses of quotation the treatment of mixed quotation is problematic (see Reimer, 2003). Here the quoted expression is used and mentioned at the same time (as in: *Quine said that quotation 'has a certain*

---

[10] To guarantee that terms are uniquely identified by a nested structure of syntactic operations, we adopt the convention that phonological sequences are described from the end. The innermost concatenation refers to the sequence of the final two phonemes of an expression. Thus $D \frown (O \frown G)$ is a well formed term while $(D \frown O) \frown G$ is not.

anomalous feature'). However, many other problems often brought forward against the phonological theory are less severe.[11]

Aside from being compositional, the phonological analysis has considerable further advantages over holophrastic analyses. Whereas holophrastic analyses only allow for the quotation of expressions with an existing underlying term, the phonological analysis immediately explains why non-words can be quoted, as in:

> *Hans always says 'shnoff' when he speaks of snow.* (29)

Holophrastic analyses furthermore have problems with explaining the quotation of syntactically ambiguous expressions. For, the syntactic operation of quotation takes as arguments only syntactically disambiguated terms. The phonological analysis has no problems here because it just recurs to the phonological surface of an expression. In a holophrastic framework there is, e.g., no syntactic analysis of the following sentence that renders the sentence true:

> *'The boy watches the girl with the telescope' is ambiguous.* (30)

Of particular interest is the question whether phonological quotation as a way of description by imitation is consistent with the assumption that quotations can be recursively embedded, as in:

> *"dog" refers to 'dog', which refers to dogs.* (31)

At first glance there seems to be no difference between the imitation of the imitation of the phonological structure of an expression and the simple imitation of the phonological structure of the expression. Both sound the same and, more importantly, have the same underlying syntactic structure. One option is to bite the bullet and argue that higher order quotation is in fact a notorious problem in spoken language. Only auxiliary conventions like reading the quotation marks aloud – *quote dog unquote* vs. *quote quote dog unquote unquote* – might be a means to circumvent the problem. A second option is to appeal to speech act theory and argue that, even though the phonological and syntactic structure of the expressions of a second and a first order quotation are the same, the illocutionary force and consequently the intentions of speaker and hearer are different. The quotation marks in writing could then be interpreted as signs of an imitative illocutionary act. Their role would thus be somewhat analogous to punctuation marks like the question mark or the exclamation mark. In many languages question marks and exclamation marks can, e.g., be placed after sentences that are otherwise indistinguishable from assertions. In those cases the punctuation marks indicate nothing but the change of illocutionary force. A third, rather technical option could be the following: We stipulate as a rule that each imitation start and end with a certain period of silence – or some other phonological marker: $\_dog\_$ would be the expression of the first order quotation $D\frown(O\frown G)$; $\_\_dog\_\_$ the expression of the second order quotation $\_\frown(D\frown(O\frown(G\frown\_)))$; and so on. Notice that higher order quotations have to make explicit reference to the periods of silence in the expressions of lower-level quotations. This is because a full description of the lower-level expressions from the start to the end of the imitation has to be given. In this approach first and second order quotations now differ phonologically and syntactically. The quotation marks we use in writing could be interpreted as expressions of those periods of silence or simply as analogous tools.

Regardless of whether the phonological analysis is the optimal treatment of direct quotation in natural language, what matters in the context of this paper is that it is the only strictly compositional analysis and as such the only one that may be transferred to the domain of thought – on a par with other comparable subsymbolic analyses. If self-awareness is regarded as self-reporting second-order thought that achieves access to an inner self, it has to be treated in analogy to direct quotation. Introspective thought is a form of direct mental quotation. However, if phonological or subsymbolic quotation is the only viable analysis of quotation that may be applied to the domain of thought, we may infer that awareness of an inner self taken as a second-order thought must be a form of phonological quotation or closely analogous to it.

7. Introspective self-awareness as the awareness of inner speech

The conclusion that awareness of an inner self must be akin to phonological or otherwise subsymbolic quotation – think of representing or simulating the gestures of sign languages – has far reaching consequences. The most immediate consequence is that thoughts reported by introspective thoughts must have an accessible phonological or quasi-phonological structure (e.g., a subsymbolic gestural structure). The best and as far as I can see only candidate here is inner speech: the mental simulation of phonologically (or gesturally) structured speech. Inner speech seems to be a prerequisite for the awareness of an inner self. A second consequence is that thoughts cannot be introspectively reported if their constituents fail to have a phonological or otherwise subsymbolic accessible surface. This implies that the inner self is accessible only insofar

---

[11] It is true, for instance, that in the phonological analysis the relation between the quotation and the quoted expression is less close than in holophrastic analyses – an often heard objection (Cappelen & Lepore, 2009) – because the quoted expression is not a syntactic constituent of the quotation, but merely referred to. I take this more as a strength than as a weakness. A sentence like *'Dog' refers to dogs*, e.g., can no longer be regarded as *a priori*. Recall that the alleged aprioricity of such sentences generates various philosophical paradoxes. Take, e.g., Putnam's (1981) famous brain-in-a-vat argument in the reconstruction of Wright (1992), shortest version: *In my language 'dog' refers to dogs. In the language of a brain in a vat 'dog' does not refer to dogs. Therefore, I am not a brain in a vat.* The problem with this argument is not its validity, but the alleged aprioricity of its two premises – the first justified by disquotation, the second by the causal theory of reference. Isn't it paradoxical that I should be able to know *a priori* that I am not a brain in a vat? According to the phonological analysis, the first premise would not be *a priori* anymore, but boils down to something like the empirical statement *A word referring to dogs in my language is the sequence of the plosive voiced dental consonant, the closed-mid back vowel, and the plosive voiced velar consonant.* This sentence after some phonological scrutiny may eventually even turn out false.

as it is constituted by inner speech. If one furthermore makes the assumption that epistemic accessibility is an essential property of the inner self, as Descartes apparently did, the inner self reduces to that part of our mind that simulates phonologically or otherwise subsymbolically structured speech. One may thirdly infer that some competence of natural language is necessary to have introspective self-awareness. It is needless to say that this has numerous ontogenetic and phylogenic consequences. A fourth implication is that any impairment of our capacity to simulate inner speech should lead to an impairment regarding the inner self. There is indeed ample evidence for such a correlation (see Morin, 2005, for review). Such dysfunctions might be due to a stroke in language-related cortical areas, be caused by developmental retardation, or be grounded in psychiatric pathologies.[12]

Before we conclude, let me briefly summarize the overall argumentation of the paper:

(P1) Self-awareness is a second-order state either in the domain of experience or in the domain of thought.
(P2) If self-awareness is a second-order experience, and experience is phenomenally transparent, then self-awareness fails to achieve awareness of an inner self and is as such empty.
(P3) If self-awareness is a second-order thought, then it is either (a) identified with a thought that is analogous to indirect quotation or (b) with one that is analogous to direct quotation.
(P4) In the case of (a) self-awareness is an instance of extrospection, fails to achieve awareness of an inner self and is as such empty.
(P5) Direct quotation is either analyzed holophrastically or phonologically (or otherwise subsymbolically).
(P6) Holophrastic analyses of direct quotation fail to be strictly compositional whereas the phonological (or any likewise subsymbolic) analysis of quotation is strictly compositional.
(P7) Thought is strictly compositional.
(P8) Experience is phenomenally transparent.
(C1) Hence: self-awareness is either empty as awareness of an inner self or analogous to phonological (or otherwise subsymbolic) quotation.
(C2) Corollary: the inner self is only insofar accessible to awareness as it has an accessible phonological (or otherwise subsymbolic) structure, as apparently only inner speech does.
(C3) Corollary: if being the object of self-awareness is essential to the inner self, then it apparently reduces to nothing but a stream of inner speech.

I am aware that the argument of the paper explores only some portion of the relevant logical space. One might, e.g., question the exhaustiveness of the dichotomy in P1. Are experiences and thoughts – both understood in a broad sense – the only major categories of mental states? What about emotions?[13] Even though there is no room for a closer discussion here, I would suggest distinguishing at least two groups: We have the group of low-level emotions like hunger and pain, which qualify as proprioceptive experiences and might be regarded as phenomenally transparent (see the foregoing treatment of pain). We also have the group of high-level emotions like hope and disappointment. These most likely are predominantly composed of thought-like states that involve beliefs, desires, and expectations. Here the compositionality requirement applies. In the context of introspective self-awareness and in the light of the argument in this paper, it should be of major concern whether the two groups exhaust the class of emotions or whether there are emotions that are neither experience-like nor thought-like.

One may of course also raise more principled objections and question the compositionality of thought or the transparency of experience. It is a remarkable fact, though, that in the argument the property of phenomenal transparency does a similar job for experience as the property of compositionality does for thought. Despite potential points of criticism, each of the premises P1 to P8 seems to be rather well justified to me in its own light. This is especially important since the conclusions strike me as rather far-reaching.

What about Descartes's idea of an inner self? We have largely, however, not completely discarded this idea. Much of what Descartes suggested as belonging to the inner self can no longer be regarded as a part of it if being the object of introspective awareness is essential to the inner self. Still, there is a remainder: linguistically structured inner speech. This conclusion, of course, kicks off a whole cascade of follow-up questions: Do we have to postulate an author behind inner speech that remains unobserved in respects other than the phonological/subsymbolic structure we are introspectively aware of? Is there an issue of ownership? Does the phonological/subsymbolic structure of which one becomes aware have to be attributed to oneself in order to constitute the inner self of one's own? May this attribution of ownership eventually even go wrong so that the speech-like structure is attributed to nobody or to somebody else? Do features internal to inner speech suffice to establish the unity of an inner self? Speech-internal features that indicate unity might be semantic such as coherence, syntactic as, e.g., anaphora and other means to interweave sentences into a discourse, and finally phonological and phonetic such as accent or voice. Similar questions arise for the uniqueness of the inner self. I am optimistic that the issues of authorship, ownership, unity, and uniqueness can in principle be dealt with in an inner speech account of the self. We may hence hope to

---

[12] A good candidate for a correlation between inner speech deficits and an impaired inner self in psychiatry are some forms of schizophrenia as pointed out by Hoffman (1986). See Stephens and Graham (2000) for discussion.

[13] For a discussion of the phenomenal transparency of emotions see Metzinger (2003). Imaginations also constitute a possible exception to transparency, for discussion see Werning (2004a).

save at least some fragment of Descartes's idea of an inner self without falling back into the Cartesian metaphysics of a thinking substance and related conceptions of the "I".

References

Beckermann, A. (2009). Es gibt kein Ich, doch es gibt mich. In M. Fürst, W. Gombocz, & C. Hiebaum (Eds.), *Gehirne und Personen–Beiträge zum 8. Internationalen Kongress der Österreichischen Gesellschaft für Philosophie in Graz* (pp. 1–17). Frankfurt, Germany: Ontos.
Botvinick, M., & Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature, 391*, 756.
Brown, J. W. (2004). The illusory and the real. *Mind and Matter, 2*, 37–59.
Cappelen, H., & Lepore, E. (2009). Quotation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Stanford, CA.
Davidson, D. (1984). Quotation. In *Truth and interpretation* (pp. 79–92). Oxford: Clarendon Press.
Dennett, D. C. (1992). The self as a center of narrative gravity. In F. Kessel, P. Cole, & D. Johnson (Eds.), *Self and consciousness: Multiple perspectives*. Hillsdale, NJ: Erlbaum.
Descartes, R. (1904). Meditationes de prima philosophia. In C. Adam & P. Tannery (Eds.), *Oeuvres de Descartes* (Vol. VII, pp. 1–90). Paris (Year of original: 1641).
Dretske, F. (2003a). How do you know you are not a zombie? In B. Gertler (Ed.), *Privileged access. Philosophical accounts of self-knowledge* (pp. 1–14). Burlington: Ashgate Publishing.
Dretske, F. (2003b). Sensation and perception. In Y. H. Gunther (Ed.), *Essays on nonconceptual content*. Cambridge, MA: MIT Press (Year of original: 1981).
Fernández, J. (2003). Privileged access naturalized. *The Philosophical Quarterly, 53*, 352–372.
Fernández, J. (2005). Privileged access revisited. *The Philosophical Quarterly, 55*, 102–105.
Fodor, J. A. (2001). Language, thought and compositionality. *Mind & Language, 16*(1), 1–15.
Gallagher, S. (2005). *How the body shapes the mind*. Oxford: Oxford University Press.
Geach, P. (1970). Quotation and quantification. In *Logic matters* (pp. 205–208). Oxford: Basil Blackwell.
Greenberg, M. (2005). A new map of theories of mental content. *Nous, 39*, 299–320.
Harman, G. (1990). The intrinsic quality of experience. *Philosophical Perspectives, 4*, 31–51.
Heidegger, M. (1977). The origin of the work of art. In *Basic writings* (pp. 139–211). New York: Harper and Row (Year of original: 1935).
Hodges, W. (2001). Formal features of compositionality. *Journal of Logic, Language and Information, 10*, 7–28.
Hoffman, R. E. (1986). Verbal hallucinations and language production processes in schizophrenia. *Behavioral and Brain Sciences, 9*, 503–517.
Kaplan, D. (1989). Demonstratives. In *Themes from Kaplan* (pp. 481–563). Oxford: Oxford University Press.
Lenggenhager, B., Tadi, T., Metzinger, T., & Blanke, O. (2007). Video ergo sum: Manipulating bodily self-consciousness. *Science, 317*, 1096–1099.
McDowell, J. (1996). *Mind and world*. Cambridge, MA: Harvard University Press.
Metzinger, T. (2003). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences, 2*, 353–393.
Moore, G. E. (1903). The refutation of idealism. *Mind, 12*, 433–453.
Morin, A. (2005). Possible links between self-awareness and inner speech: Theoretical background, underlying mechanisms, and empirical evidence. *Journal of Consciousness Studies, 12*, 115–134.
Pagin, P., & Westerståhl, D. (in press). Compositionality. In K. von Heusinger, C. Maienborn, & P. Portner (Eds.), *Semantics. An international handbook of natural language meaning*. Berlin: Mouton de Gruyter.
Potts, C. (1997). The dimensions of quotation. In C. Barker & P. Jacobsen (Eds.), *Direct compositionality* (pp. 405–431). Oxford: Oxford University Press.
Putnam, H. (1981). Brains in a vat. In *Reason, truth, and history* (pp. 1–21). Cambridge: Cambridge University Press.
Reimer, M. (2003). Too counter-intuitive to believe? Pragmatic accounts of mixed quotation. *Belgian Journal of Linguistics, 17*, 167–186.
Ricoeur, P. (1992). *Oneself as another*. Chicago: University of Chicago Press.
Stephens, G. L., & Graham, G. (2000). *When self-consciousness breaks: Alien voices and inserted thoughts*. Cambridge, MA: MIT Press.
Werning, M. (2004a). Self-awareness and imagination. In J. Sàágua (Ed.), *The explanation of human interpretation* (pp. 369–377). Lisbon: Colibri.
Werning, M. (2004b). Compositionality, context, categories and the indeterminacy of translation. *Erkenntnis, 60*, 145–178.
Werning, M. (2005). Right and wrong reasons for compositionality. In M. Werning, E. Machery, & G. Schurz (Eds.), *The compositionality of meaning and content (Vol. I: Foundational Issues,* (pp. 285–309). Frankfurt: Ontos Verlag.
Wright, C. (1992). On Putnam's proof that we are not brains-in-a-vat. *Proceedings of the Aristotelian Society, 92*, 67–94.