BEHAVIORAL AND BRAIN SCIENCES (2004) **27,** 289–312
Printed in the United States of America

Continuing Commentary

Commentary on **Ned Block (1995). On a confusion about a function of consciousness. BBS 18(2): 227–287.**

Abstract of the original article: Consciousness is a mongrel concept: there are a number of very different "consciousnesses." Phenomenal consciousness is experience; the phenomenally conscious aspect of a state is what it is like to be in that state. The mark of access-consciousness, by contrast, is availability for use in reasoning and rationally guiding speech and action. These concepts are often partly or totally conflated, with bad results. This target article uses as an example a form of reasoning about a function of "consciousness" based on the phenomenon of blindsight. Some information about stimuli in the blind field is represented in the brains of blindsight patients, as shown by their correct "guesses." They cannot harness this information in the service of action, however, and this is said to show that a function of phenomenal consciousness is somehow to enable information represented in the brain to guide action. But stimuli in the blind field are *both* access-unconscious and phenomenally unconscious. The fallacy is: an obvious function of the machinery of access-consciousness is illicitly transferred to phenomenal consciousness.

Sidestepping the semantics of "consciousness"

Michael V. Antony
Department of Philosophy, University of Haifa, Mount Carmel, Haifa 31905, Israel. **antony@research.haifa.ac.il**

Abstract: Block explains the conflation of phenomenal consciousness and access consciousness by appeal to the ambiguity of the term "consciousness." However, the nature of ambiguity is not at all clear, and the thesis that "consciousness" is ambiguous between phenomenal consciousness and access consciousness is far from obvious. Moreover, the conflation can be explained without supposing that the term is ambiguous. Block's argument can thus be strengthened by avoiding controversial issues in the semantics of "consciousness."

There is a widespread tendency among researchers of consciousness to address the semantics of the term "consciousness" (and its cognates) when investigating the mental phenomenon, consciousness. Such terminological discussions, in my view, are typically poorly motivated, add little to the inquiry, and confuse matters more than anything else. Ned Block, in his important and influential target article (Block 1995t), also touches on the meaning of "consciousness." He proposes that the term is ambiguous (sect. 4.2.2, para. 5), with senses corresponding to phenomenal consciousness (P) and access consciousness (A), among other senses less central to his aims. In contrast with many other researchers, Block motivates his semantic discussion of "consciousness." His purpose is to expose a fallacy which he claims arises when researchers reason about consciousness (sect. 1). According to his diagnosis, the fallacy results from the conflation of P and A, and he explains the conflation by appeal to the ambiguity of "consciousness": "An ambiguous word often corresponds to an ambiguous mental representation, one that functions in thought as a unitary entity and thereby misleads" (sect. 4.2.2, para. 8).

While Block takes it upon himself to motivate his semantic discussion, I believe the reasons he provides are not quite satisfactory. For, as I shall argue, conflations between P and A can be explained without supposing that "consciousness" is ambiguous, indeed without entering into the semantics of "consciousness" at all. By avoiding controversial semantic claims regarding ambiguity, Block can strengthen his argument.

If one explores the literature on ambiguity within lexical semantics, one discovers that things are in a bit of a shambles, but one also encounters serious attempts to theoretically accommodate a wide range of interesting phenomena related to ambiguity.[1] For example, a distinction is often drawn between two kinds of ambiguity: homonymy and polysemy. Homonymous words are said to have unrelated senses (like a dog's *bark* and a tree's *bark*) and correspond to distinct entries in the lexicon; whereas polysemous words have more interrelated senses (like *opening* a window and *opening* with a joke) which are listed together within lexical entries. There also appear to be phenomena that involve subtle variations of meaning but no ambiguity. Cruse (1986), for instance, speaks of *modulations of sense,* where "a single sense can be modified in an unlimited number of ways by different contexts, each context emphasizing certain semantic traits, and obscuring or suppressing others" (p. 52). One of Cruse's examples is "car" in "the car needs servicing" and "the car needs washing," where different parts of the car get emphasized or highlighted (the engine and exterior, respectively) but there is no ambiguity (p. 53). A related example drawn from Pustejovsky (1996, p. 32) concerns the word "good": it seems to express different properties in "a good car," "a good meal," and "a good knife," but it would be rash to infer from this that "good" is ambiguous, since the list of such expressions containing "good" can be extended indefinitely.

I think a reasonable case can be made for the claim that something like modulation occurs with "consciousness" – that there is a single, complex phenomenon, different features of which get highlighted by uses of "consciousness" across different linguistic contexts (sometimes phenomenal aspects, sometimes functional or cognitive aspects, etc.), but that the term "consciousness" itself is not ambiguous between P and A. In any event, it should be clear that anyone who wishes to convincingly argue that "consciousness" is ambiguous between P and A must enter into these rather messy semantic issues and exclude interpretations that treat "consciousness" as univocal. Block does not do that sufficiently, to my mind; nor do others who discuss the semantics of "consciousness." Instead, it seems, new senses of "consciousness" are forever being offered, without constraint. But surely semantics is not that easy! (Notice, by the way, that so-called "stipulative definitions" do not in themselves issue in new meanings for natural language expressions. If a theorist stipulates that "consciousness" means *milkshake,* the English word "consciousness" does not thereby acquire

 0140-525X/04 $12.50

a new sense, not even a new "technical" sense. At a minimum, a new public practice or use is required.)

Block suggests that there are several equally legitimate ways to characterize ambiguity, and he favors doing so "in terms of conflation: if there can be conflation, we have ambiguity" (sect. 4.2.2, para. 7). However, if one takes semantics seriously as a scientific enterprise, then meaning presumably will have some nature and one will not be free to characterize ambiguity however one chooses. And even if there is some indeterminacy, some leeway for alternative characterizations, it is hard to avoid seeing Block's construal of ambiguity in terms of conflation as ad hoc, given that it is conflation (of P and A) that he wishes to explain by appeal to ambiguity.

Block, however, can avoid these tangled semantic issues, for conflation does not require ambiguity. He is of course right that ambiguous words often involve ambiguous mental representations (or, perhaps more accurately, distinct mental representations), and that such representations can give rise to conflation. However, all he really needs are the representations – distinct representations or distinct "elements" of complex representations – such that it becomes possible to unknowingly slide from one representation (or representational element) to another. While ambiguity suffices for that, it is unnecessary since mental representations or concepts are cut more finely than are word meanings. Consider commonsense and scientific concepts of water. Because they are different concepts, conflation is in principle possible; but "water" is univocal.[2]

Conflation is also possible where there is modulation (Cruse 1986). Consider how the sense of "full" is subtly modified across contexts in spite of being univocal:

(1) a full bookshelf [no room across the shelf]
(2) a full auditorium [all seats occupied]
(3) a full balloon [stretched near capacity by a gas or liquid inside it]
(4) a full swimming pool [nearly all its volume occupied by water]
(5) a full swimming pool [contains a maximum number of people, as determined by safety regulations, comfort, etc.]

Examples (4) and (5) furnish us with a means of demonstrating how modulation can result in conflation. Imagine a swimming pool that is scheduled to open on June 1, but by that date contains no water (and hence no people). And suppose an employee at the pool overhears the manager complaining on June 1 about the pool not being full. Finally, imagine the employee telling a friend how the manager is upset because the pool was not filled with swimmers on opening day, whereas all the manager really cared about is that it be filled with water (since the manager's salary is independent of how many people are in the pool). In this example, both "kinds of fullness" are missing from the pool and, conflating the two, the employee fallaciously infers that the manager is upset because the pool was not filled with people. This closely parallels the fallacious reasoning about consciousness that concerns Block. However, the example shows that such reasoning, as well as the conflation on which it is based, can occur in the absence of ambiguity, since "full" is univocal. Perhaps something similar is happening with the word "consciousness" and the distinct elements of our complex mental representation of consciousness that represent phenomenal features and cognitive/functional ones.

I believe Block would do best to explain the conflation between P and A in terms of something like "carelessly sliding over representational distinctions," without committing himself to whether such distinctions are across distinct representations or elements within a single representation. And he ought not tie himself to the claim that the English word "consciousness" is ambiguous between P and A. Even if he believes it is, his argument does not depend on it being so. Block's argument would thus be strengthened if such controversial semantic issues were avoided.[3]

ACKNOWLEDGMENT

I am grateful to Ned Block for helpful discussion concerning an earlier draft of this commentary.

NOTES

1. See, for example, Lyons (1977), Cruse (1986), and Pustejovsky (1996).

2. Lest one think that something like Chalmers's (1996, Ch. 2) distinction between primary and secondary intensions suffices to show that "water" is ambiguous, notice that that could not be a kind of ambiguity relevant to the discussion in this commentary, since that would make all words (or at least all natural kind words) ambiguous, thus robbing the question whether conflation requires ambiguity, of any interest.

3. For more extended treatment of the issues discussed in this commentary, see Antony (2001; 2002).

Superblindsight, Inverse Anton, and tweaking A-consciousness further

Oliver Kauffmann

Department of Philosophy, University of Copenhagen, Copenhagen, DK-2300 S., Denmark. **kauff@hum.ku.dk**

Abstract: It is argued that Block's thought experiment on superblindsight and "the Inverse Anton's syndrome" are not cases of A-consciousness without P-consciousness. "Weak dispositional states" should be excluded from the set of A-conscious states, and a subject's being reflectively conscious of a P-conscious state is suggested as a better candidate for A-consciousness. It is further pointed out that dreams, according to Block's own criterion but contrary to what he claims, are A-unconscious and it is argued that Block should not accept the idea that high-information representational content is an empirically sufficient condition of phenomenality in human beings.

In his target article, Block (1995t) advances a conceptual distinction between a state being A- or P-conscious. A state is A-conscious if, by virtue of one's having the state, a representation of its content is poised for use as a premise in reasoning, rational control of action, and rational control of speech. A state is P-conscious if it has the experiential properties expressed by the phrase "there is something it is like to be in that state." Block's target reasoning is the fallacy committed by "jumping from the premise that 'consciousness' is missing – without being clear about what kind of consciousness is missing – to the conclusion that P-consciousness has a certain function" (1995t, p. 242). However, he also claims that there is something importantly right in this reasoning: A and P often make their presence and absence together (1995t, p. 242). In particular, it seems difficult to find clear cases of A without P. "If indeed there can be P without A, but not A without P, this would be a remarkable result that needed explanation," Block says in his response to commentators (1995r, p. 272), inviting readers "to tinker with the definitions of 'P' and 'A' so as to make them coincide better" (1995r, p. 277).

In what follows I will point to some difficulties with Block's two best proposals for "A without P" – superblindsight and the Inverse Anton's syndrome – and argue for a further tweaking of the concept of A-consciousness.

1. Superblindsight. Is superblindsight a case of A-consciousness without P-consciousness? I don't think so. "Superblindsight" is Block's label for the thought experiment in which we imagine a blindsight patient

> prompt[ing] himself at will, guessing what is in the blind field without being told to guess. . . . Visual information from his blind field simply pops into his thoughts in the way that solutions to problems we've been worrying about pop into our thoughts, or in the way some people just know the time or which way is North without having any perceptual experience of it. (1995t, p. 233)

Block's contention is that the perceptual state which gives rise to the particular thought is A-conscious. I contend that these underlying states are neither perceptual nor A-conscious. Concerning the relation between the superblindsighter's perceptual state (S1) and the resulting thought (S2), some commentators have already argued that the superblindsighter's underlying perceptual state S1

is only A-conscious via the particular thought S2 it gives rise to. This can be understood in at least two different ways, depending on whether the P-conscious or A-conscious aspect of the superblindsighter's thought is emphasized. By focusing on the P-conscious aspect it might be claimed that in the absence of the actual content presented to the subject in that phenomenal way – as that P-conscious thought – there is no reason for attributing consciousness to him at all (Lloyd 1995; Revonsuo 1995). Without the relevant P-conscious thoughts the superblindsighter is a visual zombie to whom we may perhaps grant states with access to other states – but not consciousness. In the superblindsight case, therefore, the perceptual state cannot be said to be A-conscious in itself. But this objection does not tell us how it is possible to access other states by having P-consciousness, as forcefully pointed out by Levine (1995) and Chalmers (1997).

However, Block himself sometimes leans toward the view that P-consciousness is "the core notion" of consciousness (Block 1995t, p. 274; 1997, p. 163), which I take to mean that P is an empirically necessary condition of A. But that does not preclude A-consciousness from being something different from P-consciousness, nor does it require that the superblindsighter's underlying perceptual states must be related to P-consciousness the way they are (by having those particular P-conscious thoughts) in order to be A-conscious. The afforded relation between A and P could be weaker than that: Perhaps the superblindsighter's "capacity to articulate thoughts through phenomenal verbalizations" (Burge 1996. p. 429) is what matters. Block exemplifies this possibility with the case of a drunk person becoming unconscious:

> He may have P-conscious states both before and during his episode of unconsciousness; for example, while unconscious he may be seeing stars or having mental images of various sorts . . . roughly, I think we count the drunk as unconscious to the extent that he has no A-consciousness of the environment via P-conscious perceptions of it. The drunk is A-unconscious in a way the specification of which involves appeal to P. (1995r, p. 274)

But this suggestion doesn't really go beyond a mere empirical correlation between A- and P-consciousness. A stronger claim is to say that perhaps P really "greases the wheels of accessibility" (1995t, p. 242), a suggestion which seems contrary to Block's overall suggestion that A does the work alone. Block is sensitive to this possibility, but I think he is quite right when he says that if this is what is going on, we don't know how it is possible. Nonetheless, I will suggest a concept of A-consciousness that involves reflective P-consciousness. But before introducing that, consider what results from attaching importance to the A-conscious aspect of the superblindsighter's thought S2, thereby regarding this aspect as a necessary condition for A-consciousness of the underlying perceptual state (S1).

What is the content of S1? Block presupposes that S1 and S2 have the same (or overlapping) content, an assumption which follows directly from the definition of A-consciousness: a representation of the state's content must be poised for use. In superblindsight this content can be specified by the sentence "There is an X in the visual field" (1995t, p. 233). Assuming that the two states have the same content, Kobes (1995) argues that the resultant thought is necessary for the perceptual state becoming A-conscious since it is only via that thought that the state becomes accessible. Block anticipates this objection in his target article (1995t, Note 7, p. 245). He argues that in order to prevent a contradiction in letting a (by hypothesis) A-unconscious state cause another state with the same content, and thereby, by definition, be A-conscious – "because it is in virtue of having that state that the content it shares with the other state satisfies the three conditions" (p. 245) – the notion "in virtue of" must be refined to exclude the reading that the state "can only cause this inferential promiscuity via another state." The idea is that some states *other* than S2 must be possible as effects of S1. I still think Kobes' objection is on the right track but needs some refinement, which I will try to flesh out in what follows.

Block says that "a genuinely A-conscious perceptual content would be freely available for use in thought" (1995r, p. 276) and is "poised for voluntary or direct control" (1997, p. 159). The notions "freely available" and "voluntary" indicate that S1 can result in states other than S2. But what is meant by "direct" here? Direct compared to what? In what sense and in relation to what is S1 more direct than any other arbitrary state S*x* a person might have?

Pure dispositional states like quiescent beliefs are not poised for direct control "because the access to them requires processing," we are told (Block 1997, p. 160). Chalmers has made the same claim (cf. Chalmers 1997, p. 148). Is this criterion acceptable? How do we know and compare "the amounts of processing required" for a change from an occurrent S2-state to a poised (what I will call) "weak dispositional" S1-state and the change from S2 to a "stronger dispositional" state S*x* respectively? It won't do to say that there is no processing as long as the contents of the two states are the same; if the states are of a different type (e.g., perceptual/thoughtlike), some work is required.

Perhaps a better criterion for sorting quiescent beliefs out of the A-conscious set would be to point out that it is not in accordance with our commonsense notions of consciousness to speak of such states as being conscious in any sense. But this still leaves the notion "direct" unexplained.

Hence, my suggestion is to tweak the A-concept by excluding the set of poised "weak dispositional" freely available states like S1 from the set of A-conscious states. There possibly is a difference in accessibility between "weak" and "strong" dispositional states but not a difference to be captured by the conscious/nonconscious distinction.

Block thinks that A-consciousness captures a notion of access we find in commonsense reasoning about consciousness (1995t, p. 231; 1995r, pp. 276–77). I agree that access does play a role in this reasoning but consciousness as a property of states is not in any intuitive forthright way used as a dispositional predicate, be it in a strong or weak sense (cf. Church 1995; Rosenthal 1997). This contrasts with the notion of "creature-consciousness," which is sometimes used dispositionally. A-consciousness as availability for use must not be taken to be equivalent to mere accessibility, which would make A-consciousness "a totally dispositional concept," we are told (Block 1997, p. 160). But whatever the difference between "weak" and "strong" dispositional states is, availability is still a dispositional concept.

The problem with Block's idea of S1 as A-conscious can be put in another way: Suppose we accept the existence of an S1-type of poised state in the superblindsighter "freely available for use." In accordance with Block's definition this means that a representation of the state's content can be freely used as a premise in reasoning, action, and reporting; that is, it can be followed by different tokens of S2-states having the content of S1 in common. But "can be followed" can be understood in two different ways: either as the fact that, necessarily, some S2-state or other will follow; or as the mere possibility that some other S2-state may follow. Block clearly accepts the last reading: A person can be in an A-conscious state without being in any P- and A-conscious S2-states with the same content. But if it is only available in this sense, how do we then distinguish the S1-state from strong dispositional states? And if S1 is dispositional in the sense that we can be in S1 without being in any S2-states, we are not in accordance with common sense if we call it a conscious state. Admittedly, arguing from common sense doesn't cut much ice here. My suggestion as to how the relation between access and consciousness might be described – thereby improving Block's analysis – is that access is gained by a subject's reflective consciousness of an occurrent P-conscious content. Reflective consciousness does not imply that the subject needs to be conscious of his being conscious of the P-conscious content (in the introspective sense of the subject being conscious that he is conscious). On the other hand, since I accept Block's assumption that a person (or at least subsystems thereof) might have unaccessed P-conscious states, I do not claim that a state is conscious only if the subject is having a higher order representation

of that state. Access-consciousness, however, can be understood as the subject's being reflectively aware of a P-conscious content which thereby can be used as a premise in reasoning and/or action. The disjunctive possibility captures situations where either the contents necessary for inferential control or the contents necessary for rational bodily control are unavailable (see below). Since the superblindsighter is not reflectively aware of the perceptual state S1 (whether or not this state is P-conscious) which underlies the thought (S2), this state is not A-conscious. Perhaps the S1-states are P-conscious; a possibility which is not excluded by Block (1995t, pp. 232, 242).

I admit that A-consciousness in terms of reflective consciousness still offers no explanation as to how (reflective consciousness of) a P-conscious state gives access (*pace* Block, Chalmers 1997; Levine 1995). And it would be to commit the very target fallacy to suggest that it is because of lack of reflective P-consciousness per se that there is lack of access. It is just a matter of empirical fact that reflective awareness of a P-conscious content is a sufficient condition for a person to have access to that content.

There is a further problem with Block's conception of the superblindsighter's underlying perceptual state being A-conscious. As mentioned above, Block presupposes that the contents of S1 and S2 are identical or overlapping. But what evidence do we have for individuating the content of S1?

It has been suggested that the real blindsighter's discriminative behaviour can be seen as a result of clues about the movements of their eyes or the premotor readiness of their muscles, in which case it could be argued that the content of S1 is not perceptual as is that of S2 (see Goodale & Milner 1992; Humphrey 1992; Milner & Goodale 1995; Vision 1998). Perhaps it really is a state of one's own musculature. Block is aware of this possibility but doesn't think it plays a role: Whatever it is that allows the blindsight patient to discriminate an X from an O and a horizontal from a vertical line will do (1995t, Note 14, p. 246), which I take to mean that – at least to this extent – there is an overlap between the contents. But if for the sake of argument we accept that Milner and Goodale are right about the dorsal-ventral bifurcation of the processing of visual information beyond V1 – and Block in fact hints that this distinction possibly matches his A/P distinction (1995t, p. 233) – we also have a physiological reason for the non-possibility of superblindsight: Access does not rely on dorsal stream processing since the damage to V1 involved in "normal" blindsight essentially involves closing down the input to the ventral stream.

Remember that, according to Block, blindsight patients are not A-conscious. So, if the dorsal stream processing in these subjects functions independently of the ventral system – and here, for the sake of argument, I take for granted that Milner and Goodale (1995) have delivered the relevant ("double-dissociationistic") empirical evidence – then we can only "restore" their access to the visual information by restoring the connection from V1 to the ventral stream. The dorsal stream functions as well as it ever can in normal blindsight. So if the normal blindsighted person's discriminatory abilities rely on a normally functioning dorsal system, we have no reason to think that these patients could be "trained to prompt themselves at will" as Block, and Daniel Dennett (1995), imagine. The very connection between the damaged part of V1 and the ventral pathway would have to be reestablished.[1]

Superblindsight is an imaginative thought experiment but describes nothing really possible: It should not be forgotten that superblindsight, although a thought experiment regarding the person's behavioural capabilities, is built on the real world's ordinary blindsighter when it comes to physiology.

2. Inverse Anton. Is "Inverse Anton" a case of A-consciousness without P-consciousness? I don't think so. Hartmann et al. (1991) describe the clinical syndrome "Inverse Anton" as the condition of a person who due to brain damage denies having visual sensations but in fact still has some intact discriminating abilities concerning visual stimuli. It is the inverse condition of Anton's syndrome, which refers to a person's denial of cortical blindness – a denial supplemented by the patient's confabulating situation-appropriate visual reports. As Hartmann et al. point out, reports of Inverse Anton are rare and inadequately detailed in description; after a brief review of other reported cases they argue that only one patient has clearly deserved the label.[2]

The case story concerns a person who, despite his insistence that he had no visual sensations, was able to name objects and colours, read single words, and recognize famous faces and facial emotions with an accuracy greater than 50 when these stimuli were presented in his upper right visual field. When asked to describe how he made the identifications, the patient typically stated, "I feel it," "I feel like something is there," "it clicks," or "I feel it in my mind." Now does this empirical case exemplify pure A-consciousness without P-consciousness as Block (1996; 1997) suggests?

It should be noticed that the Inverse Anton patient's discriminating abilities are clearly superior to those of the blindsighted person and he does not need not to be prompted by use of forced choice. Further, the syndrome must not be confused with blindsight since it involves spared areas of V1, which blindsight does not. For these three reasons I think Inverse Anton is far closer to being an example of "partial visual zombiehood" than blindsight, and it deserves philosophers' proportionate attention. The fact that V1 is involved suggests that if it does make sense at all to speak of a content-bearing state underlying the Inverse Anton patient's clicking thoughts and if the above comments concerning the nature of the underlying state S1 in blindsight are true, then the state involved in Inverse Anton is perhaps a better candidate for having perceptual content than the states underlying Block's superblindsighter.

Put in a simple way: Physiologically speaking, Inverse Anton is more like superblindsight than Block's own thought experiment. And if Inverse Anton is a better candidate for superblindsight, we are perhaps better off considering it as a case of pure A without P.

But first, if the arguments above concerning the nondispositional nature of A-consciousness and its dependency on reflective awareness of P-content are valid, the only A-conscious states involved in this case are the clicking thoughts insofar as the subject is aware of them.

Second, perhaps the S1 states are really P-conscious after all. It is not an argument to the contrary that the patient denies having any visual experiences, a possibility I alluded to above: Block remarks in his discussion of blindsight that the claim that P-consciousness is missing in blindsight is just an assumption. "I decided to take the blindsight patient's word for his lack of P-consciousness. . . . Maybe this assumption is mistaken" (1995t, p. 242).[3] But does anything in the Inverse Anton case-report favour this possibility? Tested for colour-naming ability, the patient not only accurately named the colours but "qualitatively he maintained that he could 'feel' or 'hear' the color and his initial responses were consistent with such perceptions. For example, to the color orange he responded 'like sky . . . like sunset' and to green he stated 'I feel something clear' " (Hartmann et al. 1991, p. 34). This reported experiential content cannot be about the occurrent "clicking thoughts," however, since the patient denies having any. Therefore, I think it possibly could be taken to be a P-conscious content which to a certain extent can be overtly reported but only covertly experienced. As Hartmann et al. note, "these responses indicated impairment at the level of color naming as opposed to a perceptual deficit " (p. 34). Granting Block the possibility of existing instances of pure P-consciousness, this is perhaps a further example.[4]

3. P-conscious content and rich informational content. I think Block is granting Dennett (1995) too much with the idea of high-information representational content as an empirically sufficient condition of phenomenality in humans (Block 1995r, p. 273). The superblindsighter reports that there is a difference between knowing about an X in his blind field and knowing about an X by having a visual experience. "There is something it is like to experience the latter, but not the former, he says" (1995t, p. 233). To show that this really comes to a difference in kind (of conscious-

ness) and not only degree (of richness in content), Dennett rightly claims that Block "must control for richness of content," which means that we must adjust the case so that the richness of content is reported as being the same in the two fields. The problem is that it simply seems counterintuitive to take a subject's word that it wasn't like anything at all to be visually informed on there being a bright orange X, in Times Roman italics font, on a blue-green background about 2 inches high with a smudge in front of him. The superblindsighter says "that he knows these sorts of features of stimuli in his blind field even though he is just guessing and contrasts what is going on with the real visual experiences of his sighted field" (1995r, p. 273). It can be argued that every attempt on the superblindsighter's behalf to specify what is meant by a difference in experiential properties seems to be fully accountable for in terms of content. Accepting Dennett's premise of an adjusted richness in content, it seems that one cannot point at any P-property-difference between a blind and some sighted part of the subject's visual field.

The problem with the counterintuition involved here is in part due to the fact that we consider the particular perceptual state in isolation from other states; what matters is not just the subject's report on an actual state's content but his ability to track a difference in experience between different states. For all we know, if there really is a difference in this case, it is related to the damaged primary visual cortex. But then the postulated difference can be tested in a way which could satisfy both "realists" and "antirealists" regarding experiential properties: Suppose that, unbeknownst to the superblindsighter, we indulge in tinkering with his neural pathways leading from (what he claims to be) the superblind field and (what he claims to be) the experienced part of his visual field to their respective target areas in the primary visual cortex: If we switch the part of his neural pathway leading into the damaged area of the primary visual cortex with a part of a pathway feeding a not-damaged part, the result is that we, cortically speaking, "move the superblindsighter's scotoma to another part of his visual field." The question is whether the superblindsighter will notice any experiential change as a result of the switch. If there really is a difference concerning "what it is like to be in a state" beyond giving a specification of its content, the superblindsighter will be able to track and report any change in location of the (superblind) scotoma which we induce by manipulating his pathways.

Superblindsight is a thought experiment and, as indicated above, I believe we have reasons as to why we won't find any real instances of it. However, this does not debar us from using it to discuss the problem of whether high-information representational content would be an empirically sufficient condition of phenomenality in humans. "Imaginary cases are of limited value in such theoretical explorations, but this time I think the flight of fancy nicely reveals how Block mislocates the issue," Dennett triumphs (1995, p. 253). But the imaginary case leads to an imaginary path of testability, which makes it an open question whether there is an experiential difference beyond specification of content. Here Block gives way to Dennett too quickly.

4. Dreams. Revonsuo (1995) and Bogen (1997) both make the suggestion that dreaming is an example of P-consciousness without A-consciousness. Why is Block unwilling to accept this proposal? To rely on Chomsky's planning of papers during his dreams (Block 1995r, p. 275) and the phenomenon of lucid dreaming – or, for that matter, on Descartes' skills for doing arithmetic in an infallible way when dreaming – is not enough. As Block himself points out, many dreams are not that rational.

Perhaps we should say instead that dreams are not A-conscious as a *kind;* some dreams are A-conscious, while others are not. This manoeuvre would also be in accordance with Block's own characterization of A-conscious content as being system-relative and his denial of the existence of any kinds of states being intrinsically A-conscious (1995t, p. 232). But Block clearly wants to treat different kinds of dreams on an equal footing – that is, as all of them being A-conscious. He draws a parallel between dreaming and the above-mentioned example of a person being drunk and unconscious. This parallel, however, is not valid because a drunk person is not A-conscious. The drunk is unconscious to the extent that he has no A-consciousness of the environment via P-conscious perceptions of it (Block 1995t, p. 232).

But there is another problem with considering dreams as A-conscious: Block claims that during dreaming one's representations are poised to control behavior but behavioral systems are paralyzed, so there is no behavior. Dream contents are A; so they do not provide a case of P without A (1997, p. 165). The contents are available for use but cannot be used.[5] But then several of Block's examples of P-consciousness without A-consciousness turn out to be A-conscious after all, and they are precisely those examples which, in Block's own words,

> were designed to exploit the fact that access to a P-content can fail for a variety of reasons, including lack of attention and various forms of blockage. (I mentioned blockage due to repression, information processing limits, fragmentation of the self, and deactivation of centers of reasoning and planning by, for example, anesthesia.) If these cases are genuine cases of P, then they are cases of P without A, because some work would be required to access the blocked representations. Attention would have to be focused or the blockage removed. (1997, p. 160)

If Block now insists that dreams are A-conscious despite the motor output blockade, consistency demands that he also accepts several of his purported pure P-conscious cases as being A-conscious. Furthermore the above-mentioned problem with sorting out strong dispositional states from the set of A-conscious states turns up again: Why should memory-blockage or blockage due to an amount of required retrieval-work make these states A-unconscious when a motor blockage cannot do the same for dream states?[6] Perhaps a better strategy would be to give up the blockage against dreams as instances of pure P-conscious states. Or, even better, accept the above-suggested reformulation of what A-consciousness comes to, which is that some dreams are A-conscious to the extent that they are reflectively conscious. Furthermore, only in pathological cases can the contents of dreams be said to be available as premises for both inferences and action control.[7] So, by giving up the conjunctive action-constraint in Block's formulation of A-consciousness and replacing it with a disjunctive account, we can capture those phenomena where we are reflectively aware of an inferential promiscuous content (but this content cannot control action), as well as action-phenomena (e.g., sports) where the control of action is what matters and the control of reasoning is to various degrees suppressed.

ACKNOWLEDGMENT

I should like to thank Jan Riis Flor for helpful discussions.

NOTES

1. This objection is basically due to Vision (1998). Problems remain, however, with the assumption concerning the essential connection between V1 and the ventral stream. Goebel et al. (2001) have presented imaging data showing that, despite total damage to the relevant parts of V1, visual information in hemianopic subjects can reach V4 in the ventral stream. This seems to indicate that blindsight does not rely exclusively on the dorsal system. So perhaps the visual information somehow might become available for "self-prompting," thereby establishing real superblindsight after all?

2. Recently though, another Inverse Anton case was reported in Brazdil et al. (2000). (Thanks to Sune Nordentoft Lauridsen for calling this to my attention.)

3. Norton Nelkin has made the interesting suggestion that colour discrimination in blindsight could be based on the patients' phenomenal but non-apperceptive awareness of hues, which I take to be equivalent with pure P-consciousness thereof; see Nelkin (1996). For some arguments (although not conclusive) against this assumption, see Stoerig (1997).

4. I am aware that the patient's lesions are perhaps not quite compatible with this interpretation. It would be natural to suggest that a disruption of the anterior executive system would be found, but CT scannings show that the patient's frontal lobes "were largely intact" (Hartmann et al. 1991, p. 39). According to the experimenters, the patient's lesions rather indicated a disconnection of parietal lobe attentional systems from the visual information processing in the occipital lobe. But still, I am not sure

whether this fact really excludes the possibility of the patient's having unattended P-consciousness.

5. I accept for the sake of argument that dreams sometimes can be inferentially promiscuous. Inferential promiscuity does not imply that an actual pattern of inference must live up to the standards of deductive or inductive logic, according to Stich (1978), from whom Block borrows the concept of inferential promiscuity (1995t, p. 231). So, despite the fact that our dreams often exhibit a high degree of incoherence and reality distortion, they can be considered to be inferentially promiscuous.

6. In fact this problem is the very same as the one Block finds in Chalmers' attempt to handle the P-cases without A-cases as instances of P and A, when he accuses Chalmers of "trying to have his cake and eat it too" by (unwillingly) including instances of both merely potentially available information and information directly poised for access in his extended notion of A-consciousness (cf. Block 1997, p. 160).

7. I am thinking of the REM sleep behavioural disorder (RBD) where the dreaming person tries to "live out" the contents of his dreams due to a pathological abolition of the atonia, as mentioned in Revonsuo (1995). See Revonsuo et al. (2000) for some detailed case descriptions of what it is like for subjects to have these and related kinds of disorders.

There is no Author's Response to these continuing commentaries.

References

Antony, M. V. (2001) Is "consciousness" ambiguous? *Journal of Consciousness Studies* 8(2):19–44. [MVA]

(2002) Concepts of consciousness, kinds of consciousness, meanings of "consciousness." *Philosophical Studies* 109:1–16. [MVA]

Block, N. (1995t) On a confusion about a function of consciousness. *Behavioral and Brain Sciences* 18(2):227–47. [MVA, OK]

(1995r) How many concepts of consciousness? *Behavioral and Brain Sciences* 18(2):272–84. [OK]

(1996) How not to find the neural correlate of consciousness. *Trends in Neuroscience* 19(2):456–59. [OK]

(1997) Biology versus computation in the study of consciousness. *Behavioral and Brain Sciences* 20(1):159–65. [OK]

Bogen, J. E. (1997) An example of access-consciousness without phenomenal consciousness? *Behavioral and Brain Sciences* 20(1):144. [OK]

Brazdil, M., Kuba, R. & Krizova, J. (2000) Inverse Anton's syndrome – Dissociation of perception and consciousness? *Ceska A Slovenska Neurologie A Neurochirurgie* 63(3):171–74. [OK]

Burge, T. (1997) Two kinds of consciousness. In: *The nature of consciousness. Philosophical debates,* ed. N. Block, O. Flanagan & G. Güzeldere, pp. 427–33. MIT Press. [OK]

Chalmers, D. J. (1996) *The conscious mind.* Oxford University Press. [MVA]

(1997) Availability: The cognitive basis of experience. *Behavioral and Brain Sciences* 20(1):148–49. [OK]

Church, J. (1995) Fallacies or analyses? *Behavioral and Brain Sciences* 18(2):251–52. [OK]

Cruse, D. (1986) *Lexical semantics.* Cambridge University Press. [MVA]

Dennett, D. C. (1995) The path not taken. *Behavioral and Brain Sciences* 18(2):252–53. [OK]

Goebel, R., Muckli, L., Zanella, F. E., Singer, W. & Stoerig, P. (2001) Sustained extrastriate cortical activation without visual awareness revealed by fMRI studies of hemianopic patients. *Vision Research* 41(10–11):1459–74. [OK]

Goodale, M. A. & Milner A. D. (1992) Separate visual pathways for perception and action. *Trends in Neuroscience* 15(1):20–25. [OK]

Hartmann, J. A., Wolz, W. A., Roeltgen, D. P. & Loverso, F. L. (1991) Denial of visual perception. *Brain and Cognition* 16:29–40. [OK]

Humphrey, N. (1992) *A history of the mind.* Simon & Schuster. [OK]

Kobes, B. W. (1995) Access and what it is like. *Behavioral and Brain Sciences* 18(2):260. [OK]

Levine, J. (1995) Phenomenal access: A moving target. *Behavioral and Brain Sciences* 18(2):261. [OK]

Lloyd, D. (1995) Access denied. *Behavioral and Brain Sciences* 18(2):261–62. [OK]

Lyons, J. (1977) *Semantics, vol. 2.* Cambridge University Press. [MVA]

Milner, A. D. & Goodale, M. A. (1995) *The visual brain in action.* Oxford University Press. [OK]

Nelkin, N. (1996) *Consciousness and the origins of thought.* Cambridge University Press. [OK]

Pustejovsky, J. (1996) *The generative lexicon.* MIT Press. [MVA]

Revonsuo, A. (1995) Conscious and nonconscious control of action. *Behavioral and Brain Sciences* 18(2):265–66. [OK]

Revonsuo A., Johanson, M., Wedlund, J.-E. & Chaplin, J. (2000) The zombies among us: Consciousness and automatic behaviour. In: *Beyond dissociation. Interaction between dissociated implicit and explicit processing. Advances in Consciousness Research, vol. 22,* ed. Y Rossetti & A. Revonsuo. John Benjamins. [OK]

Rosenthal, D. M. (1997) Phenomenal consciousness and what it's like. *Behavioral and Brain Sciences* 20(1):156–57. [OK]

Stich, S. P. (1978) Beliefs and subdoxastic states. *Philosophy of Science* 45:499–518. [OK]

Stoerig, P. (1997) Phenomenal vision and apperception: Evidence from blindsight. *Mind and Language* 12(2):224–37. [OK]

Vision, G. (1998) Blindsight and philosophy. *Philosophical Psychology* 11(2):137–59. [OK]

Commentary on **Tim van Gelder (1998). The dynamical hypothesis in cognitive science. BBS 21(5):615–665.**

Abstract of the original article: According to the dominant computational approach in cognitive science, cognitive agents are digital computers; according to the alternative approach, they are dynamical systems. This target article attempts to articulate and support the dynamical hypothesis. The dynamical hypothesis has two major components: the nature hypothesis (cognitive agents *are* dynamical systems) and the knowledge hypothesis (cognitive agents can be understood dynamically). A wide range of objections to this hypothesis can be rebutted. The conclusion is that cognitive systems may well be dynamical systems, and only sustained empirical research in cognitive science will determine the extent to which that is true.

Imposed intelligibility and strong claims concerning cognitive systems

Roy Lachman

Psychology Department, University of Houston, Houston, TX 77204-5341.
rlachman@uh.edu

Abstract: The computational hypothesis was formulated with due concern for limits and is consistent with imposed intelligibility doctrines. Theories are products of scientific work that impose human classifications and formalisms on nature. The claim that "cognitive agents are dynamical systems" is untenable. Dynamical formalisms imposed on a natural system, given an approximate fit, serve as an explanatory framework and render a represented system predictable and intelligible.

Herbert A. Simon may be foremost among the founders of modern cognitive psychology. His formulation of the computational hypothesis (CH) is expressed as follows, "The computer is a member of an important family of artifacts called symbol systems, or more explicitly, physical symbol systems. Another important member of the family . . . is the human mind and brain" (Simon 1981, pp. 26–27). I do not know if this hypothesis can be accepted

as true, or whether it will be greatly modified or totally rejected. However, we all know that the CH was inspired by scholars (including Alan Newell) who were concerned with the limitations of the standard science of their day, and who were intimately involved in the development and empirical testing of psychological models of human performance and AI models of intelligent computer functionality.

Scientific work in the trenches leaves many with a deep appreciation of observed anomalies, as well as the limitations that affect both general formulations and focused models of empirical domains; both are subject to empirical limits and boundary conditions. So why would a seminal thinker like Simon formulate so bold a hypothesis? One reason is found in the politics of science. During the reign of behaviorism, the study of the higher mental processes was moribund and in serious need of revitalization. The CH and all its ancillary doctrines appealed to frustrated behavioral scientists as a new and exciting way of looking at many aspects of psychological phenomena. Second, personal experience with empirical conundra in the study of human behavior may have predisposed Simon, Newell, and others to an awareness of limits and an openness to other points of view. (See also Newell [1992] and the multiple book review of Newell's *Unified Theories of Cognition* in BBS 15[3].) My reading of Newell's and Simon's views is that they are consistent both with imposed intelligibility and the limits CH places on claims about what the entities and systems investigated ultimately turn out to be when expressed in computational or other types of theory.

I could not find the statement "the mind (or the brain) is a computer" in either Newell's or Simon's work.[1] What is consistently present is the creative assignment of class membership. Classes are something thought up by people to impose intelligibility on the contents of the observable and conceptual universe (Munitz 1986). The CH, however bold (or outrageous) it appears, contains cautions as well as significant lacunae which can, in principle, be filled. Nowhere is the computer fully elucidated, nor is the underlying theory of automata. According to the CH, the vast variety of brain and cognitive systems responsible for the enormous range of behavior of cognitive agents cannot all be explained solely by a symbol system account or any other single formulation.

How does the bounded CH compare with the formulation of the dynamical hypothesis (DH)? Van Gelder (1998t) declares that "cognitive agents are dynamical systems" (p. 615). This identity claim is insensitive to the limits of scientific theorizing, the complexities of natural systems, and to the intelligibility that needs to be imposed to generate a scientific knowledge product. The hypothesis is further extended: "For every kind of cognitive performance exhibited by a natural cognitive agent, there is some quantitative system instantiated by the agent at the highest relevant level of causal organization [that] can and should be understood by producing dynamical models." The coverage of van Gelder's hypothesis is vast; it ranges over a universe that includes most behavior of most living things. This is clear in spite of the many possible meanings of "dynamical systems" and "highest relevant level of causal organization."

Any assumption that a particular dynamical idealization and a natural system are identical, raises insurmountable problems; an identity claim assumes that an end state has been achieved in knowledge concerning some system. At our present stage of inquiry (perhaps at any stage), it is foolhardy to imagine that we have arrived. It is near certain that an alternative cognitive model will supersede those currently preferred, whether the current model is based on symbol systems or connectionism. There can be no assurance that the currently preferred scientific pictures of cognition will never undergo drastic, even revolutionary, change.

An alternative interpretation of dynamical systems is one based on the doctrine of imposed intelligibility and it may be expressed as follows: The formalisms of dynamics imposed on a natural system, if there is a close enough fit, can serve as an explanatory framework to render the represented system predictable and intelligible. The equations of any scientific formalism, including dynamics, come embedded in natural language commentary. The combination of equations and commentary is the resultant scientific knowledge that explains a system. The knowledge product is not the system; to propose otherwise can lead to the bizarre or comic. Whittaker (1942, p. 17) offers an interesting example:

> [I]t happens very often that different physical systems are represented by identical mathematical description. For example, the vibrations of a membrane which has the shape of an ellipse can be calculated by means of a differential equation known as Mathieu's equation; but this same equation is also arrived at when we study the dynamics of a circus performer, who holds an assistant balanced on a pole while he himself stands on a spherical ball rolling on the ground. If we now imagine an observer who discovers that the future course of a certain phenomenon can be predicted by Mathieu's equation, but who is unable for some reason to perceive the system which generates the phenomenon, then evidently he would be unable to tell whether the system in question is an elliptic membrane or a variety artiste.

Recent debate among philosophers concerning realism and its alternatives has been of little help to the concerns of scientists and hence of little interest. The elucidation of the relationship between the knowledge products of science and the natural systems the products represent would be of real value. It is unfortunate that so few philosophers of science now show interest in work such as Munitz's (1986) book, which illustrates reasonable approaches to that relationship.

NOTE

1. "I am sure that somewhere in print I said specifically that brains are computers" (H. A. Simon, personal communication).

Author's Response

Response to Lachman

Tim van Gelder

Department of Philosophy, University of Melbourne, Parkville, VIC 3010, Australia. **tgelder@unimelb.edu.au**

Lachman claims that the Dynamical Hypothesis (DH) is "untenable." His own position is a version of the "The DH is epistemological, not ontological," objection to the target article, which is dealt with in section R2.3 of my original response (van Gelder 1998r). Additional objections are that the coverage of the hypothesis is "vast" and that the DH presupposes we have reached the end point of scientific theorizing. Indeed, the DH is very broad, but it does not presuppose that science has ended; that's why we call it a "hypothesis."

References

Munitz, M. K. (1986) *Cosmic understanding: Philosophy and science of the universe.* Princeton University Press. [RL]

Newell, A. (1992) Précis of *Unified theories of cognition. Behavioral and Brain Sciences* 15(3):425–92. [RL]

Simon, H. A. (1981) *The sciences of the artificial, 3rd edition.* MIT Press. [RL]

van Gelder, T. (1998t) The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences* 21(5):615–28. [RL]

(1998r) Disentangling dynamics, computation, and cognition. (Author's Response to Commentators.) *Behavioral and Brain Sciences* 21(5):654–65. [rTvG]

Whittaker, E. T. (1942) *The beginning and end of the world.* Oxford University Press. [RL]

Commentary on **William J. M. Levelt, Ardi Roelofs, & Antje S. Meyer (1999). A theory of lexical access in speech production. BBS 22(1):1–75.**

Abstract of the original article: Preparing words in speech production is normally a fast and accurate process. We generate them two or three per second in fluent conversation; and overtly naming a clear picture of an object can easily be initiated within 600 msec after picture onset. The underlying process, however, is exceedingly complex. The theory reviewed in this target article analyzes this process as staged and feedforward. After a first stage of conceptual preparation, word generation proceeds through lexical selection, morphological and phonological encoding, phonetic encoding, and articulation itself. In addition, the speaker exerts some degree of output control, by monitoring of self-produced internal and overt speech. The core of the theory, ranging from lexical selection to the initiation of phonetic encoding, is captured in a computational model, called WEAVER + +. Both the theory and the computational model have been developed in interaction with reaction time experiments, particularly in picture naming or related word production paradigms, with the aim of accounting for the real-time processing in normal word production. A comprehensive review of theory, model, and experiments is presented. The model can handle some of the main observations in the domain of speech errors (the major empirical domain for most other theories of lexical access), and the theory opens new ways of approaching the cerebral organization of speech production by way of high-temporal-resolution imaging.

Syntactic representation in the lemma stratum

Holly P. Branigan and Martin J. Pickering

Department of Psychology, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh EH8 9JZ, Scotland, United Kingdom. **Holly.Branigan@ed.ac.uk Martin.Pickering@ed.ac.uk**

Abstract: Levelt, Roelofs, & Meyer (henceforth Levelt et al. 1999) propose a model of production incorporating a lemma stratum, which is concerned with the syntactic characteristics of lexical entries. We suggest that syntactic priming experiments provide evidence about how such syntactic information is represented, and that this evidence can be used to extend Levelt et al.'s model. Evidence from syntactic priming experiments also supports Levelt et al.'s conjecture that the lemma stratum is shared between the production and comprehension systems.

An important part of Levelt et al.'s (1999) impressively detailed model of lexical access in production is the distinction between the lemma stratum, which is concerned with syntactic information, and the form stratum, which is concerned with morpho-phonological information. Following an initial stage where appropriate lexical concepts are activated, lexical processing proceeds via the selection of a lemma and its associated syntactic characteristics. Selecting the lemma "escort," for example, leads to the retrieval of information that it is a transitive verb, and the setting of diacritical parameters for features such as tense, aspect, number, and so on. However, their model provides relatively little detail about the way in which such information is represented. In particular, Levelt et al. do not consider the representation of combinatorial information (i.e., information that specifies how a word may combine with other linguistic units to form larger structures). Clearly, accessing such combinatorial information is crucial for producing multiple word utterances, and so, for embedding their model of single word production into a more comprehensive model of language production (cf. commentaries by Ferreira [1999]; Gordon [1999]; and Roberts et al. [1999]).

In a recent series of experiments using a syntactic priming paradigm (e.g., Branigan et al. 2000; Pickering & Branigan 1998), we have explored how syntactic information might be represented in the lemma stratum. Syntactic priming is the phenomenon whereby previous processing of a sentence with a particular syntactic structure (e.g., a double object structure like "The boy gave the girl the apple") increases the likelihood of producing that structure in a subsequent utterance. Previous experiments have shown that this tendency cannot be attributed to thematic, lexical, or metrical factors (Bock 1986; 1989; Bock & Loebell 1990). We have argued that syntactic priming effects are informative about syntactic representation because they depend upon the processor recognising a syntactic relationship between two otherwise unrelated stimuli (Branigan et al. 1995; Pickering & Branigan 1999).

In Pickering and Branigan (1998), we proposed an extension to Levelt et al.'s model of the lemma stratum. As in their model, we suggested that lemma nodes representing the base (uninflected) form of a word are linked to nodes representing category information (e.g., noun, verb), and to nodes representing feature information (e.g., tense, aspect, number). Each category and each feature is encoded via a single node. Therefore, all verb lemmas connect to the same verb category node, to the same present tense node, and so on. In addition, we proposed a set of combinatorial nodes that encode combinatorial potential. For example, the lemma for a verb that can appear in a double object structure is linked to a node that licenses a double object structure. When the verb is selected during production of a sentence, one of the combinatorial nodes linked to it is also selected; this then guides construction of the appropriate syntactic structure. Thus, selection of the node associated with a double object structure licenses construction of that structure. We suggested that syntactic priming occurs because combinatorial nodes retain residual activation after being selected, and this increases the likelihood of their reselection in subsequent processing.

Our extension of Levelt et al.'s model predicts syntactic priming will occur from one form of a verb to another (e.g., from "gives" to "giving") because the combinatorial nodes are linked to featurally unspecified lemma nodes. It also predicts priming will occur from one verb to another (e.g., from "gives" to "shows") because the same combinatorial node is linked to all verbs that can appear in that structure. A series of experiments employing a sentence-completion technique tested these hypotheses (Pickering & Branigan 1998). Overall, we found reliable syntactic priming effects: Participants are more likely to produce a double object target completion after producing a double object prime completion, and similarly for prepositional object completions. As predicted, the magnitude of the effect is not affected by whether the prime and target involve the same or different versions of the same verb. This is strong evidence that the locus of syntactic information is indeed the (featurally unspecified) lemma, and not the (featurally specified) word form, as proposed by Caramazza and Miozzo (1997; cf. commentary by Harley [1999]). Priming also transfers from one verb to another, though the magnitude of the effect is smaller than when the same verb is repeated. In addition, we found that priming does not depend upon the exact repetition of terminal categories, which suggests that information is encoded by the combinatorial nodes in terms of abstract phrasal categories.

In subsequent experiments (Branigan et al. 2000), we have found that syntactic priming effects also occur in dialogue. Speakers are more likely to produce a structure if they have just heard that structure produced by another speaker. These results provide strong evidence that comprehension and production access

shared syntactic representations, and hence that the lemma level is shared between the comprehension and production systems, as Levelt et al. hypothesized. In fact, our findings are much stronger evidence for a shared level of syntactic representation than is Levelt et al.'s own evidence, which depends upon semantic interference – by hypothesis mediated by the lemma level – from visually presented distractors during picture naming. Our results show further how it is possible to integrate the study of comprehension and production, as Cutler and Norris (1999) espoused in their commentary, within Levelt et al.'s framework.

Overall, the syntactic priming results are in keeping with our extension of Levelt et al.'s model. These results demonstrate that this type of framework can account for not only single word production, but also aspects of multiple-word utterance production. Furthermore, they suggest that at least some aspects of Levelt et al.'s model are relevant for comprehension as well as production.

ACKNOWLEDGMENTS

This work was supported by a British Academy Postdoctoral Fellowship, a British Academy grant, and ESRC research grant no. R000237418.

Lexical access as a brain mechanism*

Friedemann Pulvermüller

MRC Cognition and Brain Sciences Unit, Cambridge CB2 2EF, United Kingdom. **friedemann.pulvermuller@mrc-cbu.cam.ac.uk**
http://www.mrc-cbu.cam.ac.uk/People/Friedemann.Pulvermuller.html

*This commentary originally appeared in the Levelt et al. treatment in BBS 22(1) (pp. 52–54).

Abstract: The following questions are addressed concerning how a theory of lexical access can be realized in the brain: (1) Can a brainlike device function without inhibitory mechanisms? (2) Where in the brain can one expect to find processes underlying access to word semantics, syntactic word properties, phonological word forms, and their phonetic gestures? (3) If large neuron ensembles are the basis of such processes, how can one expect these populations to be connected? (4) In particular, how could one-way, reciprocal, and numbered connections be realized? and, (5) How can a neuroscientific approach for multiple access to the same word in the course of the production of a sentence?

A processing model of lexical access such as the one described in detail in the target article is not necessarily a theory about brain mechanisms. Nevertheless, it may be fruitful to ask how the model can be translated into the language of neurons.

Feedback regulation is necessary! The brain is a device with extreme plasticity. Early in ontogenesis, neurons rapidly grow thousands of synapses through which they influence their neighbors and, in turn, receive influence from other neurons. These synaptic links become stronger with repeated use. Therefore, a particular brain-internal data highway that initially consists of a few fibers, may later include thousands or millions of cables with weak synaptic links, and may finally exhibit a comparably large number of high-impact connections. In this case, the same input to the system will lead early on to a minimal wave of activity, but finally lead to a disastrous breaker. A system with such an enormous variation of activity levels requires a regulation mechanism in order to function properly (Braitenberg 1978). The task of this mechanism would be to enhance or depress the global level of activity to keep it within the limits of optimal neuronal functioning.

A simple way to regulate activity in a neuronal system is to monitor activity levels of all neurons, calculate their sum, and provide an additional input to the system that is excitatory if this sum is small (to prevent extinction of excitation), but inhibitory if it is large (to prevent overactivation). Thus, a mechanism of inhibition (or disfacilitation) appears necessary in any brainlike model.

Levelt, Roelofs & Meyer (Levelt et al.) state that their model does not include inhibition (sect. 3.2.2) and the fact that it does not may be interpreted as one of their minimal assumptions – evidencing a research strategy guided by Ockham's razor. Certainly, looking at the theory in abstract space, the assumption of inhibitory links would be an additional postulate that made it less economical and therefore less attractive. However, considering the brain with its well-known intracortical and striatal inhibitory neurons that are likely to be the basis of feedback regulation (Braitenberg & Schüz 1991; Fuster 1994; Wickens 1993), it does not seem desirable to propose that inhibitory mechanisms are absent in a model meant to mirror brain functioning.

Would this mean that the theory proposed by Levelt et al. is unrealistic from the perspective of brain theory? Certainly not. Although such mechanisms are not explicitly postulated or wired into the network (and therefore do not affect activation spreading), they kick in at the level of node *selection* where Luce ratios are calculated to obtain the probability with which a preactivated representation is selected (fully activated, so to speak). The probability that a particular node is selected depends upon its actual activity value *divided by* the sum of the activation values in a particular layer. Because the calculation performed is very similar to what a regulation device would do, one may want to call this implicit, rather than explicit, inhibition (or regulation). To make it explicit in the network architecture, an addition device in series with numerous intra-layer inhibitory links would have to be introduced. Thus, the model includes inhibition – although on a rather abstract level – and this makes it more realistic from the neurobiological perspective.

Brain loci of lexical access. Where in the brain would one expect the proposed computation of lexical concept, lexical syntax (lemma), word form, and phonetics? Most likely, phonological plans and articulatory gestures are wired in primary motor and premotor cortices in the inferior frontal lobe. The percepts and motor programs to which words can refer probably correspond to activity patterns in various sensory and motor cortices and thus may involve the entire cortex or even the forebrain. More specificity is desirable here; for example, words referring to movements of one's own body are likely to have their lexical concept representations localized in motor cortices and their vicinity, while lexical concepts of words referring to objects that one usually perceives visually should probably be searched for in visual cortices in occipital and inferior temporal lobes (Pulvermüller 1996; Warrington & McCarthy 1987).

Between phonetic-phonological and lexical-semantic representations the model postulates lemmas whose purpose can be considered to be three-fold: (1) not only do they glue together the meaning and form representations of a word, but, in addition, (2) they bind information about the word's articulation pattern and its sound image. Furthermore, (3) lemmas are envisaged to store syntactic knowledge associated with a word.

Intermediary neuronal units mediating between word form and semantics – the possible counterparts of lemmas have been proposed to be housed in the inferior temporal cortex (Damasio et al. 1996). The present theory would predict that lesions in the "lemma area" lead to a deficit in accessing syntactic knowledge about words (in addition to a deficit in naming). However, lesions in inferior temporal areas can lead to a category-specific naming deficit while syntactic knowledge is usually spared. Hence, it appears unlikely that lemmas are housed in the inferior temporal lobe. Is there an alternative to Damasio's suggestion?

One of the jobs of a lemma is to link the production network to the perception network (sect. 3.2.4). On the receptive side, sound waves and features of speech sounds activate neurons in the auditory cortex in the temporal lobe, and in order to store the many-many relation between acoustic phonetic features and articulatory phonetic features, it would have advantages to couple the respective neuron populations in auditory and motor cortices. Such coupling, however, is not trivial, because, for example, direct neuroanatomical connections between the primary motor and auditory cortices are rare, if they exist at all. Therefore, the connection can only be indirect and the detour to take on the articulatory-

acoustic path would probably lead through more anterior frontal and additional superior temporal areas (Pulvermüller 1992). In contrast to the primary areas, these areas are connected to each other, as can be inferred from neuroanatomical studies in macaca (Deacon 1992; Pandya & Yeterian 1985). The coupling of acoustic and articulatory information will, therefore, involve additional neurons that primarily serve the purpose of binding linguistic information. Thus, the physical basis of lemmas may be distributed neuron populations including at least neurons in inferior prefrontal areas (in Brodmann's terminology, areas 44, 45, and perhaps 46) and in the superior temporal lobe (anterior and posterior parts of area 22 and perhaps area 40). These neurons may not only be the basis of the binding of information about production and perception of language, they may well be the targets of connections linking the word form to its meaning, and, most important, their mutual connections may store syntactic information about a word. This proposal is consistent with the neurological observation that lesions in anterior or posterior perisylvian sites (but not in the inferior temporal lobe) frequently lead to a syntactic deficit called agrammatism (Pulvermüller 1995; Vanier & Caplan 1990).

Reciprocal, one-way, and numbered connections. Statistics of cortico-cortical connections suggest that two individual pyramidal neurons located side by side have a moderate (1–2%) probability of exhibiting one direct synaptic link, and only a very low probability of having two or more links (Braitenberg & Schüz 1991). Because synaptic connections are always one-way, a model for interaction of individual neurons may, therefore, favor one-way connections. Language mechanisms, however, are probably related to interactions of large neuronal populations, and if such ensembles include several thousands of neurons, chances are high that two ensembles exhibit numerous connections in both directions (Braitenberg & Schüz 1991). Hence the zero-assumption should probably be reciprocal connections between neuronal representations of cognitive entities such as lemmas, word forms, and their meanings.

The model postulates reciprocal connections between semantic and syntactic representations (Fig. 2) and for some within-layer links (see, e.g., Figs. 4, 6, and 7). Lemmas and word forms are connected through unidirectional links and within the form stratum there are directed and numbered links.

How could two large cortical neuron populations be connected in one direction, but lack the reciprocal link? Here are two possibilities: first, one-way connections could involve directed subcortical links (e.g., through the striatum). As an alternative, connections could in fact (i.e., neuroanatomically) be reciprocal, but activity flow during processing could be primarily in one direction. According to the present theory, the processes of naming include activity spreading from the conceptual to the lemma stratum, and from there to the form stratum whence backward flow of activity to the lemmas is prohibited. This could, for example, be due to early termination of the computation if the appropriate lemma is already selected before upward activity from the form stratum can influence the computation. Conceptualizing the process of selection as the full activation (ignition) of a lemma representation that leads instantaneously to equally strong activation of both conceptual and form representations of the same word, it could be stated that such ultimate activation makes additional activity flow impossible or irrelevant. A regulation mechanism (as detailed above) can be envisaged to suppress all activity in competing nodes as soon as selection has taken place (therefore accounting, for example, for the lack of priming of phonological relatives of words semantically related to the target [Levelt et al. 1991] if lemma selection occurs early). Activity flow primarily in one direction can still be accounted for in a system based on the economical assumption of reciprocal connections between large neuronal populations.

Numbered directed connections are proposed to link morpheme and phoneme nodes. Here, the brain-basis of the numbering needs to be specified. Again, there are (at least) two possibilities: First, different axonal conduction delays could cause sequential activation of phoneme nodes. This option has the disadvantage that differences in the delays would be hardwired in the network making it difficult to account for variations between speaking fast and speaking slow. The second alternative would suggest a slight modification of Levelt et al.'s model: phoneme nodes may receive input from morpheme nodes, but their sequence would be determined by connections between phoneme representations. Here, Abeles's (1991) concept of synfire chains comes to mind. A synfire chain is a collection of neurons consisting of subgroups A, B, C . . . with directed links from A to B, B to C, and so on. Each subgroup includes a small number n of neurons, $7 < n < 100$, and therefore, the assumption of one-way connections appears consistent with the statistics of cortical connectivity (Braitenberg & Schüz 1991).

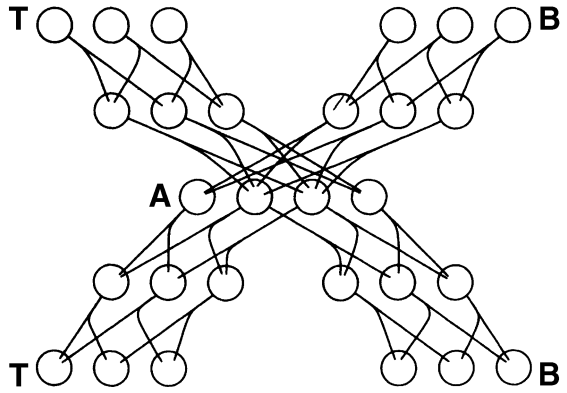

Figure 1 (Pulvermüller). Synfire chains possibly underlying the serial order of phonemes in the words "tab" and "bat." Circles represent neurons, and lines, their connections (the penetrated neurons being the ones that receive activation). The A (or /æ/-sound) is shared by the two words, but its neuronal counterparts are not identical – they have overlapping representations, the non-overlapping neurons (leftmost and rightmost neurons in middle row) processing information about the sequence in which the phoneme occurs ("context-sensitive neurons"). Also the syllable-initial and syllable-final phonemes have distinct representations. If all neurons have a threshold of 3 and receive 1 input from their respective lemma node, selection of one of the word-initial phoneme representations (uppermost triplets) leads to a well-defined activation sequence spreading through the respective chain (but not through the competitor chain). (Modified from Braitenberg & Pulvermüller 1992.)

Because phonemes can occur in variable contexts, it is not sufficient to assume that phoneme representations are the elements corresponding to the neuronal subgroups of the synfire chains in the phonological machinery (Lashley 1951). In order to distinguish the phonemes in "bat" and "tab," it is necessary to postulate that not phonemes, but phonemes-in-context are the elements of representation. Thus, the representation of a /æ/ following a /b/ and followed by a /t/ would be distinct from that of an /æ/ followed by a /b/ and preceded by a /t/ (cf. Wickelgren 1969). In addition, it has advantages to distinguish syllable-initial, central, and syllable-final phonemes, as suggested in the target article. The two /æ/s occurring in the words /tæb/ and /bæt/ could be neuronally organized as sketched in Figure 1. The selection of one of the chains would be determined (1) by activating input to all to-be-selected context-sensitive phonemes and (2) by strong input to the first neuronal element that initializes the chain. This proposal opens the possibility of determining the speed with which activity runs through the synfire chain by the amount of activation from the lemma and morpheme representations to context-sensitive phoneme representations.

Predictions about neurobiological mechanisms of language may be helpful for planning experiments in cognitive neuroscience and for interpreting their results. However, these considerations are at present necessarily preliminary, as pointed out in the target article, not only because the proposals may be falsified by future research, but also because they leave so many questions unanswered. For example, how is it possible to model multiple occurrences of a particular word (same form, same syntax, same meaning) in a given sentence? A not so attractive possibility would be that there are multiple representations for every word type in the processing model or its neurobiological counterpart. Other solutions may make the models much more complicated. Although it is clear that we can, at present, only scratch the surface of lexical processes in the brain, Levelt et al.'s target article clearly evidences that the insights obtained so far are worth the scientific enterprise.

ACKNOWLEDGMENTS

This work is supported by grants from the Deutsche Forschungsgemeinschaft. I am grateful to Valentino Braitenberg and to Bettina Mohr for comments on an earlier version of the text.

Authors' Response

BBS Note: **The original manuscript of this Response article was received on January 14, 2000.**

Relations of lexical access to neural implementation and syntactic encoding

Willem J. M. Levelt,[a] Antje S. Meyer,[b] and Ardi Roelofs[a]

[a]*Max Planck Institute for Psycholinguistics, 6500 AH Nijmegen, The Netherlands;* [b]*School of Psychology, University of Birmingham, Birmingham, B15 2TT, United Kingdom.* **pim@mpi.nl A.S.Meyer@bham.ac.uk ardi@mpi.nl http://www.mpi.nl http://www.bham.ac.uk/psychology/research_03/lang_cog.htm**

Abstract: How can one conceive of the neuronal implementation of the processing model we proposed in our target article? In his commentary (Pulvermüller 1999, reprinted here in this issue), **Pulvermüller** makes various proposals concerning the underlying neural mechanisms and their potential localizations in the brain. These proposals demonstrate the compatibility of our processing model and current neuroscience. We add further evidence on details of localization based on a recent meta-analysis of neuroimaging studies of word production (Indefrey & Levelt 2000). We also express some minor disagreements with respect to Pulvermüller's interpretation of the "lemma" notion, and concerning his neural modeling of phonological code retrieval. **Branigan & Pickering** discuss important aspects of syntactic encoding, which was not the topic of the target article. We discuss their well-taken proposal that multiple syntactic frames for a single verb lemma are represented as independent nodes, which can be shared with other verbs, such as accounting for syntactic priming in speech production. We also discuss how, in principle, the alternative multiple-frame-multiple-lemma account can be tested empirically. The available evidence does not seem to support that account.

Pulvermüller discusses possible neural mechanisms for the implementation of our computational model of lexical access (Pulvermüller 1999, reprinted here). His starting point is clear and correct: The processing model is not a theory about brain mechanisms. It is not a "'neural model" or anything of the sort. It is a psychological processing model formalized in terms of a rather classical spreading activation architecture. The issue of the model's potential neurological underpinnings is of great importance. On the one hand, the model should not be incompatible with existing neuroscience. For instance, WEAVER's chronometric properties should not violate known neurological limitations. Providing potential neural mechanisms for implementing fragments of the model amounts to providing existence proofs for compatibility. On the other hand, the theory of access can be a guide or tool for exploring the patterns of cerebral activations obtained in neuroimaging studies of word production, which involve tasks ranging from picture naming and verb generation to word and nonword reading. A rather coherent pattern of mappings between processing mechanisms in the theory and brain localizations emerges from study.

The first issue addressed by **Pulvermüller** concerns inhibition. Clearly, neuronal functioning would get disrupted without inhibitory regulation. How can this be compatible with the absence of inhibitory connection in the WEAVER model? Here Pulvermüller correctly observes that WEAVER incorporates an equivalent of inhibition, namely, in the lexical competition governed by Luce's rule.

Next, **Pulvermüller** addresses the issue of the brain loci corresponding to various computations in the model and we agree with several of his proposals. To shortcut somewhat, we refer to the above-mentioned meta-analysis by Indefrey and Levelt (2000). There each word production task used in the literature was analyzed as the combination of a "core" process and a "lead-in" process. A core process is any consecutive subset of stages in the target article's theory, ranging from conceptual preparation to articulation. The lead-in process for a given task is the task-specific initiation of these core processes. For instance, picture naming has visual object recognition as its lead-in process, followed by a core process consisting of all stages of word production. Word reading has visual word recognition as the lead-in process. It is followed by core processes from phonological code retrieval, via syllabification down to articulation. Nonword reading has some form of grapheme-phoneme conversion as the lead-in process. There is no phonological word code retrieval here; the core process begins with syllabification and is completed with articulation.

By comparing observed cerebral activations between critical pairs of task, and using a statistical criterion, the various core processes in the theory could be related to smaller or larger foci. For instance, the critical difference between word and nonword reading resides in accessing a word's phonological code. The meta-analysis indicates Wernicke's area as being involved in this core operation. Similarly, the studies indicate that syllabification involves the left inferior frontal gyrus, whereas phonetic encoding and articulation show the expected bilateral involvement of ventral sensorimotor areas. The tasks used in the imaging literature did not allow us to distinguish between conceptual preparation and lemma access. The statistically common region in the imaging studies relating to this pair of processes turned out to be in the midpart of the left middle temporal gyrus. However, the subtraction logic of the meta-analysis would necessarily miss the variability in cortical representation for different semantic fields, such as tools, vegetables, and animals. **Pulvermüller** correctly points to this issue, which has become a hot topic in imaging and patient studies of

word processing (see Martin 1998 for a review). It complicates the search for the localization of lemma-related operations.

In one point, **Pulvermüller** overstates the role of lemmas. In our theory, lemmas do not have a direct role in binding the word's articulation pattern and sound image. We do assume lemmas are shared between production and perception of speech, but on the production side their direct link is to one or more morphemes (i.e., abstract phonological codes), not to articulation patterns (see Fig. 2 of the target article). The articulation pattern is the product of phonological encoding, phonetic encoding, and articulatory motor action; it has a quite variable, indirect relation to lemmas. It is therefore not necessary to relate lemmas to an extensive network ranging all the way from auditory to primary motor cortices, as Pulvermüller suggests.

A further issue addressed by **Pulvermüller** concerns the ways in which one-way connections in the processing model can be neurologically implemented. This is an important issue. It is convincingly argued in the commentary that any pair of cell assemblies must involve bilateral connections. But the model contains several one-way connections, in particular, those leading from lemmas to word forms (morphemes/phonological codes). If the corresponding linguistic operations involve different, but connected, regions, then why does one region's activation not affect the other region's operations? According to Pulvermüller, existing feedback between brain regions need not have behavioral consequences. For instance, the fast operation of lemma selection may be completed before the region is reactivated by feedback from a phonology-dedicated region. Although it is satisfying to see there is no threatening incompatibility here either, we would not like to shortcut the issue this way. Psychologically, it would predict evidence for feedback in cases where lemma selection is slow (e.g., when there is strong lemma competition). There is no evidence this is in fact the case. Neurologically, it seems to imply that reciprocal connections serve activation. They may as well serve inhibition or far more complex forms of control, such as the equivalent of the verification operation in WEAVER, which serves binding (see sect. 3.2.3 in the target article).

Finally, **Pulvermüller** considers possible neural mechanisms for realizing "numbered connections" in WEAVER. In the model, the segments in a retrieved phonological code are numbered. For instance, the code for *dense* consists of numbered segments /d/, /e/, /n/, and /s/, where the numbering specifies the position of the segments in the word. The same segments are numbered differently in the code for *send.* Pulvermüller proposes to handle this by means of synfire chains (Abeles 1991). He rejects the simplest version of this, that is, chains linking the neuronal representations of the phonemes /d/, /e/, /n/, and /s/ as d→e→n→s for the word *dense,* and as s→e→n→d for the word *send.* Our reason for rejecting this would be that the experimental evidence reported in section 6.4.1 of the target article supports the notion that all of a code's segments are simultaneously, not sequentially, activated. Pulvermüller's stated reason is: "not phonemes, but phonemes-in-context are the elements of representation." So, for example, the /e/ phoneme will be slightly different in *dense* and in *send.* Although there is good phonetic evidence for this type of difference, it cannot be an argument for proposing the more complex synfire representation given in Pulvermüller's Figure 1, where phonemic representations are slightly different in different contexts. A first problem is that Pulvermüller's synfire chain produces sequential activation of a code's phonological segments. However, as mentioned, this is not what we find in our experiments. A second problem is that such representations will hamper the variable phonological encoding the model must allow for. Take the phonological encoding of *send.* If the speaker formulates the utterance *What shall I send?,* the speaker will encode /send/ as the final syllable of utterance. But if the speaker prepares the utterance *To whom will I send it?,* the final syllables will be /sen-dit/. The phoneme /d/ ends up as a syllable-final in the first case, but as a syllable-initial in the second case. But in both cases it emerges as a segment in the same phonological code for *send.* This shows that segments in the phonological code itself must be context-neutral, not context-sensitive as Pulvermüller proposes. If the retrieved segment /d/ would be context-sensitive – namely, one that is clustered with /n/ and syllable finally – it could not possibly end up in the syllable-initial position of /dit/. In our model, the phonetic context sensitivities that Pulvermüller observes are handled at a later stage, namely, after the phonological syllables have been computed. It is the stage of phonetic encoding discussed in section 7 of the target article. Therefore, it seems to us that more work needs to be done to develop a potential neurological account of our numbered phonemic representations.

Branigan & Pickering correctly point out that our model does not capture syntactic integration, which is, evidently, an important part of language production. However, as we stated in several places, including the title, the target article was never intended to capture syntactic processing.

Branigan & Pickering's proposal concerning the representation of grammatical information about verbs appears to be fully compatible with our view. In particular, we agree that nodes representing lexical grammatical information should be shared between words (see our treatment of the representation of grammatical gender in sect. 5.4 of the target article). Evidently, much more theoretical and empirical work is needed to gain a fuller understanding of the way syntactic information is represented and used. An open representational issue is, for instance, whether alternator verbs like "give" are represented in one lemma with two sets of syntactic nodes, as Branigan & Pickering propose, or as two separate lemmas permitting exactly one frame each, as proposed by Levelt (1989), following Bresnan's "lexical rule" analysis (Bresnan 1982). In Branigan & Pickering's own account (with which we sympathize), the syntactic priming results obtained since Bock's (1986) original study and including the recent strong findings by Pickering and Branigan (1998), cannot distinguish between these theoretical alternatives. In both cases each syntactic frame is represented by an independent syntactic node, accessible to all lemmas that share that frame. Priming results from "reusing" such a node.

There are, however, theoretical reasons for adopting the one-lemma-multiple-frames type of representation. Most verbs have multiple lexical frames, as is increasingly apparent from parsing studies of large text bases (e.g., see Bangalore & Joshi 1999). In many cases these multiple frames do not correspond to multiple verb meanings; hence they are not cases of homonym. Our account of homonyms in section 6.1.3 of the target article assigns multiple lemmas to multiple lexical concepts; homonyms only share their

morphological word form node. A multiple lemma account of a verb's (or other category's) multiple syntactic frames would create an enormous proliferation of lemmas that share the same lexical concept and the same word form. This is not attractive theoretically. It can also be tested empirically. A multiple lemma account predicts lemma competition, given Luce's rule for lemma selection (sect. 5.1 of the target article): the more co-activated lemma nodes for a given verb (or other category), the slower the selection of any one of them. This type of lemma competition is exactly the one we suggested (sect. 5.3.5) for the case of *eyes* (= gaze) versus *eyes* (= plural of sense organ), the co-activation of their lemmas leading to relatively slow selection of either of them. The test for a multiple lemma account of multiple syntactic frames would be to compare selection latencies for verbs (or nouns) that vary in number of frames, but are comparable in all other respects. If no corresponding difference in selection latencies shows up, the multiple lemma account is without support. In fact, the only available evidence (Ferreira 1996) points to faster rather than slower access for multiple frame verbs.

Whatever the solution will be, both accounts require a mechanism for choosing among alternative frames. In the multiple lemma account, this is primarily a choice among lemmas. If this choice is not conceptually driven, how does it function? No particular proposals have been made so far. In the single lemma account, the choice is one among co-activated syntax nodes. Will the choice exclusively depend on the relative accessibility of the alternative frames? This would not be in the spirit of WEAVER. There is always a verification operation to check whether a potential selection is the appropriate one. In the case of the choice of syntactic frame, this verification may involve a check of the availability of the relevant arguments for the frame at hand.

Branigan & Pickering suggest that the results of their syntactic priming experiments offer stronger support for the assumption that lemmas are shared between speech production and comprehension than the results of picture-word interference experiments. We fail to see in which respect the evidence can be viewed as stronger than ours, but it is certainly an excellent additional argument in favor of our proposal.

References

[Note: The letter "r" before author's initials stands for CC Response article references]

Abeles, M. (1991) *Corticonics – Neural circuits of the cerebral cortex.* Cambridge University Press. [rWJML, FP]

Bangalore, S. & Joshi, A. K. (1999) Supertagging: An approach to almost parsing. *Computational Linguistics* 25:237–65. [rWJML]

Bock, J. K. (1986) Syntactic persistence in language production. *Cognitive Psychology* 18:355–87. [HPB, rWJML]

(1989) Closed class immanence in sentence production. *Cognition* 31:163–86. [HPB]

Bock, J. K. & Loebell, H. (1990) Framing sentences. *Cognition* 35:1–39. [HPB]

Braitenberg, V. (1978) Cell assemblies in the cerebral cortex. In: *Theoretical approaches to complex systems. Lecture notes in biomathematics, vol. 21,* ed. R. Heim & G. Palm. Springer. [FP]

Braitenberg, V. & Pulvermüller, F. (1992) Entwurf einer neurologischen Theorie der Sprache. *Naturwissenschaften* 79:103–17. [FP]

Braitenberg, V. & Schüz, A. (1991) *Anatomy of the cortex. Statistics and geometry.* Springer. (2nd edition). [FP]

Branigan, H. P., Pickering, M. J. & Cleland, A. A. (2000) Syntactic coordination in dialogue. *Cognition* 75:B13–B25. [HPB]

Branigan, H. P., Pickering, M. J., Liversedge, S. P., Stewart, A. J. & Urbach, T. P. (1995) Syntactic priming: Investigating the mental representation of language. *Journal of Psycholinguistic Research* 24:489–506. [HPB]

Bresnan, J., ed. (1982) *The mental representation of grammatical relations.* MIT Press. [rWJML]

Caramazza, A. & Miozzo, M. (1997) The relation between syntactic and phonological knowledge in lexical access: Evidence from the "tip-of-the-tongue" phenomenon. *Cognition* 64:309–43. [HPB]

Cutler, A. & Norris, D. (1999) Sharpening Ockham's razor. *Behavioral and Brain Sciences* 22(1):40–41. [HPB]

Damasio, H., Grabowski, T. J., Tranel, D., Hichwa, R. D. & Damasio, A. R. (1996) A neural basis for lexical retrieval. *Nature* 380:499–505. [FP]

Deacon, T. W. (1992) Cortical connections of the inferior arcuate sulcus cortex in the macaque brain. *Brain Research* 573:8–26. [FP]

Ferreira, F. (1999) Prosody and word production. *Behavioral and Brain Sciences* 22(1):43–44. [HPB]

Ferreira, V. (1996) Is it better to give than to donate? Syntactic flexibility in language production. *Journal of Memory and Language* 35:724–55. [rWJML]

Fuster, J. M. (1994) *Memory in the cerebral cortex. An empirical approach to neural networks in the human and nonhuman primate.* MIT Press. [FP]

Gordon, P. C. (1999) Naming versus referring in the selection of words. *Behavioral and Brain Sciences* 22(1):44. [HPB]

Harley, T. A. (1999) Will one-stage and no feedback suffice in lexicalization? *Behavioral and Brain Sciences* 22(1):45. [HPB]

Indefrey, P. & Levelt, W. J. M. (2000) The neural correlates of language production. In: *The new cognitive neurosciences,* ed. M. Gazzaniga. MIT Press. [rWJML]

Lashley, K. S. (1951) The problem of serial order in behavior. In: *Cerebral mechanisms in behavior: The Hixxon symposium,* ed. L. A. Jeffress. Wiley. [FP]

Levelt, W. J. M. (1989) *Speaking: From intention to articulation.* MIT Press. [rWJML]

Levelt, W. J. M., Roelofs, A. & Meyer, A. S. (1999) A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22(1):1–75. [HPB]

Levelt, W. J. M., Schriefers, H., Vorberg, D., Meyer, A. S., Pechmann, T. & Havinga, J. (1991) The time course of lexical access in speech production: A study of picture naming. *Psychological Review* 98:122–42. [FP]

Martin, A. (1998) Organization of semantic knowledge and the origin of words in the brain. In: *The origin and diversification of language,* ed. N. G. Jablonski & L. C. Aiello, pp. 69–88. California Academy of Sciences. [rWJML]

Pandya, D. N. & Yeterian, E. H. (1985) Architecture and connections of cortical association areas. In: *Cerebral cortex. Vol. 4: Association and auditory cortices,* ed. A. Peters & E. G. Jones. Plenum Press. [FP]

Pickering, M. J. & Branigan, H. P. (1998) The representation of verbs: Evidence from syntactic persistence in written language production. *Journal of Memory and Language* 39:633–51. [HPB, rWJML]

Pickering, M. J. & Branigan, H. P. (1999) Syntactic priming in language production. *Trends in Cognitive Sciences* 3:136–41. [HPB]

Pulvermüller, F. (1992) Constituents of a neurological theory of language. *Concepts in Neuroscience* 3:157–200. [FP]

(1995) Aggramatism: Behavioral description and neurobiological explanation. *Journal of Cognitive Neuroscience* 7:165–81. [FP]

(1996) Hebb's concept of cell assemblies and the psychophysiology of word processing. *Psychophysiology* 33:317–33. [FP]

(1999) Lexical access as a brain mechanism. *Behavioral and Brain Sciences* 22(1):52–54. (Reprinted in this issue of *BBS*). [rWJML]

Roberts, B., Kalish, M., Hird, K. & Kirsner, K. (1999) Decontextualised data IN, decontextualised theory OUT. *Behavioral and Brain Sciences* 22(1):54–55. [HPB]

Vanier, M. & Caplan, D. (1990) CT-scan correlates of agrammatism. In: *Agrammatic aphasia: A cross-language narrative sourcebook, vol. 1,* ed. L. Menn & L. K. Obler. John Benjamins. [FP]

Warrington, E. K. & McCarthy, R. A. (1987) Categories of knowledge: Further fractionations and an attempted integration. *Brain* 110:1273–96. [FP]

Wickelgren, W. A. (1969) Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review* 76:1–15. [FP]

Wickens, J. R. (1993) *A theory of the striatum.* Pergamon Press. [FP]

Commentary on **Gerard O'Brien & Jonathan Opie (1999). A connectionist theory of phenomenal experience. BBS 22(1):127–196.**

Abstract of the original article: When cognitive scientists apply computational theory to the problem of phenomenal consciousness, as many have been doing recently, there are two fundamentally distinct approaches available. Consciousness is to be explained either in terms of the nature of the representational vehicles the brain deploys or in terms of the computational processes defined over these vehicles. We call versions of these two approaches *vehicle* and *process* theories of consciousness, respectively. However, although there may be space for vehicle theories of consciousness in cognitive science, they are relatively rare. This is because of the influence exerted, on the one hand, by a large body of research that purports to show that the explicit representation of information in the brain and conscious experience are dissociable, and on the other, by the classical computational theory of mind – the theory that takes human cognition to be a species of symbol manipulation. Two recent developments in cognitive science combine to suggest that a reappraisal of this situation is in order. First, a number of theorists have recently been highly critical of the experimental methodologies used in the dissociation studies – so critical, in fact, that it is no longer reasonable to assume that the dissociability of conscious experience and explicit representation has been adequately demonstrated. Second, classicism, as a theory of human cognition, is no longer as dominant in cognitive science as it once was. It now has a lively competitor in the form of connectionism; and connectionism, unlike classicism, does have the computational resources to support a robust vehicle theory of consciousness. In this target article we develop and defend this connectionist vehicle theory of consciousness. It takes the form of the following simple empirical hypothesis: *phenomenal experience consists of the explicit representation of information in neurally realized parallel distributed processing (PDP) networks.* This hypothesis leads us to reassess some common wisdom about consciousness, but, we argue, in fruitful and ultimately plausible ways.

Explicitness and nonconnectionist vehicle theories of consciousness

Fernando Martínez-Manrique

Departimento Filosofía, Universidad de Granada, 18011 Granada, Spain.
fernan_martinez@yahoo.com

Abstract: O'Brien & Opie's connectionist vehicle theory of consciousness is heavily dependent on their notion of explicitness as (1) structural and (2) necessary and sufficient for consciousness. These assumptions unnecessarily constrain their position: the authors are forced to find an intrinsic property of patterns that accounts for the distinction between conscious and unconscious states. Their candidate property, stability, does not capture this distinction. Yet, I show that we can drop assumptions (1) and (2) and still develop a vehicle theory of consciousness. This alternative is better served by models that incorporate both connectionist and symbolic representations.

Representational theories can account for consciousness either in terms of the representational vehicles underlying conscious states, or in terms of the computational processes that operate upon the vehicles. In contrast with the dominant process theories, O'Brien & Opie (1999; henceforth O&O) pursue a vehicle theory of consciousness in a connectionist framework. They identify conscious states with explicit representations, and argue that the latter are realized as stable patterns of activation in networks. In this commentary I focus on the relationship between explicitness and consciousness to suggest the possibility of a vehicle theory that is not purely connectionist.

As other commentators (Clapin, Schröder; cf. *BBS* 22[1], 1999) remarked, there are two views on explicitness (Kirsh 1990). In the structural view, information is explicit when it has definite location and meaning; in the process view, explicitness forms a continuum according to the accessibility of information. These views should not be conflated with the two theories of consciousness. In fact, "structural/process explicitness" and "process/vehicle consciousness" can be combined as independent dimensions. We obtain four positions:

1. Structural explicitness and process consciousness
2. Structural explicitness and vehicle consciousness
3. Process explicitness and process consciousness
4. Process explicitness and vehicle consciousness.

Structural explicitness and process consciousness represent the traditional approaches, so it is not surprising that their combination in position (1) results in classical cognitive science. O&O explore possibility (2). They differ from classicism in their endorsement of vehicle consciousness, but they maintain a structural view on explicitness. Their rejection of classicism as a candidate vehicle theory of consciousness is a consequence of the conjunction of structural explicitness with a second assumption: that explicitness is necessary and sufficient for consciousness. As symbolic representations are always structurally explicit, it follows that their contents are always conscious. Hence, classicism cannot ground the difference between conscious and unconscious states on the property of explicitness. However, both assumptions can be dropped while we maintain a vehicle theory of consciousness.

Clapin objected that explicitness in networks only makes sense in terms of availability of information (process explicitness). Thus, if conscious states are identified with explicit states, it follows that consciousness is also dependent on availability (process consciousness). Hence O&O's theory would occupy position (3) above, not position (2). O&O might counter this objection claiming that the property of stability is the intrinsic, structural feature of networks that sustains the explicit/implicit distinction. In defense of this view they affirm that "prior to stabilization there are no objects physically present in these networks whose intrinsic structural properties can stand in [a structurally isomorphic] relation to elements of the target domain" (p. 181).

That answer, however, means only that stabilization is a way of "fixing" the representation, not that stability is an intrinsic component of the representation itself. Consider two identical patterns, one stable and the other transient: if there is structural isomorphism in the former, there is no cogent reason to deny it in the latter. Compare it with maps, a paradigm of structural isomorphism. The stable pattern is similar to the final map, and the transient pattern with one of the previous sketches. If the final map got the isomorphism right, then an identical sketch must preserve the same isomorphism. There is nothing intrinsic in the structure of the respective maps/patterns that sustains a principled distinction. If there were such a distinction, then the classicist could adopt the same strategy: there are stable and non-stable symbolic representations (the latter being, say, symbols that are constructed on the fly and then erased), and only the former are explicit and conscious. Surely O&O do not want to say that these representations are structurally different: they are both symbolic and it is their properties qua symbols that are structurally relevant. Similarly, in the connectionist case what matters is the intrinsic structure of patterns qua patterns, regardless of their stability.

If we drop structural explicitness, then to avoid falling into posi-

tion (3) we must also drop the assumption of identity between explicitness and consciousness. As several commentators suggested (Church, Cleeremans & Jimenez, Dennett & Westbury, Kurthen, McDermott, Van Gulick, and Wolters & Phaf; cf. *BBS* 22[1], 1999), explicitness could be necessary but not sufficient for consciousness. O&O's only support for the identification of explicitness and consciousness comes from their reappraisal of the dissociation studies. However, this is possibly the most questionable point in their paper. Lacking a final verdict on the issue, it seems that their persistence in identifying both properties is due to their thinking that "it is clearly incompatible with the connectionist vehicle theory of phenomenal experience [to assume] the operation of explicitly represented information that does not figure in consciousness" (p. 187). I claim that there is no such incompatibility, insofar as we drop structural explicitness. This leads us to position (4).

First, all that a vehicle theory of consciousness demands, according to Thomas & Atkinson and Van Gulick (cf. *BBS* 22[1], 1999), is a principled distinction between kinds of representations, R and R^1, so that the intrinsic properties of a given kind make it the basis of conscious experience. Second, from a process explicitness viewpoint, the more accessible some information I is, the more explicit I will be. Third, from a vehicle consciousness perspective we can say that I becomes conscious only when it is explicit and encoded by a specific kind of representation, (say, R^1). This would fill position (4).

A connectionist version of this possibility is: (1a) Two kinds of patterns, P and P^1. (2a) Gradation of explicitness: information in weights is less accessible than information in patterns. (3a) Information in patterns is not immediately conscious; only some patterns are so, say P^1. But now we open the door to classical vehicle theories of consciousness: (1b) Two kinds of symbols, S and S^1. (2b) Gradation of explicitness: some symbolic information is more explicit by being more accessible. (3b) Only an explicit S^1 makes its contents conscious.

Both versions, however, face the same problem: how to single out an intrinsic property that provides a principled distinction between the patterns P and P^1 or the symbols S and S^1. There is an obvious place to look for such a principled structural distinction between representational kinds: the distinction itself between patterns and symbols. Suppose that we allow both kinds of representations in our system. We can fill position (4) as follows: (1c) Two kinds of representations: symbols and patterns. (2c) Gradation of explicitness: from content in weights to content in patterns, and content in accessible symbols. (3c) Content is conscious only when it is rendered into explicit symbolic format. This can require the extraction of the content from the network.

Two notes: First, a "purely vehicle" theory of consciousness need not be "purely connectionist" or "purely symbolic"; it can contain instances of both representational kinds. Second, even if the content has to be extracted for being conscious, this does not make it a process theory. It is not being extracted that makes the content conscious; it is being symbolic that makes it so. If O&O insist that extraction makes this version a process theory of consciousness, then they should equally answer the charge (Mac Aogáin, Wolters & Phaf; cf. *BBS* 22[1], 1999) that a pattern is always the product of some process.

Things are probably much more mixed up than suggested by any simple theory of consciousness. If connectionist and symbolic vehicles belong to different "representational genera" according to the contents they are capable of representing (Haugeland 1991), then they may underlie different kinds of conscious states. On the other hand, it is also dubious that a purely vehicle or a purely process theory will account for consciousness. I have argued elsewhere (Martinez & Ezquerro 1998) that intuitions from the structural and the process views should be integrated to offer an appropriate characterization of explicitness, and an analogous claim can be made with respect to vehicle and process theories of consciousness. In other words, the character of conscious experiences may depend not on what a representation is or on what it does but rather in the subtle interaction of both factors.

ACKNOWLEDGMENT

Preparation of this paper was supported by the MCyT research project BFF2002–03842.

Authors' Response

Vehicle, process, and hybrid theories of consciousness

Gerard O'Brien and Jonathan Opie

Department of Philosophy, School of Humanities, University of Adelaide, South Australia 5005, Australia. **gerard.obrien@adelaide.edu.au**
jon.opie@adelaide.edu.au
http://www.arts.adelaide.edu.au/humanities/gobrien/
http://www.arts.adelaide.edu.au/humanities/jopie/

Abstract: Martínez-Manrique contends that we overlook a possible nonconnectionist vehicle theory of consciousness. We argue that the position he develops is better understood as a hybrid vehicle/process theory. We assess this theory and in doing so clarify the commitments of both vehicle and process theories of consciousness.

In developing the connectionist vehicle theory of phenomenal experience we were mindful of two things: (1) that consciousness is, by and large, a consequence of the brain's representing activity, (2) that current theories of mental representation are heavily influenced by the classical computational theory of mind. Connectionism presents a unique opportunity to rethink consciousness because, unlike classicism, its account of cognition is framed in terms of certain structural properties of the brain. In particular, connectionism distinguishes between two structurally distinct kinds of representing vehicle: connection weight representations, and activation pattern representations. Others have noticed the possibility of identifying phenomenal experience with the relatively transient activation patterns that constantly course across the brain, while assigning connection weights the twin tasks of information storage and computational substrate (Rumelhart et al. 1986, p. 39; Smolensky 1988, p. 13; Lloyd 1991; 1995; 1996). In our target article we sought to further develop and defend this idea, conjecturing that phenomenal consciousness is identical to the vehicles of explicit representation in the brain – such vehicles being understood as stable patterns of neural activation.

Martínez-Manrique, in his useful commentary, argues that we have overlooked a possible variety of vehicle theory, one moreover that contains both connectionist and classical elements. His crucial move, in canvassing this possibility, is to exploit the distinction between structural and process conceptions of explicit representation. In our target article we develop a generic representational framework that characterizes explicit representation in structural terms. Martínez-Manrique observes that there is well-known analysis, primarily due to Kirsh (1990), according to which information is explicit if it is readily accessible by a cognitive system, and is, by degrees, less explicit if it is more difficult to access. As Martínez-Manrique admits, this is a process conception of explicit representation. But one may

recover a vehicle theory of consciousness, he thinks, if explicitness is treated as necessary but not sufficient for consciousness. An additional (vehicle) criterion might be added, to the effect that a widely available representational content will be conscious when its vehicle satisfies some intrinsic, structural constraint. This, claims Martínez-Manrique, ultimately permits a vehicle theory in which connectionist (activation pattern) and classical (symbolic) representations both play a part.

At the outset we must say that **Martínez-Manrique**'s analysis of the space of possible theories seems to us seriously flawed. Contrary to what he claims, one *cannot* coherently combine a vehicle theory of consciousness with a process conception of explicit representation. A vehicle theory of consciousness seeks to explain phenomenal experience in terms of the *intrinsic* nature of the brain's explicit representing vehicles – in terms of what these vehicles *are* rather than what they *do.* A process conception of explicitness holds that information is explicitly represented in a cognitive system when it can be easily accessed. But the ease with which a representational content can be accessed is not solely or even largely determined by the intrinsic properties of the vehicle that carries it; it is determined by the nature of the cognitive system in which that vehicle is embedded. Consequently, there just is no coherent formulation of a vehicle theory of consciousness which adopts a process conception of explicitness: one cannot hope to explain phenomenal consciousness in terms of intrinsic properties of the brain's explicit representing vehicles when explicitness is determined largely by properties extrinsic to these vehicles. We thus hold to our conclusion, drawn in our target article, that only connectionism has the resources to develop a plausible vehicle theory of consciousness.

Given this, perhaps a better interpretation of **Martínez-Manrique**'s commentary is not that there is a nonconnectionist vehicle theory we have overlooked but that there is a way of combining structural and process criteria within a single account – a maneuver which, in effect, generates a hybrid vehicle/process theory. Martínez-Manrique's ultimate suggestion is that a content is conscious "only when it is rendered into explicit symbolic format" (para.8). Being symbolic is the vehicle criterion. What is the process criterion? In typical process accounts a representational content is taken to be conscious when its vehicle is subject to relations of widespread informational access – that is, when it has rich and widespread information processing *effects* on the brain's ongoing operations. However, as we explained previously (O'Brien & Opie 1999, pp. 176–77), any hybrid account that followed this line would violate one of the deepest intuitions we have about consciousness: that *conscious experience makes a difference.* If a symbolic content must give rise to widespread information processing effects in order to enter consciousness, its being conscious cannot be the cause of those effects. But this is not Martínez-Manrique's strategy. Rather than focusing on informational *access,* his process criterion is informational *accessibility:* representational contents are conscious when they are encoded symbolically and can readily be accessed and put to use in the service of cognition. And it might be argued that this change of focus renders his hybrid vehicle/process theory consistent with the causal potency of consciousness.

One obvious problem with any theory that makes informational accessibility, rather than informational access, criterial for consciousness is that it runs the risk of being empirically implausible. Nothing could be clearer than the fact that we have at our fingertips a vast store of unconscious but readily accessible information. **Martínez-Manrique**'s proposal can skirt over this difficulty, however, because it holds that accessibility is insufficient for consciousness; consciousness also requires the satisfaction of a structural (vehicle) criterion. Our worry with this hybrid vehicle/process theory is different, but just as straightforward. We think it unmotivated and unparsimonious. It is unmotivated because, although it is clear why one might seek to explain consciousness by identifying it with either the intrinsic properties of the brain's representing vehicles (in doing so one connects consciousness with the very entities that drive human cognition) or the information processing effects of these representing vehicles (in doing so one connects consciousness with the process of accessing the information these vehicles carry), it is unclear why one would seek to explain consciousness in terms of the fact that certain representational contents are more readily accessible than others. And it is unparsimonious because it accounts for consciousness in terms of both intrinsic and extrinsic properties of the brain's representing vehicles when simpler theories that restrict themselves to one or other class of properties have yet to be fully explored.

In this vein, it is useful to consider why **Martínez-Manrique** so quickly dismisses our connectionist vehicle theory. He does so because he thinks connectionism is incapable of distinguishing conscious representing vehicles from their unconscious counterparts by recourse to a structural criterion of explicitness. And Martínez-Manrique reaches this conclusion by interpreting the *stability* of an activation pattern representation as a *temporal,* rather than a *structural,* property of a neural network. We think Martínez-Manrique is wrong about this. As we were at pains to point out in our original "Authors' Response" (O'Brien & Opie 1999r, pp. 181), there is a widespread misunderstanding of the significance of stability in connectionist networks that issues from a failure to distinguish between the behavior of real neural networks and the properties of their digital simulations. Since this error persists, we will conclude our discussion by briefly revisiting this issue.

In a simulation, a neural network's activity is modeled as an array of *numerical activation values,* which are periodically updated by algorithms that model the network's internal processes. Simulated relaxation search thus proceeds via a sequence of determinate numerical arrays, giving the impression that prior to stabilization a neural network jumps between specific points in its activation space, and hence generates a sequence of short-lived activation patterns before settling into a longer lasting pattern. This is the picture **Martínez-Manrique** has in mind when he claims that there is no intrinsic structural distinction among the "transient" patterns that precede the production of a "stable" pattern, and hence no structural criterion which can ground a distinction between unconscious and conscious states (para. 3). But this picture is misleading. Whenever one employs a numerical value to describe a continuously variable physical property, one is imposing an instantaneous value on this property. Since neural spikes are discrete events, neural spiking rates *do not have instantaneous values;* the notion of a rate, in this case, only makes sense relative to some time window. In a real network, stabilization is a process in which constituent neurons adjust the absolute timing of their spikes until a determinate firing rate

is achieved. Prior to stabilization, neural networks do not jump around between points in activation space. Stabilization is the process whereby a network first generates a determinate activation pattern, and thereby *arrives* at a point in activation space.

So a real neural network does not generate a pattern of activation, and thus a determinate representational content, until it achieves some measure of stability. Consequently, there is no distinction between "stable" and "transient" activation patterns. Stable activation patterns are physical objects, objects moreover that are structurally distinct from a neural network's configuration of connection weights. And it is this distinction, between activation pattern representation and connection weight representation, that according to our vehicle theory marks the boundary between the conscious and the unconscious.

References

[The letter "r" before author's initials indicates Response article references]

Haugeland, J. (1991) Representational genera. In: *Philosophy and connectionist theory,* ed. W. Ramsey, S. P. Stich & D. E. Rumelhart. Erlbaum. [FM-M]

Kirsh, D. (1990) When is information explicitly represented? In: *Information, language and cognition,* ed. P. Hanson. University of British Columbia Press. [FM-M, rGO]

Lloyd, D. (1991) Leaping to conclusions: Connectionism, consciousness, and the computational mind. In: *Connectionism and the philosophy of mind,* ed. T. Horgan & J. Tienson. Kluwer. [rGO]

(1995) Consciousness: A connectionist manifesto. *Minds and Machines* 5:161–85. [rGO]

(1996) Consciousness, connectionism, and cognitive neuroscience: A meeting of the minds. *Philosophical Psychology* 9:61–79. [rGO]

Martínez, F. & Ezquerro, J. (1998) Explicitness with psychological ground. *Minds and Machines* 8:353–74. [FM-M]

O'Brien, G. & Opie, J. (1999) A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences* 22(1):127–96. [FM-M, rGO]

O'Brien, G. & Opie, J. (1999r) Putting content into a vehicle theory of consciousness. (Author's Response to Open Peer Commentary.) *Behavioral and Brain Sciences* 22(1):175–96. [rGO]

Rumelhart, D. E., Smolensky, P., McClelland, J. L. & Hinton, G. E. (1986) Schemata and sequential thought processes in PDP models. In: *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 2: Psychological and Biological Models,* ed. J. L. McClelland & E. E. Rumelhart. MIT Press. [rGO]

Smolensky, P. (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11:1–23. [rGO]

Commentary on **Anne Campbell (1999). Staying alive: Evolution, culture, and women's intrasexual aggression. BBS 22(2):203–252.**

Abstract of the original article: Females' tendency to place a high value on protecting their own lives enhanced their reproductive success in the environment of evolutionary adaptation because infant survival depended more upon maternal than on paternal care and defence. The evolved mechanism by which the costs of aggression (and other forms of risk taking) are weighted more heavily for females may be a lower threshold for fear in situations which pose a direct threat of bodily injury. Females' concern with personal survival also has implications for sex differences in dominance hierarchies because the risks associated with hierarchy formation in non-bonded exogamous females are not off-set by increased reproductive success. Hence among females, disputes do not carry implications for status with them as they do among males, but are chiefly connected with the acquisition and defence of scarce resources. Consequently, female competition is more likely to take the form of indirect aggression or low-level direct combat than among males. Under patriarchy, men have held the power to propagate images and attributions which are favourable to the continuance of their control. Women's aggression has been viewed as a gender-incongruent aberration or dismissed as evidence of irrationality. These cultural interpretations have "enhanced" evolutionarily based sex differences by a process of imposition which stigmatises the expression of aggression by females and causes women to offer exculpatory (rather than justificatory) accounts of their own aggression.

Hierarchy disruption: Women and men

János M. Réthelyi and Mária S. Kopp

Institute of Behavioral Sciences, Semmelweis University, Budapest, 1089, Hungary. **retjan@net.sote.hu kopmar@net.sote.hu**
www.behsci.sote.hu

Abstract: The application of evolutionary perspectives to analyzing sex differences in aggressive behavior and dominance hierarchies has been found useful in multiple areas. We draw attention to the parallel of gender differences in the worsening health status of restructuring societies. Drastic socio-economic changes are interpreted as examples of hierarchy disruption, having differential psychological and behavioral impact on women and men, and leading to different changes in health status.

Campbell's (1999) target article about gender differences in aggression and status-seeking behavior describes a convincing body of evidence and presents a plausible evolutionary explanation. The target article and the commentaries raise a number of questions concerning the consequences and practical implementations of an evolutionary theory. We propose that several new findings in the areas of epidemiology and health psychology yield parallel results that fit well with Campbell's model. The phenomenon of health status deterioration in restructuring societies, primarily those of Central and Eastern Europe, and the until-now not convincingly explained gender differences in health deterioration are results that could serve as a bridge between a behaviorally oriented evolutionary model and large-scale epidemiological findings. Reading the article and the following debate was a profound intellectual experience; the recognition of parallel results between different fields was even more exciting.

Socio-economic changes following political transition in the countries of Central and Eastern Europe have influenced people's lives in a variety of ways. Among these phenomena, one of the most striking is the declining health status of these societies (Feachem 1994). The dynamics of the process show different characteristics in different countries according to the chronological nature of the political changes. In Hungary, deterioration began in the early 1970s at a constant slow grade, and male life expectancy decreased by three years between 1970 and 1995, parallel with political softening and the beginning of economic polarization (Bobak & Marmot 1996; Kopp 2000). As a more severe

example, male life expectancy in Russia fell by six years between 1990 and 1994 (Notzon et al. 1998). Paradoxically, women have not been affected as severely as men by these processes of deterioration, giving rise to a higher gender gap in life expectancy (12.1 years in Russia) and mortality. Gender ratios in mortality of the middle aged have risen threefold in several Eastern European countries (Hungarian Central Statistical Office 1999). According to these epidemiological results, women are better at staying alive. One must ask, what were the toxic effects that induced the fast deterioration of health status and the greater impact on men than on women?

The link between dominance and resource holding in humans can be described in several ways: by means of social status, education, income, occupation, and political influence. These are exactly the factors which the political and socio-economic changes turned upside-down, giving rise to a general loss of control and predictability. Hence, we consider our hypothetical model of hierarchy disruption useful for analyzing the epidemiological phenomena registered recently.

A large body of evidence supports the inverse association between socio-economic status, and morbidity and mortality (Marmot et al. 1991). Worsening health status and rising mortality in connection with socio-economic changes have been similarly thoroughly studied, as has the gender-relatedness of these phenomena (Kopp et al. 1995; Mackenbach et al. 1999; Weidner 1998). In accordance with the literature, our own results from 1988 and 1995 – two turning points during the socio-economic changes – indicate that income showed a strengthening connection to self-reported morbidity in men, measured as the number of sick days per annum, but only to a much lesser degree in women (Kopp et al. 2000; Réthelyi et al. 2002). Men seem to be more susceptible to hierarchy disruption and the loss of hierarchy status.

Parallel findings in primatology are meaningful. From a biological point of view, the political and socio-economic changes may have similarities to patterns referred to analogously as hierarchy disruption, which have been observed in baboons living in patriarchal dominance hierarchies (Sapolsky 1990a; 1990b). Observations among male baboons indicate that higher rank position goes together with protective physiological profiles for stress-related illnesses connected with lower levels of basal cortisol and faster cortisol normalization. However, not rank itself but the sense of control and predictability are the factors that determine physiologic reactions. Dominant males at the time of newly formed hierarchies do not enjoy the beneficial effect of high rank until the new order is settled. Studies regarding female dominance hierarchies in Cynomolgus macaques in connection with coronary artery atherosclerosis found that social subordination increases the development of atherosclerosis in experimental settings. Social isolation, however, had an even greater atherogenic effect on female macaques in similar experimental settings (Shively et al. 1998).

Returning to our original question, we must consider possible psychological mediators of hierarchy disruption. According to our results mentioned earlier, depression is an important mediator between income and self-reported morbidity in men, but not in women. This association might seem paradoxical because women report generally more depression. However, they also report more adaptive coping strategies, and are able to recognize depression and more willingly take effective steps to counter depression, anxiety, and pain of any kind, in forms of health-care utilization (Unruh 1996), a fact cited by Campbell as well. Social support and cohesion are other protective factors which women make more use of (Knox et al. 1998). Besides their important role in health psychology, the evolutionary importance of social support and cohesion in connection with child rearing and human socialization seems plausible, fitting well in Campbell's model. Such a framework is comparable with the results of modern epidemiology. Growing evidence supports the hypothesis that the worsening health status and the evident gender gap in health decline can be explained only by a combination of traditional risk factors and psychosocial factors. Standard risk factors for noncommunicable diseases such as smoking, diet, alcohol consumption, and obesity do not differ sufficiently in Eastern and Western countries to explain the striking differences in health status. However, there are striking differences in psychosocial risk factors such as depression, exhaustion, social support, hostility, and adaptive coping strategies (Kristenson et al. 1998).

In her response to the commentaries, Campbell addresses questions of dominance hierarchies in democracy and capitalism. From an epidemiological point of view, history is teaching us the lesson that neither an ideologically based egalitarianism (i.e., socialism), nor a change to a democratic system, reduced status seeking.

In summary, we suggest an evolutionary mechanism of tradeoffs between the possible costs and benefits of status-seeking behavior and those of social cohesion and integration, which are most apparent at times of hierarchy disruption (Kopp & Réthelyi 2004). Further research on socio-economic factors and health should bring a better understanding of causal relationships and even offer possibilities of social and medical intervention.

Editors' Note: There is no Author's Response to this commentary.

References

Bobak, M. & Marmot, M. (1996) East-West mortality divide and its potential explanations: Proposed research agenda. *British Medical Journal* 312:421–25. [JMR]

Campbell, A. (1999) Staying alive: Evolution, culture and women's intrasexual aggression. *Behavioral and Brain Sciences* 22(2):203–52. [JMR]

Feachem, R. (1994) Health decline in Eastern Europe. *Nature* 367:313–14. [JMR]

Hungarian Central Statistical Office (1999) English supplement of the *Demographic yearbook,* ed. P. Jozan & Á. Meszaros. Hungarian Central Statistical Office. [JMR]

Knox, S. S., Siegmund, K. D., Weidner, G., Ellison, R. C., Adelman, A. & Paton, C. (1998) Hostility, social support, and coronary hearth disease in the National Heart, Lung, and Blood Institute family heart study. *American Journal of Cardiology* 82:1192–96. [JMR]

Kopp, M. S. (2000) Cultural transition. In: *Encyclopedia of stress,* ed. G. Fink. Academic Press. [JMR]

Kopp, M. S. & Réthelyi, J. (2004) Where psychology meets physiology: Chronic stress and premature mortality – the Central-Eastern European health paradox. *Brain Research Bulletin* 62: 351–67. [JMR]

Kopp, M. S., Skrabski, Á. & Szedmak, S. (1995) Socioeconomic factors, severity of depressive symptomatology, and sickness absence rate in the Hungarian population. *Journal of Psychosomatic Research* 39:1019–29. [JMR]

(2000) Psychosocial risk factors, inequality, and self-rated morbidity in a changing society. *Social Science and Medicine* 51:1350–61.

Krtistenson, M., Kucinskiene, Z., Bergdahl, B., Calkauskas, H., Urmonas, V. & Orth-Gom'r, K. (1998) Increased psychosocial strain in Lithuanian versus Swedish men: The LiVicordia study. *Psychosomatic Medicine* 60:277–82. [JMR]

Mackenbach, J. P., Kunst, A. E., Groenhof, F., Borgan, J., Costa, G., Faggiano, F., Jozan, P., Lainsalu, M., Martikainen, P., Rychtarikova, J. & Valkonen, T. (1999) Socioeconomic inequalities in mortality among women and among men: An international study. *American Journal of Public Health* 89:1800–807. [JMR]

Marmot, M. G., Smith, G. D., Stansfeld, S., Patel, C., North, F., Head, J., White, I., Brunner, E. & Feeney, A. (1991) Health inequalities among British civil servants: The Whitehall II study. *Lancet* 333:387–93. [JMR]

Notzon, F. C., Komarov, Y. M., Ermakov, S. P., Sempos, C. T., Marks, J. S. & Sempos, E. V. (1998) Causes of declining life expectancy in Russia. *Journal of the American Medical Association* 10:793–800. [JMR]

Réthelyi, J. M., Purebl, G. & Kopp, M. S. (2002) Sociodemographic and behavioral correlates of depression in Hungarian men and women. In: *Heart disease: Environment, stress, and gender. NATO Science Series, Life and Behavioural Sciences, vol. 327,* ed. G. Weidner, M. Kopp & M. Kristenson, pp. 114–120. IOS Press. [JMR]

Sapolsky, R. M. (1990a) Adrenocortical function, social rank, and personality among wild baboons. *Biological Psychiatry* 28:862–78. [JMR]

(1990b) Stress in the wild. *Scientific American* 262(1):116–23. [JMR]

Shively, A. C., Watson, S. L., Williams, J. K., Adams, M. R. (1998) Social stress, reproductive hormones, and coronary heart disease risk in primates. In: *Women, stress, and heart disease,* ed. K. Orth-Gom'r, M. Chesney & N. K. Wenger. Erlbaum. [JMR]

Unruh, A. M. (1996) Gender variations in clinical pain experience. *Pain* 65:123–67. [JMR]

Weidner, G. (1998) Gender gap in health decline in East Europe. *Nature* 395:835. [JMR]

Commentary on **Friedemann Pulvermüller (1999). Words in the brain's language. BBS 22(2)253–336.**

Abstract of the original article: If the cortex is an associative memory, strongly connected cell assemblies will form when neurons in different cortical areas are frequently active at the same time. The cortical distributions of these assemblies must be a consequence of where in the cortex correlated neuronal activity occurred during learning. An assembly can be considered a functional unit exhibiting activity states such as full activation ("ignition") after appropriate sensory stimulation (possibly related to perception) and continuous reverberation of excitation within the assembly (a putative memory process). This has implications for cortical topographies and activity dynamics of cell assemblies forming during language acquisition, in particular for those representing words. Cortical topographies of assemblies should be related to aspects of the meaning of the words they represent, and physiological signs of cell assembly ignition should be followed by possible indicators of reverberation. The following postulates are discussed in detail: (1) assemblies resembling phonological word forms are strongly lateralized and distributed over perisylvian cortices; (2) assemblies representing highly abstract words such as grammatical function words are also strongly lateralized and restricted to these perisylvian regions; (3) assemblies representing concrete content words include additional neurons in both hemispheres; (4) assemblies representing words referring to visual stimuli include neurons in visual cortices; and (5) assemblies representing words referring to actions include neurons in motor cortices. Two main sources of evidence are used to evaluate these proposals: (a) imaging studies focusing on localizing word processing in the brain, based on stimulus-triggered event-related potentials (ERPs), positron emission tomography (PET), and functional magnetic resonance imaging (fMRI), and (b) studies of the temporal dynamics of fast activity changes in the brain, as revealed by high-frequency responses recorded in the electroencephalogram (EEG) and magnetoencephalogram (MEG). These data provide evidence for processing differences between words and matched meaningless pseudowords, and between word classes, such as concrete content and abstract function words, and words evoking visual or motor associations. There is evidence for early word class-specific spreading of neuronal activity and for equally specific high-frequency responses occurring later. These results support a neurobiological model of language in the Hebbian tradition. Competing large-scale neuronal theories of language are discussed in light of the data summarized. Neurobiological perspectives on the problem of serial order of words in syntactic strings are considered in closing.

Perceptual fluency and lexical access for function versus content words

Sidney J. Segalowitz and Korri Lane

Department of Psychology, Brock University, St. Catharines, Ontario, L2S 3A1, Canada. **sid.segalowitz@brocku.ca** **korrilane@yahoo.com**
http://www.psyc.brocku.ca/people/segalowitz.htm

Abstract: By examining single-word reading times (in full sentences read for meaning), we show that (1) function words are accessed faster than content words, independent of perceptual characteristics; (2) previous failures to show this involved problems of frequency range and task used; and (3) these differences in lexical access are related to perceptual fluency. We relate these findings to issues in the literature on event-related potentials (ERPs) and neurolinguistics.

Pulvermüller (1999) posits that lexical access for function words involves the perisylvian region whereas lexical access for content words additionally involves other cortical areas related to the specific meanings. Function word cell assemblies should produce faster lexical access times, because they are more concise in the geographical sense and possibly because functions whose representations are restricted to this area are deemed to be more automatized (Whitaker 1983). However, the experimental literature on function word and content word lexical access times does not support this. Pulvermüller et al.'s (1995) own data show that lexical decisions are slower for function words than for content words. We (and many others) have found this too: lexical decisions for function words took more that 40 msec longer than for nouns and verbs (which did not differ from each other), $F(2, 34) = 21.9, p < .001$ (Segalowitz & Chevalier, unpublished data).

Some researchers have suggested that the lexical decision paradigm is not an appropriate one for comparing function and content words on access times. Taft (1990) showed that lexical decisions are slower for words that do not comfortably stand alone, whether of the function type or the content type, and Schmauder (1996) found that function words and content words show the same lexical decision times when they are embedded in sentences that are read for meaning. Some support for a faster access time for function words is presented in Neville et al.'s (1992) ERP finding of a distinctive negative component at 280 msec for function words and at 350 msec for content words. The result was not found to be due to word frequency (although there were range restrictions) or word length; however, repetition within the paradigm, and word predictability, were not explored. (The original object of the study was to examine variations in the congruity of the last word with respect to sentence meaningfulness.) In order to examine lexical access of words read for meaning while controlling word characteristics, we presented sentences from Neville et al. (1992) and Schmauder (1996) one word at a time (500 msec duration, 1200 msec SOA) to subjects who read them aloud for sentence meaning. We then scored the reading times for each word, not including the first and the last word of each sentence or the few words where subjects' articulation did not distinguish adjacent words (Segalowitz & Lane 2000).

We obtained similar results whether we analyzed frequency (high, medium, low) by word type in a standard ANOVA procedure (see Table 1), or whether we treated words as cases in a regression by standardizing each subject's reading times (RTs) and averaging across subjects: We found unique variance contributions to RT from length (shorter words were faster), $t = 3.1$, $p < .005$; frequency (higher frequency words were faster), $t = 3.1$, $p < .005$; and word type (function words were faster by 23 ms), $t = 2.3$, $p < .025$, in addition to the common variance. This shows for the first time that functions words are indeed accessed faster than are content words in meaningful contexts independent of these other characteristics. As expected, function words are of higher word frequency and shorter length on average. They also repeated more of-

Table 1. *Average reading times (msec) for function words and content words at different levels of word frequency. Word frequency criteria are indicated as occurrences out of a million (Kucera & Francis 1967).*

	SuperHigh >10000	High 883–10000	Medium 125–877	Low 0–122
Content Words	—	454	470	488
Function Words	457	453	455	506
Average reading times for first occurrences only				
Content Words	—	454	473	492
Function Words	446	448	460	506

ten within the 188 sentences, but when we partialled out word repetition as well as word length, we obtained the same results.

In addition, we found the Word Type *x* Frequency interaction to be significant ($F(1, 1526) = 21.2, p < .0001$), indicating that the frequency effect (high frequency words being accessed more quickly than low frequency words) is different across word types. Since many words (especially function words) are repeated, we also examined only the first presentations and obtained the same results. However, as Gordon and Caramazza (1982) pointed out, the interaction is strongly related to the confound of word type with frequency, for although content words show a near-linear frequency effect as expected, function words show an increase in RT only at the lowest frequencies. Therefore, depending on the frequency range of function words used, the disparity between word types in frequency effect can be manipulated – any range with a bottom frequency cutoff up to 310/million produced a significant word type by frequency interaction.

Ours are the first fully supportive behavioral data we know of for the privileged access and this was found using meaningful sentence contexts. This is consistent with Pulvermüller's (1999) data which suggest a more concise storage pattern for function words, and Neville et al.'s (1992) finding of an earlier ERP component for function words. However, meaningful sentence contexts confound many factors. Function words may show a special link with the left anterior region because they are accessed more automatically on account of higher perpetual fluency, more experience reading them (higher word frequency in the language), shorter length, or even greater repetition within the study. We found that reading time differences between word types were not related to word length or repetition. But this link could also be to the result of greater predictability of the common function of words. To test this last possibility we gathered from a new set of subjects Cloze judgments (ability to predict the word from the sentence context leading up to it) of one word from each sentence. As expected, high-frequency function words were clearly more predictable (see Table 2), as were function words, which were concentrated in the highest frequency range.

Table 2. *Predictabililty values (percentage of subjects correctly guessing the stimulus word from its preceding sentence) for function and content words at different levels of word frequency gathered in Cloze procedure. Frequency values are out of 1,000,000 printed words (Kucera & Francis 1967).*

	SuperHigh >10000	High 883–10000	Medium 125–877	Low 0–122
Content Words	—	17.3	15.5	10.1
Function Words	45.6	25.1	9.3	2.0

Of particular interest is the finding that predictability (Cloze values) and word frequency (log frequency occurrence in the language) each account for the significant variance in reading times ($p < .0001$), and after this variance is removed, neither word type nor the frequency by word type interaction is significant. In other words, from our data we would conclude that the difference in reading times between word classes is due to factors relating to perceptual fluency. By extension, the electrocortical effects of word class are a reflection of these characteristics, especially in natural contexts of reading sentences for meaning. However, in the real world, this is where words occur, and function words indeed occur with greater frequency and are more predictable. Thus, the brain mechanisms responding to lexical access in meaningful contexts should differentiate function and content words because processing them lexically involves different levels of perceptual fluency.

ACKNOWLEDGMENT

This work was partly supported by a grant from NSERC to the first author.

Authors' Response

Determinants of ignition times: Topographies of cell assemblies and the activation delays they imply

Friedemann Pulvermüller[a] and Bettina Mohr[b]

[a]*MRC Cognition and Brain Sciences Unit, Cambridge CB2 2EF, United Kingdom;* [b]*APU, School of Applied Sciences, Department of Psychology, Cambridge CB1 1PT, United Kingdom.*

friedemann.pulvermuller@mrc-cbu.cam.ac.uk b.mohr@apu.ac.uk
http://www.mrc-cbu.cam.ac.uk/Common/People/people-pages/Friedemann.Pulvermuller.html
http://www.apu.ac.uk/appsci/psychol/staff/bmohr.htm

Abstract: The cell assembly model of language posits that words are laid down in the cortex by discrete sets of neurons distributed over specific parts of the brain. The strong internal links of these "word webs" may not only bind articulatory and acoustic knowledge of a lexical item, they may also link word and meaning; for example, by connecting neuron populations related to word forms to those of actions and perceptions to which the words refer. Therefore, the cortical activation elicited by words should reflect aspects of word meaning, a postulate that has received strong support from recent work using neurophysiological and metabolic imaging. **Segalowitz & Lane** make the point that this neurobiological model can also be used to predict reaction times in behavioral experiments, using the behavioral distinction between content and function words as an example. We acclaim their view, but warn that response times might be related to different mechanisms at the neuronal level, including the cortical distribution and internal connectivity of cell assemblies along with their mutual connections in the grammatical (syntactic and semantic) network.

R1. Cell assemblies with distinct topographies binding words and their meaning

Laws governing neuronal function, such as the correlation learning principle, and the knowledge about cortical connectivity can be used to predict cortical circuits involved in language processing (Pulvermüller 1999; 2002). This approach is explanatory because it deduces the where and when of cortical processing from biological principles. It predicts that acoustic word form knowledge and articulatory word form knowledge are bound together by

distributed cortical systems spread out over the perisylvian language cortex and strongly lateralized to the left language-dominant hemisphere. In contrast, referential meaning, the dynamic links of word forms to actions and perceptions of objects in the world, should materialize as cortico-cortical networks binding neuron populations in the left-lateralized perisylvian language system and in the even more widespread areas involved in acting and perceiving objects. One aspect of word meaning, reference to objects and actions, would therefore be mapped onto the cortical distribution of word-related cell assemblies distributed over both hemispheres. Words that are not related to objects or actions (most typical examples are the grammatical function words and regular inflectional affixes) would have discrete word webs spread out over the perisylvian areas and strongly lateralized to the left. Among the referring expressions, lexical items that refer to objects and actions should be mapped onto neural systems extending into sensory (e.g., visual) and motor cortical fields, respectively. The large semantic word categories, such as animal versus tool words or object versus action words, would therefore have their equivalent in the different cortical distributions of the cell assemblies involved.

This view explains neuropsychological findings about category-specific semantic networks (Humphreys & Forde 2001; Shallice 1988; Warrington & Shallice 1984) along with imaging results showing topographically specific processes for semantic word categories in the intact human brain (Chao et al. 1999; Oliveri et al. 2004; Pulvermüller et al. 1996). Taking this approach further, quite fine-grained category distinctions are possible – for example, between action words referring to different body parts (Fig. R1). Because body part representations are organized topographically in motor and premotor cortex, the networks linking words to actions would reflect this somatotopy, so that the meaning of action words could actually be read from the activation of the motor strip (Hauk et al. 2004; Pulvermüller et al. 2001; Shtyrov et al. 2004). Clearly, if a word refers to a leg action (e.g., "walk"), the network connecting word form knowledge (laid down in perisylvian areas) to the leg motor program (in dorsal motor and premotor cortex) might be more widespread than in the case of a mouth- and face-related word (e.g., "talk"). Therefore, everything else being equal, the activation time of the former might be slower than that of the latter (Pulvermüller et al. 2001).

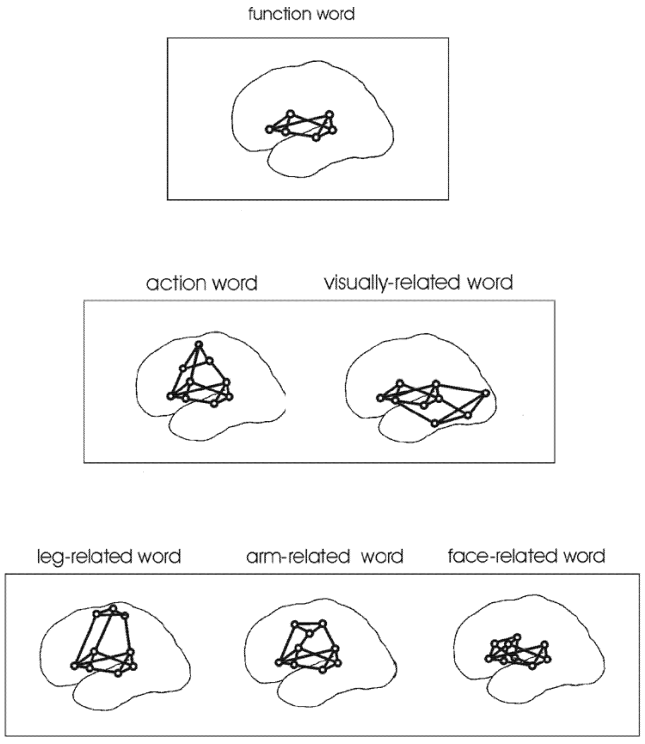

Figure R1. Left-hemispheric parts of cell assemblies that may underlie the processing of nonreferential morphemes, including function words and inflectional affixes (top), words referring to actions and visually perceivable objects (middle), and action words referring to leg, arm, and face actions (bottom).

Function words elicit left-lateralized focal activation consistent with the rapid ignition of a cell assembly spread out over perisylvian cortex and strongly lateralized to the left (Neville et al. 1992; Pulvermüller et al. 1995). Recent work on inflectional affixes, which, from a linguistic viewpoint, are very similar to function words, has even revealed the precise spatio-temporal structure of this perisylvian activation: Superior temporal areas become active slightly (22 msec) before activity spreads to the inferior frontal areas, thereby indicating that the activation of perisylvian networks sparked by grammatical elements follows a specific time course (Pulvermüller et al. 2003). In contrast, content words elicit additional activation outside left-perisylvian areas (e.g., Pulvermüller et al. 2004). Interestingly, activation spreading occurs 100–200 msec after the lexical element can be uniquely identified and it is present in passive tasks where subjects were asked to ignore spoken language stimuli and focus their attention elsewhere. The degree to which these processes are independent of attention suggests that the underlying mechanisms are neuronal circuits that become active automatically. When a stimulus matches the response characteristics of sufficiently many neurons of its neuronal representation, the assembly ignites instantaneously, provided that there are no equally strongly stimulated competing lexical networks (cf. Marslen-Wilson 1990).

R2. Ignition times may depend on different properties of lexical networks

Segalowitz & Lane's new findings about cortical processing time differences between content and function words (cf. Segalowitz & Lane 2000) are of great relevance for the way we think about the cortical mechanisms realizing the lexicon. They found privileged access to function words in a task where subjects had to read sentences aloud and make a semantic judgement later. They attribute the faster reading of function words compared with content words to greater perceptual fluency of the function words due to stimulus familiarity and the probability with which words are expected in a given sentence context (Cloze probability). Segalowitz & Lane are right in pointing out that the cell assembly model can provide a putative account for the differences they found. They emphasize the cortical distribution of word webs – that is, widespread distribution of cell assemblies of content words versus narrow localization of the left-hemispheric networks for function words. It seems plausible that small focal networks take less time to become active than do large widely spread-out networks.

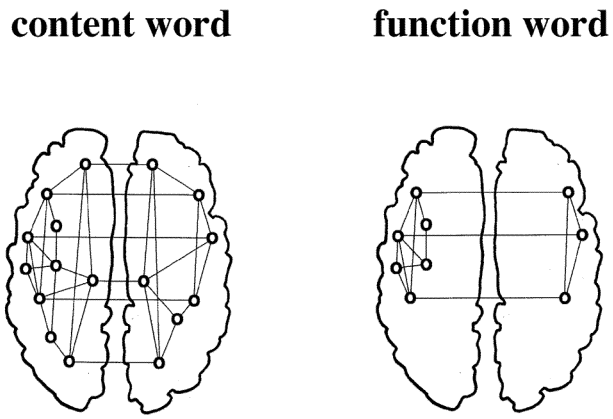

Figure R2. Differential laterality of cell assemblies spread out over both hemispheres. The networks for function words may be strongly lateralized to the language-dominant hemisphere, whereas those of content words referring to actions or objects may be lateralized to a lesser degree.

However, network size in the dominant hemisphere is but one factor determining the speed with which a distributed neuronal network ignites. Another factor may be the degree of cortical laterality of the networks (Fig. R2). As mentioned, the focal networks for function words may be strongly lateralized, whereas the widely distributed ones characterizing content words may be more equally balanced over the hemispheres (Mohr et al. 1994). For visual word recognition, this means that the information through the left visual field to the right hemisphere (Monaghan et al. 2004) is less effective in the processing of function words than it is for content word processing, a difference which should work to the advantage of content words. Whereas the cell assembly model allows for unambiguous predictions on the involvement of cortical areas and, of course, the hemispheres processing words and sentences, its implications for a general reaction time difference between the major word classes seems debatable.

Segalowitz & Lane's new findings suggest a possible fruitful target of future neurocomputational studies: Could it be that the differential floor effects of the word classes – the fact that function word response times asymptote already at lower frequencies (<310/million) than those of content words – be related to assembly size? Clarifying this issue would require simulation studies focusing on the relationship between assembly size, internal connectivity, and ignition times.

R3. Ignition times in the syntactic and semantic network

In the cell assembly framework, as in any cognitive model focusing on the issue, there are further obvious differences between function and content words. The correlation learning principle implies more strongly connected neuron sets for high frequency words than for low frequency words. Activity spreads rapidly in an assembly with strong internal connections, but spreads more slowly in a loosely linked neuron population because of the longer temporal summation times involved. Therefore, assemblies representing high frequency words may generally ignite faster than those representing rare words, a difference advantageous to highly frequent function words. In the cell assembly model, context influences ignition time in two principal ways: through cell assemblies overlapping with each other (e.g., if two words with similar meaning share neurons in their semantic network parts), and through links between word-related networks, that is, through neuronal sets specialized in syntactic processing. If, as Segalowitz and Lane (2000) showed, the Cloze probability – the likelihood with which subjects correctly guess the next lexical item when given sentence fragments – is greater for function words than for content words, this may be the result of the joint effect of priming through semantic overlap between cell assemblies and syntactic binding networks connecting sets of word webs (Pulvermüller 2003). Evidently, the speed with which a word-related cell assembly becomes active depends on its internal connections and the degree to which the network is active already before a stimulus word occurs. Internal connectivity and preactivation through priming would determine the speed with which a word can be recognized or read – what **Segalowitz & Lane** define as *perceptual fluency*.

In sum, global reaction time differences between content and function words, whether obtained in or out of context, seem to be difficult to interpret and are most likely related to multiple psychobiological mechanisms. A strong neurobiological model of language spelling out language processes at the psychological level and connecting them to neuron circuits allows for tentative explanations of general response time data, is more specific for neuropsychological studies (e.g., visual half-field research), and makes new strong predictions on the brain areas and neurophysiological dynamics of the networks involved.

ACKNOWLEDGMENTS

This work was supported by the Medical Research Council (UK) and by the European Community under the Information Society Technologies Programme (IST-2001-35282).

References

[The letter "r" before author's initials stands for CC Response article references]

Bradley, D. C., Garrett, M. F. & Zurif, E. B. (1985) Syntactic deficits in Broca's aphasia. In: *Biological studies of mental processes*, ed. D. Caplan, pp. 269–86. MIT Press. [SJS]

Chao, L. L., Haxby, J. V. & Martin, A. (1999) Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience* 2(10):913–19. [rFP]

Gordon, B. & Carmazza, A. (1982) Lexical decision for open- and closed-class words: Failure to replicate differential frequency sensitivity. *Brain and Language* 15:143–60. [SJS]

Hauk, O., Johnsrude, I. & Pulvermüller, F. (2004) Somatotopic representation of action words in the motor and premotor cortex. *Neuron* 41:301–307. [rFP]

Humphreys, G. W. & Forde, E. M. (2001) Hierarchies, similarity, and interactivity in object recognition: "Category-specific" neuropsychological deficits. *Behavioral and Brain Sciences* 24(3):453–509. [rFP]

Kucera, H. & Francis, W. N. (1967) Computational analysis of present-day American English. Brown University Press. [SJS]

Marslen-Wilson, W. (1990) Activation, competition, and frequency in lexical access. In: *Cognitive models of speech processing: Psycholinguistic and computational perspectives*, ed. G. Altmann. MIT Press. [rFP]

Mohr, B., Pulvermüller, F. & Zaidel, E. (1994) Lexical decision after left, right and bilateral presentation of content words, function words and non-words: Evidence for interhemispheric interaction. *Neuropsychologia* 32:105–24. [rFP]

Monaghan, P., Shillcock, R. & McDonald, S. (2004) Hemispheric asymmetries in the split-fovea model of semantic processing. *Brain and Language* 88(3):339–54. [rFP]
Neville, H. J., Mills, D. L. & Lawson, D. S. (1992) Fractionating language: Different neural subsystems with different sensitive periods. *Cerebral Cortex* 2:244–58. [rFP, SJS]
Oliveri, M., Finocchiaro, C., Shapiro, K., Gangitano, M., Caramazza, A. & Pascual-Leone, A. (2004) All talk and no action: A transcranial magnetic stimulation study of motor cortex activation during action word production. *Journal of Cognitive Neuroscience* 16(3):374–81. [rFP]
Pulvermüller, F. (1999) Words in the brain's language. *Behavioral and Brain Sciences* 22(2): 253–336. [rFP, SJS]
(2002) A brain perspective on language mechanisms: From discrete neuronal ensembles to serial order. *Progress in Neurobiology* 67:85–111. [rFP]
(2003) *The neuroscience of language.* Cambridge University Press. [rFP]
Pulvermüller, F., Hummel, F. & Härle, M. (2001) Walking or talking?: Behavioral and neurophysiological correlates of action verb processing. *Brain and Language* 78(2):143–68. [rFP]
Pulvermüller, F., Lutzenberger, W. & Birbaumer, N. (1995) Electrocortical distinction of vocabulary types. *Electroencephalography and Clinical Neurophysiology* 94:357–70. [rFP, SJS]
Pulvermüller, F., Preissl, H., Lutzenberger, W. & Birbaumer, N. (1996) Brain rhythms of language: Nouns versus verbs. *European Journal of Neuroscience* 8:937–41. [rFP]
Pulvermüller, F., Shtyrov, Y. & Ilmoniemi, R. J. (2003) Spatio-temporal patterns of neural language processing: An MEG study using Minimum-Norm Current Estimates. *Neuroimage* 20:1020–25. [rFP]
Pulvermüller, F., Shtyrov, Y., Kujala, T. & Naatanen, R. (2004) Word-specific cortical activity as revealed by the mismatch negativity. *Psychophysiology* 41(1):106–12. [rFP]
Schmauder, A. R. (1996) Ability to stand alone and processing of open-class and closed-class words: Isolation versus context. *Journal of Psycholinguistic Research* 25:443–81. [SJS]
Segalowitz, S. J. & Lane, K. C. (2000) Lexical access of function versus content words. *Brain and Language* 75(3):376–89. [rFP, SJS]
Shallice, T. (1988) *From neuropsychology to mental structure.* Cambridge University Press. [rFP]
Shtyrov, Y., Hauk, O. & Pulvermüller, F. (2004) Distributed neuronal networks for encoding category-specific semantic information: The mismatch negativity to action words. *European Journal of Neuroscience* 19(4):1083–92. [rFP]
Taft, M. (1990) Lexical processing of functionally constrained words. *Journal of Memory and Language* 29:245–57. [SJS]
Warrington, E. K. & Shallice, T. (1984) Category specific semantic impairments. *Brain* 107:829–54. [rFP]
Whitaker, H. A. (1983) Towards a brain model of automization: A short essay. In: *Memory and control of action,* ed. R. A. Magill, pp. 199–214. North-Holland. [SJS]

Commentary on **Ian Gold & Daniel Stoljar (1999). A neuron doctrine in the philosophy of neuroscience. BBS 22(5):809–869.**

Abstract of the original article: Many neuroscientists and philosophers endorse a view about the explanatory reach of neuroscience (which we will call the *neuron doctrine*) to the effect that the framework for understanding the mind will be developed by neuroscience; or, as we will put it, that a successful theory of the mind will be solely neuroscientific. It is a consequence of this view that the sciences of the mind that cannot be expressed by means of neuroscientific concepts alone count as indirect sciences that will be discarded as neuroscience matures. This consequence is what makes the doctrine substantive, indeed, radical. We ask, first, what the neuron doctrine means and, second, whether it is true. In answer to the first question, we distinguish two versions of the doctrine. One version, the *trivial* neuron doctrine, turns out to be uncontroversial but unsubstantive because it fails to have the consequence that the nonneuroscientific sciences of the mind will eventually be discarded. A second version, the *radical* neuron doctrine, *does* have this consequence, but, unlike the first doctrine, is highly controversial. We argue that the neuron doctrine appears to be both substantive and uncontroversial only as a result of a conflation of these two versions. We then consider whether the radical doctrine is true. We present and evaluate three arguments for it, based either on general scientific and philosophical considerations or on the details of neuroscience itself, arguing that all three fail. We conclude that the evidence fails to support the radical neuron doctrine.

Could the neural ABC explain the mind?

Maurice K. D. Schouten[a] and Huib Looren de Jong[b]
[a]*Faculty of Philosophy, Tilburg University, 5000 LE Tilburg, The Netherlands;* [b]*Department of Psychology, Vrije Universiteit, 1081 BT Amsterdam, The Netherlands.* **m.k.d.schouten@uvt.nl** **h.looren.de.jong@psy.vu.nl**

Abstract: Gold & Stoljar are right in rejecting the radical neuron doctrine, but we argue that their distinction between determination and explanation is not principled enough to support their conclusion. We claim that the notions of multiple supervenience and screening-off offer a more precise construal of the dissociation between explanation and determination that lies at the heart of the antireductionist position.

Neuron doctrine: Trivial and radical. Gold & Stoljar (1999; henceforth G&S) distinguish two varieties of neuron doctrine. One option they consider is the radical neuron doctrine (RND), a view in which explanation and determination (wrongly, in their opinion) hang together: neural properties determine and ipso facto explain psychological properties. The alternative is the trivial neuron doctrine (TND), which states that although mental properties are *determined* by neurobiological properties, they are not necessarily also *explained* by them.

We agree that the difference between the trivial and radical neuron doctrines lies in separating determination and explanation. However, we also believe that G&S's distinction between explanation and determination is not principled enough to support the TND-RND distinction. G&S analyse Kandel's experiments as support for their antireductionist position. However, Kandel is a self-confessed reductionist, and is interpreted in this light by philosophers like Schaffner (1993) and Bickle (1998). Kandel and colleagues claim that the neural plasticity paradigm offers "surprising reductionist possibilities" (Kandel et al. 1995, p. 389). They employ the metaphor of a molecular or neural alphabet (see also Hawkins & Kandel 1984): What is necessary and sufficient to understanding learning and memory in all of their varieties is a full specification of the letters of this alphabet and of the possible combinations in which these letters can be strung together. Contrary to what G&S claim, Kandel's view implies that the molecular and neural letters in the end not only determine, but also explain and replace the psychological phenomena of learning and memory. What is suggested here is reductionism of a very special sort, namely, *combinatorial reductionism.* Thus, without a principled distinction between determination and explanation, the TND may well collapse into the RND. Two arguments are discussed below to suggest that determination and explanation can still be kept apart.

Multiple supervenience. There is a well-known argument that higher-level sciences, such as biology and psychology, deal in functional explanations. The notion of function allows generalisations, and even laws, abstracting over lower-level mechanisms that would be wildly heterogeneous from a physical perspective. The

function of spatial navigation, for example, offers a grouping that may be implemented in many different substrates. It may however still be claimed that a function in a particular organism is *determined* by a specific substrate and that the substrate also *explains* the function, so that the gates to reductionism (and the RND) are still wide open (Kim 1998). We suggest the argument from multiple supervenience (MS) to block the interference from determinism to reductionism (see also, Schouten & Looren de Jong 1999). The psychoneural supervenience thesis states that psychological properties are *determined* by neural properties. However, a single neural property determines not just a single higher-level (dispositional) property, but a multitude of them. To pick out from this set of supervenient properties the one particular property that is explanatorily salient (often a functional property), requires information that goes beyond what is present in a full specification of the microlevel details. What is explanatorily interesting about the neural "letters" that compose the "words" of learning and memory is something that cannot be read off from the neural alphabet alone. In order to make this selection of causally relevant properties from the supervenience base, a higher-level perspective is needed. MS thus honors determinism of psychological properties and yet it grants an irreducible role to higher-level, often functional explanation.

It may be objected, however, that functional ascriptions will turn out to be nothing more than heuristics. That is, when all the data are in and all the mechanisms are known, they drop out of the scientific picture as mere convenient fictions. When neuroscience would finish its job, all the explanatory weight shifts to the letters of the neural or molecular ABC. This suggests that the proffered functional higher-level explanation refers to something that may be causally irrelevant; hence, these higher-level explanations may not count as bona fide explanations. So, a further argument is required in order to uphold a more objective and qualitative distinction between level of explanations.

Screening off. We believe such an additional argument, required to establish the objective, and not merely heuristic, relevance of higher-level explanations, can be found in the application of the so-called screening-off (S-O) rule. Stated in the language of conditional probabilities, the formula for one cause M screening off another cause P from its outcome R is

$$M \text{ screens off } P \text{ from } R \text{ iff}$$
$$\Pr(R \mid P \,\&\, M) = \Pr(R \mid M) \neq \Pr(R \mid P).$$

To put it colloquially, a property M screens off a property P if adding P does not improve the prediction or explanation of outcome R, whereas M does improve the prediction or explanation of R. Brandon (1990) employed this screening-off relation to substantiate the claim that phenotypes (M) and not the genes or the genotypes behind the phenotypes (P) are relevant in explaining reproductive success (R). Because the genotype is asymmetrically dependent on the phenotype with respect to natural selection (for its effect on R, P depends upon M being present or absent, whereas M contributes to R irrespective of P's being present or absent), it is the phenotype that offers the best causal explanation of reproductive success (Brandon 1990, pp. 83–85). The phenotypic level has a causal efficacy and explanatory legitimacy of its own, even if the phenotype is determined by the genotype (among other things). Identifying phenotypic traits is not a merely heuristic, free-for-all, essentially void kind of explanation, but rather, it taps real causal factors in an organism's chances of survival. In the same way, behavior may be explained in terms of information processing, screening off the underlying neural processes that determine it. Although cognition is *dependent* on neurons, it cannot be exhaustively *explained* in neuronal terms.

Screening-off thus serves as a criterion to distinguish causally relevant levels of explanation from causally irrelevant ones; it affords a criterion for "picking levels" (McClamrock 1995). While admitting that cognition is determined by neurons (and ultimately by atoms, molecules, bosons, superstrings, and what have you), we can maintain that higher functional levels genuinely and irreducibly explain behavior that is determined by lower neural levels. G&S are right in rejecting RND, but we submit that the notions of multiple supervenience and screening-off offer a more precise construal of the dissociation between explanation and determination that lies at the heart of the antireductionist position.

Editor's Note: There is no Author's Response for this commentary.

References

Bickle, J. (1998) *Psychoneural reduction: The new wave.* MIT Press. [MKDS]

Brandon, R. N. (1990) *Adaptation and environment.* Princeton University Press. [MKDS]

Hawkins, R. D. & Kandel, E. R. (1984) Is there a cell-biological alphabet for simple forms of learning? *Psychological Review* 91:375–91. [MKDS]

Kandel, E. R., Schwartz, J. H. & Jessell, T. M. (1995) *Essentials of neural science and behavior.* Appleton & Lange. [MKDS]

Kim, J. (1998) *Mind in a physical world: An essay on the mind-body problem and mental causation.* MIT Press. [MKDS]

McClamrock, R. (1995) *Existential cognition: Computational minds in the world.* The University of Chicago Press. [MKDS]

Schaffner, K. F. (1993) *Discovery and explanation in biology and medicine.* The University of Chicago Press. [MKDS]

Schouten, M. K. D. & Looren de Jong, H. (1999) Reduction, elimination, and levels: The case of the LTP-learning link. *Philosophical Psychology* 12:237–62. [MKDS]