Meaning and the extended mind
Jan Koster

1. Introduction

Cognitive science and the philosophy of mind are suffering from what Merlin Donald (2000) called "the myth of the isolated mind." This myth is characterized by some form of mind/brain identity. In theoretical linguistics, for instance, it is often assumed that there is a core domain referred to as "the faculty of language in the narrow sense" (FLN) that can be seen as a strictly individual property of humans (Hauser *et al*. 2002) . Its description is furthermore seen as a characterization of the brain at a certain level of abstraction that, ultimately, must be explained by biology ("biolinguistics" in the sense of Lenneberg 1967 and Jenkins 2000). Although advocates of mind/brain identity and biolinguistics are often not committed to the tenets of sociobiology (and its successor "evolutionary psychology"), they share the reductionist stance of those disciplines to various degrees.[1] This reductionism is crucially dependent on the flawed conception of human nature embodied by mind/brain identity and the myth of the isolated mind.

Mind/brain identity is not a good idea for a number of reasons. With respect to mind, the brain is something both too broad and too narrow. It is too broad because the brain serves a whole variety of functions not related to the cognition of content and the mental at all. Examples are automated functions like the regulation of blood pressure and the maintenance of equilibrium. In this respect, nobody takes mind/brain identity literally. More interesting from the perspective under discussion, the brain is also way too narrow as a physical description of mind. Limiting the mental to what is represented in individual brains is a denial of one of the most fundamental property of humans compared to other organisms, namely that we live in symbiosis with a brain-external and shared symbolic culture. Missing the fact that humans are symbionts in this sense is like talking about fish in denial of the fact that they live in water.

Our shared symbolic culture permeates the brain in ways that make it futile to maintain that the mind is found in the skull somehow. This can be seen as follows. In general, intelligent systems, both in biology and technology, consist minimally of processing mechanisms (like the CPU in your computer) and data structures. Looking for minimal material correlates of the mind, I assume consensus about the fact that both processing and data structures are necessary conditions for the mental to take off. As for humans, most processing is done in the brain (and the central nervous system). For relevant data structures, however, it is pointless to make a principled distinction between "inside the brain" and "outside the brain." Next to processing units and working memory, substantial parts of our brain store data that could just as well be stored outside of the brain.

The boundaries between "inside" and "outside" are fluid and ever changing. When, as a non-native speaker of English, I read an English text, I often rely on my "onboard" knowledge of

[1] See Wilson (1975), Pinker (2002) and a discussion of Chomsky's position in Segerstråle (2000, 203-206). In general, Chomsky is more skeptical about the possibility of reduction as compared to, say, occasional unification. I am not saying that biolinguistics is a reductionist enterprise in the full sense because it has low expectations of unification at best. It is a semi- or quasi-reductionist way of talking about socio-cultural phenomena. Trivially, our biological nature, with its innate brain structures, is a prerequisite for all our culture.

the word meanings of English, but it also happens that I have to look into a dictionary. Whether the relevant information comes from my long-term memory (inside the brain) or from a dictionary (outside the brain) is entirely arbitrary. Moreover, whatever information for word meaning is stored in my brain, it is external to the brains of other individuals. If somebody asks me the meaning of an English word, I can provide information from my brain-internal memory, which functions for the other person in ways very similar to a dictionary.

As soon as we reflect on the obvious fact that the brain is diverse, involving both processing and data structures, we see the issue of mind/brain identity in a new light. Even the strongest believer in mind/brain identity will admit that processing is often individual in a way that data structures are not. Data structures are shareable and supra-individual, whether they are inside or outside the brain. By "shareable" I mean shareable-in-principle. Somebody can refuse to share his or her "onboard" information with you, but in the same vein, somebody can deny you access to his private library. Apart from the irrelevant modes and conditions of access, there is no significant difference between data structures that feed the mental from inside or from outside the brain. With our brains, we are each other's libraries and archives, so to speak, which makes it possible to have a common culture, even in pre-literate societies. It would be strange to say that the Homeric epics were individual psychology when they only existed in oral tradition and that they became public culture as soon as they were written up.

Even these simple considerations show, in my opinion, that mind/brain identity is untenable. Brains are individual properties of humans in a way that minds are not. It is deeply ingrained in everyday usage to talk about "my mind" next to "my brain," but this obscures the fact that we are mental symbionts. "My mind" is a combination of individual processing and supra-individual data structures. Strip away the supra-individual part (distributed over brain-internal memory and brain-external media) and what is left can no longer be properly characterized as "mind." The processing capacities of our brain are not the mind, but only necessary conditions of mind, just as the supra-individual parts are.

Giving up mind/brain identity and the myth of the isolated mind has profound consequences for cognitive science and the philosophy of mind. This becomes particularly clear when we realize what the function is of the shared, supra-individual part of the mind. In part, at least, the supra-individual part is the record of our symbolic culture and our common memory about successful form-function assignments. There is a straightforward relation between form-function connections and the need for memories. The crucial point is that form-function relations are non-deterministic and therefore unpredictable. Function is never intrinsic to form but always extrinsic to it, i.e., a matter of context. The relation between forms and their functions is not one-one, but many-many. What works has to be found out in practice and, if useful, to be preserved with the help of memories.

Physical structures can acquire a function during biological evolution thanks to processes described as "tinkering" by Jacob (1975). The very notion of "tinkering" is an implicit denial of the intrinsic functionality of the structures being used. Structures can get better adapted to their functions by natural selection. However, no matter how much adaptation, the relation between form and function remains extrinsic. In other words, what is functional in one context is not functional in another context. An expert can tell from an animal's teeth what it eats, which creates a "panglossian" illusion of form-function transparency.[2] However, without knowledge of the relevant nutritional contexts, the functional shape of teeth remains

[2] For the term "panglossian" see Gould and Lewontin (1979).

completely opaque. The function cannot be "read" from the form. Moreover, we know from real life situations that teeth can be used for more functions than chewing food. Another fact highlighting the non-intrinsic nature of the form-function relation is what Gould and Vrba (1979) called "exaptation," in the 19th century known as *Functionswechsel*: a structure adapted to one function can be used in another environmental context for (and further adapted to) another function.[3] A standard example is wings for flying, which originally evolved as flaps for thermo-regulation.

Altogether, an organism can be seen as a huge collection of functional forms. Since forms have no intrinsic function, a memory is needed to preserve successful connections. In the organic world, most (but not all) memory functions involve DNA (see Jablonka and Lamb (2005) for the various kinds of memory in the living world).

In the case of humans, we have a fundamentally different situation in that we can assign functions to form at will. John Searle (1995) called this "agentive function assignment." An example is a stone used as a paperweight. This kind of assignment is apt to illustrate the many-many character of the form-function relation: we can use stones for many purposes, as much as we can use many different objects as paperweights. In general, assigning functions to physical structures, in daily life mundanely called "application," consists of the invention of new functional contexts for physical structures. This is even true for designed structures, like the parts of the evolutionary psychologist's Swiss army knife. None of the parts of such a knife is functionally transparent without extrinsic contexts. Thus, a knife's cork screw is only functional as such in a world in which cork is used for bottles, by no means a necessity. In a world without cork and bottles, the function of the cork screw would be opaque. Moreover, we could invent numerous other contexts, thereby assigning new functions to the original cork screw. Which functional context is the best cannot be said in advance of actual invention and use. The point, once more, is that there is just no intrinsic relation between form and function, no matter the amount of adaptation or design. This is the Achilles heel of all biological, {semi-) reductionist theories of cultural phemomena, including sociobiology, evolutionary psychology and biolinguistics.

One major function of our shared cultural memory is to make up for the accidental and non-deterministic nature of the form-function relationship: our form of life takes great advantage of remembering which combinations are useful. Apes and other animals also show some capacity for agentive function assignment, as in the use of stones for the cracking of nuts, but the cultural memory to preserve this habit seems to be limited to imitative behavior (Jablonka and Lamb, 2005). Only humans have the capacity to keep a symbolic record, which itself is based on the multiple application of information stored in our brains and elsewhere.

The significance of this for the philosophy of mind is that mental content cannot be seen in isolation of our shared cultural record. As physical structures *never* have an intrinsic function, it is meaningless to say that some part of our brain has an intrinsic cultural function. As was mentioned above, parts of our brain work like an organ, for instance those parts that regulate blood pressure. The form-function connection in that case automatically arises during ontogeny, as part of the execution of the program derived from our DNA. The capacity for computation with recursion might, at least partially, originate in a similar way. However, none of that can be called language or even a precursor of language. It can only become part of language thanks to a purely human invention, namely the invention of words. The invention

[3] For the term *Functionswechsel* see Russell 1982 [1916], 306-307.

of words is a form of agentive function assignment, which gives a function to otherwise functionless biological structures. Language, in other words, only exists thanks to function assignment (to biological structures) from outside, in this case coming from our shared cultural memory. The application casually referred to as the "the language faculty" sharply differs in this respect from organs like the heart or the liver. Talk about a "language organ" is metaphorical at best.

This makes language, even in the narrowest sense, a form of technology or culture rather than a form of biology, no matter to what degree it makes use of innate biological structures. At a strictly biological level, i.e., in abstraction from our man-made cultural record, nothing in the brain can be called "linguistic" or being predestined for language. It is an error, therefore, to attribute "universal grammar" as an initial state of language to the brain of a pre-verbal child. Something deserving the name language or grammar only emerges in the course of an acculturation process, in which the child symbiotically gets attuned to a community's shared cultural memory. It is this shared memory that stores the words necessary for language.

Note that neither innateness nor mental plasticity is at issue here. Even if language uses completely innate structural facilities with zero plasticity, these structures only become linguistically functional by acculturation, i.e., by their function assignment in relation to the cultural memory external to those structures. This can be further illustrated with a non-linguistic example. If somebody plays the trumpet, he crucially uses his lungs, just as when somebody uses his lungs for respiration. The former kind of functionality only exists thanks to a man-made cultural context, while the latter function automatically emerges during human ontogeny. Trumpet playing is usually seen as part of our culture, while respiration is seen as part of mammalian biology. This is so in spite of the fact that both functions use the same innate structures. It would be fairly absurd to see an anatomical description of the lungs as part of biomusicology and an account of "trumpet playing in the narrow sense." Similarly, it is questionable to talk about "biolinguistics" (and whatever other innate capacities used in language) as "the faculty of language in the narrow sense."[4]

2. The construction of meaning

Very similar things can be said about meaning (Koster 2008). For all kinds of well-known reasons, meaning is a more obscure phenomenon than grammar, but some contours have become clear over the years. In fact, it is relatively easy to show that meaning is at least partially constituted by supra-individual information. I will limit myself here to content words, like nouns, but similar observations can be made about other types of words.

The basic error of the tradition is the belief that meanings are properties of words, in Fregean terms described with the further distinction between *Sinn* (sense) and *Bedeutung* (reference). According to the most naïve theory of this kind (sometimes dubbed "the nomenclature theory"), words are arbitrary labels for fixed concepts. Thus, it is assumed sometimes, that both English *tree* and French *arbre* refer to the same concept, often written with capitals (TREE) to indicate its metalinguistic status. A major issue in the tradition was the question

[4] See Koster (2008) for a more elaborate version of the arguments given here. Needless to say, the acquisition of language might be more biologically facilitated than trumpet playing as a matter of brain-culture co-evolution. No matter the degree of co-evolution, the fact remains that, unlike what we see with organs, the functionality of the biological structures used in language comes from invented words that are preserved in our cultural, non-individual memory.

whether such a concept (or "idea") must be interpreted ontologically (as in Plato's theory of ideas) or epistemologically (as in the traditions of European Rationalism and Empiricism).

This naïve theory of meaning was (often implicitly) challenged in contact with other cultures in the wake of the European expansion and particularly by German Romanticism (Herder, Von Humboldt), with its insistence on the cultural dependence of our conceptualizations of the world. This idea sometimes got a relativistic interpretation (as in the later work of Benjamin Lee Whorf), but it is also compatible with the Chomskyan idea that languages are differential selections from a universal toolbox.[5] The most influential elaboration of the socio-cultural thesis was the work of Ferdinand de Saussure (1916), who insisted that not only the sound forms of words are conventional but also the conceptual systems on which they are based. Thus, according to one of Saussure's examples, English has two words, *sheep* (animal) and *mutton* (meat of that animal), where French has only one word, *mouton*. Therefore, according to Saussure, *mutton* and *mouton* are not just names for things in the world, but terms with different values, derived from different systems of mutually contrasting signs. The vocabulary of a language is not a nomenclature (of things, concepts, or whatever) but, partially at least, a set of conventions that records how a culture divides up our conceptual space. Everybody who has ever translated texts from one language into another knows how profound this difference in conceptualization can be, the gap becoming wider with cultural distance.

These Saussurian considerations suffice to show that meaning cannot be reduced to individual psychology or biology. The various conceptualizations are culture-dependent and conventional. By definition, conventions are not properties of individuals but based on agreement within communities. As there is no meaning without convention, mental content cannot be seen as something strictly corresponding to some part of the individual brain. As is generally the case, mind differs from the brain in that it is based on structures that go beyond the individual. Among other things, meaning is not a property of individuals but of *individuals in crucial symbiosis with external information structures*. Meaning belongs to the culturally extended mind, not to the solipsistic mind that corresponds to the brain and nothing beyond the brain.

It should be said at the outset that this view does not commit me to the so-called meaning externalism attributed to Putnam (1975) or Burge (1979). Considerations of external reference do not play a direct role in the above, therefore a theory of meaning based on the extended mind can be seen as a form of extended internalism along the dimension internal/external in Putnam-style theories.

Although I believe that the Saussurian approach is a tremendous step ahead with respect to the nomenclature tradition and also that it suffices to refute brain-based semantic individualism, I think it is not going far enough. My basic objection is that, although Saussure makes concepts culture- and system-dependent, he leaves the notion concept-as-a-property-of-words itself intact. In my view, what is stored with respects to words in the brain is not properly described as "concepts" at all. As long as we talk about the physical structures of the brain, or about any physical structure whatsoever, there is really nothing that corresponds to concepts or meanings. In a nutshell, words can definitely be meaningful in contexts, *but only under an interpretation*. Interpretations are not properties of individual words or of any physical representation, but are acts *extrinsic* to physical structures. In that respect, "being meaningful"

[5] For an explicit reference to the toolbox idea, see Fitch *et al*. 2005).

is like "being functional" as discussed above. What interpretations contribute to word meaning must go way beyond whatever is stored with respect to individual words in the brain.

What this means in practice can best be illustrated with the key concept of polysemy, which applies even to the "flagships" of the nomenclature tradition, the words generally known as proper names:

(1) a. Schubert is difficult
b. Schubert will take 30 pages
c. Schubert is for sale
d. Schubert will be reburied next year
e. Schubert can be downloaded everywhere
f. Schubert will be burned on CD on request

In all these cases, the word *Schubert* seems to have a different meaning and a different intended reference (in the readings that first come to mind). In (1a), for instance, *Schubert* might refer to Schubert's music (or to his character) and in (1f) to some electronic representation of his music. This differential, context-based interpretation is essentially without limit and unpredictable. This is in a particularly clear way illustrated by (1f), which was not interpretable in the intended sense even 10 years ago, as the relevant technological and institutional context had not been invented yet. Our astonishing semantic flexibility of this kind, has been observed at least since Aristotle and fascinated authors as different as Paul (1882), Rechling (1935), Chomsky (2000), Moravcsik (1990) and Pustejovsky (1993).

Polysemy makes it illusory that meaning can be seen as some fixed property of words. Interpretations are truly creative applications of what is associated with words in ever new contexts. Meanings, in other words, are not fixed properties of words, but created in language use. This is also in line with some other classical attempt to deconstruct the nomenclature paradigm, the work of the later Wittgenstein (1953) (and the subsequent analytic philosophy influenced by him). However, even if we "don't look for the meaning but for the use," this use must be based on something. It's not just so that anything goes. One traditional puzzle of polysemy is what the various interpretations in (1) have in common. Some people will be inclined to say that the common factor is a meaning or a concept after all. But this will not do for several reasons.

First of all, attempts to characterize this common meaning usually fail, no matter whether it is given in the form of definitions, paraphrases, meaning postulates or features. One problem is that the knowledge relevant for the proper use of words differs enormously from person to person. Take the following example:

(2) The heavy water costs a lot of energy

The word *water* can refer here not necessarily to the substance "water" but to some container that is filled with this substance and that is such that it would be a burden to carry it. However, in a context involving theoretical physics, this sentence could be about the presence of the isotope deuterium in the substance. In science as well as in daily life, the interpretation of a word depends on the theories which are applied to it and these theories differ from person to person and from moment to moment. Verbal descriptions of word meaning are futile, because they give only a fraction of the vast information available for the interpretation of a word. It is not even clear if it makes sense to limit this information to what can be

characterized by words. Our familiarity with water, after all, also involves tactile and other sensory information.

More important in the current context is that it is pointless to construe whatever is used for the interpretation of words as a property of individuals. Putnam (1975) aptly described the community nature of word-related knowledge as "the linguistic division of labor." It is just a fact about our use of English words that "heavy" in relation to water can refer to the relative abundance of deuterium isotopes. This remains a fact of English, even if individual knowledge of physics would die out. In a world with no theoretical physics represented in individual brains anymore, the meaning of "heavy water" could probably be reconstructed by studying old books.

It is obvious, then, that information relevant for word interpretation is not an individual property but something distributed over a community and its external memory records. In each individual brain, only a fraction of the relevant information is stored. Successful communication depends on the degree to which individual storage overlaps among individuals.

There is a second, more fundamental reason to reject the idea that what the various uses of a word have in common is stored in the brain as "meanings" or "concepts." The only form of memory storage we know of is physical, i.e., in the form of coded information. For words, that would mean storage in the form of information coded in the neural circuitry of the brain and in the various external media we use, like books, computer memories, etc. What is characteristic of coded information is what it shares with all physical structures: it is not functional, let alone meaningful by itself. As discussed above, physical structures can only become functional thanks to a wider context in which functions are assigned, either non-agentively or agentively. If this is clear for anything, it is clear for coded information. The patterns burned on CDs or DVDs are neither audio nor video nor text without *external* interpretation. External interpretation can involve machine states only, as with the robots that manufacture cars (with intentionality ultimately coming from humans), but it can involve as much as our hardly understood forms of human understanding.

How does this apply to words like "Schubert" or "water?" What is stored with Schubert in individual brains and other media cannot be the meaning(s) of "Schubert" but only "dead" information coded in neural tissue (we assume) and in media like books and computer memories. It only becomes something deserving the name "meaning" if brought to life by interpretation.

The kind of interpretation that leads to word meaning involves many mysteries that I will not go into here, but it is sufficiently understood to reject the idea that meanings are properties of individual words. The relevant point is that the interpretation of coded information *always adds information*. This is even true for something relatively simple as music storage on vinyl records, tapes or CDs. What the media contain as information is hugely incomplete as a representation of the intended output. Even something as basic as the speed of the music is not stored on the media but dependent on the rotation speed of the interpreting players. It is similarly pointless and metaphorical at best to say that DNA represents organisms. Like all information carriers, DNA is only something in combination with context and external interpretation, beginning with the surrounding living cell and constantly modified in the course of organismic development. Not even the way proteins fold during development is coded in the DNA itself.

What is true for all coded information must also true for information storage associated with words: it only becomes "meaning" by external interpretation, in which information is added, perhaps massively so. External interpretation may involve unknown quantities of the rest of somebody's knowledge and also factors that nobody understands at all, like the emergence of "types" rather than "tokens" (a Platonic residue) and the kind of awareness that distinguishes human meanings from the information structures guiding the behavior of robots and zombies (a homuncular residue).[6]

If meanings cannot be properties of individual words, we may wonder why the idea of word meaning is so tenacious. It seems to me that we are fooled by introspection, so to speak. The main reason why we think we are aware of word meanings is what we seem to experience through introspection. But introspection can plausibly be seen as a form of "use," that is, as an act that crucially involves our faculties of interpretation. During introspection, we generate possible contexts that give an extrinsic interpretation to the information associated with the interpreted words. That information itself is not directly accessible.

All in all, I conclude that an analysis of the nature of word meaning refutes the myth of the isolated mind and mind/brain identity with respect to mental content. Meanings are constructed by the interpretation of physical structures, like the information associated with words. Information relevant for concept formation for words is distributed over parts of the individual brains of a community, external memory records, and the rich information-adding contributions of interpretive processes. Only the interpretive processes can, it seems, be construed as strictly individual contributions of brains. As for the data structures that are processed, no principled boundary exists between the individual and the non-individual. But nobody would call our interpretive faculties stripped from all data structures a "mind." In fact, in isolation, interpretive faculties are as meaningless as data structures. Interpretation and coded information are mutually dependent, relational concepts.

3. Fodor's critique of the extended mind thesis

The thesis I am defending here, has become known as the Extended Mind Thesis (EMT). It has received a recent boost by Clark and Chalmers (1998) and a book-length exposition by Andy Clark (2008), which was critically reviewed by Fodor (2009). In the next section, I will also briefly discuss the even more radical views of Noë (2009). I do not consider Fodor's objections against the EMT convincing and will come to his arguments in a minute.

I myself have been advocating some version of the EMT since the 1980s (see for instance Koster 1988 [1993], 1989, 1990).[7] The EMT, I believe, is implicit in the computational approach to language and mind as has been proposed by Noam Chomsky:

> We are like a Turing machine in the sense that although we have a finite control unit for a brain, nevertheless we can use indefinite amounts of memory that are given *externally* to perform more and more complicated computations. (Chomsky, Huybregts, and van Riemsdijk 1982, 14) [emphasis added –JK].

[6] For the Platonic residue in word meaning, see Koster (2005).
[7] See Malik (2000, 331) for another convincing discussion of the extended mind concept.

I combined this idea with aspects of Popper's world 3 concept, particularly as developed in opposition to Brouwer's solipsistic intuitionistic approach to mathematics (Popper 1972). I agreed with Popper's emphasis on the partial socio-cultural, supra-individual nature of knowledge but I strongly disagreed with his idea of the autonomy of world 3 and his further suggestion that his theory provided an alternative to classical metaphysics, either of the ontological, "Platonic" or of the epistemological, "Kantian" kind.

World 3, according to Popper, is the *autonomous* world of culture and includes libraries and other records of "knowledge in the objective sense." This does not make sense. As will be clear from the previous discussion, records with coded information (like the books of a library) contain nothing significant in abstraction from extrinsic interpretation, i.c., the interpreting subject. To say that libraries contain knowledge is only true in a metaphorical sense but actually only makes sense with implicit or explicit reference to an interpreting subject. The mental content of libraries is only derived mental content, for which the contribution of individuals is a necessary condition. As concluded above, the mental only exists as a property of the *combined* contributions of individual interpreters and data structures, the former being individual by definition, the latter being such that there is no sharp distinction between the individual and the non-individual. My own views are closer to those of Donald (1991), who worked out the interactive version of the EMT in some detail.

Furthermore, Popper's non-individual epistemology fails just as much as an alternative to classical epistemology as the intuitionism of Brouwer that he rejects. Unless one believes that anything goes, possible mental constructions, either socially or individually conceived, are heavily constrained by reality. Historically speaking, there are only a few flavors as to the nature of this ultimate reality: eternal being (ontology in the sense of Plato and Aristotle), epistemology (Kant) or non-eternal being created *ex nihilo* by God's will (medieval Nominalism and its Islamic antecedents). Neither Brouwer, nor Popper (1972) and Lakatos (1976) add anything new to this traditional repertoire as far as I can see, in spite of their anti-Platonic rhetoric. In fact, Popper's approach to metaphysics is, not unlike Logical Positivism, just another manifestation of the long standing idealistic trend in European philosophy to blur the distinction between our knowledge (epistemology) and the reality this knowledge is about (ontology).

Now the EMT is getting some new momentum, it is worthwhile to have a look at Fodor's recent objections (Fodor 2009). The core of Fodor's critique is about a concrete example given by Clark and Chambers (1998). This example discusses the case of Otto, an Alzheimer's patient, who relies on his notebook to remember things (for instance where to find the Museum of Modern Art in New York City). Another person, Inga, does not have a notebook but remembers from her own "onboard" memory that the museum is at 53rd Street. According to Clark and Chambers, the notebook is on a par with Inga's inner memory records and can be seen as a case of the extended mind. According to Fodor, there is an essential difference between the two cases.

Fodor has several objections but his main points can be summarized under the following three headings:

(3) a. the part-whole argument
b. the argument of derived vs. underived mental content
c. the mode of retrieval argument

It is easy to list differences between Otto's case and Inga's case, but it is important to recall that the discussion is ultimately about mind/brain identity. In all his arguments, Fodor seems to ignore that whatever happens inside the brain is equally diverse and also that differences among brain-internal events can just as easily be listed as differences between Otto's case and Inga's case. In other words, I will argue, Fodor is applying double standards.

Discussion is somewhat complicated by the fact that it is not entirely clear what we should understand by "the mind." Exact definitions are notoriously difficult and everyday usage is not always of much help either. But let us assume that we minimally agree on the idea that the mind is the collection of interactions that create mental content. As discussed above, this collection involves at least the interaction of processing units and data structures. In that sense it is appropriate to say that the mind has parts.

Processing units and data structures are relational notions, i.e., they have no cognitive or informational significance independent of each other. That means that they are only part of the mind during actual interactions, just as the data patterns on CDs or DVDs are only audio or video information when combined with external "processing" devices. Outside of actual interactions, our media contain potential audio or video information at best. Interestingly, the same applies to the processing devices: they are nothing, just arbitrary matter without a functional context provided by external data structures.

We should keep the relational nature of the parts of computational systems in mind when we evaluate Fodor's arguments against the EMT. Discussing the case in which Otto relies on his external notebook and Inga on her internal memory, Fodor says the following:

> The notebook is thus part of an 'external circuit' that is part of Otto's mind; and Otto's mind (including the notebook) is part of an external circuit that is part of Inga's mind. Now 'part of' is transitive: if A is part of B, and B is part of C, then A is part of C. So it looks as though the notebook that's part of Otto's mind is also part of Inga's. So it looks as though if Otto loses his notebook, Inga loses part of her mind. Could that be literally true? Somehow, I don't think it sounds right.

No, of course it does not sound right, because the argument fails to make a distinction between potential parts of mind and actual parts of mind. Thus, it would be odd to say that a president of the United States has a rich mind because he lives close to the Library of Congress. For the same reason, it is odd to say that Otto's notebook is part of Inga's mind when she is not actually using it. We live in an ocean of information, all potential extension of our minds. Whatever is available only becomes an actual extension of mind under conditions of proper access, when actual interactions take place or are within immediate reach somehow.

Note that exactly the same is true of information internal to Inga's brain: when it is not accessed (or worse, not accessible anymore due to brain damage), it is not part of Inga's mind. Thus, it would be equally odd to say that a president of the United States had a rich mind when, due to an accident, he had lost all access to his long-term memory. As soon as we give up double standards, there is no good reason anymore to make a principled distinction between information stored in libraries and information stored in brains. Webster's Encyclopedic Unabridged Dictionary of the English Language is not part of my mind, but I extend my mind with parts of it whenever I look up a word in it while trying to understand an English text. The same with cognitive skills like doing arithmetic. When we say somebody is good at arithmetic, we do not mean to say that he is only good at mental arithmetic and that

using pencil and paper does not count. Come to think of it, it is fairly absurd to exclude "external circuits" from the concept of mind.

Double standards can also be seen in Fodor's argument based on the distinction between "derived" and "underived" mental content:

> 'Underived' content (to borrow John Searle's term) is the mark of the mental; underived content is what minds and only minds have. Since the content of Otto's notebook is derived (i.e. it's derived from Otto's thoughts and his intentions with a 't'), the intensionality of its entries does not argue for its being part of Otto's mind. So the intensionality of notebooks can be granted by someone who doesn't think that notebooks are the sorts of thing that could be parts of minds.

This is another case of double standards because it ignores the fact that what is inside the brain is also diverse along the dimension in question. So, if we agree that the content of Otto's notebook is derived, the same must be the case with what Inga has in her long-term memory. Neither what is found in Otto's notebook nor what is stored in Inga's brain has independent content. If there is potential content, it must be of the derived kind in exactly the same way. So, what is the extrinsic factor that gives content to data structures? Fodor calls this mysterious entity "the mind" but that is begging the question, particularly when we talk about the physical structures of the brain. Nobody has ever succeeded in isolating a part of the brain that provides autonomous, underived content. That there are brain processes that qualify as such is a hope that Fodor and Searle seem to share, but it is irrelevant for the issue at hand. There is simply no argument that the distinction derived/underived corresponds with the distinction external/internal-to-the-brain.

Recall that we concluded earlier that neither memory data nor external devices processing these data have content independent of their actual interaction. In the terminology of the current discussion this would mean that both only have derived content (whis is potential content outside actual interactions). If we assume that both data structures and processing devices are necessary for the creation of mental content, a mythical embodiment of underived content (in abstraction of data structures) would literally be impossible. Saying that there is a source of underived content is, at this point, not more enlightening than the ancient Neoplatonic contention that ideas are emanating from The Mind of The One. Our understanding does not go further than the presumption that mental content is an emergent property of the interaction of structures which taken in isolation have no content whatsoever. Searle (1992) suggests that our total lack of understanding in this regard is a problem typical of computational theories of the mental. However, the mystery of the emergence of mental content does not disappear when we switch our hopes from computation to physics. Not too surprisingly, Searle's recommendations in this regard have led to zero results so far. I will briefly return to this matter in the next section.

Be this as it may, it is clear that Fodor's distinction between derived and underived content does not properly distinguish Otto's case from Inga's case and can therefore not be seen as a valid argument against the EMT.

Fodor's third category of arguments has to do with the differences in mode of retrieval between the Otto case (notebook) and the Inga case (internal memory). Fodor claims that the derived nature of the content of Otto's notebook depends on the fact that he has to think about it. He then continues about the content of Inga's memories:

> This is markedly untrue of mental things. Inga doesn't have to think about (or, in any literal sense, 'consult') her memories; she just has them and proceeds on her way in light of them.

This is followed by a remark about what Fodor thinks distinguishes the two cases in terms of mode of retrieval:

> It's untendentious that Otto's consulting 'outside' memories presupposes his having inside memories. But, on pain of regress, Inga's consulting inside memories about where the museum is can't require her first to consult other inside memories about whether she remembers where the museum is. That story won't fly; it can't even get off the ground.

It is hard to believe that Fodor means what he says here, because it is obviously false that in general someone just "has" his inside memories without having to consult or evoke other memories. As everyone knows from personal experience, it can be very difficult to retrieve inside memories. It can involve dependence on retrieval of other memories, but also other people's memories or external information. It can even involve the triggering potential of arbitrary objects, like the famous *madeleines* of Marcel Proust. More double standards, in other words.

Everything considered, then, I conclude that Fodor's arguments against the EMT are not convincing.

4. Delimiting the extended mind

So far, I have adopted a version of the EMT that extends mental computation with data structures outside the brain. This approach still assumes, of course, that certain parts of the mind can be successfully modeled by computational theories, like those advocated by Hubel and Wiesel (2004) and Marr (1982) for vision and by Noam Chomsky for language (1986, 2000). I have not been concerned with the world external to computation and I therefore agree with Clark and Chalmers (1998) that the EMT thus conceived is a kind of extended internalism rather than a form of the externalism discussed in the semantic literature since Putnam (1975). Recently, Noë (2009) has proposed a more radical version of the EMT that extends the concept of mind beyond the boundaries of computation, explicitly adopting Putnam-style externalism (*op. cit*., pp. 89-91). Noë calls this the "embodied, situated approach to mind" (p. 186), according to which "[m]eaningful thought arises only for the whole animal dynamically engaged with its environment […]" (p. 8). This dynamic engagement, in short, involves the world beyond computation.

As a matter of fact, Noë rejects computational theories of mind altogether (*op. cit*., p. 164), adopting arguments to that effect given earlier by Searle (1992). The gist of those arguments, in terms of our discussion so far, is that computation only involves derived content, while an adequate theory of mind must somehow account for *un*derived content. As an alternative, briefly discussed above, Searle proposed to look for a deeper understanding of the physical properties of the brain. As mentioned in the preceding section, this recommendation has not led to any fruitful research whatsoever, as we are entirely clueless about the question how physical structures could cause mental content.

More generally, criticizing computational theories of (aspects of) the mind is beside the point as long as these theories generate insight and no better theories are available. Furthermore, nobody believes that computation is all there is to the mind. A fair assessment of the current

state of the art is that we have relatively successful theories of limited parts of the mind (that only involve derived content) and that the ultimate origins of content ("underived content") is a complete mystery. This unknown source is what I referred to earlier on as "the homuncular residue," a term I prefer over the more common "homuncular fallacy," because the latter term suggests too much that we have solved a problem as soon as we exorcise the homunculus (the ultimate disappointment in this genre was Dennett 1992).

Interestingly, Noë rejects Searle's alternative for computational theories (i.e., deeper physical understanding of the brain) with arguments that can also be held against his own suggestions (Noë 2009, 164):

> But this is exactly the wrong conclusion to draw from the fact that brains don't think by computing: they don't, but not because they think some other way. Brains don't think.

It is perhaps not unreasonable to say that humans think rather than their brains and that their brains are only tools, like computers. But Noë's solution is no solution at all. At best he extends the arsenal of structures with *derived* content that we use while thinking. According to Noë, the crucial step forward is giving up the idea of the mind in terms of the internal states of individuals. Instead we are invited to see the mind in terms of "…our dynamic interaction with the things around us" (p. 164).

But that leaves the problem exactly where it was, even if it is true that we have to adopt some version of the EMT. What is problematic about Noë's account is that he erroneously suggests that the problem of underived content is caused by mind/brain identity, by seeing the mind as limited to the brain in computational terms (cognitive science) or to the individual brain causing thoughts by non-computational means (Searle). The problem is much more fundamental, namely that we have no idea at all how physical structures can create underived mental content. Assuming that Noë's radical externalist version of the EMT preserves the physical basis of mind (now involving *both* the brain and parts of the world), the solution of the mind-body problem at this level remains as remote as it has ever been.

There is another general weakness in Noë's argumentation. Almost all his examples are about visual perception. Since visual perception is at the mind-world interface almost by definition, it is to be expected that this area provides the best examples of dynamic brain-world interaction. Noë recognizes that some forms of consciousness cannot exactly be construed as immediate interaction with the world, for instance dreams. Obviously, dreams are a hard nut to crack because they involve a much degraded and different form of consciousness, but some kind of awareness nevertheless, and certainly no direct involvement of the world. Noë does not get any further than the *ad hoc* assumption that consciousness in dreams is different and *derived* from normal consciousness, which does involve direct dynamic interaction with the world. But as soon as we allow one species of consciousness that does not directly involve the world, the original hypothesis loses its generality.

The problems are much more profound in the case of language and meaning. Noë refers to Putnam's externalism only in passing. He, in fact, takes it for granted without much argument, suggesting that it is on a par with his examples about visual perception and ignoring the substantial critical literature about Putnam-style externalism.

Apart from language acquisition and the extended computationalism of the EMT version advocated above, the direct involvement of the world in the construction of meaning is much

less obvious than in the case of visual perception. There are, for instance, numerous words (like those of our logical vocabulary or the characters of fiction) that have no reference in the world at all. In novels, more so than in our dreams, we evoke complete worlds in full consciousness and understanding, without any direct involvement of the world whatsoever. Or take the cases of polysemy illustrated with the name *Schubert* above. Thanks to our introspective interpretive powers, we can interpret words together with the necessary potential contexts, without any direct involvement of the world. In short, language is just not at all like visual perception in its degree of intimacy with the external world. Blurring such important distinctions considerably weakens Noë's overall conception of mind. In fact, just assuming that dynamically incorporating parts of the external world goes anywhere in solving the problem of mind is empiricism of the crudest and most naïve kind.

Given these objections, I am not prepared at this point to adopt a version of the EMT that, for language at least, goes further than extended computation, keeping the rest of the external world at bay.

5. Summary and conclusion

Even if we limit ourselves to the cognitively involved parts of the brain, mind/brain identity is an untenable, (semi-)reductionist thesis. Biology cannot be reduced to physics and human culture and thought cannot be reduced to biology. Reductionism is false because physical structures have no inherent function. Function is extrinsic to form and requires a context. For example, the function of the heart can only be determined by the wider organic context that makes it meaningful to say that the function of the heart is to pump blood. This functionality is compatible with physics, but mere physical laws of cause and effect are not going to tell you whether side-effects, like the sounds produced by the heart, are as functional as its pumping of blood (Kauffman 2007), or functional at all.

Memories, like DNA, are necessary (in the "construction" of organisms) to preserve those physical phenomena that serve a purpose in certain contexts. Similarly, supra-individual, cultural memories are necessary to preserve those fruits of our agentive function assignment that have been useful in human life. Language is a case in point. It has no existence at a strictly individual, biological level. It crucially involves human inventions, words, that give a function to otherwise functionless biological structures, just as wind instruments give a new function to the lungs in invented musical contexts. It is for such reasons that mind/brain identity breaks down in the case of language: language only exists thanks to a shared culture, with information structures that are preserved external to individual brains. For this reason alone, some version of the extended mind thesis (EMT) must be adopted.

The EMT is crucial for most aspects of language, but most obviously so in the case of meaning. Whatever information is stored with individual words in the brain is, like all physical structures, without any inherent function and therefore not properly described as "meaning" or conceptual content. As in all known cases of functional physical structure, the function must be assigned from outside, in this case in part also from capacities of interpretation of unknown origin. The interpretive mind can generate infinitely many and often novel contexts that give a function (also known as "meaning") to the "dead" information stored with individual words. This we demonstrated with the phenomenon of polysemy. The rough material for meaningful computations, the information stored with words, is thoroughly entangled with convention and distributed over a community. Word meaning, therefore, is

another argument in favor of the EMT and an argument against a strictly individual conception of language and mental content in general.

Fodor's arguments against the EMT were shown to be based on double standards, on denying to brain-external data structures what was granted to certain brain-internal structures. What is particularly interesting in Fodor's arguments is the alleged distinction in terms of derived and underived mental content. This distinction was concluded to be irrelevant for the issues at hand, because all *known* constituents of cognition, both inside and outside the brain, only deal with derived content in Fodor's sense. The ultimate causes of meaning, the sources of underived mental content, are a complete mystery.

In this respect, it does not bring us any further to make a distinction among computational theories (standard cognitive science), deep and unknown physical theories of the brain (Searle 1992), or a version of the EMT that adopts direct dynamic interaction with the world (Noë 2009). The problem is that we cannot even begin to imagine how physical structures involving parts both of the brain and the world outside the brain, can cause underived mental content. By mental content I mean thought beyond the dead structures in the brains of zombies, i.e., thought we are aware of, with qualia and universals.

It seems to me that the hope to understand underived mental content through a deeper understanding of physics is an illusion. Those who follow this line of thought are typically thinking about the frontier of physics, particularly about the world of quantum physics (see for instance Penrose 1994). Currently unknown physics (as in the case of Searle 1992) might seem a safe bet, as the hopeless inadequacy of our current knowledge of the physical nature of the brain is immediately obvious. However, what was said above about the functionless nature of physical structures applies to *all* physics, including the niceties of Penrose's quantum gravity. If we want a reductionist perspective, the problem is that the world of the mind is at least two steps removed from the world of physics. First, physical structures became functional in the emergence of life. This is the non-agentive functionality preserved via DNA. Second, there was the emergence of agentive functionality in the animal kingdom, first preserved by the "culture" of imitative behavior and eventually culminating in human, fully symbolic culture. Meaning and the mental essentially involve agentive functionality, particularly the agentive capacity to give new interpretive contexts to information stored in our brains and the media of our culture.

What this comes down to is that there is no meaning outside of human creativity and history, particularly the history of human culture, for which the DNA-based history of life was a prerequisite. In other words, looking for the solution of the problem of underived mental content in future elementary physics (or any physics whatsoever) might be looking for things at the wrong end of the cosmic spectrum. No matter how profound our theories of physics, meaning can only be found higher up on the ladder, at the level of human agentive acts and the preservation of their fruits in human history.

Meaning, it seems, is an emergent property not of the micro-world alone but minimally also of the macro-world that encompasses the complexities of human life and history. Personally, I suspect that grasping the nature of underived mental content is beyond human cognitive capacity, a mystery in the sense of Chomsky (1975, ch. 4), also known as "cognitive closure" (see, for instance, McGinn 2002). I have no reason to doubt the idea that the mind supervenes upon physical structure, but it is as if, with the emergence of mind, we have entered an entirely new world that cannot be properly understood in terms of the 3rd-person perspective

of the empirical sciences. This has for a long time been claimed for subjective experience ("qualia"), but the same can be said about other very familiar inhabitants of our mind, the universals (ideas or types) in Plato's sense. That we see things in the world as instantiations ("tokens") of general notions ("types") is so elementary that it can be understood by a child. But even this most intimate and familiar aspect of mental content has resisted understanding for more than 2000 years of philosophical reflection. Terms like "generalization" label the problem rather than giving the solution and any imaginable physical realization of such a process would only produce more particulars instead of universals (see Koster 2005 for discussion). Since universals are an essential part of meaning, I do not see real understanding of *un*derived mental content forthcoming in the foreseeable future.

In the meantime, there is no reason to give up the project of exploring computational theories for *derived* mental content, be it that these theories, to my mind at least, should no longer make a principled distinction between brain-internal and brain-external data structures.

March 10, 2009

Bibliography

Burger, Tyler

1979 Individualism and the Mental. *Midwest Studies in Philosophy* 4, 73-121.

Chomsky, Noam

1975 *Reflections on Language*. New York: Pantheon Books.

1986 *Knowledge of Language*. New York: Praeger.

2000 *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.

Chomsky, Noam, Marinus Huybregts and Hendrik van Riemsdijk

1982 *Noam Chomsky on the Generative Enterprise: A Discussion with Riny Huybregts and Henk van Riemsdijk*. Dordrecht: Foris.

Clark, Andy

2008 *Supersizing the Mind: Embodiment, Action and Cognitive Extension*. Oxford: Oxford University Press.

Clark, Andy and David Chalmers

1998 The Extended Mind. *Analysis* 58, 7-19.

Donald, Merlin

1991 *Origins of the Modern Mind*. Cambridge, Mass.: Harvard University Press.

2000 The Central Role of Culture in Cognitive Evolution: A Reflection on the Myth of the 'Isolated Mind'. In L. Nucci, Ed., *Culture, Thought and Development*. Philadelphia: Lawrence Erlbaum Associates, 19-38.

Dennett, Daniel

1992 *Consciousnouss Explained*. Boston: Little Brown.

Fitch, Tecumseh W., Marc D. Hauser and Noam Chomsky

2005 The Evolution of the Language Faculty: Clarifications and Implications. *Cognition* 97, 179–210.

Fodor, Jerry
2009 Where is My Mind? Review of Clark (2008). *London Review of Books*, 12 Feb.
Gould, Stephen Jay and Richard C. Lewontin
1979 The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. *Proceedings of the Royal Society of London*. B 205 (1161): 581–598.
Gould, Stephen Jay and Elisabeth Vrba
1982 Exaptation: A Missing Term in the Science of Form. *Paleobiology* 8: 4-15.
Hauser, Marc D., Noam Chomsky, and Tecumseh W. Fitch
2002 The Faculty of Language: What Is it, Who Has it, and How Did it Evolve? *Science* 298, 1569-1579.
Hubel, David H. and Torsten N. Wiesel,
2004 *Brain and Visual Perception: The Story of a 25-Year Collaboration*. New York: Oxford University Press.
Jablonka, Eva and Marion J. Lamb
2005 *Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral and Symbolic Variation in the History of Life.* Cambridge, Mass.: MIT Press.
Jacob, François
1982 *The Possible and the Actual*. New York: Pantheon Books.
Jenkins, Lyle
2000 *Biolinguistics*. Cambridge., Mass.: MIT Press.
Kauffman, Stuart
2007 Beyond Reductionism: Reinventing the Sacred. *Zygon* 42, 903-914.
Koster, Jan
1988 Language and Epistemology. *Groningen Papers in Theoretical and Applied Linguistics TENK Nr. 6.* Published in French as Koster (1993). (Also at: http://www.let.rug.nl/koster/old papers.html).
1989 How Natural is Natural Language. In J.E. Fenstad *et al*., Eds., *Logic, Methodology, and Philosophy of Science VIII*. Amsterdam: Elsevier.
1990 Pork without Pigs. In Joan Mascaró and Marina Nespor, Eds., *Grammar in Progress. GLOW Essays for Henk van Riemsdijk*. Dordrecht: Foris, 305-315. (Also at: http://www.let.rug.nl/koster/old papers.html).
1993 Langage et Épistémologie [Translation of Koster 1988 into French by Nicolas Ruwet]. *Recherches Linguistiques* 22, 59-74.
2005 Is Linguistics a Natural Science? In Hans Broekhuis, Norbert Corver, Riny Huybregts, Ursula Kleinhenz and Jan Koster, Eds., *Organizing Grammar: Linguistic Studies in Honor of Henk van Riemsdijk*. Berlin: Mouton De Gruyter, Berlin, 350-358.
2008 Ceaseless, Unpredictable Creativity: Language as Technology. Ms., Groningen: University of Groningen (http://www.let.rug.nl/koster/1999.htm).
Lakatos, Imre
1976 *Proofs and Refutations: The Logic of Mathematical Discovery*. Cambridge: Cambridge University Press.
Lenneberg, Eric
1967 *Biological Foundations of Language*. New York: John Wiley & Sons, Inc.
Malik, Kenan
2000 *Man, Beast, and Zombie: What Science Can Tell Us About Human Nature*. New Brunswick, N.J.: Rutgers University Press.

Marr, David
1982 *Vision.* San Francisco: W.H. Freeman.
McGinn, Colin
2002 *The Making of a Philosopher: My Journey through Twentieth-century Philosophy*. New York: Perennial.
Moravcsik, Julius
1990 *Thought and Language*. London: Routledge.
Noë, Alva
2009 *Out of Our Heads: Why You Are Not Your Brain, and Other Lessons from the Biology of Consciousness*. New York: Hill and Wang.
Paul, Hermann
1880 *Prinzipien der Sprachgeschichte*. [Tübingen 1975: Niemeyer].
Penrose, Roger
1994 *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford: Oxford University Press.
Pinker, Steven
2002 *The Blank Slate: The Modern Denial of Human Nature*. New York: Viking Penguin.
Popper, Karl R.
1972 Epistemology Without a Knowing Subject. In Karl R. Popper, *Objective Knowledge: An Evolutionary Approach*. Oxford: Oxford University Press, 106-152.
Pustejovsky, James, Ed.,
1993 *Semantics and the Lexicon*. Dordrecht: Kluwer.
The University of Chicago Press.
Putnam, Hilary
1975 The Meaning of 'Meaning'. In Hilary Putnam, *Mind, Language, and Reality. Philosophical Papers, Volume 2.* Cambridge: Cambridge University Press.
Reichling, Anton
1935 *Het Woord: Een Studie omtrent de Grondslagen van Taal en Taalgebruik.* Nijmegen: Berkhout.
Russell, Edward Stuart
1982 *Form and Function: A Contribution to the History of Animal Morphology*. Chicago: The University of Chicago Press [originally appeared in 1916 at London: John Murray (Publishers) Ltd].
Saussure, Ferdinand de
1916 *Cours de linguistique générale*. C. Bally and A. Sechehaye, Eds. Lausanne and Paris: Payot.
Searle, John R.
1992 *The Rediscovery of Mind*. Cambridge, Mass.: MIT Press.
1995 *The Construction of Social Reality*. London: Penguin Books.
Segerstråle, Ullica
2000 *Defenders of the Truth: The Battle for Science in the Sociobiology Debate and Beyond.* Oxford: Oxford University Press.
Wilson, Edward O.
1975 *Sociobiology: The New Synthesis*. Cambridge, Mass.: The Belknap Press of Harvard University Press.
Wittgenstein, Ludwig
1953 *Philosophical Investigations*. Oxford: Basil Blackwell.