Prolegomena 13 (2) 2014: 233–251

Externalism and Empirical Research Programs in Semantics

JOŠKO ŽANIĆ

Department of Linguistics, University of Zadar, Obala kralja Petra Krešimira IV.2, 23000 Zadar
josko_zanic@yahoo.com

ORIGINAL SCIENTIFIC ARTICLE / RECEIVED: 06–07–14 ACCEPTED: 30–09–14

> …presumably, we do sometimes think of Venus and, presumably, we do so in virtue of a causal relation between it and us. But there's no practical hope of making science out of this relation.
>
> Fodor (1980)

ABSTRACT: The paper considers (causal) semantic externalism as a potential basis for an empirical research program in semantics and claims that externalism has not and cannot deliver in this respect. Externalism claims that content, at least for certain classes of expressions or concepts, is, at least in part, determined or individuated by factors external to the individual, the latter usually being cashed out as causal relations to the environment; internalism, on the other hand, claims that content is fully determined by factors internal to the individual. Externalism is criticized in its diachronic and its synchronic variety, and it is concluded that, for the purposes of a feasible empirical research program in semantics, organism-environment relations should not be seen as constitutive of content, but only as potential props for eliciting content, which should be seen as a mental/neural structure. The problem behind all forms of externalism is diagnosed as a certain misapplication or abuse of what I have termed the interpretative scheme, the assumption of which seems to be a necessary precondition for doing semantic description.

KEYWORDS: Content, externalism, internalism, research program, semantics.

1. Ways to study meaning and preconditions thereof

What is it to study meaning? The question could be posed with the intent to 'characterize the abstract form of a [semantic] theory' (Katz & Fodor 1964: 479; cf. also Dummett 1975, 1976). However, there is no particular reason

to suppose that there is a single form that a theory of meaning must conform to, nor does an overview of the current situation in theorizing about meaning lead one to suppose that any kind of general consensus as to this form is forthcoming. Rather, one should assume that there will be several such forms attempting to describe meaning in language, and I will begin this paper by giving a brief overview of the most prominent ones.

To study meaning has to involve, at some point, breaking out of the 'circle of language'. Unless this step is taken, what one gets is merely a translational semantics, one that issues in statements of the following sort: '"A" means the same as "B"' – statements that substitute one expression for another, giving no deeper insight into the nature of meaning (Davidson's 1984 theory of meaning is characterized by Dummett 1975 as being of this kind, although the verdict is later revoked[1]). But now, breaking out of the circle of language can be done in several ways, and this seems to be the point where, depending on how it is done, theories of meaning will assume different forms. There seem to be three basic ways of doing this (perhaps they exhaust the possibilities, but I will not claim this).

One way to break out of the circle of language is to attempt to connect linguistic expressions to objects and states of affairs 'in the world' (or, rather, model of the world). This will result in formal semantic theories of familiar varieties (cf. Chierchia & McConnell-Ginet 2000, Larson & Segal 1995). These theories will begin with axioms specifying the extensions of the primitive terms of the language (the extension being the set of things to which a term applies), and will then, for any arbitrary sentence of the language, spell out its truth-conditions (namely, states of affairs the obtaining of which would make it true) by composing them out of the relations among extensions of the primitive terms used in the sentence.

Another way to break out of the circle of language is to connect linguistic expressions to concepts. The goal of these kinds of theories (e. g. Jackendoff 1983, Croft & Cruse 2004) is to elucidate meaning in terms of analyzing the concepts that are activated when certain linguistic expressions are used or encountered. A full analysis of meaning will then be given as an account of the conceptual structure (a mental phenomenon) that underlies our use of language.

Yet another way to break out of the circle of language is to attempt to connect linguistic expressions to practices. These are the use theories of meaning (for one of the most recent formulations cf. Horwich 2005). These theories will attempt a reduction of meaning-properties of linguistic expressions to use-properties: they will account for the meaning of an expression in terms of its law of use, a regularity or know-how.

[1] In the Appendix, cf. p. 128.

These, then, are the several research programs in semantics (taken to be clusters of theories sharing fundamental assumptions), attempting to describe meaning in natural language.

The question that now poses itself is: what is the basis on which we judge whether a certain account of meaning is successful or not? This basis is, like in any science, successful prediction. Amongst the things that a semantic theory has to predict are the following: for isolated words, categorization responses of speakers; for words in sentences, syntactic behavior; for pairs (or sets) of sentences, logico-semantic relations such as entailment or contradiction. What is sought after are confirmed empirical predictions.

So, these different descriptions of meaning will be tested against data provided by speakers, of the sorts enumerated above.

According to many semanticists, the logico-semantic relations between sentences are one of the most important things that a semantic theory has to get right. If we assume, for the purposes of illustration, that speakers are prone to take the sentence 'Fritz is a cat' as entailing the sentence 'Fritz is an animal', we might sketch how the several theories (research programs) outlined above might attempt to account for this.The formal theories, if they construe the extension of 'cat' to be a subset of the extension of 'animal', will predict the entailment relation correctly, based on set-theoretic relations just mentioned. The conceptualist theories will account for the entailment by either taking the concept ANIMAL to be part of the concept CAT[2], or by taking the latter concept to be subordinated to the former in a conceptual taxonomy. Finally, the use theories might account for the entailment relation by taking the sentence 'A cat is an animal' to belong to the set of sentences specified by the law of use for the word as being meaning-constituting.

Now, my goal in this paper is not to adjudicate between these several research programs (the example above obviously leaves them on a par). It is rather to inquire whether a more ambitious goal with regard to the description of meaning might be attainable – one inspired by a position in the philosophy of language and mind called semantic externalism. Putnam ends his seminal 1975 paper by proposing a 'normal form for the description of meaning' (269), where one of the components is the description of the extension (which, as he famously claims, isn't determined by what's 'in the head'). The main question I want to pose in this paper is this: is this additional requirement on the project of describing meaning, adumbrated by externalism, feasible?

Before I proceed to the main issue, I would like to point out the following: any description program in semantics presupposes what I will term

[2] I will adopt the following standard graphical convention: words are referred to by putting them in quotes, whereas concepts are referred to by words in capital letters.

the interpretative scheme[3] as its necessary background. In assigning meaning/content to utterances, the semanticist has to assume an external standpoint: she has to have independent recourse to objects and events that are being talked about in the utterances that are the object of analysis. In formal semantics, the semanticist has to build a model of the world, in order to be able to assign to expressions their extensions, and this model is given from the standpoint of the semanticist (as interpreter, external observer), couched in her concepts. In conceptualist semantics, the semanticist attempts to analyze the mental structure of the speakers/hearers, the concepts activated upon using/encountering certain expressions, but has to retain independent access to the objects that the concepts are concepts of, in order to be able to say what the concepts are concepts of (even though a speaker/hearer only has access to cats, as it were, via her concept of cats, the semanticist can assume the external observer position and differentiate between the concept being analyzed, and its referents). In use-theoretic semantics, even though, as Horwich (2005: 78) puts it, 'there is no need for a word's law of use to relate occurrences of that word to members of its extension', the semanticist still has to be able to talk about the members of the extension, and identify them independently.

So the interpretative scheme is a necessary precondition for doing scientific semantic description. In the third section, I will suggest that the problem with an externalistically-based description program is a certain misapplication or abuse of the interpretative scheme.

2. On externalism

Externalism and internalism in current philosophy of language and mind are usually construed as metaphysical positions, positions on what determines or individuates (linguistic, mental) content[4], on what content supervenes on. Externalism claims that content, at least for certain classes of expressions or concepts, is, at least in part, determined or individuated by factors external to the individual, that content supervenes on the conjunction of internal and external factors, the latter usually being cashed out as causal relations to the environment[5]. Internalism, on the other hand, claims that content is fully

[3] Using the term 'the interpretative scheme' might call to mind the influential proposals of Dennett (1987) and Davidson (2001) regarding such matters. However, my notion of the interpretative scheme is much less theoretically loaded and I will outline it briefly above.

[4] For the purposes of this paper, I won't draw any kind of important distinction between linguistic and mental content.

[5] I will not in this article discuss Burge's (1979) social externalism, founded on the 'arthritis' thought experiment. I am here concerned with causal externalism only.

determined by factors internal to the individual, that it supervenes exclusively on internal factors (cf. Kallestrup 2012, also Schantz 2004).[6]

I propose to assess these positions with regard to the question which seems a better candidate to form the basis of a successful empirical research program in semantics. It might be claimed that this practical concern is irrelevant to the metaphysical *truth* of externalism or internalism. Metaphysical truth is, however, notoriously hard to come by, metaphysical conclusions being fragile and, more often than not, inconclusive. So, it might be useful to adopt, at least for the time being, this practical standpoint, and to inquire into the prospects of these positions as bases of empirical research programs. Indeed, externalism and internalism are sometimes treated explicitly as research methodologies ('… internalism and externalism are fully general research methodologies …, Hinzen 2006: 122). I will treat them, however, as potential *bases* for research programs, and attempt to investigate which seems to be the better candidate for the job. The outcome, if correct, might not speak directly to the truth of these positions, but that is not my concern here – the concern is rather to pick the more promising candidate for the job at hand (however, cf. Mendola 2008 for the most elaborate criticism known to me of externalist doctrines with regard to their metaphysical content).

I suppose internalism is the default position. Not, to be sure, in the sense of statistical prevalence in current theorizing, where externalism seems to be the dominant position, one that has acquired the status of the standard view. By 'default position' I mean one that seems prima facie the most natural or intuitive. Indeed, when Putnam (1975) and Kripke (1980) set forth their externalist arguments that turned the tide in the contemporary philosophy of language and mind, they were *challenging* a view that was taken for granted, that no one before them thought to question. The rest is history, as they say.

But I don't think the history has been a happy one. What I will argue here is that, although externalism made some incisive points about language and mind that no one can afford to disregard, it is hopeless as the basis of empirical research into meaning (although it was originally envisioned as such – cf. the reference to Putnam's 'normal form for the description of meaning' above). In the almost forty years since it was proposed, externalism has, as far as I can tell, not inspired any kind of semantic research that was able to yield interesting empirical results. Internalistically-based research programs have, on the other hand, delivered quite a lot, it seems. Detailed descriptions of large classes of expressions are available, cf. Larson & Segal (1995) in the formal tradition or Jackendoff (1990) in the conceptualist tradition. Putnam does allow for, in his normal form for the description of meaning, describing what's in the head (syntactic markers, semantic markers, and 'stereotypes')

[6] Externalism and internalism are also positions in epistemology, but that is for the most part an independent debate. Cf. Goldberg (2007).

– but it seems that this is all that can be hoped to be successfully describable in doing semantics. Externalism fails precisely where its most characteristic aspect comes into play, namely a request to study organism-environment interaction.

Internalism claims that whatever is described whilst studying meaning must ultimately be cashed out as something that is in the head, a neural structure. Externalism claims that a full description of meaning will at least partly be concerned with describing causal relations between the organism and its environment. This idea of causal relations between the organism and the environment being relevant to content can take different forms, however, and so externalism can appear in different varieties.

On the most general level, however, causal semantic externalism seems to come in one of two forms. Either it is claimed that certain causal contact between an organism, a user of a term, and the environment, that occurred at some point in time (in the past, of course), has fixed the meaning of the term, and forever holds it fixed (renewed causal contact may strengthen or confirm this fixing, but is inessential); or it is claimed that the causal relation between the organism and the environment that is relevant to meaning is causal covariance that occurs under certain circumstances. I will term the first kind of externalism *diachronic* externalism and the second *synchronic* externalism.

Diachronic externalism is the standard Kripke-Putnam story (a version of it, with troublesome consequences for externalism itself, is Davidson's 2001 externalism). The most prominent contemporary version of synchronic externalism is Fodor's (1990b, 1994, 1998, 2008). There are of course many other versions of semantic externalism on the market, e. g. those of Dretske (1981) or the teleological theories of Millikan and Papineau (for an overview cf. Loewer 1997), but they seem to fall into one of the two categories given above (or are a combination of the two), with the most prominent versions being those of Putnam-Kripke and Fodor, respectively.

I will claim that a semantic research program based on either diachronic or synchronic externalism is not feasible. I will attempt to show that the only feasible research program[7] in semantics goes *from the outside in* – namely it

[7] By 'research program' I have in mind here a program more general than the three programs outlined in part one, a program that can accommodate any one of them, as long as they study meaning as an internal aspect of the mind. I see no obstacle to construing formal semantics in such a way (the Larson & Segal 1995 version lends itself most easily to such a construal). As for the use theories, at least the Horwich (2005) version that I am relying on here as representative seems to be amenable to a mentalistic construal. If the meaning of a word-type is, on that account, constituted by a law of use for that word, which dictates the acceptance conditions of certain specified sentences containing the word, whereas accepting a sentence is understood as a having that sentence in one's 'belief box', which is a psychological phenomenon, there is no obstacle as seeing this approach to meaning as also being basically mentalistic.

treats the environment only as a prop, as a way of getting at what's in the heads of speakers/thinkers, and not as in any way constitutive of content.

2.1. Diachronic externalism

Diachronic externalism, as established by Putnam and Kripke, and given textbook form by Devitt & Sterelny (1999), is the position that, at least for some of our terms, at the point of their introduction into the language there occurred a reference-fixing, a content-conferring event, that determined the content of those terms. So, in the case of names, there was presumably a 'dubbing ceremony' where the name was causally *grounded* in its bearer. This happened in the following fashion: the object being named, in being perceived by those present at the ceremony, causally affected them, caused them to have certain thoughts about itself, and thereby the name was 'locked' to its bearer. Later the name was passed on to others who borrowed it from the original dubbers, or from those who in turn borrowed it from the original dubbers, etc., and so a causal-historical link was established leading back to the original object and the *ur-event* of its naming. This ur-event, given a certain point in time when the name is used, could have occurred years or centuries or millennia prior to that point, and all knowledge about the original name-bearer could have, at that point, been lost or distorted, but nevertheless using the name will result in successful reference to the object in virtue of the causal link leading back to the original event (and on the condition that the name is used with the intention to refer to the same thing that the original dubbers referred to).

This position was motivated by criticism of the earlier description theories of the reference of names, according to which what determines the reference/content of a name is a description or cluster of descriptions that speakers associate with it and that uniquely identifies its referent. In his famous examples with Jonah and Gödel, Kripke showed that such identifying descriptions are neither necessary nor sufficient for successful reference. Even in the absence of correct identifying descriptions, and even in the case that the descriptions associated with a name in fact identify someone other than the intended name-bearer, the name can be used to refer successfully to its intended bearer, in virtue of the link connecting the name-user to the name-bearer (the referent).

Descriptions were introduced into the semantics of names as a way of cashing out the Fregean notion of sense. Frege (1892) famously argued that names have to have senses because there has to be an aspect of the meaning of co-referential names that differentiates between them, and since the reference is in this case the same, it has to be something else – viz. sense as a mode of presentation of the referent. It is sense that is used to account for the informa-

tiveness of identity statements where coreferential names flank the identity sign. The causal theory of reference recognizes this issue, and preserves the notion of sense, but construes it as the property of referring to the object by a certain type of causal link. So, on this account, co-referential names will be causally linked to the same object, but by different causal links, and therefore have different senses.

In addition to names, this causal theory of content is also standardly applied to natural kind terms, such as 'water' or 'gold'. The externalist approach to natural kind terms was of course established by Putnam's famous Twin Earth thought experiment. Imagine a distant planet in our universe that is a molecule-for-molecule duplicate of Earth, except for the fact that what is called 'water' on Twin Earth, and what has the same phenomenal properties as water (the liquid that appears here on Earth), doesn't have the chemical structure H_2O but rather XYZ. Then, picking a point in time on both planets before chemistry discovered the structure of the respective liquids, we can say that the mental states of the utterers of the word(s) 'water' that are associated with this (phonological) word as used on Earth and Twin Earth, and that carry the knowledge of the phenomenal properties of the respective liquids, are the same on both planets, but nevertheless the extension of the terms is different, because the stuff referred to is different on the two planets. So, given that causal contact responsible for fixing the extension involved, on Earth, samples of H_2O, but, on Twin Earth, samples of XYZ, 'water' as used on Earth and Twin Earth refers to different stuffs.

The externalist account of content for natural kind terms, inspired by the Twin Earth thought experiment, is this: there has to have occurred an initial content-determining event, that locked the term onto the respective natural kind, but, as opposed to the case of names, there are complications due to the fact that a natural kind term is a general, rather than singular term, and so applies to each of a class of entities (viz. those that comprise the natural kind). In this case, the term will refer to whatever has the same underlying nature or essence as the paradigmatic samples of the kind that the original contact was with. The notion of 'paradigmatic' samples is brought in to ensure that, even though some things (or stuff) that superficially resembled the members of the intended kind, but were in fact not instances of the kind, might have been present at the original dubbing, they wouldn't be included in the extension of the term. Whereas the notion of paradigmatic samples ensures the 'only members of the kind' condition on the extension of the term, the notion of underlying nature or internal structure is supposed to ensure the 'all members of the kind' condition. This nature needn't be known at the time of the grounding of the term in the kind, but it is what ensures the right extension, and is in principle discoverable by science at some later point.

Is there any prospect of an empirical research program that would be engendered by the above insights? I don't think so. The basic problem with turning the above externalist claims into an empirical research program is that the level of description remains completely unclear. At what level is the allegedly content-conferring causal contact supposed to be investigated and described? Two levels come to mind: the macrolevel of everyday objects, and the microlevel of biochemistry and physics. But is a research program based on either capable of delivering any results? Consider a name, like 'Aristotle'. The reference-fixing event belongs to the distant past. We could tell, on the macrolevel, some kind of story of how Aristotle's parents named him 'Aristotle', how this involved him causally affecting them because of a perceptual connection, how this conferred content on the name, etc., but this is only a *story*, and a largely speculative one at that (due to lack of data), out of which no science can be made. Telling this story will not deliver any kind of interesting formalizable empirical results with regard to the description of meaning.

What are the prospects on the microlevel? Should we aim to find, via some super-duper advanced form of physics, the individual photons that bounced off Aristotle and hit his parent's retinas during the naming ceremony? Should we attempt to reconstruct the exact properties of these retinas? Should we search for Aristotle's DNA, in order to have an exact genetic 'fingerprint' of the referent of the name? This seems to be an impossible task, doomed from the start.

Consider now a natural kind term, such as 'kangaroo'. I've never seen a real-life one, and some other person (e. g. Michael Devitt) presumably has. What accounts for our ability to refer to kangaroos? The externalist will claim: the history of some Englishman or Irishman borrowing the term from a native of Australia who borrowed it from some other native, etc., all the way back to an original aboriginal who baptized the said animals (actually, the etymology of the term contains a misunderstanding between the settlers and the natives, but let's assume, for the sake of argument, that there was no such misunderstanding). But again, telling this story (needless to add, full of gaps and speculation) won't produce any kind of interesting formalizable, empirical addition to the scientific description of meaning.Therefore, the search for the presumed reference-fixing event will be irrelevant, and consequently a semantic description program based on diachronic externalism will be pointless.

A diachronic externalist, like Davidson, will claim that 'what our words mean is fixed in part by the circumstances in which we learned, and used, the words' (2001b: 29), bringing *individual* history into the story. But, if we followed this idea, there would be no consistent entity to describe: each individual's learning history is different, involving a different object (e.g. a

different kangaroo) so we would be forced to attempt to provide descriptions of idiosyncratic meanings, by trying to reconstruct these individual histories (and again, the level at which such a project is supposed to be carried out is unclear).

Of course, diachronic externalists allow for *multiple grounding* (cf. Devitt & Sterelny 1999: 75). This is possible with names and natural kind terms whose referents are still present to be interacted with. In these cases, it can be claimed that the sporadic causal interaction with the referent repeatedly grounds the term in the referent. However, in that case, there is no need to search for an original reference-fixing event, because the interaction can be directly observed. And so diachronic externalism, as a basis for an empirical research program, collapses into synchronic externalism: whatever promise externalism might hold as a basis for a successful empirical description program in semantics, it seems to lie in synchronic externalism. I will argue, however, that a research program based on synchronic externalism is also hopeless.

2.2. Synchronic externalism

If historical study is not the way to go, could it be that an externalistically-inspired research program in semantics has to proceed by studying presently observable organism-environment interaction? More specifically, is (at least an aspect of) meaning to be found in causal co-variance between aspects of the organism and aspects of the environment?

This is Fodor's approach. He (1994: Appendix B) takes Davidson's Swampman argument (2001b: 19–20) to constitute something of a reductio of diachronic externalism, and proposes a synchronic co-variational theory instead.

The basic idea is this: the concept WATER refers to water, not because an initial tokening of it *was* caused by water, but because tokenings of it *would* be caused by water under appropriate circumstances. Concepts (words)[8] have the content they do because of reliable causal co-variance between tokenings of the concept and instantiations of the property it 'locks onto'. Fodor allows for cases where a tokening of the concept isn't caused by an instantiation of the property; e.g. someone is looking at water but isn't thinking about it. Fodor also allows for cases where a tokening is caused of a concept that is not the 'right one', viz. not the one that is locked to the property; e.g. someone might be looking at water, but be caused by it to think of trees or whatever. Fodor can allow for these kinds of cases because what matters is the *reliability*

[8] Fodor mostly talks about concepts, so I will do the same in this subsection.

of the causal relation, in light of which exceptions can be tolerated. To put it another way, content depends on nomic relations among properties – so, if being water and being disposed to cause WATER-tokenings are reliably (nomically) connected, as Fodor claims they are, then it won't matter if there are occasional failures of this connection.

A concept might also occur to someone without there being external stimulation of the kind that standardly causes tokenings of it; e.g. someone might think of water because he is thirsty and there is no water around, so in this case the tokening of the concept WATER is caused by certain bodily states in the *absence* of water, rather than presence of it. But again, Fodor needn't be worried by this, because the fact that WATER-tokenings sometimes occur without the presence of water doesn't refute the reliable connection between being water and causing WATER-tokenings.

What Fodor does consider to be a difficulty is the so-called disjunction problem. Namely, since WATER-tokenings are sometimes (by an error in percept-categorization) caused by something that isn't water, e.g. by transparent jelly, how come WATER doesn't mean *water-or-transparent-jelly*? Fodor's reply is asymmetric dependence: he claims that transparent jelly causing WATER-tokenings asymmetrically depends on water causing WATER-tokenings. What that means is this: if water didn't cause WATER-tokenings, then neither would transparent jelly, but not vice versa. 'False tokens are metaphysically dependent on true ones', as he puts it (1990b: 91).

In contrast to the more standard version of externalism presented in the previous subsection, Fodor's externalism has several distinguishing characteristics (in addition to its synchronic character). First, it is *comprehensive*, in the sense of being meant to apply not only to names and natural kind terms, but to all terms in the language, i.e. to all concepts. Second, it is *extreme*, in the sense of claiming that 'content is constituted, exhaustively, by symbol-world relations' (1998: 14). So, whereas Putnam (1975) allowed for semantic features of terms that belonged 'in the head' (stereotypes and semantic markers) and were the proper subject matter of psycholinguistics, Fodor claims that mental processes are exclusively syntactic and so semantics has nothing to do with psychology. Semantics, according to Fodor, is the study of symbol-world relations, which exhaust meaning (content), and are given as reliable causal co-variance. Finally, Fodor's externalism is *counterfactual*, in the sense of construing content as being determined, not by what *did* happen, but by what *would* happen under certain circumstances.

The main problem with any such causal co-variational theory is of course that any particular event that is singled out as a cause of some other event is part of a causal network where many (indeed, infinitely many, cf. Field 2003) causes acted together to bring about the latter event. So, singling out *the*

cause is a kind of foregrounding, focusing of attention, which also depends on construal in terms of grain-size (are we talking about causes at the level of events involving medium-sized physical objects or microparticles?). How to get *the* cause out of this, which is meant to be the content of a term(concept)? Fodor (2008) proposes to do this by adapting Davidson's (2001) notion of triangulation. According to Davidson, there is no way to tell what in the environment a creature is responding to (what the contents of its thoughts are), or indeed whether it's responding to something in the environment rather than to a surface irritation of its sense-receptors, unless there is a similar creature present in the same environment, responding to it, and observing (interpreting) the first. Only by locating the common factor ('triangulating') causing similar responses in the two creatures, by seeing where the causal chains intersect, can the content of thought be determined. Fodor embraces triangulation as a means of solving the 'which cause' problem, but construes it counterfactually. So, it is not necessary for the determination of content, he claims, that there be an actual second creature present, an interpreter – it suffices to consider what *would* cause the original creature to have a thought of the same type if it were located a bit differently in the same environment. Again, the point of intersection of causal chains will give us the objective content, on this story.

Can synchronic externalism of the Fodor variety provide interesting descriptions of content? Not likely, in my opinion. Again, what kind of research program could these insights be turned into? Should we go about investigating the content of CAT or WATER by checking whether the presence of cats or water in the (perceptible) environment correlates with subjects' thinking about cats or water? On the one hand, this would turn out to be a pointless exercise. The only way to test the hypothesis would be to present experimental subjects with a cat, or a sample of water, and ask them what they are thinking about. Given the situation and the salient fact of being led to focus on an entity, most of them would probably be led to say 'I'm thinking about a cat' or 'I'm thinking about water'. But what we have here is a highly artificial situation, where subjects are basically asked to apply terms to entities in the environment (and prefix these terms with 'I'm thinking about'), and where the results are extremely predictable and uninformative. It's a trivial truth that competent speakers of English are able to apply the word 'cat' (the concept CAT) to cats, so what insight could we ever gain from such an exercise?

Fodor could of course claim that his theory is therefore obviously confirmed. But one should be wary of theories that are confirmed before experimentation has even started.

A research program based on synchronic externalism seems pointless, therefore. In addition, it seems to lead to dead ends and wrong predictions.

First, the dead ends. If content is seen as reliable causal co-variance (cum asymmetric dependence), what about concepts of fictional entities, such as UNICORN? Fodor proposes (1990b: 100–1) to deal with this problem by invoking nomic relations among *uninstantiated* properties: so the property of being a unicorn is, he claims, nomologically linked with the property of being a cause of UNICORN-tokens even if there aren't any unicorns. But how would we ever test something like this? How would we go about looking for relations among properties of something that doesn't exist? Fodor can claim that the concept UNICORN locks onto the property of being a unicorn by virtue of a hypothetical causal co-variance relation, but this sounds very much like a just-so statement, rather than an empirical claim. The notion of 'locking onto' (or 'resonating to', cf. 1998: 137) seems to be a vague metaphor that isn't likely to lead to any kind of empirical research program. The only kind of research program with regard to the concept UNICORN that would seem to be able to deliver interesting results (or any results) would be to use pictures of unicorns, the word 'unicorn', etc. to activate a concept in the research subjects' minds, in order to then attempt to discover the structure of this concept and how it is instantiated in the brain. One could use examples involving both concepts UNICORN and HORSE to this end, e.g. hypothesise that these two concepts share most of their structure (and therefore are instantiated in the brain in closely analogous fashion) except for a feature specifying that the former concept is one of a fictional entity, and then attempt to test these claims on the experimental subjects' verbal and non-verbal responses. However, this research program goes from the outside in: it uses certain environmental cues (not necessarily ones that would be considered to be content by Fodor's theory) to get at what is in the head, with only that what is in the head being considered to be content, and therefore the object of inquiry.

Another example of a dead end that Fodor's approach leads to is this. On his account, a social concept like PUNCTUALITY won't exhibit any kind of interesting difference of kind from a concept like CAT: the former concept will lock onto the property of being punctual, the latter concept will lock onto the property of being a cat, and that's that. But is this really all that can be said about these concepts?It would seem that being punctual is a property that depends on there being a community of people, certain norms and expectations being in force in this community, there existing certain forms of interaction that these norms and expectations apply to, there being minds with certain psychological properties, etc. Being a cat depends on none of this. Shouldn't our concepts somehow reflect these differences? Don't they in fact reflect them? We would need to frame hypotheses about the structure of a concept like PUNCTUALITY (e. g. that it involves a concept of a norm as a constituent of sorts), and then test them by way of psychological experi-

ments. But again, our hypotheses will have the from-the-outside-in form, namely the object of inquiry will be what's in the head and the environment will only serve as a source of cues.

Now for the wrong predictions that the theory makes. Since content is given, on Fodor's account, by a concept locking onto a property whose instantiations reliably cause tokenings of the concept, then, for any concept, there has to be a single property that it expresses (locks onto). But what is this property for a concept such as GAME? As Wittgenstein's (1953) famous example showed, there is no necessary property that something has to possess in order to be a game. Rather, GAME is a cluster concept (Jackendoff 2002: 352–6), such that there is no property that something *must* have in order to be a game, no necessary property, although various combinations of properties are sufficient for something to be a game. Again it seems that an empirical research program will have to look into the form the concept GAME has in the mind, viz. how a cluster concept such as this one is actually encoded in the mind/brain, whereas Fodor's approach cannot account for the content of this concept. Fodor's approach will face this kind of problem for any such family resemblance concept, and it would seem to be the case that many of our concepts are such.

Among other wrong predictions that Fodor's approach engenders is the prediction that words like 'cat' and 'cathood' will be synonyms, express the same concept, or at least coextensive concepts. For, if the concept CAT locks onto the property of being a cat and is reliably caused to be tokened by instantiations of this property, what's the story for CATHOOD? It seems that it has to be the same story: CATHOOD-tokens are also caused by instantiations of being a cat and therefore CATHOOD locks onto the same property. Fodor's way of dealing with examples like 'Jocasta' and 'Oedipus' mother' is to say that they indeed have the same content but differ *syntactically* at the level of language of thought. However, the 'cat'/'cathood' problem is the reverse of this: these are words/concepts that don't seem to apply to the same things (CAT applies to cats and CATHOOD applies to a property, viz. one of being a cat), but the theory predicts that they do. If they do lock onto the same property, and therefore are the same concept (or coextensive concepts), they should be able to enter into the same predication relations, i. e. be mutually substitutable. But this isn't the case: replacing 'cat' in 'A cat is an animal' by 'cathood' will yield an anomalous sentence/thought. This kind of problem can be generated for any concrete noun and an abstract term formed from it. The way to successfully differentiate between CAT and CATHOOD would seem to lie in accounting for the way abstract concepts function in the mind as opposed to concrete ones, but this explanation will not involve systematic recourse to organism-environment relations (yes, systematically, there are

situations where we can say 'that X' if X is a concrete concept but not if it is an abstract one – but this is an explanandum rather than an explanans).

Also, there is another kind of wrong prediction plaguing Fodor's account, concerning particularly asymmetric dependence, namely the following. Uttering the word 'cat' will reliably covary with the tokening of the concept CAT in the mind of the hearer. There is no reason to believe that this causal relation asymmetrically depends on the relation between cats and the CAT concept. They seem to be independent relations: for instance, if the cats-CAT connection were somehow broken, so that cats didn't cause the tokening of this concept any more, the other relation could still be preserved. But then the concept CAT, and any other concept that can be expressed in words is ambiguous: its content is both the property that it expresses and the word(-type) that can be used to express *it*.But it is highly implausible that the content of a concept is the word used to express it: this seems to have the relation the wrong way round. Asymmetric dependence will encounter this problem for any concept and a word for it.

Finally, a methodological point concerning Fodor's recourse to triangulation. The content of a thought is supposed to be pinpointed by figuring out where causal chains leading from the world to an observer and her counterfactual counterpart intersect when the observer and the counterpart are tokening the same concept. However, since it is a presupposition of this method that subjects have to be similar (Davidson 2001c: 119) – in order for them to have the same reaction to similar phenomena – the option is always open that we will find something else that they share, at the organism-internal level. Is there something like this? Since Fodor and Davidson restrict triangulation to perceptual situations, we can invoke something that is necessary for recognition of referents of concepts, namely certain visual prototypes of perceptible entities encoded in the mind. In order to be able to recognize cats in the environment, and to apply the concept CAT to them, there has to exist some kind of a visual representation of a typical cat in our minds (Jackendoff 2002, relying on the work of Marr, calls this level of cognitive structure *spatial structure*). Since the concept is presumably abstract and non-visual, there has to be some other kind of mental structure that enables us to recognize instances of it in the environment, a certain flexible kind of visual encoding that stores schematic representations of perceptible entities. But then, whenever two people can be said to have been caused by observing a cat in the environment to think of cats, the visual structure will also have been activated, because the causal chain has to pass through it. Of course it is natural to say (and think) that 'cat' refers to cats, and not to internal schematic representations of cats. But as far as cognitive science is concerned, it seems clear that the relevant object of investigation, in addition to the concept itself, will be this schematic

representation and its instantiation in the brain, and not real cats out there in the world.

I conclude that synchronic externalism doesn't seem to be able to form the basis of a promising empirical research program in semantics. Even if in some metaphysical sense content is 'out there', it seems clear what the object of study for a cognitive scientist (semanticist) should be. She could choose to study cats, water, doorknobs, etc. (whatever any word can be used to refer to outside the mind), or she can choose to study the mental structure that the mind uses to encode information about these entities, and how this structure is encoded in the brain. The way seems pretty clear – and, again, it goes from the outside in.

3. Causation and Content

I promised at the end of section one a diagnosis of the problem with externalism based on a certain misapplication of what I termed the interpretative scheme. I claimed that this scheme is a necessary precondition for doing semantic description. It involves having independent recourse to the objects and events talked about in the sentences that are the object of analysis. So, in doing semantics, one has to assume an external standpoint with regard to the discourse being analyzed.

However, assuming an external standpoint is not the same as assuming an external*ist* standpoint. It seems to me that externalism is an attempt to turn this unavoidable external position into a way towards a full(er) description of meaning. However, this seems to be a non-starter.

As I tried to show in the preceding discussion, there is no systematic, tractable, external relation between expressions/concepts and external entities. Attempts to turn externalistic insights into research programs turn out to be either pointless or impossible.

The key difference between externalism and internalism is, I think, in the way that the role of the basic causal relation C → E is construed in specifying content. Externalism places content (exclusively or partially) in the cause, in the external objects that brought about a certain mental effect. On the externalist picture, part of the content of the concept WATER is real water that caused (or *would* cause) the subject to entertain certain thoughts. However, as I tried to show, an attempt to turn this view into a scientific description program fails. One could always decide to turn semantics into a theory of everything, to require of it to describe anything and everything that we can talk about, i. e. simply *everything* – but I take it that it is clear that this isn't a feasible project, to put it mildly.

On an internalist account, things stand very differently. Content is placed exclusively in the effect (E), in mental/neural structures activated

while speaking, listening, and thinking. All sorts of causal relations between organism and environment obtain, of course, but they are not seen as *constitutive* of content. While trying to discover the structure of our concept of water, or cats, and its neural underpinnings, one can rely on all sorts of ways of activating this structure. One can show pictures to the experimental subject, one can show real entities, one can talk about water or cats, etc. These are all ways to cause the concept WATER (or CAT) to be activated in the mind, but none of these ways are taken to be constitutive of content, and therefore there is no pressure to investigate scientifically this alleged constitutive relation. They are just ways of *eliciting* content.

What to do with H_20 and XYZ, the example with which causal externalism started? From the perspective of doing semantics, such examples could be seen in two ways. Either they can be construed as an invitation to do some chemistry – but, most semanticists probably won't be interested in taking this route. Or they could be seen as showing something about our essentialist intuitions (although this is unclear – cf. Segal 2000: 127–8), viz. that we take the hidden internal structure of substances as more pertinent to their classification than superficial appearance. But the science of this is a science of an aspect of the mind/brain. Examples such as the Twin Earth thought experiment can be used to tease out the said intuitions and what they reveal about this aspect of the mind/brain. But then the route of a science based on such experiments goes in exactly the opposite direction from what Putnam intended – not from the inside out, from the mental concept to the study of the extension, but from the outside in. It goes from experimental situations where such examples are presented to experimental subjects to aspects of their minds and brains.

References

Burge, T. 1979. 'Individualism and the Mental', *Midwest Studies in Philosophy* 4, 73–121.

Chierchia, G. and S. McConnell-Ginet. 2000. *Meaning and Grammar* (2nd edition) (Cambridge, MA: The MIT Press).

Croft, W. and D. A. Cruse. 2004. *Cognitive Linguistics* (Cambridge: Cambridge University Press).

Davidson, D. 1984. *Inquiries into Truth and Interpretation* (Oxford: Clarendon Press).

——. 2001a. *Subjective, Intersubjective, Objective* (Oxford: Clarendon Press).

——. 1987/2001b. 'Knowing One's Own Mind', in Davidson (2001a: 15–38).

——. 1992/2001c. 'The Second Person', in Davidson (2001a: 107–121).

Dennett, D. 1987. *The Intentional Stance* (Cambridge, MA: The MIT Press/A Bradford Book).

Devitt, M. and K. Sterelny. 1999. *Language and Reality* (2nd edition) (Oxford: Blackwell Publishers).

Dretske, F. 1981. *Knowledge and the Flow of Information* (Cambridge, MA: The MIT Press).

Dummett, M. A. E. 1975. 'What is a Theory of Meaning?', in S. Guttenplan (ed.), *Mind and Language* (Oxford: Clarendon Press), 97–138.

——. 1976. 'What is a Theory of Meaning? (II)', in G. Evans and J. McDowell (eds.), *Truth and Meaning* (Oxford: Clarendon Press), 67–137.

Field, H. 2003. 'Causation in a Physical World', in M. Loux and D. Zimmerman (eds.), *The Oxford Handbook of Metaphysics* (Oxford: Oxford University Press), 435–460.

Fodor, J. A. 1980. 'Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology', *Behavioral and Brain Sciences* 3, 63–73.

——. 1990a. 'A Theory of Content, I: the Problem', in J. Fodor, *A Theory of Content and Other Essays* (Cambridge, MA: The MIT Press/A Bradford Book), 51–87.

——. 1990b. 'A Theory of Content, II: the Theory', in J. Fodor, *A Theory of Content and Other Essays* (Cambridge, MA: The MIT Press/A Bradford Book), 89–136.

——. 1994. *The Elm and the Expert* (Cambridge, MA: The MIT Press/A Bradford Book).

——. 1998. *Concepts* (Oxford: Clarendon Press).

——. 2008. *LOT 2* (Oxford: Clarendon Press).

Frege, G. 1892/1952. 'On Sense and Reference', in M. Black and P. Geach (eds.), *Translations from the Philosophical Writings of Gottlob Frege* (Oxford: Basil Blackwell), 56–78.

Goldberg, S. C. (ed.) 2007. *Internalism and Externalism in Semantics and Epistemology* (Oxford: Oxford University Press).

Hinzen, W. 2006.'External and Internal Aspects in the Semantics of Names', in Marvan (2006: 121–141).

Horwich, P. 2005. *Reflections on Meaning* (Oxford: Clarendon Press).

Jackendoff, R. 1983. *Semantics and Cognition* (Cambridge, MA: The MIT Press).

——. 1990. *Semantic Structures* (Cambridge, MA: The MIT Press).

——. 2002. *Foundations of Language* (Oxford: Oxford University Press).

Kallestrup, J. 2012. *Semantic Externalism* (London and New York: Routledge).

Katz. J. J. and J. A. Fodor. 1964. 'The Structure of a Semantic Theory', in J. A. Fodor and J. J. Katz (eds.), *The Structure of Language* (New Jersey: Prentice-Hall), 479–518.

Kripke, S. 1980. *Naming and Necessity* (Cambridge, MA: Harvard University Press).

Larson, R. and G. Segal. 1995. *Knowledge of Meaning* (Cambridge, MA: The MIT Press/A Bradford Book).

Loewer, B. 1997. 'A Guide to Naturalizing Meaning', in B. Hale and C. Wright (eds.), *A Companion to the Philosophy of Language* (Oxford: Blackwell), 108–126.

Marvan, T. (ed.) 2006. *What Determines Content? The Internalism/Externalism Dispute* (Newcastle: Cambridge Scholars Press).

Mendola, J. 2008. *Anti-Externalism* (Oxford: Oxford University Press).

Putnam, H. 1975. 'The Meaning of "Meaning"', in H. Putnam, *Mind, Language and Reality: Philosophical Papers, vol. 2* (Cambridge: Cambridge University Press), 215–271.

Schantz, R. (ed.) 2004. *The Externalist Challenge* (Berlin: de Gruyter).

Segal, G. 2000. *A Slim Book about Narrow Content* (Cambridge, MA: The MIT Press).

Wittgenstein, L. 1953. *Philosophical Investigations*, trans. G. E. M. Anscombe (Oxford: Basil Blackwell).