Symbol Grounding - the Emperor's New Theory of Meaning?

Morten H. Christiansen*
Center for Research on Concepts and Cognition
Indiana University
510 North Fess Street
Bloomington, Indiana 47408
morten@cogsci.indiana.edu
and
Centre for Cognitive Science
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW

Nick Chater
Neural Networks Research Group
Department of Psychology
University of Edinburgh
7 George Square
Edinburgh EH8 9JZ
Scotland, U.K.
nicholas@cogsci.ed.ac.uk

Abstract

What is the relationship between cognitive theories of symbol grounding and philosophical theories of meaning? In this paper we argue that, although often considered to be fundamentally distinct, the two are actually very similar. Both set out to explain how non-referring atomic tokens or states of a system can acquire status as semantic primitives within that system. In view of this close relationship, we consider what attempts to solve these problems can gain from each other. We argue that, at least presently, work on symbol grounding is not likely to have an impact on philosophical theories of meaning. On the other hand, we suggest that the symbol grounding theorists have a lot to learn from their philosophical counterparts. In particular, the former must address the problems that have been identified in attempting to formulate philosophical theories of reference.

Introduction

In recent cognitive science, there seems to be a consensus that cognitive systems are best understood as complex information processing systems situated in and continually adapting to their immediate environment. Insofar as these systems are endowed with internal representations, such representations are typically assumed to be environmentally grounded. The problem of what it means for such internal states to represent a category of things in the external world has been the subject of much discussion in the philosophy of language and mind. Recently the same problem has gained much attention within cognitive science under a new and apparently different guise. Whereas the problem is known within philosophy as the problem of intentionality (or the problem of meaning), the cognitive science version has been named the *symbol grounding problem.*

The problem of symbol grounding arose from the philosophical criticism of AI found in the Searle's (1980) Chinese room argument. This parable purports to show that a symbol-processing system cannot be the right substrate for cognition, because the system's most basic constituents have no *intrinsic* meaning. Such a system would in principle be able to pass as a lifetime Chinese pen-pal without having *any* understanding of Chinese whatsoever. The bottom-line of the parable is that the very same symbol manipulation can be simulated (externally or internally) by an agent who does not understand any Chinese at all.

Putting the validity of the Chinese room argument aside, the conclusion that symbols have no intrinsic meaning is an unobjectionable one. Besides being an integral part of the definition of a physical symbol system (Newell & Simon, 1976), the arbitrariness of the internal constitution of symbolic expressions is one of their most noted features.[1]

Indeed, this conclusion is closely related to semantic externalism—the doctrine that meaning must be specified in terms of the relationship between the agent and environment, rather than in terms of the properties internal to the agent. Internalist semanticists (like Searle) might object that this relationship is irrelevant since meanings are *in the head* only. However, semantic internalism suffers from the *"matching problem* (McGinn, 1989); that is, the problem concerning how the taxonomy of meanings in-the-head originally (e.g., through learning) came to match up with the taxonomy of things in-the-world. It is far from obvious that this

*Morten Christiansen is supported by a CRCC Research Assistantship and by award No. V920053 from the Danish Research Academy. Nick Chater is partially supported by grant No. MRC SPG 9024590 from the Joint Councils Initiative in Cognitive Science/HCI.

[1]See, for instance, Putman (1981) for general arguments that the intrinsic properties of representations do not in specify, or even constrain, their meanings.

problem can be solved without specifying some sort of referential function that secures the right connection between internal meanings and the external things they are about (i.e., are content of). But this is similar to what the extenalist doctrine is all about. Furthermore, as this doctrine has been discussed extensively within the philosophy of language and mind, we draw the attention to the outcome of those discussions in our treatment of the symbol grounding project.

The structure of the paper is as follows. First, we argue for the relevance of philosophical theories of meaning to symbol grounding. The former has raised a number of problems that the latter will need to address. Next, we therefore give a brief outline of the most relevant of these problems with emphasis on the grounding of connectionist representations. These problems involve matters concerning categorization errors and categories with no members. Finally, we turn our attention to symbol grounding in Stevan Harnad's sense, arguing that this approach—despite its *prima facie* empirical nature—is prone to the problems which dog recent philosophical theories of meaning.

Why Theories of Meaning Matter to Cognitive Science

A central problem in the philosophy of language is to provide a theory of meaning for natural languages. This philosophical problem abstracts away from the details of particular human languages (i.e., it abstracts away from questions of linguistics). Rather, it aims to provide an account of what it is for a language, or perhaps particular statements or utterances, to have a specified meaning.

It is generally assumed that the project of providing such a theory of meaning can be broken into at least two subprojects: specifying what it is for a particular atomic expression (generally individual words or morphemes) in a language to have a particular meaning—this is the *theory of reference*; and determining how the meanings of complex expressions can be derived from the meanings of their parts and their mode of composition—this is *compositional semantics*. This distinction seems to be paralleled in Harnad's (1990, 1992, 1993) hybrid analog/symbolic model of cognition. He, too, distinguishes between specifying how basic tokens can be grounded externally to a system; and determining how these representations can be systematically composed to produce complex, semantically interpretable compounds. The theory of reference therefore seems closely related to Harnad's (1990) first subproject: "How is symbol meaning to be grounded in something other that just more meaningless symbols? This is the symbol grounding problem" (p. 340).

A superficial difference between symbol grounding and the traditional philosophical problem is that symbol grounding is concerned not with external natural languages, but with internal mental systems of representation. However, this difference is not an important one. Indeed, in recent philosophy of language and mind, the orthodox position is that the meaning of external natural language is derivative on the content of the internal states (inspired by a Gricean analysis of communication). For example, Fodor (1987, 1990) treats the problem of meaning as a matter of explicating the meaning of formulae of an internal system of mental representations. Furthermore, there has been a recent and vast philosophical literature concerned with the meaning of representationally atomic terms which is fundamentally concerned with the problem of reference for internal states (for example, Dretske 1981; McGinn, 1989; Millikan 1984; Putnam, 1981). It appears, then, that symbol grounding is a new name for an older problem—the problem of providing a theory of reference for atomic formulae of a system of internal representation. This problem is itself closely connected to the much older problem of establishing a theory of meaning for external natural language.

In this connection it might be objected (e.g., by Harnad, personal communication) that symbol grounding is not just a philosophical problem rediscovered, but that it is a separate, empirical problem. Furthermore, it might be assumed that this empirical problem can be resolved prior to the resolution of the philosophical problem. This position is difficult to understand. The symbol grounding problem, as stated, is precisely the problem of providing a theory of how atomic symbols can refer; it does not concern what one might most naturally think of empirical issues in this context—the problem of finding out the meaning associated with particular representations in particular organisms. Indeed, the latter problem seems to presuppose that a solution to the former problem is solved, since they require the application of a theory of reference for representational primitives—a theory we currently lack.

In addition, a similar "empirical" retort could be made to any philosophical problem. The scientist can always say: "I'm not concerned with the philosophical problems of aesthetics or ethics, but with a science of what is beautiful or good". What makes the philosophical problems so compelling is that they represent fundamental *conceptual* difficulties which arise in attempts to theorize about these domains, scientifically or otherwise. That is, scientific theorizing runs headlong into, and cannot simply ignore, philosophical issues. Specifically, we have no idea what a scientific theory of meaning might look like, precisely because of the profundity of the philosophical problems that we face.

Thus, appealing to the *prima facie* empirical nature of the symbol grounding project won't help.

The relationship between the symbol grounding problem and the problem of meaning from an externalist perspective is sufficiently close to warrant that the former must meet the same challenges as the latter (and *vice versa*). Consequently, we can construe Dretske's (1981) information theoretic view of content, Fodor's (1987, 1990) criticisms of the "crude causal theory" of meaning, and Millikan's (1984) teleological theory of content as concerned precisely with the problem of symbol grounding—though, of course they do not use this terminology. In the next section, we therefore discuss the problems facing symbol grounding in this light.

Problems with Externalist Semantics

Within cognitive science the advent of connectionism has given rise to much enthusiasm regarding the search for a solution to the symbol grounding problem. This optimism has been expressed by cognitive scientists in terms of statements such as, for example, "Analog sensory projections are the inputs to neural nets that must learn to connect some of the projections with some symbols (their category names) and some of them with other symbols (the names of other interconfusable categories) by finding and using invariant features in them that will subserve correct categorization performance" (Harnad, 1993; p. 8),[2] and "networks are self-organizing systems that learn to represent the important features of their environment" (Cottrell, 1987: p. 68). It has even permeated into philosophy: "connectionism offers significant resources for explaining how representations are *about* other phenomena and so possess *intentionality*" (Bechtel, 1989: p. 553).

The reason for this enthusiasm is to be found in the ability of distributed neural networks (via learning) to develop internal representations that in interesting ways mirror the structure in the externally given input. However, the internal states of present day connectionist networks are no more "grounded" than their symbolic counterparts (also cf., e.g., Cliff, 1990). Crucially, the distributed representations in question are only non-arbitrary in relation to the structure of the given input representations, not in relation to what the latter are representations *of*; i.e., the entities they supposedly refer to in the outside world. Since the input representations provided by the programmer are typically pre-structured and of a highly abstract nature, it is always possible to give a network's input representations a different interpretation, thus changing the projected content of the internal distributed representations. This has been mirrored empirically by the fact that only a few experiments have been carried out with "real" sensory-type data (in sense of not having been pre-processed by the programmer), and then usually without success.[3]

Nevertheless, seeking to endow a system of any kind with meaning involves implicitly making assumptions about what it is for a state of a particular system to represent. For example, without a theory of meaning, whether explicit or implicit, it is impossible to view networks *systematically* as possessing or developing *representations* at all. More generally, seeing a connectionist network, or any other system, as a *computational device* at all, is dependent on being able to ascribe meaning to the states of the system (at least as far as any interesting, non-trivial sense of "computational device" is concerned). Otherwise its internal states are not appropriately viewed as *processing information*, but simply as passing through sequence of states; the network will be viewed simply as an informational "black box", where only inputs and outputs are interpreted, and those by fiat.

In connectionist "symbol" grounding meaning is attached to the states of a network on the basis of what those states can be said to "connect" or *correlate* with. More generally, connectionist units or patterns of activation are viewed as picking out categories, with which they correlate, and which specify their meaning. Thus the network is viewed as acquiring the corresponding concept (or, what Harnad refers to as an associated "category name"). In the following we therefore focus on connectionist models which plausibly can be viewed as involving category or concept learning. However, there are serious philosophical problems concerning not only connectionist "symbol" grounding but also, more generally, the adequacy of any correlational semantics as the basis of a theory of meaning. We will address these problems in the next two subsections and emphasize their impact on the symbol grounding project.

The Problem of Error

One of the fundamental challenges to the project of symbol grounding is allowing for the possibility of categorization error. People routinely make both false positive and false negative errors. Mistaking a pattern of shadows at night for a person is an instance of the former; failing to see a dark figure in the bushes is a case of the latter. However, without some added machinery connectionist "symbol" grounding models are unable to account for the possibility that we have the concept PERSON but still

[2] It should be noted that Harnad stresses that connectionist networks are only one candidate for the part of an invariants extracting learning mechanism in his 3-way analog/invariants-extractor/symbolic model.

[3] For an example of such failure, a more detailed discussion of connectionist "symbol" grounding and the problems of externalist semantics as well as references to specific connectionist models, see Christiansen & Chater 1992.

make such mistakes. The content of this concept (equally, the meaning of the corresponding state, or for Harnad, 1990, 1992, 1993, what internal state(s) a category name is associated with) is determined by what it correlates with. The fact of error—i.e., the fact that the concept does not pick out all and only objects that are people—shows that something is not quite right. That is, the PERSON concept in question does not consist of the sensory invariants necessary for correct categorization of persons.

To meet this challenge it might be suggested that content is fixed during the *learning* of the concept, instead of being determined by subsequent performance, outside the learning period, when mistakes may occur (Dretske, 1981). The idea is that the correlation holds within the learning period (fixing the content correctly), but not necessarily afterwards (allowing for the possibility of error). Unfortunately, the relevant property can never be determined by the training set alone; even if the learner is given perfect feedback about which of a set of things are people and which are not, forming the concept PERSON involves an inductive generalization from a finite set of instances. Which concept has been formed cannot therefore be determined from the correlation observed in the training set alone, since all manner of different properties will fit that training set, but differ elsewhere, such as the pathological PERSON-OR-PATTERN-OF-SHADOWS or PERSON-NOT-IN-CAMOUFLAGE. Rather, which of these concepts has been formed is determined by how the system has *generalized* from the training set—that is, how it would respond to stimuli outside the training set. Subsequent errors, after the learning period has been completed (assuming that some such boundary can be enforced), demonstrate that generalization has been imperfect; the correlation is violated and the concept has not been learned after all. Thus, appeal to learning fails to reconcile the possibility of learning a concept (or, in Harnad's terms, learning to assign a category name) with the possibility of occasional misclassification.

Another possible suggestion is that while errors may occur in difficult cases (perhaps when the stimulus is degraded in some way), the correlation that fixes content need only hold in clear cases. As with appeal to the learning period, the idea is to partition performance into two classes, one in which correlation determines the concept in play (and which is necessarily error-free) and a second class in which the correlation need not be maintained, thus allowing for errors. Unfortunately, however, as Fodor (1990) forcefully argues, what counts as a clear case cannot be specified independent of the concept being learned. It could, for instance, be argued that in our previous example of miscategorizing a pattern of shadows as a person, the error is due to suboptimal conditions; that is, nighttime is suboptimal for detecting persons. During the daytime it is conceivable that no such errors would occur, thus making daytime optimal for recognizing persons. However, optimality could equally be invoked to argue that the internal state supposedly corresponding to PERSON is, in fact, a PERSON-OR-PATTERN-OF-SHADOWS concept. The reason for this being that the latter correlates properly at night (due to optimal conditions), but is prone to miscategorizations in the daytime (due to suboptimal conditions only allowing persons to be seen). The general moral is that there is no known independent, non-circular distinction between "good" and "bad" cases, in terms of which the problem of error can be dissolved.

The Problem of Non-Existing Entities

A further problem for the symbol grounding project is explaining the origin of the meaning of categories with no instances such PHLOGISTON, UNICORN and, rather more arcane cases such as Harnad's (1992) PEEKABOO UNICORN ("a horse with a horn that vanishes without a trace whenever senses or measuring instruments are trained on it", p. 12–13). These categories have no instances and hence can neither be causally implicated in producing, nor correlated with, internal states (see, for example, Fodor, 1987). These symbols cannot be "grounded" by some state of a network which comes to correlate the presence of peekaboo unicorns in the environment; for these are *never* present in the environment—they don't exist. The only proposed solution for a causal/correlational view—which, by the way, also is the solution that Harnad (1992) endorses—is that the meaning of non-existents is composed out of the meaning of more primitive terms, which do exist. So, the story goes, the meaning of *peekaboo unicorn* can be found as the right composition of "horse", "horn", "vanish", "trace", "measure", "instrument", and "train", if each of these categories are grounded.

This view presupposes that terms for things which do not exist can be defined in terms of things that do. This position appears to have the rather radical consequence that *every* term must have a definition. For suppose that a term *X* does not have a definition; then it must refer to something real; hence *X*s must exist. Thus a semantic fact (concerning definability) appears to be revealing about a metaphysical fact (whether there are *X*s). On the face of it, this means that we could learn what there is in the world, simply by examining language, which seems to be absurd. So it seems that we must conclude that *every* term must be definable in terms of other terms. The thesis that some terms have good definitions is highly controversial; the thesis that *all* terms do is so radical that it has not, to our knowledge, ever been advanced.

Harnad's Symbol Grounding Model

As we have seen, Harnad's (1990, 1992, 1993) proposed solution to the symbol grounding problem essentially involves the notion that internal states of an agent are sensitive to environmental invariants. In this way, they have their meaning—with respect to the internal workings of that agent—causally specified in terms of those invariants. In particular, he advocates the idea of *"iconic" representations* as tranduced projections of distal objects impinging on an organism's sensory surfaces. These correspond to the configuration of the sensory array at any given time, albeit in a somewhat reduced form. They are, as such, not stable representations stored in long-term memory, but rather a kind of perceptual "snapshot". Categories, or *"categorial representations*, can be reduced from such iconic representations by extracting their invariant features. Harnad (1990) stresses the non-symbolic nature of both iconic and categorial representations: "The former are analog copies of the sensory projection, preserving its 'shape' faithfully; the latter are icons that have been selectively filtered to preserve only some of the features of the shape of the sensory projection: those that reliably distinguish members from non-members of a category" (p. 342). Through the association of a category with a "name" (in the form of an arbitrary symbol), the representation of the former becomes available for further processing. This allows a non-standard "dedicated" symbol system to construct structured representations (e.g., complex concepts).

This position, associating arbitrary names (or *concepts*) with categories characterized by sensory invariants, has a long pedigree. It goes back to the behaviourists assumption that the meaning of verbal behaviour could be expressed in terms of contingencies between the production of that behaviour and states of the world (e.g., Skinner, 1957). More recently, the ecological approach to perception (most notably, Gibson, 1979) has stressed that perception involves picking up invariants in the perceptual array. In a similar way, information based semantics (e.g., Dretske, 1981) has attempted to identify invariant features of the environment which are systematically linked to the internal states of an organism. Lastly, a similar "correlational" view is presupposed in the literature on "animal concepts" (for an overview, see Chater & Heyes in submission) and in the interpretation of the activity of real neurons (e.g., Schurg-Pfeiffer & Ewert, 1981).

All of these approaches share a fundamental assumption with Harnad's approach to symbol grounding: they assume that there are certain aspects of the external world—invariants, classes of stimuli, or whatever—which systematically co-occur, or correlate with, an internal (or external) state, and that this systematic state-world relation underwrites the interpretation of the state as having the ontent associated with that aspect of the external world. That is, all are species of correlational approaches to the theory of reference. The meaning of an internal symbol is guaranteed by its correlation with the corresponding invariants or stimulus class in the world. Hence, we suggest, all these approaches are prone to the problems currently plaguing externalist semantics based on correlation.

It will not suffice to reply that the problems of error and non-existent entities are armchair criticism with no empirical support. That is, simply to exhort the scientist to search for the sensory invariants associated with atomic symbols will not make the problem of misrepresentation disappear. Neither will the problem of non-existing things be solved merely by looking for definitions that will provide meaning for (complex) symbols referring to empty categories. Besides, as we shall see below, there is much empirical evidence which underlines, rather than undermine, the two philosophical criticisms outlined above.

Regarding the first philosophical point, Harnad (1992) contends that all-or-none categories exist, for which necessary and sufficient conditions for membership can be found. That is, it is possible to extract sensory invariants for categories like BIRD, and in this way overcome the problem of erroneous inductive generalization. However, this is incongruous with much recent cognitive psychology research into adult human categorization. The bulk of these studies suggests that our categories are inherently graded and unstable—even when it comes to birds (see, e.g., numerous papers in Neisser, 1987—or, Christiansen, in preparation, for an overview). These results have received further support from numerous developmental studies of categorization behaviour. For example, Fivush (1987) found that infants do not categorize objects according to perceptual features or attributes, but according to how they are *used* in particular contexts and to what their *common function* is. In other words, human categorization behaviour seems to involve more than just categories based entirely on bottom-up extraction of sensory invariants.[4] This is further supported by many of the adult categorization studies which indicate that categorization involves not only bottom-up processing, but also a significant amount of top-down processing. This may constitute another problem for Harnad's approach because of its sole committment to the former: "Grounding, by its very nature, is something that is better done bottom-up ..."

[4] It is also worth noticing that artificial intelligence approaches to visual object recognition and speech perception have long since abandonned the search for the magic "perceptual invariants" in terms of which high level categories are grounded (see Marr, 1982, for discussion).

(1993: p. 6).[5]

Regarding the second philosophical point, a general consensus in current psychological theories of concepts is that no words (or *common-sense* concepts) have a definitional structure (again, see Neisser, 1987, for an overview). In addition, a number studies have shown that concept combination does not appear to follow standard symbolic compositionality (for a discussion, see Lyon & Chater, 1990). These results might be accounted for in Harnad's model depending on what his notion of "dedicated symbol system" entails—but the lack of definitions for even complex concepts seems to block this way out.

Conclusion

In this paper we have tried to caution against unduly optimistic claims about connectionist solutions to the symbol grounding problem. Philosophical considerations about externalist semantics provide valuable benchmarks against which models of symbol grounding should be assessed. Consequently, we are not saying that connectionist symbol grounding is *a priori* impossible, only that there are a number of challenges which need to be met before such models can become candidates for *true* grounding of internal representations. These benchmark problems, of which we have mentioned the most pressing above, concerns such matters as the problem of error and the representation of categories with no instances. Unless these problems can be addressed, the project of grounding representations by appeal to connectionism (or any other kind of "analog" representational device), for all its scientific ring, seems no more likely to succeed than all those old-style philosophical theories of reference. Thus, it still remains to be seen whether symbol grounding is more than just the emperor's new theory of meaning.

Acknowledgements

We would like to thank Dave Chalmers for comments and suggestions regarding the first draft of this paper.

References

Chater, N. & Heyes, C. (in submission) Animal Concepts: Content and Discontent.

Christiansen, M. & Chater, N. (1992) Connectionism, Learning and meaning, *Connection Science*, 4, 227–252.

Christiansen, M. (in preparation) Beyond Localist Concept Representation. Ms. Centre for Cognitive Science, University of Edinburgh.

[5]However, this top-down input might be exactly what is needed to constrain the process of inductive genralization from indeterminate sensory input to categories.

Cliff, D. T. (1990) Computational Neuroethology: A Provisional Manifesto. Technical Report No. CSRP-162, School of Cognitive and Computing Sciences, University of Sussex, Brighton.

Cottrell, G. W. (1987) Toward a Connectionist Semantics. *Theoretical Issues in Natural Language Processing*, 3, Association for Computational Linguistics, University of New Mexico.

Dretske, F. I. (1981) *Knowledge and the flow of information*. Cambridge, Mass: MIT Press.

Fivush, R. (1987) Scripts and categories: Interrelationships in development. In U. Neisser (Ed.), *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*. Cambridge: Cambridge University Press, Pp. 234–254.

Fodor, J. A. (1987) *Psychosemantics*. Cambridge, Mass.: MIT Press.

Fodor, J. A. (1990) *A Theory of Content and Other Essays*. Cambridge, Mass.: MIT Press.

Gibson, J. J. (1979) *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Harnad, S. (1990) The Symbol Grounding Problem. *Physica D*, 42, 335–346.

Harnad, S. (1992) Connecting Object to Symbol in Modeling Cognition. In A. Clark & R. Lutz (Eds.), *Connectionism in Context*. Springer-Verlag.

Harnad, S. (1993) Grounding Symbols in the Analog World with Neural Nets. *Think* (special Issue on Machine Learning)

Lyon, K., & Chater, N. (1990) Localist and Globalist Approaches to Concepts. In K.J. Gilhooly, *et al. Lines of Thinking*, Volume 1. Chichester: John Wiley.

Marr, D. (1982) *Vision*. San Francisco: Freeman.

McGinn, C. (1989) *Mental Content*. Oxford: Basil Blackwell.

Millikan, R.G. (1984) *Language, Thought and Other Biological Categories*. Cambridge: Cambridge University Press.

Newell, A. and Simon, H.A. (1976) Computer science as empirical inquiry. *Communications of the ACM*, 19, 113–126. Reprinted in M. Boden (Ed.) *The Philosophy of Artificial Intelligence*. Oxford: Oxford University Press.

Neisser, U. (Ed.) (1987) *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*. Cambridge: Cambridge University Press.

Putman, H. (1981) *Reason, Truth and History*. Chigaco: Chicago University Press.

Schurg-Pfeiffer, E. & Ewert, J.P. (1981) . Investigation of neurons involved in the analysis of Gestalt prey features in the frog Rana temporaria. *Journal of Comparative Physiology*, 141, 139–158.

Searle, J. R. (1980) Minds, Brains and Programs. *Behavioural and Brain Sciences*, 3, 417–424. Reprinted in M. Boden (Ed.) *The Philosophy of Artificial Intelligence*. Oxford: Oxford University Press.

Skinner, B. F. (1957) *Verbal Behavior*. Metheun: London/New York.