Harnad, S. (1996) The Origin of Words: A Psychophysical Hypothesis In Velichkovsky B & Rumbaugh, D. (Eds.) " Communicating Meaning: Evolution and Development of Language. NJ: Erlbaum: pp 27-44.

THE ORIGIN OF WORDS: A PSYCHOPHYSICAL HYPOTHESIS

Stevan Harnad Department of Psychology
University of Southampton
Highfield, Southampton
SO17 1BJ UNITED KINGDOM
harnad@soton.ac.uk
harnad@princeton.edu
phone: +44 1703 592582
fax: +44 1703 594597

http://cogsci.soton.ac.uk/~harnad/
http://www.princeton.edu/~harnad/
ftp://ftp.princeton.edu/pub/harnad/
ftp://cogsci.ecs.soton.ac.uk/pub/harnad/

ABSTRACT: It is hypothesized that words originated as the names of perceptual categories and that two forms of representation underlying perceptual categorization -- iconic and categorical representations -- served to *ground* a third, symbolic, form of representation. The third form of representation made it possible to name and describe our environment, chiefly in terms of categories, their memberships, and their invariant features. Symbolic representations can be shared because they are intertranslatable. Both categorization and translation are approximate rather than exact, but the approximation can be made as close as we wish. This is the central property of that universal mechanism for sharing descriptions that we call natural language.

In speculating about the origins of language we do well to remind ourselves just what we are pondering the origins of: For some, a language is something so general that just about every form of human activity qualifies: music, dance, even emotional expression (Agawu 1991, Goodman 1968, Pribram 1971). For others, it is a very specific and complex mental organ that allows us to produce and recognize grammatically correct sentences (Chomsky 1980). I would like to take a third road and consider language to be only that form of human activity that is intertranslatable with English (or any other language), plus whatever mental capacity one must have in order to produce and understand it. The intertranslatability criterion, however, though rather powerful, is still too vague and general. So let me add that one of the principal features of language is that it allows us to categorize the world and its parts in what appear to be an infinity of different ways, among them possibly a way that comes close to the way the world really is.

1. Translation and Categorization

So in pondering the origins of language, we are pondering the origins of an intertranslatable form of classifying ability. It is an ability that allows us to say: That is an apple; an apple is a round, red fruit, etc. Now this view of language is dangerously reminiscent of positions that are reputed to have been discredited by philosophers -- by Wittgenstein (1953) and Quine (1960), for example. Wittgenstein was at pains to show us that the "Look, this is an X and that is a Y" model of language is wrong, or woefully simplistic: What matters is not what words stand for but how they are used by a speech community. Quine even held that the "X" in "Look, that is a rabbit" (uttered while pointing to a rabbit) is so hopelessly ambiguous that it could mean just about anything to anybody: rabbit parts, rabbit stages, unique instants, or what have you. There is simply no way of arriving at the fact of the matter -- or perhaps no fact of the matter to arrive at.

How then is one to defend the "glossable-classificatory" view of language being proposed here in the face of such prominent criticism? Well, there is always a point of retreat that one can safely repair to as long as one is willing to abandon realism about word meaning: There may be no way of settling on the fact about what people mean when they say "Look, that is an X," but we can certainly describe the regularities in the external conditions under which they tend to do so, and the requisite internal conditions that would make it possible for them to do so under those external conditions. This position is not *behaviorism*, for it is very much concerned with what is going on inside the head. A behaviorist can never explain *how* an organism manages to classify its inputs as it does; he must take that success for granted. All he can tell you is what kind of a history of rewards and punishments shaped the organism to do so, given that it can and does do so (Catania & Harnad 1988).

So it is a form of *cognitivism* that is being proposed here (Harnad 1982): People use language to classify the world in a shared and modifiable way. The internal structures that allow them to do so are the physical substrate of language, and hypotheses about the origins of language are hypotheses about the origins of those structures, so used. There is room for *functionalism* here too (Fodor 1975, Pylyshyn 1984): The most important property may not be the specific physical realization of the structure underlying language, but its functional principles, which may be physically realizable in many different ways. The question "what is language?" becomes the question "what functional substrate can generate language's expressive power?" and this in turn becomes (according to what we have just agreed) "what functional substrate can generate our glossable classifying ability?"

Let us return to Quine's underdetermined rabbit -- which he chose to call, in an undetermined language, "Gavagai." Gavagai is meant to stand holophrastically for our expression: "Look, that's a rabbit." According to the glossability criterion, the two phrases must be intertranslatable. Now let me inject an important qualifying note right away: Intertranslatability is never exact; it is only approximate. However, the approximation can be made as close as one desires -- not necessarily holophrastically, perhaps using a profligate quantity of words, but with the resultant meaning coming as close as need be, reducing uncertainty to whatever level satisfies the demands of the shared external communicative context for the time being (Steklis & Harnad 1976). (People presumably communicate in order to *inform* one another, and to inform is to reduce uncertainty about competing possibilities among which a choice must be made). It is an interesting and suggestive parallel fact that categorization, like translation, is provisional and approximate rather than exact (Harnad 1987a).

Consider the first of Quine's variant readings, "undetached rabbit parts": On this reading, Gavagai could mean: "Look, that's undetached rabbit parts." But of course all that is needed to disambiguate the two is a larger sample of classification problems. For a language with an expressive power that allows full intertranslatability with English must be able to capture the difference between the external circumstances in which we are speaking of rabbits and those in which we are speaking of undetached rabbit parts. "Rabbit," for example, is no good for distinguishing detached from undetached rabbits: They're both rabbits, as far as that goes. "Detached rabbit," on the other hand, is a closer approximation, but now we are unpacking English's holophrastic side: "Rabbit" is indifferent to the distinction between intact and disassembled rabbit conditions. In English, we need two words to mark that difference. But if in Gavagese the holophrastic "Gavagai" really means "Look, that's undetached rabbit parts," then (to meet our stern criterion of intertranslatability) there will have to be another lexical item in Gavagese for "Look, that's detached rabbit parts," "Look, that's rabbit parts," "Look, that's a part," and "Look that's a rabbit." Now one can certainly continue to play this game holophrastically ("Bavagai," "Travagai," etc.), and the more synthetic languages such as German and Innuit (Pullum 1989) certainly go further in this direction than, say, English or Chinese do. But there are limits to what it is practical to do in this holistic way, and most languages seem to have elected instead to go analytic, coining small, detached portable words to mark important classes, and making combinations of them in the form of phrases and propositions to mark complex or composite conditions.

The point does not depend on practicality, however, for whether it does so analytically, synthetically, or even entirely holophrastically, a language must provide the resources for marking distinctly all the categories we distinguish (in English, say). Now Quine could argue that even with all distinctions marked there can always be higher-order ambiguities we have not yet thought of. But then at that point one must revert to the approximationism mentioned earlier: All that is needed is that language have the resources to mark all potential distinctions as they arise; pre-emptive ambiguities with respect to inchoate future distinctions (such as Goodman's [1954] green vs. "grue") do not count as underdetermination for they are differences that do not yet make a difference[1].

It is also a rather vague conjecture that a language as a whole is open to multiple interpretations -- say, English as it is, versus "Fenglish," in which the meanings of "true" and "false" are swapped and all other meanings are suitably adjusted so that everything remains coherent: If one said "`That is a rabbit' is true" in Fenglish, "true" would mean what false means in English, but only because Fenglish "is" means what English "isn't" means, "rabbit" means "non-rabbit," and so on. Yet in order to have English and Fenglish speakers continue to discourse with one another coherently, in the same world of objects and events, without ever suspecting that their words don't mean the same thing, so many adjustments seem to be needed that to conjecture that the deception is even *possible* may be equivalent to assuming that formal "duals" of meaning exist (like the duals of logic and mathematics, where it can be proved that certain formal operations can be systematically swapped under a transformation in such a way as to yield coherent dual interpretations)[2]. Such a strong conjecture calls for a proof, and as far as I know, no one has offered a proof of the existence of semantic duals in language.

Apart from the absence of a formal proof, another reason for suspecting that coherent dual interpretations of languages may not be possible is that the systematic adjustments would have to go beyond linguistic meaning. They would have to encompass perception too, and would thereby inherit the problems of the "inverted spectrum" conjectures (e.g., Cole 1990): Could you and I be walking around the same world speaking and behaving identically, even though I see the sky as blue and the earth as green, whereas you see the sky as green and the earth as blue? Again, if our classifications are always approximate, it may be a long time before we discover the difference[3]. But if there is ever a difference, it will disambiguate us forever. And until then there's no difference between identity and approximate identity -- or at least no difference that makes a difference -- no uncertainty on which any actual outcome depends.

2. The Problem of Grounding Word Meanings in Perceptual Categories

The problem of the underdetermination of meaning has taken us rather far afield from our original intention merely to say informally what it is that a language-origin theory is a theory of the origin of. However, the fact that conjectures about semantic duals turn out to be related to conjectures about perceptual duals is not, I think, coincidental; and it also happens to be closely related to the hypothesis to be put forward in this paper. For to contemplate swapping the meaning of words is also to contemplate swapping experiences. True/false is a rather abstract distinction, but blue/green is just about as concrete as a distinction can get. Is there a way to *ground* the former in the latter -- to ground abstract semantic categories in concrete perceptual categories, and thereby to ground the meanings of the *names* of abstract categories (the words denoting them) in the meanings of the names of concrete categories? This is the kind of theory of the origin of words and word meanings that will be put forward here. And although the theory is primarily a bottom-up psychophysical model for the representation of word meaning, it has some rather straightforward implications for the origin and nature of language.

Psychophysics is the branch of psychology that is concerned with our perceptual capacity: (1) What stimuli can we detect? (2) What stimuli can we tell apart (discriminate)? (3) What stimuli can we identify (categorize)? The first two questions pertain primarily to the sensitivity of our sense receptors, although limits on our ability to make sensory discriminations (and to extend them with instruments) will also influence our ability to make conceptual and semantic distinctions. The third question, about identification or categorization, however, coincides squarely with a large segment of our linguistic capacity: the naming of sensory categories.

The connection between language and perception is at the heart of the "Whorf Hypothesis" in linguistics and anthropology, a conjecture that has had a chequered history. The hypothesis is that language influences (or perhaps even determines) our view of reality. To state it less vaguely: the way things look to us (and what things we believe really exist) depends on how we name and describe them in our language. Whorf's original example concerned the Hopi language, which apparently lacks a future tense. He accordingly inferred that the Hopi lacked a concept of the future. It turns out that Whorf was wrong in that case, partly because of an imperfect understanding of the Hopi language, and partly because the lack of a concept of the future seems to be too radical a deficit to attribute to a human culture living in an Einsteinian universe, given all the ways the temporal dimension impinges ineluctably on human life.

Having adopted the intertranslatability constraint we might already have suspected that something was amiss in Whorf's inferences, because English ought to be fully intertranslatable with Hopi. Hence, whatever *concepts* an Englishman might have, a Hopi should likewise be eligible to have. The Hopi language might lack, for example, the vocabulary for discussing quantum mechanics or general relativity, but this lexical deficit is trivial, and should be remediable by providing the requisite information and instruction *in Hopi* (albeit perhaps with the help of a few coinages or semantic extensions for the sake of convenience and economy). The same should be true of the concept of future (Steklis & Harnad 1976).

Whorf might have replied that it was not that he thought the Hopi could not acquire a concept of the future (perhaps even in Hopi), but simply that they did not have one at the time. Unfortunately, it is likely (on account of the universal temporal contingencies mentioned earlier) that they did. But let us suppose that they might not have had one -- although only in the sense that they (and most people without an advanced education) likewise do not have the concepts of quantum mechanics or general relativity. In that form, the Whorf hypothesis is really only a rather obvious statement about the relation between one's lexicon and one's conceptual repertoire: We tend to have names for the kinds of things that we think there are and that we tend to talk about; if the existence of new things is pointed out, we can always baptize them with a new name, not thereby changing our language, but only extending its lexicon. Hence, on the face of it, the real causal story seems to be the reverse of the Whorf Hypothesis: Reality influences language, which was presumably the commonsense view in the first place.

I think there may be more to the Whorf Hypothesis than this, however, so let us pursue it a bit further: The second specific case in which the hypothesis has been investigated is that of color terms (Berlin & Kay 1969)[4]. The prediction was that the visible spectrum was subdivisible in many different ways, and that the qualities of the colors and the differences among them should be influenced by the way we partition the spectrum into the named color categories of the language we speak. Berlin & Kay studied color terms in different cultures; they found that whereas languages did differ in how and where they subdivided the spectrum (although the differences were not quite as radical as one might have hoped), the effects on color *perception* seemed minimal, if there were any effects at all. Our color perception -- and hence the quality of the colors we can identify and discriminate -- is determined largely by the physiology of our color receptors, which is for all practical purposes identical across cultures (and languages) (Boynton 1979).

3. Categorical Perception

The universality of color perception would appear to represent another defeat for the Whorf Hypothesis -- with reality (this time internal rather than external reality) again influencing language, rather than the reverse. However, a closer look at the actual processes and mechanisms involved suggests that there may also be some Whorfian effects in the predicted direction (language on perception, rather than perception on language) in color perception after all. The area of research in which these subtler effects have been investigated is a specialized subfield of psychophysics called "categorical perception" (Harnad 1987). It has been found that although the boundaries of color categories are governed primarily by the physiology of the color receptor system, their exact location can be modulated by experience with seeing and naming colors. Boundaries can be moved somewhat; they display some plasticity, and secondary boundaries can perhaps be created on the basis of subcategorization and naming alone (Bornstein 1987).

How qualitative these effects are is open to different interpretations; but they are certainly *quantitative,* in the sense that equal-sized physical differences are more easily discriminated across named category boundaries than within them. This is really a Whorfian manifestation of perceptual learning theory's old interest in the "acquired distinctiveness" and "acquired similarity" of cues: The idea was that two stimuli would look more alike if they had the same name and more different if they had different names (Lawrence 1950; Gibson 1969).

It is the subject of another book (Harnad 1987) how categorical perception (CP) works in detail. An even more closely investigated case of CP than color categories is the perception of speech sounds. Another theory related to the Whorf Hypothesis -- the "emic/etic" distinction in phonology and anthropology -- turns out to be closely related to the work on the categorical perception of phonemes. The emic/etic distinction is a somewhat metaphorical extension of the phonemic/phonetic distinction in phonology (Pike 1982). The speech sounds of a given language vary in many ways. Only some of these differences signal a difference in meaning in the language. These are called "phonemic" differences. The rest of the differences are "phonetic" differences -- non-signalling differences that are less salient, less readily perceived and less easily produced than the phonemic differences.

In CP terms, phonemic differences are differences *across* a phoneme category boundary and phonetic differences are differences *within* a phoneme category. By analogy with color terms, the difference between blue and green is an "emic" difference, whereas differences among different (unnamed) shades of green are "etic." The metaphoric extension underlying emic/etic theory is that the emic distinctions are the salient ones in a culture, the ones that have been underwritten by language, whereas the etic distinctions have not (or have not *yet*) become bounded, named categories.

So one way to reconstruct the Whorf Hypothesis is this: Naming a category generates an "emic" (perhaps qualitative) distinction along an "etic" continuum of (quantitative) differences. Language does not create the etic differences. Those are furnished by our sensory apparatus; however, it does create the emic distinctions, and these become the salient ones, the ones that govern our (provisional) view of reality.

Let us recall, at this midpoint of our discussion, the informal criteria we adopted for language at the outset: A language is an approximately intertranslatable system for approximately categorizing the world. We have spoken a little about categorization and approximation. What about intertranslatability? Names of perceptual categories are trivially intertranslatable. All one need do is coin a gloss: "rabbit" = "gavagai." But naming categories is not all there is to language. Even if language's principle function is conceded to be classificational, not all classification takes the form "This is an X," or "X" = "Y." Language can also describe properties, chief among them being category membership itself.

4. Naming and Describing

If I say "Rabbits are white" I am actually making a statement about category membership: The members of the category "rabbits" are members of the category "white." [5] I want to conjecture here that any assertion in any language can be reformulated as a statement about category membership.

(For example, the preceding sentence would assert that members of the category "assertion in any language" are members of the category "statement about category membership.") Other speech acts are simply assertions with special markers, such as interrogative or imperative. If this conjecture is correct then it follows that, after *naming*, the second critical linguistic function is *describing*: A description always takes the form of a statement about category membership. Note that even an "ostensive" (pointing) statement such as the holophrastic "Gavagai" is actually a description, "This is a rabbit," where the deictic "this" refers to the member of a singular category, namely, "the thing I am pointing to right now." The assertion is that "that thing" is a member of the category "rabbit." Hence ostensive statements are simply special cases of descriptions in which the item for which the deictic term (this, that) stands is present and available to the senses.

So although it is true that names must precede descriptions in the sense that they provide the atomic terms of a description, it is also true that some (possibly holophrastic) ostensive assertion must be primitive in all category naming. No chicken/egg worries need arise here. If all assertions are statements about category membership, and the assertion is the minimal linguistic utterance, then all one need note is that there are two kinds of assertions: ostensive and descriptive (or "de re" and "de dicto"). Both say something of the form "The members of X are members of Y," but in the ostensive case X is a sensory event that is available and can be pointed to (or, for the realist, an object that is present and can be pointed at by means of the sensory event) and Y is simply "things that are called Y"; whereas in the descriptive case X is "things that are called X" and Y is "things that are called Y."

At this point my psychophysical hypothesis about the origin of words can be presented explicitly: Words originate by ostensive experience with concrete sensory categories. This "grounds" them psychophysically. They can then enter into descriptions of higher categories, including abstract ones. Here is a simple example that I have used before (Harnad 1990):

> (1) That is a "horse" (naming by ostension).
>
> (2) That is "stripes" (naming by ostension).
>
> (3) A "zebra" is a "horse" with "stripes" (description).

All the important features of language's remarkable expressive power and its grounding in prior sensory experience are captured by this exceedingly simple example. "Horse" is named on the basis of direct sensory acquaintance. "Stripes" likewise. Then these grounded terms can be used to ground a new term, "zebra," by description alone, and so on.[6]

Is this just an empiricist (sense-datum) theory of meaning, and hence vulnerable to the many existing objections[7] to such theories? In order to avoid those objections, my hypothesis has actually been put forward in the form of a black-box theory of object-sorting and word-use rather than a theory of perception and meaning. Without getting involved in side-issues, this means that this kind of theory can only hope to explain word-use behavior that is Turing-indistinguishable from the use of meaningful language. There is always the possibility that it only describes the inner workings of a mindless, meaningless robot that simply acts exactly as if it were speaking meaningfully. That's a limitation I am happy to accept (Harnad 1991).

But I have not yet said much of inner workings. Nor is "ostension" a very satisfactory account of the learning of concrete categories from sensory experience. What must be noted in order to see that ostension is a far from trivial candidate for the grounding of words is that the problem of *category acquisition* is itself far from trivial. It is at least as general as the problem of induction and pattern recognition. What a category-learning device must be able to do is to sample a finite number of sensory instances of "X"'s and thereafter name X's correctly -- for all the "X"'s that human beings can name. This is something that no man-made pattern-learning device can even come close to doing at the present time. Nor do I claim to have solved the categorization problem. I have merely proposed a representational model that has some features I think the ultimately successful model will have to have too. These features are described more fully elsewhere (Harnad 1987a, 1992; Harnad et al. 1991, 1996). For present purposes, the brief description that follows should be sufficient.

5. Symbolic Representations

Learning to categorize all X's will require three kinds of internal representation. The first, "iconic representations," are analogs of the physical patterns that concrete objects project onto the surfaces of our sensory receptors. These representations are used primarily to discriminate stimuli that occur simultaneously or in rapid succession. They allow judgments to be made about whether two sensory projections or traces are the same or different, and if different, about the degree of difference between them.

The second kinds of representation, "categorical representations," do not preserve the analog shape of the sensory projections; they preserve and encode only the invariant sensory properties *shared* by all the members of a concrete perceptual category. These invariant properties are learned by sampling positive and negative instances of the category in question (i.e., members of the category and its complement: the set of alternatives with which the members could be confused [8] and finding the features that will correctly sort that particular sample as well as future samples. Such features are converged on by a learning algorithm [9] that generates successful categorization. The invariant features are always provisional, however, and the categorization always approximate, as the context of confusable alternatives could always be widened.[10]

The names for the categories that have iconic and categorical representations then go on to furnish the primitives for the third kind of representation: "symbolic representations." These include the primitive category names (including names of invariant features) and combinations of names in the form of propositions about category membership. Most of the names correspond to the lexicon of a natural language; their combinations take the form of sentences (descriptions).

This three-level representational system is grounded bottom-up in psychophysical categories. "Top-down" influences occur through CP: Similarity judgments are not mediated purely by iconic representations (or perhaps iconic representations are not pure): Belonging to the same category -- i.e., having the same name -- makes things look more similar, and belonging to a different category (different name) makes them look more distinct. Most of the symbolic component consists of internal translation: "An X is a Y," "A Y is a Z," etc. It is the primitive symbols, which are grounded in nonsymbolic representations -- iconic and categorical ones -- that prevent this symbolic circle from being vicious. [11] Two different systems of grounded symbolic representation are in principle intertranslatable and their respective groundings can be tested against ostensive experience in the real world.

6. The Origin of Words

So my hypothesis about the origin of words is really a hypothesis about the origin of symbolic categories: They originate in sensory categories, and are grounded in the iconic and categorical representations that make it possible for you to pick out those sensory categories. Abstraction occurs first with the extraction of invariant sensory features that takes place in concrete perceptual category learning, and then proceeds to abstract higher-order categories by symbolic description. The grounding hypothesis in turn has some implications for how language may have begun. It suggests that language evolved along lines similar to the way it is learned: Concrete categories were assigned names because of a collective utility that shared names had for the community as a whole.

Many categories come to mind that it would have been useful to name and describe: kinfolk, tribesmen, enemies, foods, predators, weather conditions, tools, places, discomforts, dangers. Simply naming these categories and sharing and using the names would seem to be its own obvious reward under easily imaginable primitive conditions benefitting from the sharing of information (reducing uncertainty) about future contingencies: Names could at first have been shared "iconic" responses (both verbal and gestural, as suggested by the "bow-wow" [imitation] and "yo-he-ho" [motor-correlate] theories of language origin; see Harnad et al. 1976). These iconic names could then have taken the usual path to arbitrariness as their linguistic function, because of its powerful consequences, took precedence over their original imitative, instrumental and expressive origins. Name concatenation would be a natural development (as would the converse process of dismantling holophrastic expressions), particularly with object/name serving as the first model in ostension. "That `cat'" and "That `dog'" lead quite naturally to the superordinate category statement: "`Cat', `dog,': `animal'".[12]

Intertranslatability did not, I suggest, require a separate development, since translation is a lexical/lexical matter for the most part, and that is what the "internal translation" performed by the symbolic representations really is. Once there is a representational system that allows things to be reliably sorted into named categories, with their names going on to figure in (symbolic) descriptions of still more categories, their membership and their invariant features, then intertranslatability is trivial: It depends only on a shared grounding, which can always be checked and adjusted to as close an approximation as necessary for successful coherent interaction and joint operations on objects, for example, sorting them into categories. Given a "common ground" of primitive categories and the capacity to name and describe them, all the rest can be settled verbally (as in Kenneth Pike's "magic show," in which, on stage, before an audience, he learns to converse in a language he has never before encountered, by interacting with a native speaker who speaks only that language).

Could two symbolic representational systems fail to be intertranslatable? If two organisms were capable of human-scale categorization, naming and describing performance, could a Quinean indeterminacy nevertheless leave them lost in Babel-like confusion and uncertainty about whether they were really understanding one another or just talking at cross-purposes? [13] I think not. The approximateness of categorizing and describing may be a liability, but the capacity to revise and tighten the approximation as closely as necessary (as dictated by "consequences") through further ostension and discourse seems a much more powerful countervailing asset, and a universal one.

What a theory of the origin of language cannot resolve, of course, is the mind/body problem. One source of indeterminacy accordingly remains (Harnad 1993b, 1996): There is no way to know (because of the inverted-spectrum conjecture, for example) that the qualitative sensory experience in which our words are grounded is a shared experience -- or even that anyone other than oneself has any qualitative experience at all, rather than merely going mindlessly through the behavioral motions. That, however, is a distinction on which this hypothesis must remain approximate, at least until it can be shown to have differential consequences for how we sort, name and describe things and then go on to share our names and descriptions with one another through that universally glossable classificatory system we call language.

References

Agawu, V. Kofi (1991) Playing with signs: A semiotic interpretation of classic music. Princeton, NJ: Princeton University Press.

Berlin, B. & Kay, P. (1969) Basic color terms: Their universality and evolution. Berkeley: University of California Press

Bornstein, M. H. (1987) Perceptual Categories in Vision and Audition. In: Harnad (1987)

Boynton, R. M. (1979) Human color vision. New York: Holt, Rinehart, Winston

Catania, A.C. & Harnad, S. (eds.) (1988) The Selection of Behavior. The Operant Behaviorism of BF Skinner: Comments and Consequences. New York: Cambridge University Press.

Chomsky, N. (1980) Rules and representations. Behavioral and Brain Sciences 3: 1-61.

Cole, D. (1990) Functionalism and inverted spectra. Synthese 82: 207-222.

Fodor, J. A. (1975) The language of thought. New York: Thomas Y. Crowell.

Gibson, E. J. (1969) Principles of perceptual learning and development. Engelwood Cliffs NJ: Prentice Hall

Goodman, N. (1954) Fact, fiction and forecast. University of London: Athlone Press

Goodman, Nelson (1968) Languages of art: An approach to a theory of symbols. Indianapolis: Bobbs-Merrill

Hanson & Burr (1990) What connectionist models learn: Learning and Representation in connectionist networks. Behavioral and Brain Sciences 13: 471-518.

Harnad, S. (1982) Neoconstructivism: A unifying theme for the cognitive sciences. In: Language, mind and brain (T. Simon & R. Scholes, eds., Hillsdale NJ: Erlbaum), 1 - 11.

Harnad, S. (ed.) (1987) Categorical Perception: The Groundwork of Cognition. New York: Cambridge University Press.

Harnad, S. (1987a) The induction and representation of categories. In: Harnad 1987.

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

Harnad, S. (1991) Other bodies, Other minds: A machine incarnation of an old philosophical problem. Minds and Machines 1: 43-54.

Harnad, S. (1992) Connecting Object to Symbol in Modeling Cognition. In: A. Clarke and R. Lutz (Eds) Connectionism in Context Springer Verlag.

Harnad, S. (1993a) Grounding Symbols in the Analog World with Neural Nets. Think 2: 12 - 78 (Special Issue on "Connectionism versus Symbolism" D.M.W. Powers & P.A. Flach, eds.).

Harnad, S. (1993b) Turing Indistinguishability and the Blind Watchmaker. Presented at Conference on "Evolution and the Human Sciences" London School of Economics Centre for the Philosophy of the Natural and Social Sciences 24 - 26 June 1993.

Harnad, S. (1993c) Grounding Symbolic Capacity in Robotic Capacity. In: Steels, L. and R. Brooks (eds.) The "artificial life" route to "artificial intelligence." Building Situated Embodied Agents. New Haven: Lawrence Erlbaum

Harnad, S, (1996) Does the Mind Piggy-Back on Robotic and Symbolic Capacity? To appear in: H. Morowitz (ed.) "The Mind, the Brain, and Complex Adaptive Systems.

Harnad, S., Hanson, S.J. & Lubin, J. (1991) Categorical Perception and the Evolution of Supervised Learning in Neural Nets. In: Working Papers of the AAAI Spring Symposium on Machine Learning of Natural Language and Ontology (DW Powers & L Reeker, Eds.) pp. 65-74. Presented at Symposium on Symbol Grounding: Problems and Practice, Stanford University, March 1991; also reprinted as Document D91-09, Deutsches Forschungszentrum fur Kuenstliche Intelligenz GmbH Kaiserslautern FRG.

Harnad, S. Hanson, S.J. & Lubin, J. (1995) Learned Categorical Perception in Neural Nets: Implications for Symbol Grounding. In: V. Honavar & L. Uhr (eds) Symbol Processing and Connectionist Network Models in Artificial Intelligence and Cognitive Modelling: Steps Toward Principled Integration. (in press)

Harnad, S., Steklis, H. D. & Lancaster, J. B. (eds.) (1976) Origins and Evolution of Language and Speech. Annals of the New York Academy of Sciences 280.

Kuhn, T. (1970) The structure of scientific revolutions. Chicago: University of Chicago Press

Lawrence, D. H. (1950) Acquired distinctiveness of cues: II. Selective association in a constant stimulus situation. Journal of Experimental Psychology 40: 175 - 188.

Pike, Kenneth L. (1982) Linguistic concepts: An introduction to tagmemics. Lincoln: University of Nebraska Press.

Pribram, Karl H. (1971) Languages of the brain. Englewood Cliffs, N.J.: Prentice-Hall.

Pullum, G. K. (1989) The great eskimo vocabulary hoax. Natural Language and Linguistic Theory 7: 275-281.

Pullum, Geoffrey K. (1991) The great Eskimo vocabulary hoax, and other irreverent essays on the study of language. Chicago: University of Chicago Press.

Pylyshyn, Z. W. (1984) Computation and cognition. Cambridge MA: MIT/Bradford.

Quine, W. V. O. (1960) Word and object. MIT Press.

Steklis, H.D. & Harnad, S. (1976) From hand to mouth: Some critical stages in the evolution of language. In: Harnad et al. 1976, 445 - 455.

Whorf, B. L. (1956) Language, thought and reality. MIT Press.

Wittgenstein, L. (1953) Philosophical investigations. New York: Macmillan

Footnotes

[1] Behaviorism is right in at least one respect. Our experience shapes us to mark the differences that make a difference for us (Harnad 1987a). This, I take it, is the core of truth in Skinner's notion of "selection by consequences" (Catania & Harnad 1988). To put it information-theoretically: Linguistic communication can only resolve the actual uncertainties one has encountered so far, not all potential uncertainties yet to be encountered. For one thing, language and cognition can itself always *generate* new uncertainties by formulating questions and distinctions that no one has encountered or thought of (hence been "uncertain" about) before.

[2] For example, in the propositional calculus, and/not and or/not are duals, and in group theory, +/0 and x/1 are. In both cases the interpretations of all the formulas can be systematically swapped in such a way as to preserve truth values.

[3] Saul Kripke (unpublished) has pointed out that the similarity structure of the 2-dimensional color wheel would have to be gerrymandered quite radically in order to make all of our perceptual similarity judgments come out identically; hence there may be an unproved psychophysical dual conjecture (blue-green vs. green-blue) implicit here too.

[4] A related case that is often mentioned, though not the data on which it is based, is Eskimo snow terms, but unfortunately that seems to have turned out to be a Whorfian Canard too (Pullum 1991).

[5] Whether *all* are members or *some* are is simply another higher-order category-membership matter, this time having to do with the membership of the categories "some," and "all." It may be that one or the other of these, like negation and two-valued logic, has to be an innately given primitive in any categorization system, but these are formal details on which I am not expert and fortunately need not commit myself for present purposes.

[6] Note that I am not claiming that "horse," "stripes," and "zebra" are actually learned this way; only that knowledge by description must be grounded in *some* primitive concrete categories named on the

basis of direct acquaintance in this way.

[7] Chief among these is the "vanishing intersections" objection, to the effect that most categories have no common sensory properties (and perhaps no common properties at all). I will return to this in my discussion of sampled variation below; see footnote 10.

[8] Intertranslatability is what makes category learning difficult. If category invariance were not underdetermined -- if categories wore all their invariant features "on their sleeves," so to speak -- then category learning would be trivial, and so would all linguistic communication, because there would be no appreciable uncertainty to reduce. To "inform" someone would simply be to point to something he already knew and to give it an arbitrary label, for future common reference. In reality, however, categories are underdetermined (nontrivially interconfusable); hence both ostension and description are informative, because they induce us to resolve the confusion (provisionally) by revising our representations.

[9] A possible candidate for this learning algorithm may emerge from contemporary connectionistic research (Hanson & Burr 1990; Harnad 1991, 1993a,c; Harnad et al. 1991, 1995) or related work on statistics and general induction.

[10] According to the "vanishing intersection" critique, there exist no shared features for most of our categories. But then how do we nonetheless succeed in sorting their members? The assumption made in this paper is that where there is successful categorization performance, there *must* be invariant features. The grounding process, however, being recursive and potentially highly embedded, can quickly reach levels of abstraction that are quite remote from their sensory invariants, which may not even be consciously accessible. Moreover, most invariants are also provisional, the categories they pick out being based on a finite sample of interconfusable members and nonmembers. With a widening of the sample, however, it is not that the invariants vanish, but rather that the approximation becomes tighter, sometimes by replacing or revising some or all of the prior invariants, but always by subsuming them as a special case -- if, that is, successful categorization performance continues to be possible. Otherwise it is the *category* rather than the invariance that has vanished.

[11] As an example of a symbolic circle that *is* vicious, consider the hopeless task of a nonspeaker of Chinese trying to learn the meaning of Chinese from nothing but a Chinese/Chinese dictionary (Harnad 1990). (The fact that cryptographers -- say, the decipherers of cuneiform writing or the Enigma code -- seem to be able to break out of such a circle owes itself, if one thinks about it, to the fact that they already speak at least one grounded language which is intertranslatable with the target language. The rest is just the exploitation of shared statistical regularities.)

[12] Note that much of this seems explicable as operant "shaping by consequences" -- categorization and naming being, after all, operant responses -- except that the critical component, the underlying representational system that makes it all possible, is anything but behavioristic. Nevertheless, shaping by consequences no doubt played a role in the origin of language, just as it does in its development (Catania & Harnad 1988).

[13] Or worse, could a radical Kuhnian gap of incommensurability (Kuhn 1970) separate some symbolic systems from others?