Harnad, S. (2003) The Symbol Grounding Problem. *Encyclopedia of Cognitive Science*.Nature Publishing Group/Macmillan.

The Symbol Grounding Problem

Stevan Harnad
Centre de Neuroscience de la Cognition (CNC)
Universite du Quebec a Montreal
CP 8888 Succursale Centre-Ville
Montreal, Quebec, Canada H3C 3P8
harnad@uqam.ca
http://cogsci.soton.ac.uk/~harnad/

SUMMARY: The Symbol Grounding Problem is related to the problem of how words get their meanings, and of what meanings are. The problem of meaning is in turn related to the problem of consciousness, or how it is that mental states are meaningful.

KEYWORDS: categorization, computation, consciousness, language, meaning, perceptio, sensorimotor function, reference, robotics, Turing test

Words and meanings. We know since Frege that the thing that a word refers to (its referent) is not the same as its meaning. This is most clearly illustrated using the proper names of concrete individuals (but it is also true of names of kinds of things and of abstract properties): (1) "Tony Blair," (2) "the UK's current prime minister," and (3) "Cheri Blair's husband" all have the same referent, but not the same meaning.

Some have suggested that the meaning of a (referring) word is the rule or features one must use in order to pick out its referent. In that respect, (2) and (3) come closer to wearing their meanings on their sleeves, because they seem to be explicitly stating a rule for picking out their referents (find whoever is the UK's current PM, or whoever is Cheri's current husband). But that does not settle the matter, because there's still the problem of the meaning of the components of the rule ("UK," "current," "PM," "Cheri," "husband"), and how to pick *them* out.

Perhaps "Tony Blair" (or better still, just "Tony") does not have this component problem, because it points straight to its referent, but how? If the meaning is the rule for picking out the referent, what is that rule, when we come down to non-decomposable components?

It is probably unreasonable to expect us to know the rule, explicitly at least. Our brains need to have the "know-how" to follow the rule, and actually pick out the intended referent, but they need not know how they do it consciously. We can leave it to cognitive science and neuroscience to find out and then explain how.

The means of picking out referents. So if we take a word's meaning to be the means of picking out its referent, then meanings are in our brains. If we use "meaning" in a wider sense, we may want to say that meanings include both the referents themselves and the means of picking them out. So if a word (say, "Tony–Blair") is located inside an entity, then its meaning consists of both the means that that entity uses to pick out its referent, and the referent itself: a big causal nexus between a head, a word inside it, an object outside it, and whatever "processing" is required to connect the inner word to the outer object.

But what if the "entity" in which a word is located is not a head but a piece of paper? What is its meaning then? Surely all the (referring) words on this page, for example, have meanings, just as they have referents.

Consciousness. Here is where the problem of consciousness rears its head. For there would be no connection at all between scratches on paper and any intended referents if there were no minds mediating those intentions, via their internal means of picking out those referents.

So the meaning of a word in a page is "ungrounded," whereas the meaning of a word in a head is "grounded" (by the means that cognitive neuroscience will eventually reveal to us), and thereby mediates between the word on the page and its referent.

Computation. What about the meaning of a word inside a computer? Is it like the word on the page or like the word in the head? This is where the Symbol Grounding Problem comes in. Is a dynamic process transpiring in a computer more like the static paper page, or more like another dynamical system, the brain?

There is a school of thought according to which the computer is more like the brain –– or rather, the brain is more like the computer: According to this view, called "computationalism," that future theory about how the brain picks out its referents, the theory that cognitive neuroscience will eventually arrive at, will be a purely computational one (Pylyshyn 1984). A computational theory is a theory at the software level; it is essentially a computer program. And software is "implementation–independent." That means that whatever it is that a program is doing, it will do the same thing no matter what hardware it is executed on. The physical details of the implementation are irrelevant to the computation; any hardware that can run the computation will do.

The Turing Test. A computer can execute any computation. Hence once computationalism finds the right computer program, the same one that our brain is running when there is meaning transpiring in our heads, then meaning will be transpiring in that computer too.

How will we know that we have the right computer program? It will have to be able to pass the Turing Test (TT) (Turing 1950). That means it will have to be capable of corresponding with any human being for a lifetime as a pen–pal, without ever being in any way distinguishable from a real pen–pal.

Searle's Chinese Room Argument. It was in order to show that computationalism is incorrect that Searle (1980) formulated his celebrated "Chinese Room Argument," in which he pointed out

that if the Turing Test were conducted in Chinese, then he himself, Searle (who does not understand Chinese), could execute the same program that the computer was executing without knowing what any of the words he was processing meant. So if there's no meaning going on inside him when he is implementing the program, there's no meaning going on inside the computer when it is the one implementing the program either, computation being implementation–independent.

How does Searle know that there is no meaning going on when he is executing the TT–passing program? Exactly the same way he knows whether there is or is not meaning going on inside his head under any other conditions: He understands the words of English, whereas the Chinese symbols that he is manipulating according to the program's rules mean nothing to him. And there is no one else in there for them to mean anything to. They are like the ungrounded words on a page, not the grounded words in a head.

Note that in pointing out that the Chinese words would be meaningless to him under those conditions, Searle has appealed to consciousness. Otherwise one could argue that there *would* be meaning going on in his head under those conditions, but he would simply not be aware of it. This is called the "System Reply," and Searle rightly rejects it as simply a reiteration, in the face of negative evidence, of the very thesis that is on trial in his thought–experiment: Are words in a running computation like the ungrounded words on a page, meaningless without the mediation of brains, or are they like the grounded words in brains?

In this either/or question, the (still undefined) word "ungrounded" has implicitly relied on the difference between inert words on a page and consciously meaningful words in our heads. And Searle is reminding us that under these conditions (the Chinese TT), the words in his head would not be consciously meaningful, hence they would still be as ungrounded as the inert words on a page.

So if Searle is right, that (1) both the words on a page and those in any running computer–program (including a TT–passing computer program) are meaningless in and of themselves, and hence that (2) whatever it is that the brain is doing to generate meaning, it can't be just implementation–independent computation, then what *is* the brain doing to generate meaning (Harnad 2001a)?

Formal Symbols. To answer this question we have to formulate the symbol grounding problem (Harnad 1990):

First we have to define "symbol": A symbol is any object that is part of a symbol system. (The notion of symbol in isolation is not a useful one.) A symbol system is a set of symbols and rules for manipulating them on the basis of their shapes (not their meanings). The symbols are systematically interpretable as having meanings, but their shape is arbitrary in relation to their meaning.

A numeral is as good an example as any: Numerals (e.g., "1," "2," "3,") are part of a symbol system (arithmetic) consisting of formal rules for combining them into well formed strings. "2" means what we mean by "two", but its shape in no way resembles "two–ness." The symbol

system is systematically interpretable as making true statements about numbers (e.g. "1 + 1 = 2").

It is critical to understand that the symbol–manipulation rules are based on shape rather than meaning (the symbols are treated as primitive and undefined, insofar as the rules are concerned), yet the symbols and their rule–based combinations are all meaningfully interpretable. It should be evident in the case of formal arithmetic, that although the symbols make sense, that sense is in our heads and not in the symbol system. The numerals in a running desk calculator are as meaningless as the numerals on a page of hand–calculations. Only in our minds do they take on meaning (Harnad 1994).

This is not to deprecate the property of systematic interpretability: We select and design formal symbol systems (algorithms) precisely because we want to know and use their systematic properties; the systematic correspondence between scratches on paper and quantities in the universe is a remarkable and extremely powerful property. But it is not the same as meaning, which is a property of certain things going on in our heads.

Natural Language and the Language of Thought. Another symbol system is natural language. On paper, or in a computer, it too is just a formal symbol system, manipulable by rules based on the arbitrary shapes of words. In the brain, meaningless strings of squiggles become meaningful thoughts. I am not going to be able to say what had to be added in the brain to make them meaningful, but I will suggest one property, and point to a second.

One property that the symbols on static paper or even in a dynamic computer lack that symbols in a brain possess is the capacity to pick out their referents. This is what we were discussing earlier, and it is what the hitherto undefined term "grounding" refers to. A symbol system alone, whether static or dynamic, cannot have this capacity, because picking out referents is not just a computational property; it is a dynamical (implementation–*de*pendent) property.

To be grounded, the symbol system would have to be augmented with nonsymbolic, sensorimotor capacities –– the capacity to interact autonomously with that world of objects, events, properties and states that its symbols are systematically interpretable (by us) as referring to. It would have to be able to pick out the referents of its symbols, and its sensorimotor interactions with the world would have to fit coherently with the symbols' interpretations.

The symbols, in other words, need to be connected directly to (i.e., grounded in) their referents; the connection must not be dependent only on the connections made by the brains of external interpreters like us. The symbol system alone, without this capacity for direct grounding, is not a viable candidate for being whatever it is that is really going on in our brains (Cangelosi & Harnad 2001).

Robotics. The necessity of groundedness, in other words, takes us from the level of the pen–pal Turing Test, which is purely symbolic (computational), to the robotic Turing Test, which is hybrid symbolic/sensorimotor (Harnad 2000). Meaning is grounded in the robotic capacity to detect, identify, and act upon the things that words and sentences refer to (see entry for **Categorical Perception**).

But if groundedness is a necessary condition for meaning, is it a sufficient one? Not necessarily, for it is possible that even a robot that could pass the Turing Test, "living" amongst the rest of us indistinguishably for a lifetime, would fail to have in its head what Searle has in his: It could be a Zombie, with no one home, feeling feelings, meaning meanings.

And that's the second property, consciousness, toward which I wish merely to point, rather than to suggest what functional capacities it must correspond to (I have no idea what those might be -- I rather think it is impossible for consciousness to have any independent functional role except on pain of telekinetic dualism). Maybe robotic TT capacity is enough to guarantee it, maybe not. In any case, there is no way we can hope to be any the wiser (Harnad 2001b).

REFERENCES

Cangelosi, A. & Harnad, S. (2001) The Adaptive Advantage of Symbolic Theft Over Sensorimotor Toil:Grounding Language in Perceptual Categories. Evolution of Communication 4(1) 117–142. http://cogprints.soton.ac.uk/documents/disk0/00/00/20/36/index.html

Cangelosi, A.; Greco, A.; Harnad, S. From robotic toil to symbolic theft: grounding transfer from entry–level to higher–level categories. Connection Science, June 2000, vol.12, (no.2):143–62. http://cogprints.soton.ac.uk/documents/disk0/00/00/16/47/index.html

Frege, G. (1952/1892). On sense and reference. In P. Geach and M. Black, Eds., Translations of the Philosophical Writings of Gottlob Frege. Oxford: Blackwell

Harnad, S. (1990) The Symbol Grounding Problem. Physica D 42: 335–346. http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad90.sgproblem.html

Harnad, S. (1994) Computation Is Just Interpretable Symbol Manipulation: Cognition Isn't. Special Issue on "What Is Computation" Minds and Machines 4:379–390 http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad94.computation.cognition.html

Harnad, S. (2000) Minds, Machines and Turing: The Indistinguishability of Indistinguishables, Journal of Logic, Language, and Information 9(4): 425–445. (special issue on "Alan Turing and Artificial Intelligence") http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad00.turing.html

Harnad, S. (2001a) What's Wrong and Right About Searle's Chinese Room Argument? In: M. Bishop & J. Preston (eds.) Essays on Searle's Chinese Room Argument. Oxford University Press. h ttp://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad00.searle.html

Harnad, S. (2001b) No Easy Way Out. The Sciences 41(2) 36–42. http://cogprints.soton.ac.uk/documents/disk0/00/00/16/24/index.html

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

Searle, John. R. (1980) Minds, brains, and programs. Behavioral and Brain Sciences 3 (3): 417–457 http://www.bbsonline.org/documents/a/00/00/04/84/index.html

Turing, A.M. (1950) Computing Machinery and Intelligence. Mind 49 433–460 [Reprinted in Minds and machines. A. Anderson (ed.), Engelwood Cliffs NJ: Prentice Hall, 1964.]
http://cogprints.ecs.soton.ac.uk/archive/00000499/

FURTHER READING

Barsalou LW Perceptual symbol systems BEHAV BRAIN SCI 22: (4) 577–+ AUG 1999

Bartell, B. & Cottrell, G. W. (1991). A model of symbol grounding in a temporal environment. In *AAAI Spring Symposium Workshop on Connectionism and Natural Language Processing*: 142–147.

Chris Malcolm and Tim Smithers. Symbol grounding via a hybrid architecture in an autonomous assembly system. In Pattie Maes, editor, Designing Autonomous Agents, pages 145––168. MIT Press, 1990.

Cummins, Robert (1996). Why There Is No Symbol Grounding Problem, Chapter 9 of: Representations, Targets, and Attitudes. Mit Press.

Freeman, W.J. A neurobiological interpretation of semiotics: meaning, representation, and information. Information Sciences, May 2000, vol.124, (no.1–4):93–102.

Glenberg, AM; Robertson, DA. Symbol grounding and meaning: A comparison of high–dimensional and embodied theories of meaning. JOURNAL OF MEMORY AND LANGUAGE, 2000 OCT, V43 N3:379–401.

Grumbach, A. Grounding symbols into perceptions. Artificial Intelligence Review, April 1996, vol.10, (no.1–2):131–46.

Jackson SA, Sharkey NE Grounding computational engines ARTIF INTELL REV 10: (1–2) 65–82 APR 1996

Jackson, S.A.; Sharkey, N.E. Grounding computational engines. Artificial Intelligence Review, April 1996, vol.10, (no.1–2):65–82.

Jung, D.; Zelinsky, A. Grounded symbolic communication between heterogeneous cooperating robots. Autonomous Robots, June 2000, vol.8, (no.3):269–92.

Jung, D.; Zelinsky, A. Grounded symbolic communication between heterogeneous cooperating robots. Autonomous Robots, June 2000, vol.8, (no.3):269–92.

MacDorman, K. F. (1998). Feature learning, multiresolution analysis, and symbol grounding. A peer commentary on Schyns, Goldstone, and Thibaut's 'The development of features in object concepts'. Behavioral and Brain Sciences.

MacDorman, KF. Feature learning, multiresolution analysis, and symbol grounding. BEHAVIORAL AND BRAIN SCIENCES, 1998 FEB, V21 N1:32+. Pub type: Editorial.

McKevitt, P. From Chinese rooms to Irish rooms: new words on visions for language. Artificial Intelligence Review, April 1996, vol.10, (no.1–2):49–63.

Plunkett, Kim; Sinha, Chris; Moller, Martin F.; Strandsby, Ole. Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. Connection Science: Journal of Neural Computing, Artificial Intelligence & Cognitive Research, 1992, v4 (n3–4):293–312.

Prem, E. (1995). Symbol grounding and transcendental logic. In Niklasson L. & Boden M. (eds.), Current Trends in Connectionism, Lawrence Erlbaum, Hillsdale, NJ, pp. 271–282.

Sun, Ron. Symbol grounding: A new look at an old idea. Philosophical Psychology, 2000 Jun, v13 (n2):149–172.

Takeuchi, I.; Furuhashi, T. Acquisition of manipulative grounded symbols for integration of symbolic processing and stimulus–reaction type parallel processing. Advanced Robotics, 1998, vol.12, (no.3):271–87.

Tani, J. (1996) Does Dynamics Solve the Symbol Grounding Problem of Robots? An Experiment in Navigation Learning. Learning in Robots and Animals – Working Notes. AISB'96 workshop, Brighton, UK.

Thompson E. Symbol grounding: a bridge from artificial life, to artificial intelligence. Brain and Cognition, 1997 Jun, 34(1):48–71.

Thompson, E. Symbol grounding: A bridge from artificial life to artificial intelligence. BRAIN AND COGNITION, 1997 JUN, V34 N1:48–71.