Words, Symbols, and Symbol Grounding

Georg Dorffner, Erich Prem, Harald Trost
Dept. of Medical Cybernetics and Artificial Intelligence
University of Vienna
and
Austrian Research Institute for Artificial Intelligence

Abstract

In this paper we present a definition of 'symbol' in cognitive science which is designed to clear some obvious misunderstandings in discussions around "symbolic" vs. "sub-symbolic" approaches. We discuss this definition in the light of three different frames of reference (i.e. three different views, namely the intelligent agent's, an observer's, and a meta-observer's). Then we show the implications of these views for cognitive science and artificial intelligence (AI) and discuss whether the most conspicous "symbols" in cognition – words in a language – can fulfill the ideals behind their definition.

1 Introduction

In cognitive science, many fundamental discussions have centered around the notion of 'symbol' and its role in cognitive modeling. Recently a dichotomy of paradigms, often labeled 'symbolic' vs. 'subsymbolic,' has arisen, mainly due to the upsurge of connectionist theory. In the subsequent philosophical exchange about the epistemological bases of both paradigms, however, arguments sometimes seem to have lost sight of why and where the notion 'symbol' has entered cognitive science in the first place.

In this paper we first give an overview over the discussions around symbols and present our own solutions to solving some obvious misunderstandings. In particular, we suggest that in cognitive science it makes best sense to look at a 'symbol' in its semiotic definition of being a sign. In cognitive modeling – as we want to argue – we need only look at symbols that obtain their so-called "referential link" from the intelligent agent to be modeled, i.e. something that is a symbol *with respect to that agent.* In this definition, symbols are signs and symbolic processing is behavior involving signs. The most important form of this is language, meaning that cognitive science should primarily be interested in studying words (or better: morphemes) when studying symbols.

An important aspect of symbols is their *arbitrariness.* In particular, there is no inherent relationship between symbol and reference, and slight changes of the symbol's shape do not lead to slight changes of its reference. Going even further, arbitrariness entails that *any* possible sign can stand for any possible reference (an aspect most prevalent in traditional artificial intelligence). In the second part of this talk, we will discuss to what degree words of a language can fulfill this requirement of being arbitrary. It will turn out that they are rather far away from being truly arbitrary signs, yet they still belong to the observations closest to pure symbols which we can find in “everyday” human cognition.

2 The notion ‘symbol’ in cognitive science and AI

The notion ‘symbol’ has played an important role in recent discussions about different paradigms of cognitive science or artificial intelligence (AI), especially when it comes to distinguish so-called “symbolic” from “non-symbolic” or “sub-symbolic” approaches. One obvious reason for the importance of this notion was the formulation of the ***physical symbol systems hypothesis (PSSH)*** by [15]. There, a definition of intelligence was inextricably based on “physical symbols” and a system’s ability to represent and manipulate them:

> PSSH: A phsyical symbol system has the necessary and sufficient means for exhibiting intelligence.

Alongside with that, the most prevalent programming languages in AI are those that are said to be suitable for “symbol manipulation,” such as LISP or Prolog. It is worth taking a closer look at what ‘symbol’ then means in an implementation of an AI system. In all cases a symbol is a string of characters (itself represented by electrical states in the computer) with the following properties (see also [8]):

- a symbol is **discrete**, i.e. neither at the level of electric states nor at its common depiction as a string of characters is there a continuum between two symbols that can systematically be exploited.
- a symbol is assumed to **represent** objects or states in the world (their referents).
- a symbol is arbitrary, i.e. there is no similarity relation defined between different symbols such that it would map to a similarity relation between the referents. Furthermore, any given string can be used to represent any given object or state in the world. In a LISP program, for instance, it does not matter whether one uses the string ‘APPLE’ or ‘X*0001’ to represent the class of all apples. Only our familiarity with the commonly used strings that resemble words leads us to believe that there are non-arbitrary relations between symbol and referent (this phenomenon is refered to as “hermeneutic hall of mirrors” by [7]).
- “atomic” symbols can be combined to more complex ***symbol structures*** by the process of **concatenation** to represent complex relationshsips between different objects and states in the world.

- symbols are processed by algorithms which depend merely on a **symbols' shape**. For instance, an algorithm might contain a rule looking for the string 'APPLE' in a symbol structure. This will work only if the exact string 'APPLE' is found and causes no effect at the encounter of, say, 'APLE' (sic!) or 'PEAR'.

- all symbol structures must be **systematically semantically interpretable**. If a symbol like 'APPLE' is interpreted as representing the class of apples in one context (in one symbol structure containing 'APPLE') then it must be thus interpretable in all contexts (in all symbol structures containing 'APPLE'), and the same must be true for any concatentaion of two symbols as representing a relationship between objects or states in the world.

A famous criticism of AI based on this notion of 'symbol' is the Chinese Room Analogy ([18, 19]), which will not be repeated here. A consequental notion has been put forward by [8], called the "symbol grounding problem" by suggesting that processes working solely on the shape of arbitrary symbols do not reflect the way living beings deal with symbols, where the main difference is a symbol system's lack of "grounding" in non-arbitrary, non-discrete bodily experiences.

3 Sources for misunderstandings

The above definitions, taken by themselves, do not cause an epistemological problem on the cognitive scientist's part when dealing with "symbolic" AI. A firm assumption by those who believe in the necessity and sufficiency of symbols as defined above for explaining or simulating intelligence stands vis a vis some criticism (like that of Searle and Harnad) which suggests a deep problem in the processing of symbols lacking grounding (or meaning for the system itself).

Matters become more subtle when notions like 'sub-symbolic' or 'non-symbolic' enter the discussions, most prominently when connectionist models are proposed. The question that arises is "What is the main difference between a 'symbolic' and a 'sub-symbolic' approach, and what does connectionism (or other paradigm) have to say about the necessity (or sufficiency) of symbols?" In recent literature there seems to be some deep confusions as to what the answer to this question should be. [20], for instance, in his seminal paper trying to clear exactly this matter, speaks about a sub-symbolic systems as follows

> In the symbolic paradigm, the context of a symbol is manifest around it and consists of other symbols; in the subsymbolic paradigm, the context of a symbol is manifest inside it and consists of subsymbols. ([20], pp. 17)

He refers to the fact that connectionist activation states are ***not*** arbitrary and discrete in that they form a continuum with systematic relations between different, neighboring or distant, states. Nevertheless, he speaks of 'symbols'. Some time earlier, [9], who was one of the first to coin the term 'sub-symbolic', uses a similar diction when speaking of *active symbols* in a sub-symbolic state space (in one of his many analogies, he likens this to teams of ants forming in an army of millions of single ants) – see also [11]. This suggests a picture where an intelligent system still consists of symbols, but in which they arise from concerted actions at a lower layer of non-symbolic elements. But are these "symbols" still arbitrary or discrete, or do they represent?

Along a different vein we find connectionist work like that by [21]. In their "connectionist production system" arbitrary, discrete tokens (usually represented as strings, as explained above) are represented by connectionist activation states. Thus, even though the resulting network would be capable of implementing continuous state spaces like in [20], the only states that are meaningful are the discrete states that represent the discrete tokens. Since the tokens are assumed to be arbitrary, the representing connectionist states are also. For this the authors even must introduce a certain mechanism to avoid the non-arbitrary (i.e. similarity-based) connectionist processes (in particular, they carefully design weights and threshold such that overlaps between representations do not have an effect on their processing). One of the reasons behind devising such models obviously is to show to the "symbolic" community the capabilities of connectionist models to exhibit the same processes as symbol systems. Is, then, a sub-symbolic approach merely an alternative implementation of a symbols system? Or, better still, should it be?

The most radical view – often overheard in discussions about different AI paradigms – denying differences between "symbolic" and "sub-symbolic" approaches with respect to the notion of 'symbol' is the observation that any neural network can be implemented on a Turing machine, and since Turing machines are the most general symbol processors, they are by definition symbolic. Are we back to the PSSH, recognizing that neural networks in fact do fulfill its premises?

4 Our approach

In this paper, we want to suggest a view and a definition of 'symbol' that makes clear some essential differences between the two paradigms of AI. We do this not to contribute to a terminological war, but to make visible some essential aspects and to define the basis for our own approach of thinking about AI.

Before giving a definition for 'symbol' we take a brief look at how we humans are usually attributed to being "symbol processors." This sheds some light as to why and how the notion 'symbol' has entered discussions in AI and cognitive science to begin with. In other words, why should cognitive scientists be interested in symbols? Obviously there are two main aspects about human cognition that could be the reason for such an interest. First, when thinking or acting humans think or act in *categories* or *concepts*. In other words, compared to the richness of sensory stimuli our cognition is largely based on conceptions on a high level of *abstraction* or *reduction* (meaning that a large number of different stimuli are treated the same in a certain situation). For example, we usually do not react to all kinds of instances, views, and shades of apples differently, but we usually see them all as the category *apple* and base our (re-)actions on that category, such as grabbing and eating one. This does not mean that we are not able to see the differences in different situations. To the contrary, we do have representations that reflect those differences ([7] calls them *iconic*). But in high-level cognition, especially in language, the reduced and abstract category plays an important role. This is one thing that is often refered to as 'symbolic thinking (or acting)'.

Secondly, humans are able to take arbitrary signs and manipulate them no matter what they mean. [10], for the sake of discussion such abilites, lists the following example

> Consider the phrase "love ever keeps trying." Proceeding from left to right, mentally [without looking at the phrase anymore] extract the second letter from each word, and concatenate these letters in sequence. If the resulting string forms an English word, make a note of it ([10], p. 580).

Mathematics is another example, where (some) humans are able to manipulate 'x's and 'y's without any knowledge of what they refer to, merely based on their shape and their concatenations in formulas. This bears a large resemblance to the manipulation of arbitrary symbols in AI and is thus often also called 'symbol processing' (such as in [10]).

The difference between these two observations, as we want to argue, is essential and leads us to a definition where we distinguish between two entities: ***Concepts*** on one hand, and ***Symbols*** on the other.

- A ***concept*** is a categorical mental state that is formed based on an adaptive categorization process working on invariances in stimuli and situations.

- A ***symbol*** is a label or ***sign*** used by an intelligent agent to refer to one of its concepts.

Thus we introduce 'symbol' largely as it is defined in semiotics (as opposed to icons and indices, e.g. in [17, 5]). From there it inherits three important properties:

- A symbol as a sign is **discrete** (i.e. there is no continuum of symbols when they are actually used as signs)

- A symbol as a sign **refers** to an object or state in the world (i.e. at least one intelligent being uses the sign to stand for an object or state).

- A symbol as a sign is **arbitrary** (i.e. there is no inherent relationship between the symbols and the referents, any symbol can be made to stand for any object or state in the world).

Note that this laregly coincides with three of the properties of 'symbols' in AI, when one replaces 'represent' by 'refers to'.

The confusions cited above come from using the word 'symbol' for both ***symbol*** in our definition and *concept*, without recognizing their difference. This difference can be made most visible when evaluating the three above properties in the example of the concept ***apple*** (i.e. the categorical mental state one activates when cognizing about an apple) and the word (symbol, sign) 'apple':

- Arbitrariness:
  The word 'apple' is (to a certain degree, see later) arbitrary in that we could use 'pomme' or 'Apfel' or 'karakura' instead to refer to the category of all apples (we just need to learn it this way).
  The concept ***apple*** is ***not*** arbitrary, since this mental state is inextricably tied to our other mental states, most prominently to our bodily experiences with apples. We do not possess the power to choose a different mental state to fulfil the same role (other than by applying a label again).

- Reference:
  The word 'apple' refers to the category of apples by virtue of at least one human to act accordingly.
  The concept ***apple*** does ***not*** refer to anything, it is only causally tied to stimuli or situations. It would take an external observer to identify this causal relationship between states in the world and mental states and thus to identify 'reference.' For the intelligent agent itself that state ***is*** the object, not a reference to it (see also later).

- Discreteness:
  The word 'apple' is discrete in that in its immediate (e.g. acoustic) neighborhood their is no other meaningful word (such as 'apple' uttered with a higher tone or with an 'a' that is half-way between [a] and [e]. This is different for different languages, but always applies in this or a similar way within a language).
  The concept ***apple*** is somewhat discrete, but – especially as many connectionists suggest – non-discrete aspects play a role (e.g. in that we recognize ***apples with some resemblance to pears*** or the like).

In summary we see that, except for discreteness, the two entities do not share the same properties. What we have called a 'symbol' (the label or word) is most similar to what is called 'symbol' in semiotics, while what we call 'concept' is not.

5 Three views of the same thing

Before we discuss the implications of these definitions on AI and the discussions around different paradigms, we need to make things a little clearer. By defining 'symbol' and 'concept' like above there is still a large source for misunderstandings, mainly because it depends on the view or ***frame of reference*** ([1]) one is applying when describing cognition in the way we intend to. We highlight three views using figures depicting the world (with its physical states) and the intelligent agent (figure 1).

The first frame of reference is the agent's own. There are certain physical states in the world identifiable as objects (including the agent's own mental states), and other physical states identifiable as signs. A symbol is a sign with the above three properties and, for the agent, refers to the object. There is no intermediate ***concept*** in this view, since for the agent the concept is identical with the real object (it is its conception as being real).

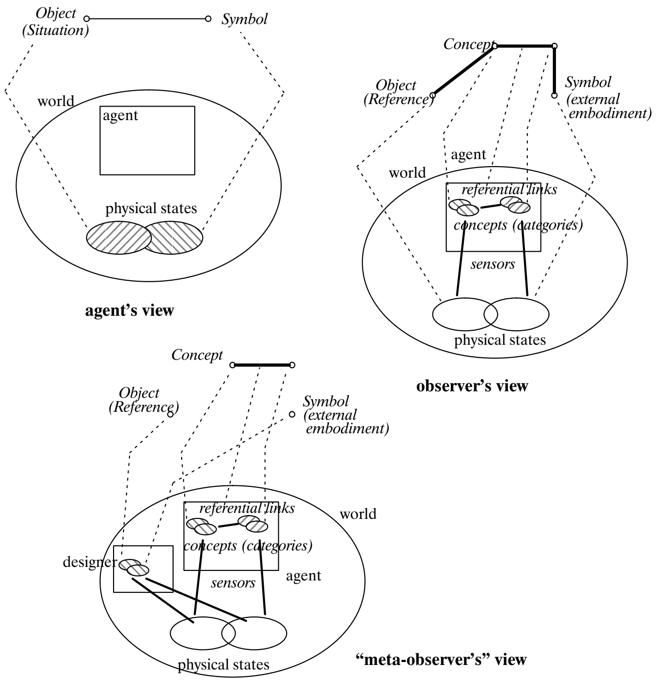

Figure 1: Three views (frames of reference) of symbols and concepts

The second frame of reference is that of an observer who objectively can take a look inside the agent's mental states. In AI this would usually be the designer of an artificial system. In this view, there are physical states which *for the agent* are recognizable as objects and others which *for the agent* are recognizable as signs. Recognition comes from categorizing stimuli (or, more generally, situations including prior mental states) via sensors (or via internal pathways). This is true for both categorizing objects or states into concepts *and* categorizing other physical objects into signs. We see that, when speaking about 'symbols' we have to distinguish between the actual physical object – the *external embodiment* – and the categorical mental state from recognizing that object as a separate entity (in the same way we recognize objects that are not signs). Neither of the two alone are symbols – external embodiments are physical states like any other phsyical state, while the resulting categorical mental states have the same properties as concepts. Thus, in order to introduce a symbol it takes *both* entities *plus* mental links – *referential links* – forming the connection between the sign and the refered concept. From this we can conclude that reference is given as a link between the categorical mental state of the sign and one of the concepts. To identify this link we (as the observer) need to take the intelligent agent into account. Thus reference cannot be defined without including the agent who implements the link. As a result, signs bear no relation to objects, other than through the intermediate categorization and linking by an agent. This roughly corresponds to the well-known ***triangle of meaning*** ([16]), the only difference being the introduction of a fourth intermediate element (the categorical mental state from recognizing the external embodiment as an object).

The third frame of reference could be called a "meta-observer's". In this view we recognize that the previous view was distorted through the fact that the observer itself brings in its own coception of the world when depicting the situation. In other words, in order to draw a line between 'object' and 'concept' on one side, and between 'external embodiment' and its categorical mental state on the other, it takes the observer to identify the object and the external embodiment. For an AI designer this would imply the explicit design of concepts such that they correspond to his or her own concepts. This, however, is a too restrictive view. The new view suggests that both concepts and the recognition of signs are subjective with respect to an agent and thus dependent on its experiences and previous history. A connection between the external and internal parts of the "triangle" can only be drawn if the agent's concepts coincide enough with the observer's, otherwise they cannot. Of course this reflects the situation between any two humans in our physical environment. Each individual is left to themselves to categorize and form concepts and learn to use signs. It is only through a sufficient overlap in the reactions to concepts and the use of signs that different individuals can interact or communicate with each other. This overlap, of course, is increased through adaptive learning during interaction and communication, but can never reach the theoretical maximum of the concepts of both individuals being identical. This view largely corresponds to the suggestions of constructivism ([14], [6]).

6 Consequences for cognitive science and AI

We claim that our definition, together with the three frames of reference, clears up the matters and confusions around "symbolic" and "sub-symbolic" AI. First of all, we see that the primary interest of cognitive science and AI should be in explaining the formation of concepts and the actions based thereon (compare [12, 13]. Only when it comes to the use of signs (and – as will be discussed a little further below – this is mainly the case when it comes to language) we need be interested in symbols. Now we have seen that only 'symbols' in our diction share the main properties that are prevalent in classical AI – mainly discreteness, arbitrariness, and reference (representation) – while concepts do not. In particular, we see that

- discreteness of symbols can be found in the fact that categorization of external embodiments leads to largely distinct mental states.

- arbitrariness of symbols can be found in the nature of referential links which permit similarity-insensitive connections.

- reference comes about through the active formation of referential links by the agent, independent of any observer or designer.

while all this (with the exception of some discreteness) cannot be observed for concepts. In classical AI the nature of our own signs when speaking about our concepts has been taken and projected onto the internal nature of concepts – an influential but wrong (or at least unsubstantiated) move according to this view. Furthermore (and this is how Searle's critique can be seen in this view) the referential links of classical AI symbols cannot be found in the AI system itself but rather in the ***designer's*** mind (he or she is the one assigning meaning to the symbol). Thus classical AI systems are based on "implanted" external embodiments (or representations of the resulting categorical states, since the categorization step is bypassed) which cannot be called symbols with respect to those systems – a rather paradoxical but enlightening result. Similar things happen in many connectionist models, especially in those like [21] that reimplement the classical picture.

The alternative approach which is suggested by our view is a paradigm which tries to account for the third frame of reference above: It should provide for means of modeling categorization (concept formation) based on rich stimuli leading to reduced but not necessarily fully discrete states, and for means of forming referential links which support the arbitrariness observed in symbols as signs. It should model this process without any further reliance on the model's designer, but solely based on the system's own experiences and adaptivity. A radical version of connectionism (e.g. [2], which is not the most radical conceivable version since it presupposes the design of components like concept formation and referential links) is one candidate for the basis of such a modeling paradigm. It could be called 'sub-conceptual' in that its buidling blocks are below the level of concepts which emerge from concerted actions of such building blocks. It could also be called 'non-symbolic' since the building blocks are not arbitrary tokens (external embodiments) as in classical AI. The term 'sub-symbolic' is too misleading in this view and is therefore not used here.

Symbol grounding in this approach becomes primarily ***concept grounding*** (i.e. the modeling of the formation of concepts grounded in experience). Since "symbols" have been identfied as consisting of the external embodiment, its categorical mental state ***and*** the referential link to a concept, they will not exist but grounded through the concepts' grounding (for examples of symbol grounding in this approach see [3, 4]).

One might ask of how combination of atomic elements (in AI through concatenation) is or should still be possible. We see that only external embodiments can be concatenated. Mental states like concepts must recombine taking their intricate relationships to other mental states into account. Of course, a radical connectionist (or other) approach must account for complex conceptual structures, but it will not be as systematic as combining external symbol tokens. Exactly how this could be realized is certainly still an open question. The question of systematic semantic interpretability no longer has to be asked, since symbols in this view no longer exist without the interpretation through an individual.

7 Can we preserve the ideals of AI symbols?

As we have seen in the above example of ***apple*** vs. ‘apple’, language is the most conspicuous place of where we can find what we have identified as symbols in cognition (arbitrary signs). Thus, in everyday cognition, words (or better still: morphemes) are symbols. The final question is whether these symbols still correspond to an AI or even semiotic ideal. In other words, do the three main properties still hold up? The answer is, only to a certain limited extent.

- Discreteness:
  In general it is true that there is no meaningful continuum between words (morphemes), but several aspects of language, such as prosody, do have analog character.
- Arbitrariness:
  Words (morphemes) are far from being fully arbitrary. If they were, any string or acoustic signal could stand for anything, and there would be no similarity between word and meaning. However, words must be pronouncable, usually follow certain additional articulatory constraints, and are limited in length. They cannot stand for any arbitrary mental state but only states that are a clear result from categorization. And onomatopoetics (the sound similarities in words like ‘splash’ and ‘spray’) shows that words can sometimes even express subcategories or share features with what they stand for (iconic prperties of words).
- Reference:
  Words do not always refer, some never do. Examples are grammatical words and language utterances as social action.

From all this we can conclude that words (morphemes) of a language are relatively bad examples of symbols, at least when these three properties are as highly valued as they were in classical AI. Other domains would be the performance of symbolic mathematics or logic, where these ideals are met a little better. However, these are exceptional cases not usually occuring in “every-day” cognition, and not even there symbols are as perfect as AI suggested.

In conclusion we can say that cognitive science and AI should primarily be interested in modeling conceptualization and cognition based on concepts. These concepts are somewhat discrete mental states but ***not*** arbitrary and do not refer. The closest thing to “symbols” in general cognition are words (morphemes) of a language, even though they do not meet an ideal definition of ‘symbol’. They should therefore be modeled as such.

Acknowledgments

We thank Prof. Trappl for his support of this research, and several reviewers of previous versions of this paper for pointing out its shortcomings. The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry for Science and Research.

References

[1] Clancey W.J. (1989): The Frame of Reference Problem in Cognitive Modeling, in *Proceedings of the Eleventh Anual Conference of the Cognitive Science Society*, Lawrence Erlbaum, Hillsdale, NJ, pp.107-114.

[2] Dorffner G.(1991): "Radical" Connectionism for Natural Language Processing, *Working Notes of the AAAI Symposium on Connectionist Natural Language Processing*; also: Austrian Research Institute for Artificial Intelligence, Report TR-91-07.

[3] Dorffner G. (1992): Taxonomies and Part-Whole Hierarchies in the Acquisition of Word Meaning – A Connectionist Model, in: *Proc. of the Annual Conf. of the Cognitive Science Society*, Erlbaum.

[4] Dorffner G., Prem E. (1993): Connectionism, Symbol Grounding, and Autonomous Agents, *Proc. of the 15th Ann.Meeting of the Cognitive Science Society*, Boulder, CO, pp. 144-148.

[5] Eco U. (1973): *Segno.* (*Zeichen.* Suhrkamp, Frankfurt am Main, 1977).

[6] Glasersfeld E.von(1987): *Wissen, Sprache und Wirklichkeit*, Vieweg, Braunschweig.

[7] Harnad S. (1990): Lost in the Hermeneutic Hall of Mirrors, *J. of Experimental and Theoretical AI*, 2(1990)pp.321-327.

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

[9] Hofstadter D.R. (1985): Waking up from the Boolean Dream, in Hofstadter D.R.(ed.), *Metamagical Themas: Questing for the Essence of Mind and Pattern*, Basic Books, New York.

[10] Hadley R.F. (1990): Connectionism, Rule Following, and Symbolic Manipulation, in *Proceedings of the Eighth National Conference on Artificial Intelligence*(AAAI-90), AAAI Press/MIT Press, Menlo Park, pp.579-586.

[11] Kaplan S., Weaver M., French R.(1990): Active Symbols and Internal Models: Towards a Cognitive Connectionism, *AI & Society*, 1(4)51-72.

[12] Lakoff G. (1987): *Woman, Fire and Dangerous Things.* University of Chicago Press, Chicago.

[13] Lakoff G.(1989): Some Empirical Results about the Nature of Concepts. *Mind and Language*, Vol.4 Nos.1 and 2 Spring/Summer 1989, pp.103–129.

[14] Maturana H.R., Varela F.J.(1980): *Autopoiesis and Cognition*, Reidel, Dordrecht.

[15] Newell A., Simon H.A. (1976): Computer Science as Empirical Inquiry: Symbols and Search, *Communications of the ACM*, 19(3)113-126.

[16] Ogden C.K., Richards I.A. (1923): *The Meaning of Meaning*, Routledge & Kegan Paul, London.

[17] Peirce C.S. (1931-58): *Collected Papers*, vols. 1-8, ed. by C. Hartshorne & P. Weiss, Harvard University Press, Cambridge.

[18] Searle J.R. (1980): Minds, Brains and Programs, *Behavioral and Brain Sciences*,3,417-457.

[19] Searle J.R. (1990): Is the Brain's Mind a Computer Program?, *Scientific American*, 1(1990)pp.20-25.

[20] Smolensky P. (1988): On the Proper Treatment of Connectionism, *Behavioral and Brain Sciences* 11(88), p.1-74.

[21] Touretzky D.S., Hinton G.E.(1985): Symbols Among the Neurons: Details of a Connectionist Interference Architecture, in *Proceedings of the 9th International Joint Conference on Artificial Intelligence* (IJCAI-85), Los Angeles, California.