A hierarchy of meaning systems based on value

Jordan Zlatev
Department of Linguistics and Phonetics, Lund University, Sweden, jordan.zlatev@ling.lu.se

Abstract
The paper presents the outlines of a general, multidisciplinary theory of meaning based on the concept of value as a biological and social category, synthesizing ideas from biologically oriented psychology, semiotics and cybernetics. The theory distinguishes between four types of meaning systems: *cue-based, association-based, icon-based* and *symbol-based*, forming an evolutionary and epigenetic hierarchy. This hierarchy is applied to phylogenetic and ontogenetic development, pointing out significant parallels between the two, involving both continuity and discontinuity between different levels, i.e. meaning systems. The theory is finally applied to the field of developing intelligent artificial autonomous systems, showing grave limitations in existing systems and suggesting guidelines for future work.

1. Introduction

Theories of meaning come in every color and persuasion: mentalist, behaviorist, (neural) reductionist, (social) constructivist, functionist, formalist, computationalist, platonist... and even eliminativist: "meaning" is a non-scientific concept that should rather be dropped! While a certain degree of perspectivism concerning a multi-faceted concept such as meaning is certainly healthy, the fragmentation of views concerning what is perhaps most central for human and other living creatures has reached catastrophic proportions. It is possibly therefore that since Harnad (1990) it has been popular to try to "ground" meaning in some common (hard) currency, so to speak. The term "grounding", however, clearly has reductionism connotations, and Harnad's own proposal of "sensorimotor grounding" ("the power of names and propositions is completely grounded in the sensorimotor interactions with the kinds of objects they designate, and the sensorimotor invariants on the basis of which the names are assigned" ibid: 23) falls in the reductionist (Lockean) pit. So on the one hand it is important to try to limit fragmentation and blatant inconsistencies where meaning is concerned, but on the other, one should resist painting all cats gray by reducing the concept of meaning to a single, preferably measurable category such as "information", "utility", "behavioral dispositions" etc.

In this paper I undertake the difficult task of developing an integrative, interdisciplinary theory of meaning that is not reductionist in the above sense. The key idea is that there is *a hierarchy of meaning systems*, which is both evolutionary and epigenetic – each preceding level is presupposed by and integrated in the one that follows, both in evolution and in ontogenetic development. But for each level of the hierarchy it holds that *meaning is a relationship between the individual and the environment, picking out the categories in the environment which are of value for the individual, and thereby defining its "life-world" or Umwelt* (von Uexküll 1982). While for plants and lower animals this life-world is physical, for social animals and especially for human beings it consists even more of social relations, practices and other conventional and normative categories. The value, and therefore the meaning, of these "higher categories" interacts with, but is not "grounded" in (or determined by) the essentially life-supporting value of the categories of the lower levels.

The background ideas that have served as influence for the endeavored theory derive from *biologically-informed psychology*: von Uexküll's (1982) "Umwelt", Gibson's (1979) "affordances", Edelman's (1992) "value-category memory", Donald's (1991) "mimesis" and Damasio's (1994) "feelings"; from *semiotics*: Vygotsky's (1978) distinction between "signal" and "sign" and the social "mediational" character of the latter, Pierce's (1931-35) "icon/index/symbol" triad and Halliday's (1975) "functional approach"; and from *cybernetics*: Weiner's (1958) "feedback loops" and "control systems". A triggering event for writing this paper was provided by a recent article by Cisek (1999), which combines several of these themes, though not the social ones:

> Animals have physiological demands which inherently distinguish some input … as 'desirable' and other input as 'undesirable'. … This distinction gives motivation to animal behavior – actions are performed in order to approach desirable input and avoid undesirable input. It is also gives *meaning* to their perceptions – some perceptions are cues describing favorable situations, others are warnings describing unfavorable ones which must be avoided. (Cisek 1999: 134, my emphasis)

The disposition of the paper is as follows: In Section 2, I present the outlines of a "meaning-as-value" theory, which distinguishes between 4 types of meaning systems on the basis of (a) the character of the internal value system, (b) categories of perception, (c) categories of communication and (d) type of learning involved. The 4 meaning systems go into one another like Russian dolls, and thus form an (evolutionary) hierarchy. The theory implies a close connection between meaning and emotion and thus has implications for a theory of consciousness, though I will not develop this theme in the present paper.

In Section 3, 4 and 5 I briefly apply the proposed theory to the three domains of phylogenesis, human ontogenesis and "robotogenesis": the development of artificial autonomous systems. In the first two cases the purpose is to provide a framework that would help us understand both the continuity and the discontinuity between different species and developmental stages in terms of different meaning-*cum*-value systems. In the third case, the theory highlights a central deficiency in existing artificial "autonomous" systems, and offers some suggestions for their future development.

A final caveat before I proceed: Since my purpose is integrative and constructive, I will not argue in detail for some statements that may be quite controversial (which would easily turn the paper into a book). It is also possible that some of the empirical claims are not consistent with the most recent results on e.g. animal communication, but if so they are not necessarily false, since our empirical knowledge on the matter of mind and meaning is very much in flux.

2. An outline of a theory of meaning as value

In this section I will be maximally schematic and propose a theory of meaning in terms of 6 theses, followed by brief explanations.

1. Meaning is a relationship between an *individual* and its *environment*, defined by the *value,* which particular *aspects* (falling into *categories*) hold for the individual.

Meaning is thus an ecological concept, it is not purely subjective ("in the head") and it is not objective ("in the world"), but characterizes the interaction between individual and environment. The term *individual* is intended to be general with respect to the terms "organism" (as used in biology), "subject" (as used in psychology) and "autonomous system" (as used in cognitive science). The individual's environment can be only physical or both physical and social. An example of a physical aspect is sunlight, an example of a social aspect is a handshake. An aspect of the environment has meaning for the individual to the extent that this aspect has value for this individual. Value can be positive or negative, and thus the aspect will be positively or negatively meaningful. If the aspect has no value for the individual, it will be meaningless. Aspects with equivalent value form categories whose members are treated identically.

2. The value of physical aspects (categories), perceived via *innate value systems,* is initially based on their role for the preservation of the life of the individual (and its kin).

For individuals with only a physical environment, aspects will have high positive or negative value (and thus be meaningful) or be neutral (and thus be meaningless) depending on whether they play any role for the individual's survival. Thus the primary form of "good", "bad" and "neutral" is individual-relative ("selfish") and species-relative. By selecting for life-forms which are viable in relation to their niches evolution provides organisms with innate value systems, which play a crucial role in sensing the (ecological) value of the environment. An innate value system can be conceived of as a system of preferences of different degrees of specificity and strength that controls the behavior and learning of the individual (Thesis 4) and is intimately connected to *emotion* (Thesis 5).

3. The value of social aspects (categories), as well as physical ones once the social ones are present, is based on their role in *conventional meaning-value systems* that need to be acquired by the individual, before they can become meaningful.

For beings living in complex societies and capable of learning social conventions, the environment is more social than physical, consisting of various social practices: rituals, rules, conventions and norms. The value of these social categories is not directly connected to survival, but to the maintenance of social cohesion and communication. Their meaning is thereby "cultural" and mediated by signs, rather than "natural" and direct. However, since the social environment and meaning does not substitute, but rather integrate the physical one, the two kinds of meaning interact. The innate value systems are still present, and probably enriched to include biases for social interaction and imitation, but are less determinative of the individual's behavior than when they operate alone.

4. Both innate and acquired meaning-value systems serve as *control systems* by directing and evaluating the individual's behavior.

On the one hand, internal value systems signal to the individual that some action needs to be taken. Thus they give rise to motivation and (degrees of) *intentionality,* in the sense of goal-directedness. On the other hand, they participate in the evaluation of the actions by assigning "reward" or "punishment", and in this way influencing future behavior. This leads to *learning* (in those beings capable of this). This applies to both the innate value systems, and to those acquired in a culture, the latter functioning as *norms* of various kinds (moral, linguistic etc.) defining "right" and "wrong". This is the basis for self-correction and self-censorship.

5. Value, and consequently meaning, is intimately connected to *emotion* and *feeling*.

The internal value systems (innate and acquired) interact tightly with the emotional system, so that in beings capable of (primary) consciousness negative value is *experienced* as negative emotions, with pain as prototype and positive value as positive emotions, with pleasure as prototype. Emotions play a central role in learning, and since it is possible to have simple meaning systems without learning, where the (expected) value is completely defined by the innate value system, it is also possible to have meaning systems *without* emotions. If we also distinguish between emotion and feeling, the latter being necessarily conscious, while the first is not (in the sense of Damasio) it is also possible to have meaning systems with learning and emotions, but without consciousness.

Table 1. *A hierarchy of meaning systems based on value*

Meaning system	Internal value system (VS)	Perception	Social interaction	Learning	Emotions and consciousness
I. Cue-based	innate VS	cues	-	-	-
II. Association-based (including I)	innate VS + tuning to environment	associations, schemas	partially learned signals	reinforcement learning, conditioning	primary emotions
III. Icon-based (including II)	innate VS, biased for social value + acquired iconic VS	conventional categories	icons, mimes	imitation	secondary emotions
IV. Symbol-based (including III)	innate VS, biased for social value and symbolic relations + acquired symbolic VS	classes, hierarchies	symbols, propositions, narratives	reflection	higher-order consciousness

6. On the basis of the concepts introduced in Theses 1-5, four different types of meaning systems can be defined, forming an *evolutionary hierarchy*.

The names given to the four meaning systems, based on their predominant category of perception/social interaction are (I) cue-based, (II) association-based, (III) icon-based and (IV) symbol-based. The first two correspond to systems of *natural meaning*, while the latter two to systems of "non-natural", conventional meaning in other semantic accounts (cf. Grice 1957). However, in the present approach they are not juxtaposed, but rather form an evolutionary hierarchy in the sense that each succeeding type in the hierarchy builds upon, and "encapsulates" the preceding ones, in a manner similar to that suggested by Donald (1991). The main characteristics of each meaning system are presented in Table 1. Most of the concepts of the presented outline of a theory of meaning are no doubt in need of further explication. In applying the theory to the three domains of phylogenesis, ontogenesis and "robotogenesis", i.e. the development of artificial autonomous systems, in the following three sections respectively, the concepts will hopefully be clarified.

3. Evolution of meaning in phylogenesis

The presented hierarchy of meaning systems can help classify existing living creatures and possibly reconstruct aspects of phylogenesis because as in evolution each higher type of system builds on the one preceding it. Furthermore the transitions between types are relatively straightforward. At the same time, the classification of Table 1 is quite schematic and is therefore problematic to apply to actual species. In this sense it is similar to a popular evolutionary classification proposed by Dennett (1996), which also happens to consist of 4 major types ("Darwinian", "Skinnerian", "Popperian" and "Gregorian" creatures). The two hierarchies, however, have rather different defining criteria: In Dennett's case the different types have different representational abilities, while in the present account, the crucial characteristic of each type is its value-based relationship to the environment.

3.1. Cue-base meaning creatures

Every evolutionary model requires a starting point and the most clear dividing line is that of life itself. Living creatures are traditionally classified into 5 "kingdoms": *Monera* (e.g. bacteria), *Protista* (single-celled organisms), *Plantae* (plants), *Fungi* (e.g. yeast) and *Animalia* (animals). The first 4 of these and the simpler animals, i.e. those with no central nervous system (e.g. flatworms) can be regarded as creatures living in an *Umwelt* of cue-based meaning: They perceive the environment in terms of fixed, pre-determined set of cues. To the extent that we can talk of "communication" between such organisms at all, it is done through pre-determined signals that are not different in kind from the cues of environment perception. There is no learning involved, and consequently no need and no basis for emotion.

One may ask: Why even consider such lowly being as bacteria and plants as capable of meaning at all? The answer is because they have innate value systems in the form of homeostatic mechanisms that help them preserve and propagate life. The chemicals consumed by bacteria and the sunrays required by plants constitute aspects of the environment that are meaningful for them in a very literal sense. The value systems of these creatures motivate them to seek out such substances and to avoid those that are harmful.

3.2. Association-based meaning creatures

The second category of the hierarchy includes a rather heterogeneous set of animals: from mollusks to reptiles, birds and most mammals, though excluding apes (and possibly cetaceans). This group no doubt can (and should) be further divided into sub-groups on the basis of substantial neural and cognitive differences, and possibly even primary consciousness, but for present purposes will be treated as a whole due to the following shared characteristics:

The innate value system of these creatures not only motivates them to act in particular ways, but also evaluates their actions and thus allows them to *learn* by forming *associations* between environmental aspects and the animal's actions (stimulus-stimulus, stimulus-response and response-stimulus). Combinations of such associations correspond to the simplest form of *schemas*. Communication between individuals is done through *signals*, which at least in the higher forms of the group are partially learned, rather than innate (e.g. bird song and vervet monkey alarm calls). In neurobiological terms, such animals have a central nervous system, and in birds and mammals a limbic system and neocortex. The *limbic system*, especially amygdala and hippocampus, appears to represent the bodily state of the organism and plays a crucial part for the formation of *primary emotions*, which are "hard-wired", integrated reactions of the whole body to value-laden physical and social stimuli (Damasio 1994). It is still unclear to what extent they are also *experienced* by the animal, and thus represent a form of "feeling-consciousness" (MacPhail 1998).

On the other hand, most animals have strong limits on what they are capable of learning: Pigeons associate sight with food, but not odor or sound and food; rats can associate size with shock, but not taste with shock etc. These restrictions make good ecological sense, and show species-specific particularities of the innate value systems. They also do not show population-specific characteristics, i.e. no "culture", no ability to learn through imitation, and no ability to learn language. It is possible to find a common denominator to these limitations: Association-based meaning creatures are not able to learn a system of *conventional*, shared meanings, which constitutes a higher-order meaning-value system. This is possibly connected to a limited ability to have *secondary emotions*, "which occur once we begin experiencing feelings and forming *systematic connections between categories of objects and situations, on the one hand, and primary emotions, on the other*" (Damasio 1994: 134, original emphasis).

3.3. Icon-based meaning creatures

Even though there is still controversy concerning the degree to which apes are capable of sign use and imitation learning in the wild it is clear that at least enculturated apes, especially chimps and bonobos, do learn to communicate with people, and even among themselves to some degree, through conventional signs, especially gestures, and gestures are a clear example of *iconic* signs, hence the name given to the meaning-system. The ability to use true *imitation learning*, as opposed to mimicry, requires the ability to imagine another's point of view, or in other words, at least a basic form of a "theory of mind" (cf. Tomasello 1999). Extensive recent experimentation has shown that apes do indeed have at least a simple form of this. They also have the most developed non-human "societies" with various degrees of intra-species, inter-population variation, i.e. the rudimentary of culture (Dunbar 1996). Such extensive learning and reasoning capacities in the social and personal domain would require the ability to entertain *secondary emotions* toward social categories and situations. This ability is supported by the prefrontal cortex and not just by the limbic system. Once acquired, largely through imitation, the social convention system takes on the role of an extended value system, supplementing the innate value system in evaluating physical and social events.

Given that the human and chimpanzee lineages have shared an ancestor some 6 million years ago, it is quite probable that *Authalopithecines* from about 4 million and especially *Homo erectus* from 1.5 million years ago which colonized the world, leaving numerous traces of material culture had even more developed communication and reasoning abilities than the apes of today. It is likely that these creatures used representational bodily gestures, i.e. *mimesis*, for both communication and a first form of self-conscious thinking (cf. Donald 1991). On the other hand their ability to use *symbolic* language was probably limited. We can see such limitations clearly in apes. For one thing, in language teaching experiments, arbitrariness seems to be problematic: Apes are better at learning gestures, a simplified form of "sign language" than arbitrary symbols. Despite a number of recent successes, above all the bonobo Kanzi (e.g. Savage-Rumbaugh and Rumbaugh 1993) the acquisition of language by apes remains controversial. Not only syntactically, which is where the debate most often focuses, but semantically. Above all, ape language in spontaneous environments is largely constrained to the instrumental ("I want") and regulatory ("do this") functions, while utterances that can be classed as instances of interpersonal ("me and you"), heuristic ("what is this?") and informative ("let me tell you something you don´t know") functions (cf. Halliday 1975) have not been convincingly shown. What apes also seem to lack is the ability to use signs not only for communication but for self-communication, i.e. reflection (though there is some evidence for this in Kanzi). They also lack cultural histories: The inter-population variation mentioned above does not go very far (Tomasello 1999).

3.4. Symbol-based meaning creatures

Language, reflection, cultural variation: these are all defining characteristics of *Homo sapiens*, also called "homo symbolicus" by Deacon (1997) because of our innate superiority in "symbolic reference", which is more arbitrary and systematic than communication based on icons and mimes. The acquisition of symbolic systems appears to rely on a strongly expanded prefrontal cortex (PFC), giving our species an important edge in working memory, inhibition and attention control. At the same time, with its connections to the limbic system and the somatosensory cortexes, the PFC is also crucial for the formation of the culturally modulated secondary emotions (Damasio 1994). Putting these two facts together leads us to understand language not as a "cold" reasoning device, but as inherently "emotional" and value-laden. This also explains its central role in the creation of higher-order meaning-value systems such as religions. Such higher-order values can introduce a detachment from the more basic life-supporting value system inherited from our predecessors. This can be seen as a form of freedom from biological determinism, as in Buddhist philosophy, e.g.:

> It is necessary to have a mind which is beyond having a self, or anything belonging to a self, a mind which is even beyond the ideas of good and bad, merit and demerit, pleasure and pain. … Dependent Origination teaches us to be careful whenever there is contact between the senses and their objects. Feeling must not be allowed to brew up or give rise to craving. (Buddhadasa Bhikku 1992: 7-8)

But instead of independence, the "power of the word" can generate an even deeper dependence on beliefs, ideas etc. (as warned by Zen Buddhists) which can even be life-destructive, as demonstrated most graphically by suicide sects. Somewhat paradoxically then, symbolic language can be either advantageous for reasoning by providing a powerful combinatory system of expressing and "manipulating" social meanings internally and thus enhancing reflection and self-consciousness, or disadvantageous by "short-circuiting" the evolutionary derived loop of evaluating situations on the basis of their "somatic" value. This double role of language can be conceptualized by assuming that it plays a central part in a simulated "as if" loop in feeling and reasoning in which the body is "bypassed". As Damasio (1994) argues, dissociating symbolic reasoning from "somatic markers" is pathological, and in the long run self-destructive.

4. Evolution of meaning in ontogenesis

The fact that there are similarities between phylogenetic and ontogenetic development has been observed at least since Haeckel (1874) stated the famous "biological law" of *recapitulation*. At the same time the notion is often criticized for being aprioristic: why should the individual develop similarly as the species, given that the driving forces and the contexts of the two processes are so different? Nevertheless, there is indeed a good reason for such recapitulation (in broad lines): the fact that ontogenetic development is *epigenetic* (Piaget 1970, Müller 1996, Zlatev 1997). That is, development proceeds through an ordered succession of stages that are guided, but not determined by the genes, where "higher" stages require more time and experience than the "lower" ones. Since evolutionary more ancient mechanisms require less experience in order to mature than evolutionary modern mechanisms, the early stages of ontogenesis will be dominated by evolutionary more ancient mechanisms, and the later stages by more modern mechanisms. In some cases, the "lower" stages may serve as an epigenetic prerequisite, a necessary step for reaching the "higher" functions.

Given this assumption, it should be possible to find more concrete ontogenetic evidence to corroborate the rather speculative phylogenetic stages presented in Section 3. In Zlatev (2001) I suggest that stages proposed for hominid evolution by Donald (1991) have approximate correlates in ontogenesis. Below I summarize evidence pertaining to the development of *social interaction* (neglecting physical interaction for the sake of brevity) in terms of the hierarchy of meaning systems presented in the last two sections.

4.1. Cue-based meaning stage?

Considering that cue-based meaning lacks learning and is entirely fixed by genetic factors while the human being begins adapting to the social environment *prior* to birth, it is inappropriate to talk of a cue-based "stage" in development. Still, it is possible to see the child's first expressions of social interaction – the cry and the smile – as essentially cue-based, innate expressions of negative, respectively positive somatic states. Unlike simple organisms whose world is entirely cue-based, however, it is clear that the newborn baby has emotions, though one may still wonder to what extent they are accompanied by "feelings".

Other early forms of interaction which could be seen as fairly automatic responses to social stimuli is the *mimicry* of mouth opening and tongue protrusion usually described as *imitation* (Meltzoff and Moore 1995), though this term is arguably better reserved for the copying of goal-directed behavior which requires an understanding of intentionality and is clearly beyond the competence of neonates (cf. Tomasello 1999).

4.2. Association-based meaning stage: 0-9 months

The first 9 months of the child's life can be seen as being dominated by association-based meaning, where the innate value system is applied to experience and the child learns to discriminate and categorize objects, people and situations. As early as 2 months, the child is capable of engaging in protoconversations (Trevarthen 1992) which demonstrate both structural (e.g. turn-taking) and "semantic" (primarily emotional) aspects of communication:

> The baby can turn affective contact on and off by fixating the mother's eyes, then looking away briefly to make an utterance, or more definitely when tired or distressed. … Infant and adult evidently meet with matching standards of emotion and emotional change that define "good" or "bad" expression and reply. (Trevarthen 1992: 107)

These "matching standards", however, are still done on the child's terms since they are based on the innate value system and not on learned criteria. Unsurprisingly "motherese" shows universal, culture-independent characteristics indicating that parents tune in to these innate criteria. However, the role of learned patterns of interaction soon increases. By 3 months the child engages in interactive games such as Peekaboo, with more or less stable "rules". By 6 months she starts to initiate acts of communication herself. Also by that time the child starts forming "proto-gestures" such as raising her arms when asking to be hugged. It is not impossible to see such acts as the birth of intentional communication and conventionality. Still, I agree with Tomasello (1999) that these bahaviors can be explained as *ritualizations* of patterns of interaction, i.e. as schemas (i.e. sequences of associations) learned from experience, motivated and learned by desires and emotions, but not yet requiring an understanding of other/self intentionality and a "theory of mind".

4.3. Icon-based meaning stage: 9-18 months

Understanding and using signs and their shared meanings, on the other hand, does require at least a first understanding of others and self as intentional beings – at the same time as sign-use probably further contributes to such understanding. There is strong evidence that such understanding begins around 9 months when a number of behaviors depending on it emerge: joint attention, imitation, social referencing, and soon later pointing (Tomasello 1999).

Other researchers have independently argued for a cognitive reorganization around that age. Among them are two who otherwise emphasize continuity in development: Trevarthen (1992) claiming the emergence of *secondary intersubjectivity* and Halliday (1975) pinpointing the start of *protolanguage* at precisely 9 months. Acredolo and Goodwyn (1996) find that the teaching of *baby signs*, iconic gestures which they recommend to parents as a means of communication with their children "before they can talk", becomes effective around that time too. A common theme in these otherwise quite different theoretical approaches is the crucial role of *imitation*. Another is that the first communicative signs are *iconic* and *indexical* rather than truly *symbolic* (cf. Peirce 1931-35). This raises the possibility that imitation, along with the biases of the innate value system that motivates and rewards it, is a necessary first step in the acquisition of (full) intersubjectivity and language. Contra Tomasello, however, it does not seem that it is "all downhill" from here. The repertory of signs the child possesses (even those trained by the methods of Acredolo and Goodwyn) remains limited to a few dozen for a long period of time. The functions (in Hallidayan terms) they are put to are restricted, and there doesn't seem to be much of a *system* in the way the signs are combined and interrelated, i.e. a *grammar*. Finally, there is no evidence of pre-linguistic signs being used in reflection. These limits are quite similar to those of apes (and by extension, to the gesture-using *Homo erectus*) as discussed in 3.3. Just as with apes, there are those who argue that the only difference between pre-linguistic and linguistic children is quantitative, or a matter of motoric skills. There is, however, strong evidence that there is a "second revolution" in ontogenesis around 18 months.

4.4. Symbol-based meaning stage: 18 months –

Sometime between 18 and 24 months nearly all children undergo a rapid development of productive vocabulary and shortly afterwards, an increased number of multi-morphemic utterances, i.e. the vocabulary and grammar spurts. If this signals some kind of discontinuity in development, as most researchers agree, then what is the cause for it? Considering that the understanding of intentions and sign-use begins earlier, it cannot be a matter of a literal "symbolic insight". An explanation compatible with Deacon's (1997) view of the PFC as responsible for acquiring a system of symbolic reference is that the growing maturation of the PFC allows attention to be focused on inter-symbolic relations, and not just on the relation between sign and referent. It is interesting (and has remained unnoticed) that Halliday (1975) also offers a similar explanation of the transition from protolanguage to conventional, grammar-based and dialogic language:

> It is as if up to a certain point the child was working his way through the history of the human race creating a language for himself to serve those needs which exist independently of language and which are an essential feature of human life at all times and in all cultures. Then ... taking over in one immense stride its two fundamental properties as a system: one, its organization on three levels, with lexicogrammatical level of wording intermediate between the meaning and the sounding, … and two, its ability to function as an independent means of human interaction, as a form of social interaction which *generates its own set of roles and role relationships*. (ibid: 32, my emphasis)

Notice that it is not the manual modality that is iconic and the vocal one that is symbolic. The same kind of qualitative difference between gesture and Sign Language can be observed, e.g. in a comparison between gestural communication, Homesign and ASL by Morford, Singleton and Goldin-Meadow (1995), who also state a similar explanation of the difference:

> This change, we believe, comes about as a result of the language user attending less to the relationship between the referent and its symbol, and more to the relationship between the symbols that make up the system. Second, there is increasing evidence of a level of arbitrary representation over time and as the size of the community of language users increase. (ibid: 325)

The fact that with the acquisition of symbolic language the child enters a new mode of thinking and not just communicating is shown in at least two groups of studies: those pertaining to the functional role of egocentric and internal speech (Vygotsky 1962, Berk 1994) and those of social origins of autobiographic memory (Nelson and Hudson 1993). The existence of different perspectives embedded in language is probably instrumental to the ability to rapidly switch between different points-of-view, and a full-fledged theory of mind (Tomasello 1999). Therefore it seems correct to pinpoint the entry into symbolic language as the main "cognitive revolution" in ontogenesis, even though there are important milestones both preceding and following it. As the same time, language can *not* be the major cause of (self-)consciousness as claimed by numerous contemporary theorists (e.g. Dennett 1991, MacPhial 1998), since its acquisition presupposes (a degree of) intersubjectivity, which presupposes consciousness.

5. Meaning and value in artificial systems

The theory of meaning outlined and applied to the evolution and ontogenesis in the preceding two sections can be applied to a classical question: What (if any) kind of artificial system would be capable of meaning? The answer seems to be: Systems with internal control, i.e. *cybernetic systems* as they where originally thought of (Weiner 1958). Opposed to systems which operate on the basis of inbuilt (as in most so-called "symbolic" models) or learned (as in most "connectionist" models) IF-THEN "rules", cybernetic systems necessarily have an internal value system. Such a system determines what in the environment is "right" for the total system (which can change depending on the system's internal "motivational state"), rather than the programmer or an external observer, thus escaping the so-called grounding problem (Harnad 1990):

> Rather than viewing behavior as 'producing the *right response* given a stimulus', we should view it as 'producing the response that results in the *right stimulus'* ... While the first viewpoint has a difficult time deciding what is 'right', the second does not. (Cisek 1999: 134, original emphasis)

An artificial system capable of meaning therefore requires: (1) a self organizing body with an intrinsic value system (*embodiment*), (2) interaction with an environment, which for all but the simplest creatures is social as well as physical (*situatedness*) and, if its meaning system is not to be fixed in advance, (3) cumulative, step-wise development (*epigenesis*). These three conditions can be seen as generalizations of the requirements for artificial systems capable of *human-like* meaning, argued for in Zlatev (2001: 161):

- *sociocultural situatedness:* the ability to engage in acts of communication and participate in social practices and 'language games' within a community;
- *naturalistic embodiment:* the possession of bodily structures giving adequate causal support for the above, e.g. organs of perception and motor activity, systems of motivation, memory and learning;
- *epigenetic development:* the development of physical, social and linguistic skills along a progression of levels so that level *n+1* competence results from level *n* competence coupled with interaction with the physical and social environment.

Note, however, that in the generalized form, embodiment does *not* require the artificial system to be physical, i.e. to be a robot (cf. also Dautenhahn 1999). As far as the internal value system is concerned, the control and learning algorithms of a *simulated* organism are no different from those of a robot. In fact, it is easier to see "agents" embedded within an artificial-life environment with natural selection as possessors of internal value systems grounded in the basic value of survival and hence inherent meaning than present-day robots. A truly autonomous robot would need to be able to seek out energy sources in the environment, and even if there are some efforts along these lines (Hashimoto p.c.) there are obvious difficulties.

Therefore in terms of *meaning*, the embodiment of present-day physical systems is not more (and is perhaps less) adequate than that of embedded Alife agents. The other requirements for achieving human-like meaning in an artificial system: situatedness and epigenesis, however, make it necessary eventually to have a physical system, i.e. a robot, because only then can the system be naturally embedded in a *human* learning environment.

An implication is that because of their somewhat complementary strengths and weaknesses, research in Alife and robotics should collaborate more closely than what seems to be the case. The following is an attempt to classify selected research dealing with "artificial meaning" with the help of the hierarchy presented in this paper. By treating simulated and robotic systems in parallel I hope to suggest that there are more similarities between them than is most often acknowledged.

5.1. Artificial cue-based meaning systems

As mentioned, most computational systems have nothing resembling internal value systems, nor any ability to distinguish between themselves and the environment, and therefore lack a basis for meaning. A "first generation" of simulated creatures that can be said to do so are those with an internal *motivational module,* guiding their behavior in a simulated environment such as the reactive BERRY creatures described by Balkenius (1995). Depending on this module, certain environmental cues will be reacted to or not, e.g. "food" will be "consumed" if "hungry", but not otherwise. Given an environment for which their "innate" motivation-perception-action systems are well-tuned, such creatures have been shown to be capable of quite complex behaviors. The crucial limitation is that they do not learn, and if they do not have automatically replenishable energy (which is biologically impossible), they quickly "die". Even whole populations of such creatures, provided with a "mutation" mechanism introducing variations in the innate tunings across generations die out quickly if the environment is "unfriendly" (Ackley and Littman 1992).

With robotic systems the situation is similar. Many (if not most) "behavior-based" robots operate on the basis of hard-coded complexes of stimulus-response pairs, possibly ordered in subsumption hierarchies. This seems to be the case with the well-known *Cog,* as implemented thus far (Brooks et al. 1999). A related project, however, that of *Kismet* (Breazeal 2000) arguably takes the step into meaning through its implemented motivational system. As the reactive BERRY creatures, but with much more sophisticated interactions with real people and toys, Kismet reacts to cues that are currently "interesting" for it. Like them, however, it doesn't learn. Therefore despite its complex, intention-*like* interactions and the fact that human subjects may treat it *as* intentional, Kismet is only a cue-based system with no real basis for *intentionality*, understood as both goal-directedness and as the ability to entertain mental states.

5.2. Artificial association-based meaning systems

Considerably more adaptability is achieved by systems in which the internal value system not only *directs* which actions are to be taken in order to reach an "innately" specified internal state on the basis of predefined preferences, but also *evaluates* the system's actions on the basis of their effectiveness in reaching the desired state, or "goal". A general computational method to achieve this is *reinforcement learning* (Barto 1992). Ackley and Littman (1992) show how such a strategy gives rise to considerably more viable populations of artificial creatures than the populations with only "evolutionary learning" (i.e. mutations and natural selection), and a combination of both perform best. Even though standard versions of reinforcement learning are "atomistic" in the sense that individual actions become associated with values ("rewards" or "punishments"), and subsequently chosen on a one-by-one basis, it is possible to augment such systems with a procedural memory, so that *sequences* of simple actions (and their corresponding perceptual states) leading to valuable outcomes are memorized as sensorimotor schemas.

The effectiveness of reinforcement learning for robots has also been shown (Lin 1993). It is conceivable that Kismet will be extended in this direction in the future, so that it e.g. learns new strategies of engaging an interlocutor's attention. Even though Edelman (1992) argues against any kinds of "learning algorithms" in modeling natural learning, the latter being based on "selection rather than instruction", even his "noetic devices" such as Darwin IV, which learn associations on the basis of innate value systems are of the same type.

What are the limits of such systems? From a technical perspective, most attention has been directed to the "temporal credit assignment problem": given that the reward (or punishment) that should be attributed to a specific action can be delayed, how is credit and blame to be attributed over time? This is an instance of the more general problem that given a sufficiently complex and dynamic environment and a rich repertoire of possible actions and perceptions, (simple) reinforcement learning becomes intractable.

The general way to try to avoid this problem is through some form of second-order control in which an "adaptive critic" improves the way in which evaluative feedback is assigned (e.g. Barto 1992). Such an improvement would require the ability to predict future events, and to remember past events, along with their associated values: a "value-category memory" (Edelman 1992). Second-order control is also the function attributed to *emotions* by e.g. Balkenius (1995: 179): "Emotion … is concerned with what the animal *should have done*. When the reward or stimulus situation of the animal is unexpected, an emotional state is activated". Despite some intriguing suggestions by Balkenius and others, there doesn't seem to be any generally accepted way for achieving such higher-order control in an artificial system yet.

5.3. Artificial icon-based meaning systems?

A truly autonomous artificial system capable of generalized learning would require some degree of "self-reflection", resting on the ability to distinguish between itself and the environment and an internal value system which is not fixed but changeable (Takadama and Shimohara 2000). Since extensive social interaction is a prerequisite for the emergence of these properties in evolution and ontogeny (e.g. Vygotsky 1978), it appears to be necessary for anything similar to appear in artificial life. An important role in this respect is arguably played by imitation and/or mimesis (Donald 1991, Tomasello 1999). The role of imitation can be compared to that of two-way bridge in which the actions of another are understood by relating them to one's own, at the same time that one can "externalize" one's body image (Zlatev 2000).

The idea of using imitation in robotics has been popular lately (Dautenhahn 1994, Hayes and Demiris 1994, Schaal 2000) but so far the discussion has mostly focused on the *mechanism* of imitation rather than on the reason for and role of imitation for an individual in a social context (cf. Dauntenhahn and Nehaniv in press). Even less has there been a discussion of how imitation-mimesis could give rise to a new level of meaning, learning, memory, and even self-consciousness in an artificial being.

This question is, of course, still largely speculative, but the theory presented in this paper suggests the following hints for an answer: It is possible that human imitation skills depend on a change to the innate value system, so that successful imitation of a con-specific is rewarding *in itself*. If so, an artificial system needs to be constructed similarly. Furthermore, communication through imitated gestures will lead to a shared system that can be learned and transmitted. Such gestures will in general be *iconic* (resembling their meanings) rather than arbitrary. Finally, the addition of a conventionalized <Gesture – Meaning> level to the value-category memory would give the beginning of semantic memory, which is possibly the key to a self-concept and to self-reflection.

If this hypothesis shows to be successful, it would provide strong evidence for a mimetic theory of the origins of conventional communication in evolution and childhood. However, such communication is still not *language*: the similarity between forms and meanings in iconic systems of communication does not allow precise distinctions to be made, and the lack of a systematic relationship between the different signs does not allow the construction of more complex structures: narrative, belief systems, theories.

5.4. Artificial symbol-based meaning?

In contradistinction to traditional AI systems, few autonomous systems even approach the question of language learning and use. This makes perfect sense from the evolutionary perspective adopted here, since linguistic meaning is the "top of the iceberg" of meaning and value systems. Without a grounding in these all "symbol manipulation" is inherently meaningless, and a matter of external attribution (Searle 1995). Since there is no artificial system with icon-based meaning yet, it follows from the present account that there will be no true symbol-based meaning either. Let me mention two prominent efforts to model linguistic/symbolic emergence, the first simulated and the second robotic, and explain briefly why they lack inherent meaning.

Hutchins and Hazlehurst (1995) present a connectionist model of a "community" of neural networks, which converge on a set of arbitrary, distinct symbols for perceptual stimuli. The model is a very good example of the power of self-organization through inter-individual interaction, but it can only serve as an illustration, or an *explication* of the proposed cultural theory of symbol emergence. There is no *actual* meaning involved in the system, since there is no inherent value of the emergent symbols, as pointed out by the authors themselves: "…the meaning of what "good" is (and how good is "good enough") can only be determined by the functional properties entailed in agents using the lexicon, something these simulations do not address" (ibid: 177).

Similarly, Steels and Vogt (1997) attempt to model "lexicon grounding" through interactions of pairs of robots taking the form of stereotypical "language games". The problem is that the "rules" of these games, i.e. the "goals" of the robots, the protocol for the interaction and most importantly the criteria for success or failure are defined prior to the experiment. Thus, the robots are not true autonomous agents; they lack value systems, and therefore lack any inherent meaning. Again, the model can be useful as an explication of certain aspects of language emergence, but not as an instantiation of a true symbol using system. The same applies for the model of Billard and Hayes (1998).

6. Conclusions

The goal of this paper was an ambitious one: to propose and outline a multidisciplinary and multilevel theory of meaning, which both allows for a "continuum of meaning" in all living beings, and provides a basis for dividing meaning systems in qualitatively different types. Furthermore, the theory can be used to guide experimentation in creating artificial systems capable of meaning. Given such a broad goal, I cannot claim to have shown any particular thesis beyond reasonable doubt. Therefore the conclusions that I list schematically below are better taken as hypotheses for further investigation. The first four apply mainly to natural systems and the last three to artificial ones.

- Small changes in the innate value system, especially in the direction of social value, can have enormous consequences for learning abilities, cognitive skills and survival: positively (*Homo sapiens* vs. Neanderthal man?) or negatively (normal vs. autistic children?).
- The intersubjectivity attained through imitation can be the crucial means of understanding conventional meaning (X means for you what it means for me), and for acquiring symbolic gesture and language.
- Semiotic mediation brings about a higher-level value-memory system, with consequences for learning, reasoning and self-consciousness.
- Through arbitrariness and greater systematicity, symbol-based meaning systems offer greater creativity and "ungrouding" from the primary value system, that can either be a blessing or a curse.
- Artificial systems, both simulated and physical, with internal value systems can be attributed intrinsic meaning to the degree that the value systems are instrumental for their self-preservation. If they are not, then all meaning is in the eye of the beholder.
- All but the simplest kind of meaning systems in living organisms imply emotion, and therefore artificial systems need to have at least a "correlate" to emotion. But it remains unclear if and how they could achieve *feeling*, the subjective dimension of meaning.
- There is currently no system, simulated or robotic, that is capable of icon-based or symbol-based meaning, i.e. of semiotic mediation. Systems which attempt these levels have underestimated the lower two, and can thus only be looked upon as models of explication.

References

Ackley, D.H., and Littman, M.L. (1992). Interactions between learning and evolution. In C.G. Langton, C. Taylor, C.D. Farmer, and S. Rasmussen (Eds.) *Artificial Life II,* Reading, MA: Addison-Wesley.

Acredolo, L. and Goodwyn, S. (1996). *Baby Signs: How to Talk with your Baby before your Baby Can Talk.* Chicago: NTB/Contemporary Publishers.

Balkenius, C. (1995). *Natural Intelligence in Artificial Creatures*, Lund University Cognitive Studies 37.

Barto, A. (1992). Reinforcement learning and adaptive critic methods. In D. A. White and D. Sofge (Eds.) *Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches*. New York: Van Nostrand Reinhold.

Berk, L. (1994). Why children talk to themselves. *Scientific American* 271 (5), 78-83.

Billard, A. and Hayes, G. (1988). Transmitting communication skills through imitation in autonomous robots. In A. Birk and J. Demiris (Eds.) *Learning Robots: A Multi-Perspective Exploration*, LNAI Series, vol. 1545, Springer-Verlag.

Breazeal, C. (2000). *Sociable Machines: Expressive Social Exchange Between Humans and Robots*. Sc.D. dissertation, Department of Electrical Engineering and Computer Science, MIT.

Brooks, R., Breazeal, C., Marjanovic, M., Scassellati, B., and Williamson M. (1999). The Cog project: Building a humanoid robot', In C.L. Nehaniv (Ed.), *Computation for Metaphors, Analogy, and Agents,* Springer-Verlag Lecture Notes in Computer Science, 1562.

Buddhadasa Bhikku (1992). *Paticca Samuppada: Practical Dependent Origination.* Chaya: Suan Mokkhabalarama.

Cisek, P. (1999). Beyond the computer metaphor: Behavior as interaction. *Journal of Consciousness Studies* Vol. 6, Nov/Dec, 125-142.

Damasio, A. (1994). *Decartes' Error, Emotion, Reason and the Human Brain*, New York: Grosset/Putnam.

Dautenhahn, K. (1994). Trying to imitate - a step towards releasing robots from social isolation, In P. Gaussier and J-D. Nicoud (Eds.) *Proceedings from Perception to Action Conference*, Lausanne, Switzerland, September 7-9, 1994.

Dautenhahn, K. (1999). Embodiment and interaction in socially intelligent life-like agents, In C. L. Nehaniv (Ed.), *Computation for Metaphors, Analogy and*

Agent, Springer-Verlag Lecture Notes in Computer Science, 1562.

Dautenhahn, K. and C.L. Nehaniv (in press). The agent-based perspective on imitation. In K. Dautenhahn & C. L. Nehaniv (Eds.) *Imitation in Animals & Artifacts*, MIT Press.

Deacon, T. (1997). *The Symbolic Species: The Co Evolution of Language and the Brain*. NY: Norton.

Dennett, D. (1991). *Consciousness Explained.* Boston: Little, Brown.

Dennett, D. (1996). *Kinds of Minds, Towards an Understanding of Consciousness*, London: Phenix.

Donald, M. (1991). *Origins of the Modern Mind. Three Stages in the Evolution of Culture and Cognition,* Cambridge, Mass.: Harvard University Press.

Dunbar, R. (1996). *Grooming, Gossip and the Evolution of Language*, London: Faber and Faber.

Edelman, G. (1992). *Bright Air, Brilliant Fire. On the Matter of the Mind*, New York: Basic Books, HarperCollins Publications.

Grice, P. (1957). Meaning. *Philosophical Review* 66, 377-88.

Harnad, S. (1990). The symbol grounding problem, *Physica D* 42, 335-346.

Haeckel, E. (1874). *The Evolution of Man: A Popular Exposition of the Principle Points of Human Ontogeny and Phylogeny*, New York: International Science Library.

Halliday, M.A.K. (1975). *Learning How to Mean.* London: Edwin Arnold Press.

Hayes, G.M. and Demiris, J. (1994). A robot controller using learning by imitation, *Proceedings of the 2nd International Symposium on Intelligent Robotic Systems*, Grenoble, France, July 1994.

Hutchins, E. and Hazlehurst, B. (1995). How to invent a lexicon: The development of shared symbols in interaction, In N. Gilbert & R. Conte (Eds.) *Artificial Societies: The computer simulation of social life,* UCL Press, 1995.

Lin, L. (1993). *Reinforcement learning for robots using neural networks*, PhD. Thesis, Carnegie Mellon University.

Macphail, E. (1998). *The Evolution of Consciousness.* Oxford: Oxford University Press.

Meltzoff, A. and Moore, M. (1995). Infants' understanding of people and things: From body imitation to folk psychology, In J. L. Bermudez, A. Marcel, and N. Eilan (Eds.), *The Body and the Self.* Cambridge, Mass.: The MIT Press.

Morford, J., Singleton, J. and Goldwin-Meadow, S. (1995). The genesis of language: How much time is needed to generate arbitrary symbols in a sign system. In K. Emmorey and J. Reilley (Eds.) *Langage, Gesture, and Space*. Lawrence Erlbaum.

Müller, R. (1996). Innateness, Autonomy, Universality? Neurological Approaches to Language, *Behavioral and Brain Sciences* 19: 611-675.

Nelson, K. and Hudson, J. (1993). The psychological and social origins of autobiographical memory, *Psychological Science* 4, 85-92.

Piaget, J. (1971). *Biology and Knowledge.* Chicago: University of Chicago Press.

Pierce, C. S. (1931-35). *The Collected Papers of Charles Sanders Peirce.* Vols. 1-4. Cambridge, Mass.: Harvard University Press.

Savage-Rumbaugh, S and Rumbaugh, D. (1993). The emergence of language. In K.R. Gibson and T. Ingold (Eds) *Tools, Language and Cognition in Human Evolution.* Cambridge: Cambridge University Press.

Schaal, S. (1999). Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences* Vol.3, 233-242.

Searle, J. (1995). The mystery of consciousness, *The New York Review of Books*, Nov. 2, 1995, 61-66.

Steels, L. and Vogt, P. (1997). Grounding adaptive language games in robotic agents. In P. Harvey and P. Husbands (Eds.) *Proceedings of the 4th European conference on artificial life 97.*

Takadama, K. and Shimohara, K. (2000). Interactive self-reflection: Aim and overview. *AROB'2000.*

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*, Cambridge, MA: Harvard University Press.

Trevarthen, C. (1992). An infant's motives for speaking and thinking in the culture. In A. Wold (Ed.) *The Dialogical Alternative, Towards a Theory of Language and Mind*. Oslo: Scandinavian University Press.

Von Uexküll, J. (1982). The theory of meaning, *Semiotica* 42 (1), 25-82.

Vygotsky, L. (1962). *Thought and Language.* Cambridge, Mass.: MIT Press.

Vygotsky, L. (1978). *Mind in Society, The Development of Higher Psychological Processes,* Cambridge, Mass.: Harvard University Press.

Weiner, N. (1958). *Cybernetics, or Control and Communication in the Animal and Machine*, Paris: Hermann.

Zlatev, J. (1997). *Situated Embodiment, Studies in the Emergence of Spatial Meaning*. Stockholm: Gotab.

Zlatev, J. (2000). The mimetic origins of self-conciousness in phylo-, onto- and robotogenesis, *Proceedings of the Third Asia-Pacific conference on simulated evolution and learning (SEAL2000)*, Nagoya, Japan.

Zlatev, J. (2001). The epigenesis of meaning in human beings, and possibly in robots. *Minds and Machines*, Vol. 11 (2), 155-195.