Self-organised communication in autonomous agents: A critical evaluation of artificial life models

Margareta Lützhöft
Department of Computer Science
University of Skövde, Box 408
S-541 28 Skövde, SWEDEN

HS-IDA-MD-00-008

Submitted by Margareta Lützhöft to the University of Skövde as a dissertation towards the degree of M.Sc. by examination and dissertation in the Department of computer science.

September, 2000

I hereby certify that all material in this dissertation which is not my own work has been identified and that no work is included for which a degree has already been conferred to me.

_____

Margareta Lützhöft

Abstract

This dissertation aims to provide a critical evaluation of artificial life (A-Life) models of communication in autonomous agents. In particular the focus will be on the issue of *self-organisation*, which is often argued to be one of the characteristic features distinguishing A-life from other approaches. To ground the arguments, a background of the study of communication within artificial intelligence is provided. This is followed by a comprehensive review of A-Life research on communication between autonomous agents, which is evaluated by breaking down self-organisation into the following sub-questions. Is communication self-organised or hard-coded? What do signals mean to the agents, and how should an external examiner interpret them? Is there any spatial or temporal displacement, or do agents only communicate about their present situation? It is shown that there is very little self-organised communication, as yet, when examined on these grounds, and that most models only look at communication as relatively independent from other behaviours. As a conclusion, it is suggested to use integrated co-evolution of behaviours, including communication, in the spirit of the enactive cognitive science paradigm, and by using incremental evolution combined with learning.

Acknowledgements

I would like to thank my advisor, Tom Ziemke, for motivation and inspiration, and my family – Bengt and Ebba, and friends – especially Tarja and Karina, for their endurance and support.

Table of contents

Table of Figures

1 Introduction

The study of communication within cognitive science and artificial intelligence has, as the fields in general, moved through two paradigm shifts; from traditional or good old-fashioned AI – GOFAI as Haugeland (1985) calls it – to connectionism and recently further on to an emerging theory of mind called the situated, embodied or enactive perspective (Clark, 1997; Franklin, 1995; Varela, Thompson & Rosch, 1991). While GOFAI and connectionism have engaged in communication research primarily focused on *human* language, much work during the 90s within Artificial Life (A-Life), an approach that adopts the ideas of the situated perspective, has focused on autonomous agents, and the emergence and use of these agents' own self-organised communication. It is not quite clear how to accomplish *self-organised* communication in A-Life, but the following quote indicates that it is an important part of the study of communication within A-Life, as is also implied by the title of this work.

> "In the spirit of the bottom-up approach, these communication systems must be developed by the robots [agents] themselves and not designed and programmed in by an external observer. They must also be grounded in the sensori-motor experiences of the robot as opposed to being disembodied, with the input given by a human experimenter and the output again interpreted by the human observer" (Steels & Vogt, 1997, p 474).

The extent to which the designer provides the input and interprets the output will be reviewed in the following chapters, but for now it is sufficient to know that this is always the case in GOFAI and connectionist models, whereas A-Life models more often use some degree of self-organisation. Self-organised communication thus means that the communication system used has not been imposed by an external designer, as in the case of GOFAI and connectionism, but has been developed/evolved by the agents or the species in interaction with their environment (cf. the Steels quote above). Artificial Life is a fairly new area of research, and the central dogma for A-Life is not "life as it is" but "life as it could be" (Langton, 1995). A-Life explores existing phenomena and phenomena as they could have been, mainly using autonomous agents, such as computer

simulations or robots. Humans and other animals, as well as many artificial systems, can be seen as autonomous agents, which means that they function independently of a (potential) designer, such that an agent's behaviour is dependent on its experience. Autonomy also implies being situated in an environment, as in "being there" (Clark, 1997).

The possibility of sharing information about the environment is the main reason why communication has a supportive function for agents trying to cope with their world. This information may be unavailable or unknown to some, or all, of the agents for some reason, such as the information being hidden, too far away or in the 'mind' (internal state) of another agent, or simply not anticipated as significant by the designer of the system. Mataric (1998), for instance, argues that communication is helpful to agents that cannot sense all information in the environment, but does not assume that communication will arise because of this. There are, according to Kirsh (1996), at least three strategies an agent might use, when facing a problem to be solved or a difficult situation, for instance, to improve its fitness or 'cope better' with its world. The agent can migrate to other surroundings, adapt the environment (e.g. make a path through difficult terrain), or adapt *to* the environment. The last strategy can be broken down further into different types of behaviour. This will be discussed later in this thesis, but one candidate for such behaviour is clearly communication.

Communication is often conceived of as an ability that supports co-operation in those A-Life models where autonomous agents are used. In order to justify this view, we may look at living agents, where it is obvious that using different forms of communication promotes co-operation which then enables organisms to better cope with their environment than they would do on their own. This makes it reasonable to assume that this is true for A-Life agents as well. In natural systems, communication and co-operation are self-organised in the sense that they have evolved or been learned (as the belief may be) but the important point is that they are not provided by an external designer as in GOFAI and connectionist studies of communication. Many studies performed are in fact inspired by different types of communication in various natural systems, which often constitutes the motivation behind the experiments. Some researchers claim to replicate or simulate findings from biology, some employ a linguistic

approach, and yet others have AI or cognitive science as their point of departure. This dissertation aims to provide a critical review of these many models, looking at the different approaches and techniques, as well as examining a selection of experiments closer, to investigate in which ways the models afford and constrain self-organised communication.

Different definitions of communication will be discussed in the next section, but for now a very simple and basic assumption is that we need one or more agents as 'speakers', an external world, and some sign or signal, not to mention at least one possible receiver. We should also consider whether or not the sender and the receiver should have a common aim, since otherwise it is doubtful if there is communication (or a signalling system) at all (cf. Bullock, 2000). This is not a definition of what communication *is*, but what might be considered the minimum requirement to achieve it. Various types of communication in natural systems could be considered to have a continuity, that in some way mirrors the 'evolution of communication', an idea that at first glance seems natural. This implies a continuity between on the one hand (to name but a few systems, with examples from Hauser, 1996) mating signals in birds, vervet monkey alarm calls, and non-human primate vocalisations (no order implied) and on the other hand what we consider the most complex of all communication systems, human language.

As the following chosen excerpts will show, this 'continuity idea' is controversial. On the one hand, Hurford (2000) uncompromisingly states that it can be agreed that nothing of the complex structure of modern languages has its ancestry in animal communication systems and is backed up by Pinker (1994) who argues that there is no connection between the two systems, since they (among many other convincing arguments) even originate from different areas of the brain. These non-believers are followed by Noble (2000) who more cautiously remarks that human language may or may not be continuous with animal communication and Gärdenfors (1995) who notes that no animal communication systems have been found that use grammar or compose more than two words into 'sentences'. On the other hand, Aitchison (1998) claims that there are degrees of continuity between animal communication and language, where different aspects of language are more or less continuous. Many researchers within A-Life believe in this continuity, explicitly stated or more implicitly shown in their way of creating models, and

Parisi (1997) claims that we must shift our attention (from abstract studies) to the emergence of language from animal communication. If we study how communication can arise in these natural systems, researchers in A-Life may be inspired by the findings and apply ideas from animal communication (which is not uncommon, cf. later chapters) to their research, regardless of there being a continuity between the systems or not. The largest problem with this diversity of views is in fact not the 'continuity debate', but that it is complicated to compare the different experiments.

There are simulations that examine only biological-evolutionary issues, such as the honest signalling of viability (Noble, 1999), which explores the 'showing of fitness' by males to find mates. Such work is outside the scope of this review, even if they may serve as an inspiration, as does animal signalling in general. These simulations are among the borderline cases mentioned earlier since the signallers and receivers did not co-evolve to achieve a common aim (Bullock, 2000). Another line will be drawn at experiments evolving strictly human-oriented linguistic models such as those of Kirby and Hurford (1997), looking to explain the origins of language constraints, i.e., why humans do not acquire dysfunctional languages (that for instance are hard to parse), or simulating the origins of syntax in human language, by using linguistic evolution and no biological evolution (Kirby, 1999).

There are a few surveys written on the subject of the evolution of language or communication by using A-Life or related models. For instance, Parisi (1997) wished to bridge the gap between animal and human communication, Steels (1997b) worked with modelling the origins of human language, where language itself was viewed as a dynamical system, and de Jong (2000) investigated the autonomous formation of concepts in artificial agents. This dissertation will review some of the same experiments as the preceding surveys, albeit from another angle. The focus of this dissertation is to investigate the question to what degree current A-Life models really allow agents to develop "their own language" (cf. the Steels & Vogt quote above). More specifically, the investigation of different levels of self-organisation can be broken down into the following main questions:

- Why is communication taking place, that is, is communication self-organised in the sense that it is 'chosen' as one of several possible reactions/behaviours, or are

agents ‘forced’ to communicate, and, do they benefit directly from communication or not?

- What is being communicated? This regards what the signals might mean to the agents, and the extent to which we can and should interpret the meaning of a self-evolved language.
- Is there any temporal and/or spatial detachment of the topic, i.e. can agents communicate about something that is not present?

1.1 Definitions

One commonly finds definitions for different types of *communication*, which often are useful only within the context for which they were created. This is not the place to try and resolve the matter once and for all, but instead a few common definitions of communication, and other central concepts, will be assessed as being more or less useful in the endeavour to achieve self-organised communication in autonomous agents.

One definition that has been used by researchers in A-Life (e.g. MacLennan, 1991; MacLennan & Burghardt, 1994) is Burghardt's (1970), in which he states that:

> “Communication is the phenomenon of one organism producing a signal that, when responded to by another organism, confers some advantage (or the statistical probability of it) to the signaler or his group” (p 16).

Burghardt did not intend the definition to deal with communication viewed as human language only. A central criterion is that communication must involve ‘intent’ in some manner, the intent in this case being the adaptive advantage, but he does not want it to be confused with human ‘intentional’ or ‘self-aware’ behaviour.

According to Hauser (1996), most definitions of communication use the concepts of information and signal, where information is a concept ultimately derived from Shannon (1948), later known as ‘Shannon and Weaver’s information theory’ which proposes that information reduces uncertainty; the more information, the more uncertainty is reduced. Signals are thought of as the carriers of information. But, Hauser points out, if we are discussing communication between organisms, the concepts should be explained in terms

of their functional design features. By this he means that information is a feature of an interaction between sender and receiver, and not an abstraction that can be discussed in the absence of some specific context. Signals have been designated to serve particular functions, explained in the following way: functionally referential signals are, for instance, animal calls that are not like human words although they appear to function the same way. Di Paolo (1998) also criticises the concept of information as a concrete object being transmitted, but proposes a completely different outlook, namely autopoiesis (reviewed in chapter 3).

Noble and Cliff (1996) add the concept of intentionality to the discussion, sharpening the meaning of intent so cautiously circumscribed by Burghardt. They claim that when discussing communication between real or simulated animals, we must consider if the sender and receiver are rational, intentional agents. Communication cannot be explained at levels of neurology or physics, but one might usefully talk about what an animal intends to achieve with a call, or what a particular call means. Noble and Cliff are not, however, claiming that animals have intentionality, but that attempts to investigate communication without an intentional framework will be incoherent.

Self-organisation concerns to what degree the agents are allowed to evolve/learn on their own without a designer telling them what to learn and how to talk about it (which features are interesting, and which 'signals' to use). In this case the evolution/learning concerns a 'language', but other behaviours are conceivable, including co-evolution of several behaviours. Pfeifer (1996) describes self-organisation as a phenomenon where the behaviour of an individual changes the global situation, which then influences the behaviour of the individuals. This phenomenon is an underlying principle for agents using self-supervised learning. Steels (1997b) defines self-organisation as a process where a system of elements develops global coherence with only local interactions but strong positive feedback loops, in order to cope with energy or materials entering and leaving the system, and according to Steels there is no genetic dimension. In this dissertation, self-organisation will represent situations where agents exist, in body and situation, and create, by learning or evolution, their own way of communicating about salient features of the environment to their fellow agents. This view of self-organisation has been concisely formulated by Dorffner (1997): "automatic adaptation via feedback through the

environment", and by Elman (1998) as "…the ability of a system to develop structure on its own, simply through its own natural behavior". This is closely related to the concept of *autonomy*, since both imply that agents obtain their own input, without the 'help' of an external designer.

It seems that the simple assumption made earlier, that we need an agent, an external world, and some sign or signal, and a possible receiver is close to what is required. The question of who should profit, the sender, the receiver, or both, will be discussed in chapter 4. If the emphasis lies on letting agents self-organise their own language, it may be that a definition of communication is not what we need, but a way to measure if agents do cope better when exchanging knowledge about their environment.

1.2 Approaches to the study of communication

There are some options to consider in choosing how to study communication. There is the possibility of studying communication 'live' in humans or animals, but this will not tell us anything about its origins, i.e., how it might self-organise. Therefore our study will be limited to simulated communication. Communication in natural systems requires some cognitive ability, which is why we will start by looking at different approaches to cognitive science, or the study of mind. According to Franklin (1995) the mind can be studied in a number of ways as illustrated in Figure 1.

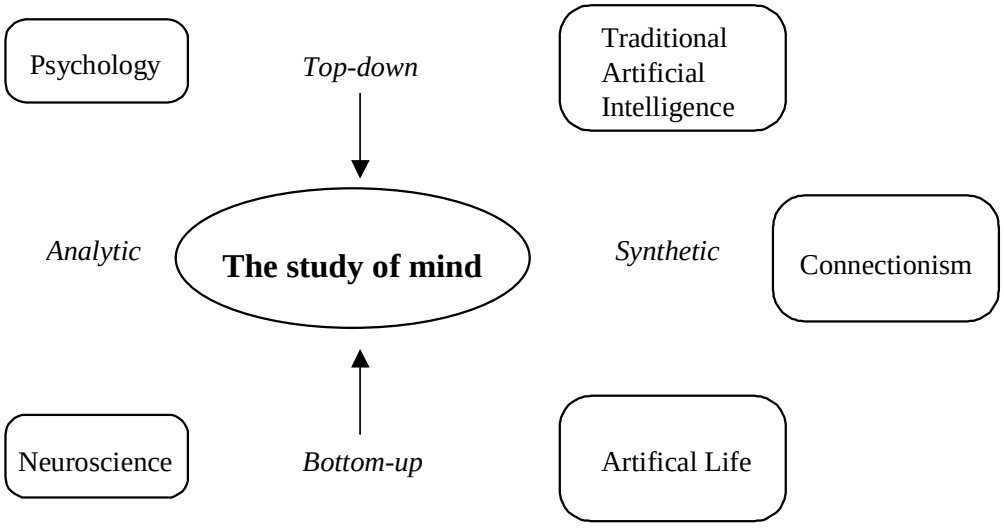

Figure 1: Different approaches to studying mind, adapted from Franklin (1995).

There is the top-down approach, which sets out with a whole ability, for instance the ability to combine words into sentences, and attempts to decompose it into smaller parts, which might concern how words and grammar are learnt (or innate, as the case may be). The bottom-up approach, on the other hand, starts with the 'parts' and/or the supposed underlying mechanisms and the aim is to construct a composite behaviour. If viewed from the sides, the analytic methods examine existing abilities, such as psychology investigating behaviour, and neuroscience studying the brain. These two methodologies will not be considered further, and the focus of the dissertation will be on the right-hand side of the figure – the synthetic approaches. Synthesis is the man-made composition of abilities or behaviours, not necessarily existing in humans, which for instance can be studied by computer simulations of human cognitive capacities.

Traditional Artificial Intelligence (GOFAI) studies of communication have mostly been applied to human language, in a field called natural language processing, see for example Schank and Abelson (1977). When designing such systems, the designer typically determines both the 'external' language and its internal representations. This means that the input consists of strings of symbols with specified meanings, and the internal representations characteristically consist of pre-programmed rules to govern the manipulation of these symbols, as is appropriate in a top-down approach. The rules might for example be algorithms for parsing, or syntactically analysing, the symbol strings. Apart from the question if the system's understanding of the structure of sentences can actually lead to its understanding of language, which will be covered in more detail in chapter 2, it is also clear that the designer influence is total, with no room for self-organisation.

Connectionism, an approach not explicitly mentioned in Franklin's original figure, is here placed on the synthetic side, in the middle, since its mechanisms are bottom-up and the tasks (usually) top-down and psychologically based. Examples of typical tasks is the simulation of how humans learn to pronounce text (Nettalk; Sejnowski & Rosenberg, 1987) or how children learn the past tense of verbs (Rumelhart & McClelland, 1986), a simulation that will be accounted for more explicitly in the background. Connectionist models allow self-organisation of the internal representations, but the external

representations, input and output, are typically human language, and thus pre-determined.

In the lower right corner we find an alternative to traditional AI and connectionism, Artificial Life. Franklin originally used the phrase ”mechanisms of mind”, and defines this concept as ”artificial systems that exhibit some properties of mind by virtue of internal mechanisms” (1995). This encompasses, among others, the study of Artificial Life, the area within which communication is explored in this dissertation. A-Life models can be considered to be more appropriate for realising self-organised communication, since they allow for self-organisation of both an external language and its internal representations. This is due to the fact that the systems often learn/evolve from scratch, in interaction with an environment, rather than focusing on human language. The motivation for studying this, as mentioned earlier, is to make it easier for autonomous agents to manage in various unknown or dynamic environments, by letting them self-organise a language relevant for the environment and task at hand. The intended contribution of this thesis is suggestions of alternative ways of achieving self-organised communication between autonomous agents by critically reviewing previous work and pointing out issues and models that constrain this objective.

In sum, this chapter has introduced the general aim, motivation, and intended contribution of this dissertation, as well as briefly discussed key concepts. The remainder of the dissertation is divided into 4 chapters: background, review, evaluation and discussion/conclusion. The background, chapter 2, will briefly review relevant work within GOFAI and connectionism, and introduce the field of Artificial Life. The review in chapter 3 presents a comprehensive selection of work within A-Life in achieving communication between autonomous agents, where some models will be examined in detail, and others described more concisely. In the evaluation chapter (4), the main issues are the questions outlined above, used to discuss the reviewed models. Chapter 5 will conclude the dissertation with a more general discussion of the approaches used, and a conclusion where the (possible) contribution of A-Life towards self-organised communication is summarised.

2 Background

This chapter opens with a historical background, reviewing and discussing work on communication and language within traditional AI and connectionism, and concludes with an introduction to A-Life models of communication.

2.1 Traditional AI – Natural language processing

Traditional AI uses a top-down approach on the 'synthetic side' of the study of mind as illustrated in Figure 1. When it comes to language and communication the emphasis has been on trying to make computers understand and speak human language – 'natural language understanding' and 'generation' respectively – which together are referred to as 'natural language processing'. A system constructed to model the human ability to understand human stories was scripts (Schank & Abelson, 1977). A script was an abstract symbolic representation, or stereotype, of what we can expect from everyday situations in people's lives, as Schank and Abelson described it: "…a structure that describes appropriate sequences of events in a particular context" (1977, p. 41).

A script would be instantiated with a story, or 'parts of' a story with some information missing, which a human would nonetheless be able to make sense of. The script most commonly used to illustrate this was the restaurant script, containing knowledge (i.e., a symbolic description) of a typical restaurant visit, which might be a sequence of events like this: customer enters, sits down, orders, is served, eats the food, pays, and leaves. If Schank's program was given as input a story about a person who goes to a restaurant, sits down, orders food, receives it, and later calls the waiter to pay, the question to the program might be: did the person actually eat the food? The program could answer correctly because in its prototypical representation of a restaurant visit (cf. above), a customer does eat the food. Thus it might be argued that the system understood the story, and also that the model told us something about the way humans understand stories.

Much of the early work in knowledge representation was tied to language and informed by linguistics (Russell & Norvig, 1995), which has consequences for both natural

language processing and knowledge representation in traditional AI systems. Since human language is both ambiguous *and* depends on having context knowledge it is more suited for communication than for knowledge representation, as Sjölander (1995) comments: "...language...has proven to be a singularly bad instrument when used for expressing of laws, norms, religion, etc". On the other hand, when looking at the usefulness of language for communication, consider this brief exchange (from Pinker, 1994):

The woman: I'm leaving you.
The man: Who is he?

This example clearly shows how much of human language is context-dependent, i.e., it is difficult to understand language without knowing anything about the world, which indicates some of the difficulties that lie in the task of making a computer understand what is actually being said. This is one of the reasons that attempts to formalise language by using logic-like representations were used, as in the traditional AI way of representing knowledge, illustrated by Figure 2. There is supposedly a mapping between the objects in the real world and the representations (symbols).

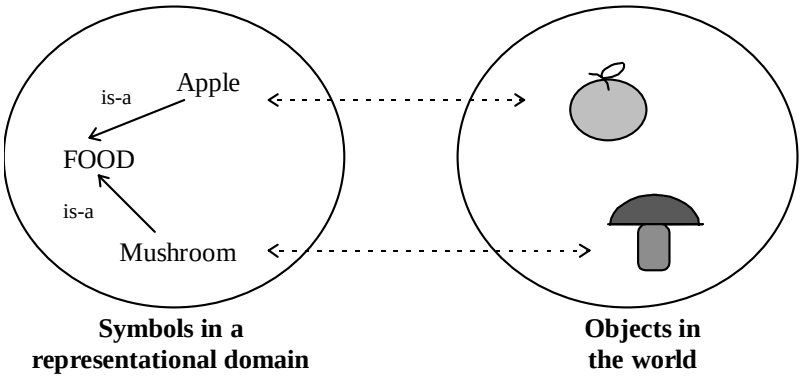

Figure 2: The notion of traditional AI representation, adapted from Dorffner (1997).

The best-known test/criterion for an intelligent machine, or, if you will, artificial intelligence, is the Turing test (Turing, 1950). The test amounts to a human communicating in natural language with an unknown agent, hidden from view. The agent may be another human, but may also be a machine. If the human cannot tell the

difference after communicating about any chosen topic, the agent (possibly a machine/computer) is deemed intelligent. This will depend on the agent's answers, both in content and wording. So whether it is considered intelligent or not will depend on how natural, or human-like, the language it uses is.

One of the first examples of such a natural language conversation system was ELIZA (Weizenbaum, 1965). ELIZA was programmed to simulate a human psychiatrist, by using certain words from the input (e.g., mother, angry), and a set of pre-programmed syntactical rules. The system did not really understand what it was told, but simply used the rules to syntactically manipulate the input and produce an output, which more often than not was the input formulated as a question, like for example, if the human 'tells' ELIZA "I dislike my mother", ELIZA might 'reply' "Why do you say you dislike your mother?". As Weizenbaum constructed the system to solve part of the task of getting computers to understand natural language, he was shocked by people's reactions to the program, among them many psychiatrists, who were ready and willing to believe that ELIZA could be used for automatic psychotherapy (Weizenbaum, 1976).

The idea that computer programs such as Schank's scripts and ELIZA actually understand stories and human language has been challenged by (among others) the Chinese room argument (CRA; Searle, 1980). Searle distinguished between two types of AI, "strong AI", which is the claim that a computer with the right program can understand and think, as opposed to "weak AI" in which the computer is a tool in the study of mind. The CRA, a thought experiment pitched against strong AI, goes approximately like this: a person is placed in a room with a batch of symbols ('a story' in Chinese) and a 'book of rules' that shows how to relate symbols (Chinese signs) to each other. He is then given a set of symbols ('questions' in Chinese) by a person outside the room, uses the rules to produce another set of symbols ('answers' in Chinese), and sends them outside the room. To the observer outside, it will seem as if the person in the room understands Chinese, since answers in Chinese were produced as a result of passing in questions in Chinese. Searle's (1980) point is that the person inside the room does not understand, since he is only "performing computational operations on formally specified elements", in other words running a computer program.

Hence, the words used in ELIZA and the scripts amount to no more than meaningless symbols, ”formally specified elements”, to the systems themselves. The responses are based on syntactic handling of the symbols, and thus there is no ‘understanding’ of the semantics. This is primarily due to a missing link between symbol and reality. This missing link is shown in Figure 3, which illustrates how representations actually are created by a designer, and thus the symbols representing objects are not causally connected to their referents in the world.

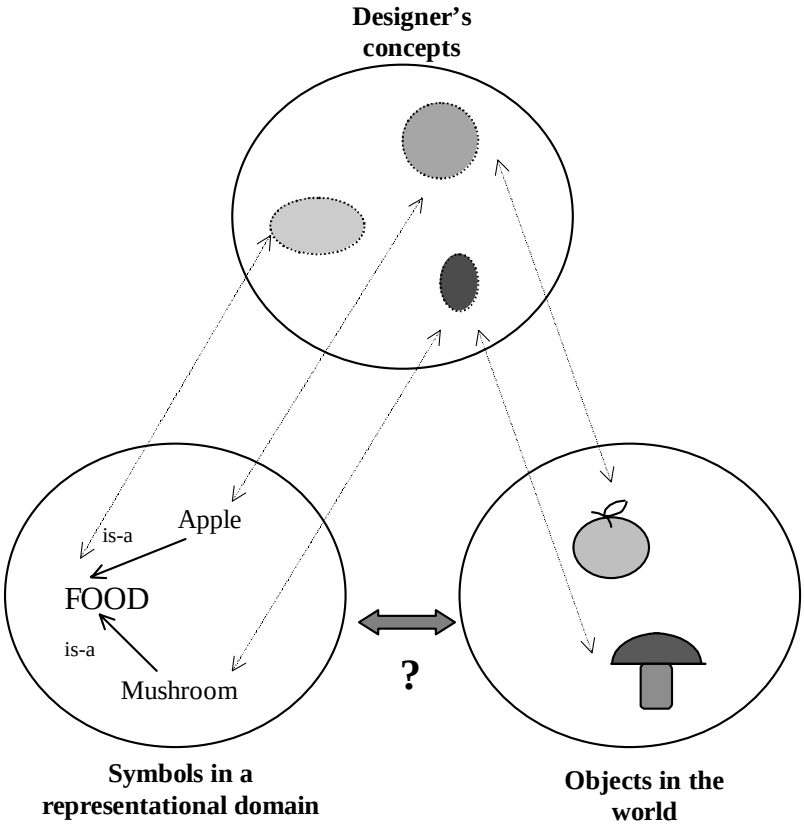

Figure 3: The problem with traditional AI representation, adapted from Dorffner (1997).

Furthermore, when these symbols are operated upon using formal rules, this leads to the following (non-) consequence, as Searle declared: ”...whatever purely formal principles you put into the computer will not be sufficient for understanding...” (1980, p. 187). Harnad (1990) repeated and rephrased Searle’s critique, introducing the expression ”the symbol grounding problem”. The main question was how to make symbols intrinsically meaningful to a system, and not ”parasitic” on the meaning in the designer’s mind (Harnad, 1990).

The designer is in fact the system's only context (cf. also Ziemke, 1997). The mapping between the world and representations is in reality bypassed through the designer, which severs the (imagined) link between symbols and world. There is no context consisting of an environment, dynamical or otherwise, and no other agents to understand or signal to, except for the external designer/observer. This is important, of course, especially in these circumstances, where we are looking to investigate communication. The simple assumption made about communication in the introduction, remember, involved speaker(s), an external world, some sign or signal, and receiver(s).

The necessary context consists of, but is not necessarily restricted to, having a body and being in the world. Traditionally, AI has ignored both the organism and reality (Ziemke & Sharkey, in press), and treated the mind as 'self-supporting' in the way of representations, so that when we have constructed our representations, we will not need feedback from our body or the outer world anymore. Weizenbaum himself was aware of the fact that context is extremely important, as he comments: "I chose this script [the psychotherapist] because it enabled me to temporarily sidestep the problem of giving the program a data base of real-world knowledge." (1976, p. 188).

Even if we could construct such a database, the problem remains; how can we ground the knowledge in the real world? This is also pointed out by Haugeland (1985, cf. above), who declares that all the hard problems of perception and action involve the interface between symbolic cognitions and non-symbolic objects and events. This can be read as a criticism both against micro-worlds and the lack of grounding. The reason that the systems seem to cope so well is that they interact by using rules on symbols in stripped-down worlds, and there is no need to interface cognitions with non-symbolic objects, since there are none. We will let Clancey (1995) summarise this critique of GOFAI: "Knowledge can be represented, but it cannot be exhaustively inventoried by statements of belief or scripts for behaving".

To end this section, and reconnect to the focus of the dissertation by way of the title, there is positively no self-organisation, of language or otherwise, and there are no autonomous agents in the language domain of classical AI. The following section will investigate to what degree connectionism addresses/solves these problems.

2.2 Connectionist modelling of natural language

Connectionism is inspired by the operation of the nervous system and many human cognitive abilities have been simulated in a way classical AI has not been able to do. Since these systems can learn, they should be able to learn to categorise their environment without the help of a designer (Pfeifer, 1996), and in this way achieve grounding (for a more complete discussion of categorisation and its relation to symbol grounding, see Harnad, 1996). The experiments discussed all relate to human language, since this is where connectionism has turned its attention where communication is concerned.

Connectionism has been proposed as an alternative way of constructing AI, by using a network of nodes (artificial neural networks or ANNs), whose connections are provided with weights that, with training, change their value. This allows the networks to learn or construct an input-output mapping, typically from a noisy data set. This ability makes them ideal for tasks that are not easily formalised in rules, e.g., human language (cf. (Haugeland, 1985; Pinker, 1994), which is not always completely 'predictable'. In connectionist modelling, the internal representations (the weights) are allowed to self-organise, during a period of training or learning, but the external representations (inputs and outputs to the network) are generally prepared and interpreted by the researcher. This is due to the fact that, as earlier mentioned, the tasks within connectionism are typically high-level, top-down processing tasks. More specifically in the case of language learning, connectionist networks have in many experiments been trained to self-organise their own internal representations from pre-given human language input, which then is interpreted as human language again on 'the output side'.

An illustrative example of such a task is the one studied by Rumelhart and McClelland (1986) who devised a neural network to learn the past tense of English verbs. They emphasise that they did not want to build a language processor that could learn the past tense from sentences used in everyday settings, but a simple learning environment that could capture three stages evident in children learning English (Rumelhart & McClelland, 1986). In the first of these stages children know just a few verbs and tend to get tenses right, whereas in stage two they use the regular past tense correctly but get some of the

verbs wrong that they used correctly in stage one. Finally, in the third stage both regular and irregular forms are used correctly. The network manages fairly well in achieving the task of mimicking children's learning of the past tense, that is, it can learn without being given explicit rules. So now we have a system that, given a prepared input, can learn from examples to produce the correct output, but the designer provides the examples. Is there a way the network can seek out some context on its own? Context in language understanding might consist of the preceding words in a sentence, which in this next experiment is made available to the network.

Elman (1990) investigated the importance of temporal structure in connectionist networks, using letter and word sequences as input. The network consists of input, hidden and output units, but also of some additional units at the input level called context units. These context units receive as input, in each time step, the activation from the hidden layer in the previous time step. In the following cycle, activation is fed to the hidden layer from input units *and* context units, which accordingly contains the values from the time step before. This technique provides the network with short-term memory, and consequently, context. The results from two experiments show that the network extracts 'words' from letter sequences and lexical classes from word sequences. Only the second experiment will be discussed here.

In the experiment, the network was presented with 10,000 2- and 3-word sentences with verbs and nouns, for example: "boy move girl eat bread dog move mouse…", with more than 27,000 words in a sequence. An analysis of the hidden units shows that the network extracts the lexical categories verbs and nouns, and also divides them into sub-classes, more like semantic categories than lexical. Nouns, for instance, are grouped into animates and inanimates, and animates into humans and animals, depending on the context in which they typically appear ('bread', for instance, appears after 'eat' but not before). Furthermore, the significance of context is demonstrated when the word "man" is substituted in all instances by a nonsense word, "zog". The network categorises the two words very similarly, and the internal representation for "zog" has the same relationships to the other words as "man" did in the earlier test, even though the new word has an input representation that the network has not been trained on. This shows

that the provided context lets the network, by itself, ‘draw out’ some categories that are very close to being syntactical and semantic.

Elman’s results add an important feature to earlier attempts in traditional AI; the representations are augmented by context. One must remember, however, that in Elman’s work (for example) there is some implicit structure in the way sentences are formed before being input, not to mention how the words are actually represented, which may help the network somewhat. The performance of a model will depend to a great deal on the input and output representations. The internal representations (symbols) are self-organising patterns of activation across hidden units, while the input and output is greatly influenced by the designer. Still, assuming we have achieved some measure of self-organisation, there is still the problem of representation grounding, to provide the system with meaning or understanding of the input. That means that although Elman’s network could be said to have learned that ‘man’ and ‘zog’ are synonymous, it certainly does not know what a man or a zog actually is.

But, in addition to naming this predicament (the symbol grounding problem), Harnad (1990) also suggests a hybrid model to solve it. The model consists of a combination of connectionism to discriminate and identify inputs, and traditional symbol manipulation in order to allow propositions to be manipulated using rules as well as being interpretable semantically. Let us further, just for the sake of argument, suppose that this would solve the grounding problem, is this not sufficient to create communicative agents? Not quite, since the agents are not in the world and have no way of perceiving it in way of a body, and hence the designer is *still* the system’s only context. Elman appreciates this challenge, and comments: ”…the network has much less information to work with than is available to real language learners” (1990, p. 201), and goes on to suggest an embedding of the linguistic task in an environment.

The lack of embodiment and situatedness is a common criticism of connectionist models, as undeniably also of the GOFAI paradigm before them. In systems such as ELIZA or the scripts, the hardest problems are already solved, which are, as Haugeland (1985) points out, to interface symbolic and nonsymbolic objects. This criticism applies to connectionist systems as well, and to interface the symbolic and nonsymbolic we need

perception and action, but since these systems already ‘live’ in symbolic worlds, there is no problem. To “make sense” (Haugeland, 1985) of the world we need to consider the ‘brain’, body, and environment as a whole, cf. Clark (1997) where our perceptions of and actions in the world are seen as inseparable parts of a learning system. Perception is not a process in which environmental data is passively gathered, but seems to be closely coupled to specific action routines. Clark (1997) exemplifies this by citing research where infants crawling down slopes learn to avoid the ones that are too steep (Thelen & Smith, 1994), but seem to lose this ability when they start to walk, and have to relearn it. Thus, the learning is not general but context-specific, and the infant’s own body and its capabilities are ‘part’ of the context. This is why neither traditional AI systems nor connectionist systems ‘understand’, or grasp the meaning of a symbol, since they are not in a situation and do not have a body with which to interact with an environment.

A parallel critical line of reasoning against GOFAI, which may be interpreted as relevant critique against connectionism as well, is taken by Dreyfus (1979). He states that a major dilemma for traditional AI systems is that they are not “always-already-in-a-situation”. Dreyfus further argues that even if all human knowledge is represented, including knowledge of all possible situations, it is represented from the outside, and the program “isn’t situated *in* any of them, and it may be impossible for the program to behave as if it were”. Analogous to this is Brooks (1991a) claiming that the abstraction process of forming a simple description (for input purposes, for example) and ‘removing’ most of the details is the essence of intelligence, and dubs this use of abstraction “abstraction as a dangerous weapon”.

Hence, traditional AI and connectionism have dealt with, at least where natural language is concerned, programs that are not situated – in situations chosen by humans. These two facts in combination: the human designer choosing how to conceptualise the objects in the world for the system to use, including the abstracting away of ‘irrelevant’ details, and the system’s not “being there” lead to exactly the dilemma shown in Figure 3. A system that has no connection to the real world except via the ‘mind’ of the designer will not have the necessary context for a system trying to ‘understand’ language, or engaging in any other cognitive process for that matter.

Parisi (1997) comments that classical connectionism (as opposed to A-Life using ANNs) views language as a phenomenon taking place inside an individual, and further accuses connectionism of not being radical enough in posing their research questions, more or less using the same agenda as traditional AI (cf. also Clark, 1997). It is evident that connectionism also makes the mistake depicted in Figure 3, by letting a designer prepare the input to the system, even if the internal representations are self-organised. How, then, should the agents actually connect to the world? Cangelosi (2000), among others, claims that the link needed between the symbols used in the models and their (semantic) referents in the environment, is sensori-motor grounding. If there is no sensori-motor grounding, this will diminish the possibility of understanding the evolution of cognition, and hence communication.

Brooks suggests a way via the "Physical Grounding Hypothesis": to build an intelligent system it is necessary to have its representations grounded in the physical world, which is accomplished by connecting it via a set of sensors and actuators (Brooks, 1990). The approach is called behaviour-based AI, as opposed to knowledge-based AI, as traditional AI has been named (e.g. Steels, 1995). This calls for "embodiment" and "situatedness", central concepts for the emerging fields in the study of mind. Brooks (1991a) indeed claims that deliberate (symbolic) reasoning is unnecessary and inappropriate for interaction with world, and representations are useless, since "the world is its own best model".

There are others who are not convinced that we can make do completely without representations. Dorffner (1997) proposes "radical connectionism" as a way to stay with connectionism, but do away with the designer (well, almost). The approach is very similar to behaviour-based AI, but also requires agents to be adaptive, and acquire their own view of the world. This will make them able to 'survive' even if there temporarily is no input from the world, which Brooks' agents will not, since their world is not only the best model but their *only* model.

Finally, traditional AI and connectionist systems are passive and goal-less; that is, they have no actual 'reason for living'. This is a principal issue with, among others, Brooks (1990, 1991, 1991b), who asserts that an agent must "have an agenda". If we want the

agent to be autonomous, it must have some goal, so that when choosing what to do next the action must be meaningful (Franklin, 1995).

Before summing up the criticism, there is one more important point to be made. Steels (1997b) reminds us that linguistics and cognitive science focus on single speakers/hearers, which is also true of traditional AI and connectionism. For communication to take place, it is at least intuitively defendable to argue that there should be someone to communicate *with*.

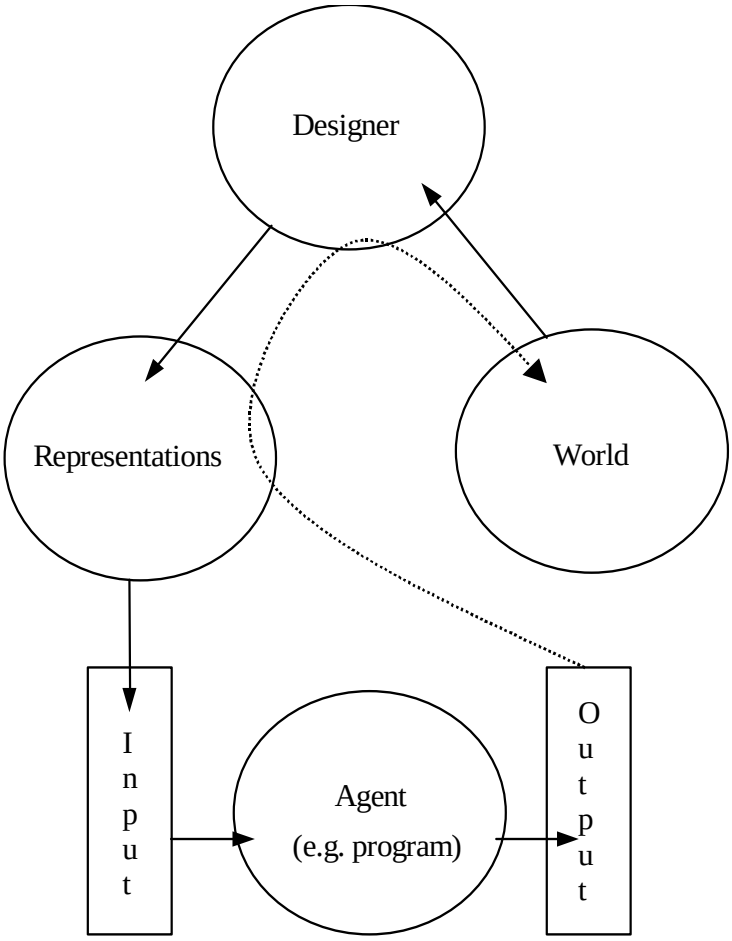

Figure 4: Designer representations revisited – how agents and their representations are related to the world by the designer in traditional AI and connectionism.

To recapitulate the critical observations from the two previous sections, dealing with both traditional AI and connectionism, we will use Figure 4. Starting with the world, at middle right, the tasks are usually concerned with *human language*, and situations or objects from the real world are conceptualised into a *representation by the designer*.

This representation is used as input to an agent and is *neither symbolically nor physically grounded*. Furthermore, the *agent is alone*, and the primary emphasis is on the *processes inside the individual*. It must be granted that some self-organisation takes place in connectionist systems. Since the agent is *not situated, is passive and without subjective goals* it also follows that it is not obliged to adapt or manage in any way, and thus there is *no autonomy*. When an output is produced, the *designer observes and interprets* (dotted line), usually using the representations as 'templates' to assert that something has been achieved (by the agent) that has some bearing on or in the real world. When we attribute "cognition" to artefacts by metaphor or analogy we extend our own intentionality, defined as the feature of mental states by which they are directed at or are about objects and states of affairs in the real world (Searle, 1980). Intentionality is in this case, instead of being an intrinsic capacity, only attributed to the agent by the designer. The next section will introduce the field of A-Life, to prepare for the review of models of communication that follow.

2.3 The A-Life approach to the study of communication

There are approaches where many of the issues criticised in section 2.1 and 2.2 are being tackled. What is more, the AI community is glancing this way. Steels (1995) describes this as AI researchers acknowledging the importance of embodied intelligence, combined with an interest in biology and A-Life research. Now we are ready to consider whether or not or not A-Life and related fields of study can provide us with the means to meet the requirements indicated by the title: self-organised communication in autonomous agents, with designer influence reduced to a minimum.

A-Life can be described as an approach using computational modelling and evolutionary computation techniques (Steels, 1997b), and in this particular case: to study the emergence of communication. Within the areas that make use of autonomous agents, sub-domains can be found. Two of these sub-domains are: 'pure' A-Life research, mostly concerned with software models, and robotics where actual embodied robots are used, in turn divided into cognitive, behaviour-oriented, or evolutionary robotics. Both of these approaches will be examined more closely in the next chapter. When used in communication research, these disciplines frequently, but by no means always, allow for self-organisation of both external language and its internal representations. As an

introduction to the field of A-Life we will look at two research statements made by well-known researchers.

MacLennan (1991) defines his area of study in the following fashion: A complete understanding of communication, language, intentionality and related mental phenomena is not possible in the foreseeable future, due to the complexities of natural life in its natural environment. A better approach than controlled experiments and discovering general laws is *synthetic ethology*; the study of synthetic life forms in a synthetic world to which they have become coupled through evolution. A few years later, MacLennan and Burghardt (1994) clarify the issue further by stating that when A-Life is used to study behavioural and social phenomena, closely coupled to an environment, then it is essentially the same as synthetic ethology.

Cliff (1991), on the other hand, concentrates on computational neuroethology. Although not explicitly mentioning communication, he refers to expressing output of simulated nervous systems as observable behaviour. He argues that the 'un-groundedness' of connectionism is replaced with (simulated) situatedness, which also automatically grounds the semantics. Meaning is supplied by embedding the network model within simulated environments, which provides feedback from motor output to sensory input without human intervention, thereby eliminating the human-in-the-loop. By removing the human from the process, the semantics are supposedly well grounded, which also Franklin (1995) believes, but Searle would not necessarily agree to.

We have now reached the lower right corner of the image (Figure 1) of different approaches to the study of mind. The two above characterisations of "mechanisms of mind", as denoted by Franklin (1995), i.e. synthetic ethology and computational neuroethology, are two sides of the same coin. Both are dedicated to artificial life, but from different viewpoints. The distinction is whether or not the approach is synthetic or virtual, which Harnad (1994) points out has consequences for the interpretation of results. Virtual, purely computational simulations are symbol systems that are systematically interpretable *as if* they were alive, whereas synthetic life, on the other hand, could be considered to be alive. MacLennan (1991) compromises on this issue by declaring that communication in his synthetic world is occurring for real, but not that the

agents are alive. Harnad's distinction between virtual and synthetic is not present in, for example, Franklin (1995) where both A-Life and robotics are regarded as bottom-up synthetic approaches.

Key features of A-Life are emergentism, evolution and goals according to Elman (1998) and for robotics embodiment and situatedness (e.g. Brooks, 1990). The last two concepts have been clarified in previous sections, and goals have been briefly discussed, but the notions of emergentism and evolution warrant an explanation. Emergence is usually defined as some effect that will arise from the interaction of all the parts of a system, or perhaps the "degree of surprise" in the researcher (Ronald, Sipper & Capcarrère, 1999). It is somewhat related to self-organisation, defined in section 1.1. This reconnects to the enactive view (Varela, Thompson & Rosch, 1991), where being in a world and having a history of interaction within it is essential; a "structural coupling" between agent and environment is emphasised.

To allow self-organisation to span generations and not just a lifetime as in the case of connectionism, evolutionary algorithms are frequently used, modelled on the processes assumed to be at work in the evolution of natural systems. At this point the radical connectionism proposed by Dorffner (1997) and the enactive paradigm have separate views, since Dorffner wants to leave some of the "pre-wiring" to the scientist, and Varela *et al.* (Varela, Thompson & Rosch, 1991) wish to eliminate the designer even further by using evolution. This issue concerns the degree of design put into a model, where a first step would be explicit design; midway we find adaptation and self-organisation where a trait develops over a lifetime, and finally there is pure evolution. The last two can be synthesised into genetic assimilation (Steels, 1997b), where there is both evolution and individual adaptation.

It is difficult to ascertain to what degree evolution and learning are combined in the real world. Therefore we do not know what the 'most natural' step is. But we need not be overly concerned with being natural as commented before, also well put by de Jong (2000) who declares that we should not exclude research that develops methods that are not present in world, they can be useful nonetheless.

This chapter has summarised important work within traditional AI and connectionism in order to see where earlier approaches have gone wrong and right. Further the idea of artificial life has been introduced, an approach with, at least at a first glance, great prospects for producing self-organised communication, even if the full potential has not been realised yet. The following chapter is a review of a large part of the work studying the self-organisation of communication within this approach.

3 A-Life models of communication

This chapter reviews a number of models and experiments aimed at achieving communication between autonomous agents in various ways. Some experiments assume that a language is already in place and the agent's task is to co-operate by using it. Other models start from scratch completely, with no lexicon, and set out to evolve a language by genetic or cultural means. A few experiments use learning only in an agent's lifetime as the mechanism to self-organise language, while others combine genetic approaches with learning, and some use cultural transmission, i.e. through artefacts created by the agents (e.g. a 'book'). This diversity is probably at least partly due to the inability to reach a consensus as to whether or not human language is continuous with animal communication, or rather: what are the origins of language? (cf. e.g. Hauser, 1996 and Pinker, 1994). Moreover, finding a way of categorising existing experiments is difficult, since they often are unrelated and use very different approaches, often tailored to the need at hand, which is why there is not much coherence in the field as yet. Therefore, in the first three sections, three state-of-the-art approaches to self-organised communication will be discussed. These three 'research groups' are considered state of the art since they have, over several years, more or less systematically conducted research in the area of communication or language by using autonomous agents in the form of simulations or robots, and as such are central to the focus of this dissertation. Section 3.4 then deals with a large part of the remaining work performed in this area, both early and contemporary.

3.1 Cangelosi and co-workers

Cangelosi has, together with several co-workers, explored the evolution of language and communication in populations of neural networks from several perspectives, e.g. the meaning of categories and words (Parisi, Denaro & Cangelosi, 1996), and the parallel evolution of producing and understanding signals (Cangelosi & Parisi, 1998). In later models symbol combination has been simulated, as a first attempt to synthesise simple syntactic rules, e.g. (Cangelosi, 1999, 2000). Cangelosi has also pursued a different direction of study, together with Greco and Harnad, in which the authors use neural

networks to simulate categorisation, symbol grounding and grounding transfer between categories, e.g. (Cangelosi, Greco & Harnad, 2000; Cangelosi & Harnad, 2000).

3.1.1 Cangelosi and Parisi, 1998

Inspired by the animal kingdom, where many species communicate information about food location and quality, this scenario simulated communication concerning the quality of a perceived food that might be edible or poisonous. A motivation behind the experiments was to explore the functional aspect of communication as an aid to categorisation as well as language evolution and its possible relevance to the evolution of cognition.

The environment and the agents:

The agents moved around in a grid-world environment, 20 x 20 cells, which contained two types of objects (denoted as edible and poisonous mushrooms). Each agent consisted of a neural network (Figure 5), with slightly different input and output depending on if the agent was a speaker or a listener.

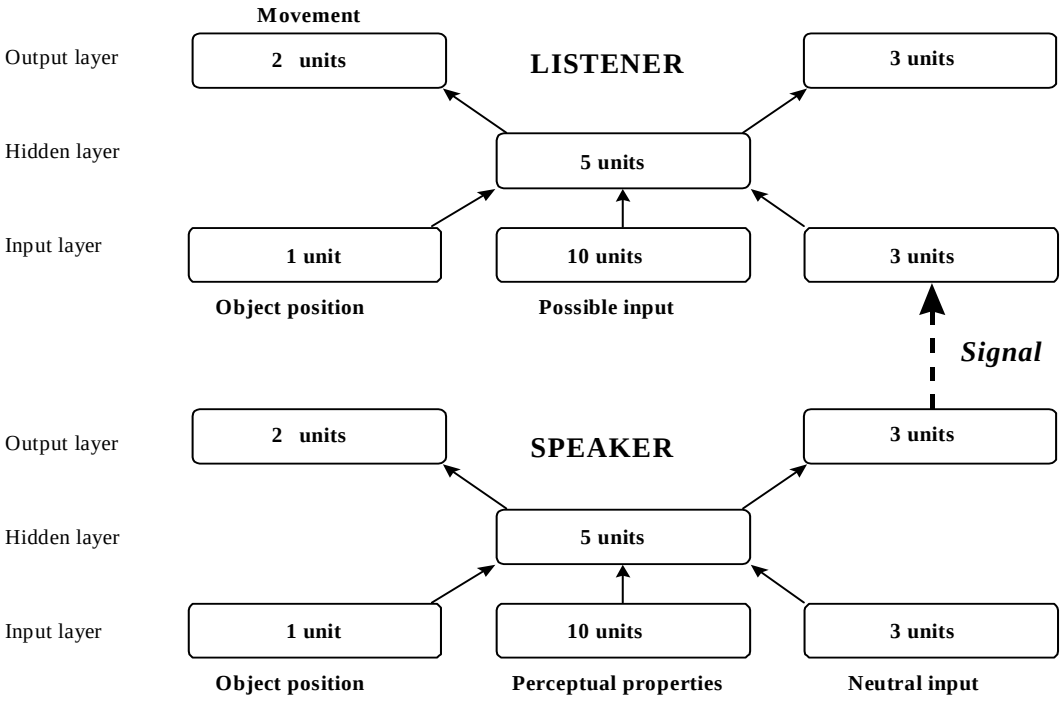

Figure 5: The neural nets used for the agents, adapted from Cangelosi & Parisi (1998).

A mushroom position (the closest) was used as input to one unit, ten input units encoded the perceptual properties of the mushrooms and the remaining three input units were used for signal receiving. There were five hidden units, two output units encoding movement, and the three remaining output units were used for signal production. The 'perceptual' input and output was binary, and the perceptual patterns within a mushroom category were slightly different in appearance, 1-bit variations on the theme 11111 00000 for edible and 00000 11111 for poisonous, which means that the difference between categories was larger than within. The position input was an angle mapped between 0 and 1, and the signals were binary (8 variations in total).

Experimental setup:

A population consisted of 100 agents, each with zero energy at 'birth', which then increased by 10 units when an edible mushroom was 'eaten' and decreased 11 points for poisonous ones. The angle indicating the position of the closest mushroom was always perceived, but the mushrooms' perceptual properties could only be seen if they were in one of the eight cells adjacent to the agent, which meant agents had to approach mushrooms to see them. The life span was the same for all agents, and reproduction depended on the amount of energy accumulated, which in turn depended on classifying objects correctly, so at the end of their life, the agents were ranked according to their energy, and the 20 best generated five offspring each. The offspring were clones of their (one) parent except for a mutation on 10% of their connection weights. The experiment ran for 1000 generations, in which behaviour emerged where agents approached and ate the edible mushrooms and avoided the poisonous ones. The important issue was the communication, however.

Method of self-organisation:

The language was transmitted by evolution only, since good discriminators survived to reproduce.

Experiments:

Three different experiments were compared; the first had a silent population, the second used an externally provided (designer) language and the third autonomously evolved a language. In the 'silent' simulation the input to the signal-receiving units was kept at 0.5

and the signal output was ignored. The agent had to approach the mushroom to be able to classify it. In the second simulation, the language did not evolve and was provided by the researchers, using the signal '100' for edible and '010' for poisonous. The third population evolved its own language, by using speakers and listeners. For each cycle, a listener was situated in the environment and perceived the position of the closest mushroom (and perhaps its perceptual properties). Another agent was randomly selected from the current generation, was given the perceptual input of the closest mushroom, irrespective of its distance, and then acted as speaker; its signal was used as input for the listener. In simulations 2 and 3 the listener could use the signal as help in categorising a mushroom it saw, or as a substitute for seeing, if the mushroom was too far away.

Results:

Three principally interesting results are reported: at the end of simulations the agents in simulation two and three have a higher amount of energy than in the first (without language); the evolved language in simulation three is generally shared by the whole population and the silent population evolves a language even though their output is ignored.

3.1.2 Parisi, Denaro and Cangelosi, 1996

To examine the hypothesis that categories and word meanings are not well-defined single entities, the same environment as in Cangelosi and Parisi (Section 3.1.1) is used. The neural net is similar, but not identical, in that it has two levels of internal units: level 1 where perception is input and level 2 where the position is sent directly from the input layer. The agents are said to categorise implicitly – when the response is 'the same' at some high level of description but the actual behaviour varies with circumstance, e.g. 'approach' and 'avoid' are the same at the higher level but the succession of outputs is almost never identical. The conclusion is that: "Categories are virtual collections of activation patterns with properties that allow the organism to respond adaptively to environmental input".

3.1.3 Cangelosi, 1999

These models are to be seen as a first approach to the evolution of syntax, by evolving word combinations, and the task was influenced by ape communication experiments (e.g. Savage-Rumbaugh & Rumbaugh, 1978).

The environment and the agents:

The environment consisted of a 100x100 cell grid, containing 1200 mushrooms. There were six categories of mushrooms; three edible (big, medium, small) and the same for toadstools, and thus there were 200 per category placed in the grid. The agents were 3-layer, feed-forward neural networks, similar to the ones used in 1998 (Section 3.1.1), but with some modifications, see Figure 6. The input layer had 29 units, divided into: 3 units for location, 18 for feature detection, and 8 for naming. There were 5 hidden units and the output layer consisted of 3 units for movement and identification, and 8 units to encode mushroom names. The symbolically labelled output units were divided into two clusters of winner-take-all units (2+6) which meant that only one unit per cluster could be active, and hence the names would consist of two symbols.

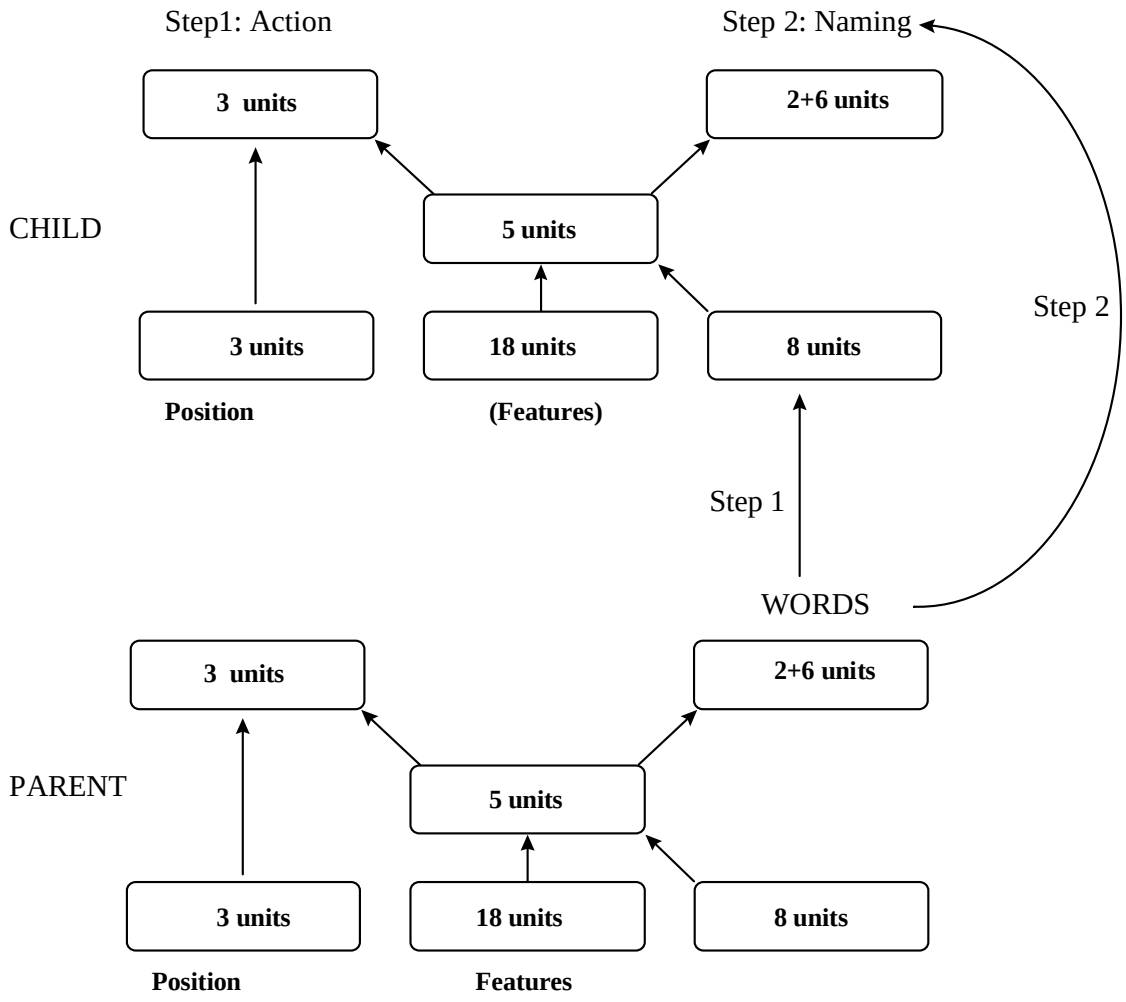

Figure 6: Parent and child neural nets, adapted from Cangelosi (1999).

Experimental setup:

Once an edible mushroom was approached, the size had to be identified to gain fitness (1 point per correct categorisation). No classification of toadstools was required, since they should be avoided. The input was a binary string of 18 features, where three bits always were set to one, the same three within a category, which acted as a common prototype for the categories. Identification was based on the level of activation on one output unit. The simulation involved 80 agents with a foraging task that in a first stage (300 generations) learned to differentiate the 6 types of mushroom without communicating. At the end of their lifetime, the 20 fittest agents reproduced and had 4 offspring each. In generation 301-400 the 80 new agents lived together with their 20 parents, and only the offspring foraged and reproduced.

Method of self-organisation:

In one time interval a child did two things; listened to the parent's symbols to decide on a categorisation/action (step 1 Figure 6), then performed a naming task (step 2 Figure 6), followed by an imitation cycle, (cf. Cangelosi, Greco & Harnad, 2000, section 3.1.4). Backpropagation was used for both naming and imitation tasks, and the 8 linguistic units were used by the parents who acted as language teachers to simulate a process of cultural transmission. During each action the parents received an 18-bit feature input and produced 2 output symbols, which then was used as input to children with random noise added to introduce some variability in the process of cultural transmission. 10 percent of the time the children also received the 18-bit input, to "facilitate the evolution of good languages as the availability of the mushroom features is rare".

Results:

The results show that after repeating the first stage (1-300 generations) for 10 populations, 9 were found to be optimal in avoiding/approaching-identifying. The 9 successful populations were then used in the second stage, in a total of 18 simulations (9 populations x 2 initial random lexicons). In 11 of the 18 runs, good languages evolve (table 1) where 'good' is defined as using at least four words or word combinations to distinguish the four essential behavioural categories (all toadstools, three edibles). The remaining 7 languages were poor due to incorrect labelling.

Table 1: The results of the experiments in Cangelosi (1999)

	Single word	**Word combination**	**Verb-object**	**Total**
Good languages	1	3	7	11
Imperfect languages	1	2	4	7

The network's symbol output clusters were used in different ways by the agents, which resulted in one population using only the four units in the first (6-unit) cluster, and still having optimal behaviour. Since the focus was declared to be on evolving word-combination rules that resemble known syntactical structures, such as the verb-object rule, populations that used combinations of the two clusters were considered more interesting. The remaining 10 of the 11 populations did use combinations of symbols, where 3 of the 10 used various two-word combinations and 7 used verb-object rules. Looking at the way the output clusters were used identifies the verb-object rule. In the two-word cluster, one 'verb' symbol was used as "avoid" for toadstools and the other as "approach" for mushrooms, and the units in the 6-word cluster were used to distinguish single objects (categories).

3.1.4 Cangelosi, Greco and Harnad, 2000

Cangelosi, Greco and Harnad (2000) then continued the work on evolving the ability to communicate and examining its supposed relation to categorisation. They wanted to examine the grounding of symbols in the real world, and suggested using a hybrid symbolic/non-symbolic approach (Harnad, 1990, cf. Section 2.1). The motivation of the model was that an agent can learn to categorise low-level categories by trial and error (toil), and then learn high-level categories in one of two ways; by 'toil' as before or by 'theft' which means that the already grounded category names are combined into propositions describing the higher-order category. This leads to a grounding transfer from the grounded categories to higher-order categories that are not grounded by the agent. The model was concerned with the implications for the evolution and learning of (human) language.

The environment and the agent:

The environment is of a set of symbols, shown to the agent in a training series. There are four low-level categories to be learnt: circle, ellipse, square, and rectangle, and two

higher-level categories: symmetric and asymmetric. An agent is a feed-forward neural net with 49 input units (a 7x7 “retina”), and 6 units for input of each of the six category names, a localist encoding. The hidden layer has five units, and the output is the same as the input layer.

Experimental setup:

A ‘population’ consists of 10 nets with random initial weights. The task consists of first learning the names of the four low-level categories by seeing all the instances of them, using supervised learning (toil), then learning the higher-order categories by being exposed to symbol strings only (theft). The agents are not communicating.

Method of self-organisation:

An agent learns by receiving error-correcting feedback (back-propagation) in all instances of the experiment, augmented by an imitation task in the stages where names were used, which meant that every naming trial was followed by an extra activation cycle to practice the name learned in the preceding cycle.

Results:

The results show that all 10 nets learn the three tasks. The categorisation was correct, and the naming task was also successful.

3.1.5 Cangelosi and Harnad, 2000

The authors hypothesise that a basis of the adaptive advantage of language is that new categories can be acquired through ”hearsay” (symbolic theft) instead of learning the new categories in real time (sensori-motor toil). It is an extension of the previous experiment (cf. Section 3.1.4), using agents situated in an environment.

The environment and the agents:

A 400-cell (20x20) environment is used. There are 40 mushrooms, divided into four categories. An agent can move around, perceive the closest mushroom, its features and its location, and ‘vocalise’. The mushroom features are binary, localist representations, as are the calls and actions. The net now looks like those of Parisi, Denaro & Cangelosi

(Section 3.1.2), except only one hidden layer is used, i.e. position is sent directly to the movement output without passing an internal layer.

Experimental setup and results:

Agents go through two life-stages; first all agents learn basic behaviours, and later to vocalise. There are two types of agent, the toiler and the thief. The toiler learns to categorise by trial and error, whereas the thief listens to the vocalisations by the toilers and uses them to categorise the mushrooms, thereby saving time and effort. The thieves are more successful than the toilers, and gradually outnumber them. This is only true when there is lots of food, however, which is discussed further in section 4.1.2.

3.2 Steels and the 'Origins of Language' Group

Steels and his group use the working hypothesis that language itself is an autonomous evolving adaptive system. The aim of their experiments is to get agents to use a self-organised language, but with all or most of the characteristics found in natural language, such as words grounded in the real world, synonymy and syntax. This is accomplished by self-organisation, which according to Steels does not include genetic evolution, using ”language games”, a Wittgenstein-inspired guessing game, where agents take turns in speaking or listening, and thus develop a common language. The agents have been both computer simulated and robots, and at present the most known project is the ”talking heads”, software agents that travel the Internet and are loaded into robot bodies at different installations. The agents evolve a shared language in interaction with each other and an environment, consisting mainly of various shapes in different colours, pasted onto boards for the agents to see (through cameras) and 'discuss', see Figure 7.

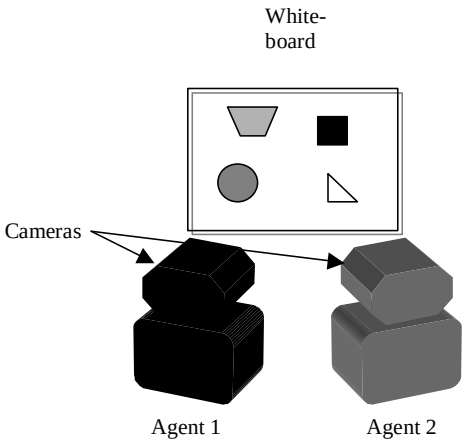

Figure 7: The “talking heads”, adapted from Steels & Kaplan (2000).

The review starts in 1996, with the simulation of the evolution of a shared adaptive lexicon (Steels, 1996a), and a self-organising spatial vocabulary (Steels, 1996b), continues to 1997 for agents to construct and share perceptual distinctions (Steels, 1997a), when he also publishes a general review on the synthetic modelling of language origins (Steels, 1997b), and reports on grounding language games in robots (Steels & Vogt, 1997). Work that entirely focuses on the evolution of syntax will not be considered here, so we continue to 1999 and 2000 when reports of the integration of many of the previous results into the talking heads experiment are published (Steels & Kaplan, 1999; Steels & Kaplan, 2000).

3.2.1 Steels, 1996a

The environment and the agents:

Assume a set of agents each of which had a set of features, see Figure 8. A feature was a pair with attributes and values, e.g., size {tall, small, medium}. A word was a sequence of letters from a finite shared alphabet, an utterance was a set of words, and word order was not important. A lexicon was a relation between feature sets and words. Coherence was achieved through self-organisation in the following way: Agents randomly coupled words to meanings, and engaged in communication. The environment consisted of the other agents, which then constituted the topic. If a word-meaning pair was successful it was preferably used in future communications.

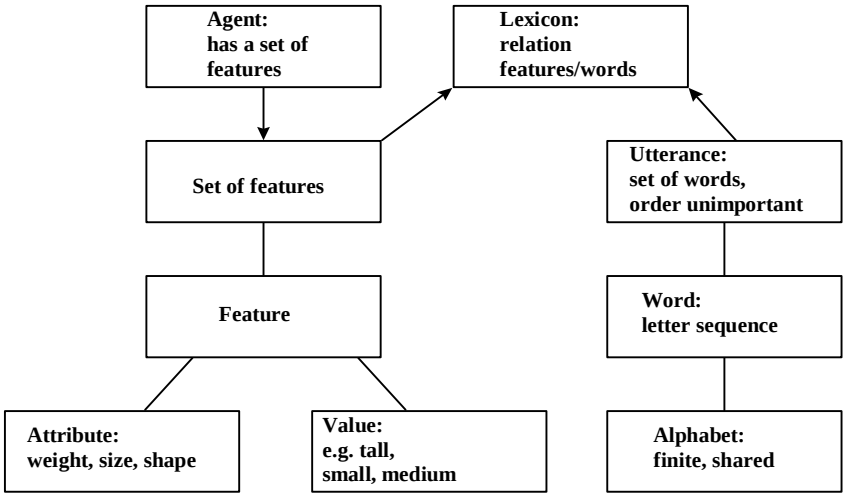

Figure 8: Relations between concepts used in Steels' models

The language game progressed as follows:

1. Making contact: A speaker and a hearer were randomly selected from a population of typically 4 or 5 agents. The remaining agents formed the 'context' from which an agent (the topic) was chosen.
2. Topic identification: The speaker selected which of the other agents would be the topic and indicated this one to the hearer.
3. Perception: Both identified which feature sets might make the topic distinctive from the others.
4. Encoding: The speaker selected one feature set and translated it into words.
5. Decoding: The hearer interpreted the word and compared it with his expectations.

What were the possible results?

1. The speaker did not have a word. If one feature distinguished the topic, but the speaker did not have this word, he could create it, and it would be used the next time.
2. The hearer did not have a word. He had a hypothesis about which features could be used, and created an association, which might be ambiguous.
3. Both knew the word:

   3a. The meanings were compatible with situation. The dialogue was a success, but there could still be incoherence. If they used different

feature sets, it would be resolved when new distinctions became important.

3b. The meanings were not compatible with the situation. The features heard by the hearer were not the ones expected, which meant no communicative success.

Results:

The results show that after a dozen conversations, the first word has been created and is used consistently by all agents. Ambiguities do appear, but are resolved, and if an agent knows two meanings, it uses the one most successful in the past. After about 4000 language games all the needed distinctions are in the lexicon. Multiple word utterances emerge when the feature set is expanded, i.e. new features (e.g. colour) are added. When a new agent is introduced, it learns the existing lexicon, but is also in itself a source of novelty, because it introduces new sets of features. "[N]o agent shares exactly the same language but the global system nevertheless manages to achieve quasi-total communicative success" is a comment by Steels, who then concludes by saying that self-organisation is effective for achieving a coherent lexicon, and as a side effect, many properties of natural languages occur, e.g., synonymy, ambiguity and multiple-word sentences.

3.2.2 Steels, 1996b

A set of experiments was conducted to show that agents could develop a vocabulary from scratch, with which they identify each other using names and spatial descriptions.

The environment and the agents:

The agents were located in a 2-dimensional grid, where they could move around. Their positions were given relative to the agents themselves, and consisted of front, side, and behind, and to further clarify the orientations the aspects left, right, and straight were used. The language game was used again, and the resulting language was a set of one-to-one mappings between meanings and words, where a meaning was either an orientation aspect or a name for an agent, and a word was a combination of letters.

Experimental setup:

A speaker spoke to another agent, referring to yet another agent (which may well be itself or the listener, as well as one of the others) using either pointing followed by naming, or giving a set of spatial descriptions, using itself as a reference point. In the following dialogue, the two agents communicated until they reached a common understanding, possibly by iterations to reduce the set of possible meanings until a consensus was reached. For the variant when spatial descriptions were used, the agents had to have at least a partially shared vocabulary of names for each other.

Results:

After 200 conversations, only a few agents are starting to agree on a name for *one* of the others (in this particular experiment, five agents were positioned in a 10x10 grid), and no spatial relations. At conversation 500, there is more coherence. There is consensus on one name and spatial relations are beginning to be used, and dialogs based on naming are mostly successful. At conversation 1000, the vocabulary is almost complete, and spatial descriptions are common. When a new agent is added, it adopts most of the vocabulary in about 300 conversations.

In order to not only construct a self-organised lexicon, but also the distinctions underlying it, the experiments that follow used a discrimination game to let agents construct their own distinctions instead of learning by examples, modelled in software simulations, and hence the agents self-organised both their view of the world and their vocabulary (Steels, 1997a), which was followed by an attempt to ground these simulations by porting them to Lego robots (cf. Section 3.2.4).

3.2.3 Steels, 1997a

To avoid giving the agents designer-conceptualised distinctions, Steels (1997a) describes how agents themselves can construct appropriate distinctions. He discusses situations where there may not be a teacher (designer) present to supply examples that will help an agent to learn a concept, and also argues that in a teaching scenario the intelligence might be considered to reside in the teacher and not originate in the learner. The construction of distinctions is accomplished by ”discrimination games” that an agent

plays by itself, followed by language games to achieve coherence in a group of agents, and has been tested in simulations as well as robots.

The environment and the agents:

The environment consisted of a set of objects or situations, sensed through sensory channels. A specific object might give rise to activation in some or all sensory channels. To each channel there was a discrimination tree (D-tree) connected, which mapped the sensory channel into discrete categories represented as features. A feature is, as before, an attribute-value pair, such that a path in the tree is the attribute and the value is the end point of the path, much like a decision tree that makes binary divisions only. The trees were constructed from scratch and depended on the task and the environment at the time. If all objects were the same size but had different colours, a colour-distinction tree would be appropriate. The discrimination game procedure was this:

1. A context was determined, which is a set of objects.
2. One object was chosen as the topic, to be discriminated from the other objects.
3. All available D-trees were used to construct the features for each object.
4. A distinctive feature set was computed, the minimal set needed to distinguish the topic.

If the last step failed due to insufficient features for discrimination, a random D-tree was picked and used, and if it turned out to be incorrect, it would be corrected in later games. On the other hand, if there was more than one possible set of features, a choice was made to minimise the set by firstly choosing the smallest set, and if there were equals, choosing the one with the most abstract features, and finally, if this turned out a draw as well, the set was chosen that contains the features most used.

Results:

When tested in software agents, it was shown that an agent does develop all discriminations necessary. Continuing to the language game, Steels points out some aspects that the agents already share, namely, they will discriminate similarly due to that the objects in the shared environment determine the features. But to enhance their common understanding of the world, a language game, similar to the one described in

section 3.2.1, is used. Some further details will enhance the understanding of the procedure, e.g. that the use and success in use of word-meaning pairs is recorded, and naturally the most used and most successful ones are preferred. The same relations as in Figure 8 above were true, and an agent started with an empty lexicon.

1. A speaker, a hearer, and a context (subset of objects) were randomly selected.
2. The speaker selected one object to be the topic and indicated this one to the hearer.
3. Both identified possible feature sets that might make the topic distinctive from the others, using their perceptions and the D-trees.
4. If there was one (or more) set/s of distinguishing features, the speaker selected one and encoded it into words.
5. The hearer decoded the expression and compared it with his expectations.

The results show that a vocabulary of word-meaning pairs develops alongside the feature repertoire. The system is open in 'all' respects, feature and language formation, as well as letting new agents in. Steels concludes by claiming that both the origins (constructing) and the acquisition (learning) of new concepts and words can be explained by the same mechanism. Mechanisms are selectionist but not genetic; operate on structures *inside* agents, which results in *cultural* evolution.

3.2.4 Steels and Vogt, 1997

Steels and Vogt (1997) report results from grounding experiments in Lego robots. The communication is assumed to be in place, and the focus lies on the grounding problem. The naming game is the same as above, but one feature is added – feedback, which means that the steps now (very briefly) were:

1. Making contact
2. Topic identification
3. Perception
4. Encoding
5. Decoding

6. Feedback

The environment and the agents:

The Lego robots were programmed with a behaviour-oriented architecture, and a radio link, which (when it delivered a message) guaranteed that there were no errors in the message, and thus no noise. There were several robots moving around freely as well as static objects in the environment.

Experimental setup:

The language game started when a robot “decided” to become a speaker when detecting another robot (by adopting that behaviour mode). The speaker sent a signal to the chosen listener and they aligned themselves so that they faced each other. To achieve a common view of the environment, both robots rotated on the spot, scanned the environment and stopped when facing each other again. The speaker then chose a random object from this ‘scanning session’ and turned toward it, indicating it to the listener (the speaker could actually *be* the topic, in which case it did not move). The rest of the game proceeded as in the software experiments, with the possibility of success or failure, except for the added feedback component.

Results:

The general success rate was about 75% and in the specific game reported, after 45 games the robots recognise *each other* as the topic 31% of the time, although recognition of other object was reported as ”still low” but increasing.

3.2.5 Steels and Kaplan, 1999, 2000

Findings from the ”talking heads” project are reported in Steels and Kaplan (1999, 2000). The focus is on single words and how they get their meaning, or ”semiotic dynamics” (there are also reports concerning syntax, but they are outside the scope of this dissertation). This is accomplished by using visually grounded agents (cf. Figure 7) that bootstrap their ontology and construct a shared lexicon with no prior design or human intervention. The project has run twice on the Internet, the second time was the summer 2000.

Two problems are identified: how the meaning of a word is acquired, and the grounding problem. According to Steels and Kaplan, simulated agents have often been assumed (by other researchers) to have access to each other's meaning, so it was just a question of learning the associations between word and meaning, and getting feedback on the success. Realistically, the authors remark, we get feedback on communicative success, not on meaning. This means that the agents must acquire word-meaning and meaning-object relations, that are compatible with the word-object situations they observe, see Figure 9, as word-meaning relations are not observable.

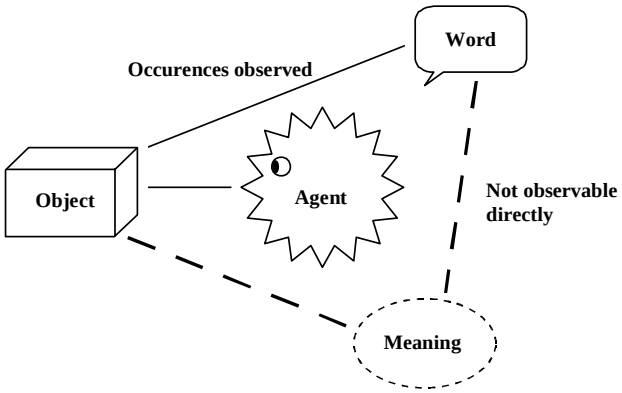

Figure 9: Agents acquire word-meaning and meaning-object relations, concepts used in Steels & Kaplan (2000).

According to Steels and Kaplan, the grounding problem tends to get harder with embodied and situated agents, since their perception depends on their viewpoint and thus also their categorisation[1]. There might, for example, be problems with deciding on whose left and right is the norm. The success of a language is supposed to depend on whether or not agents can abstract away from the "contingencies of viewpoints and situations".

The environment and the agents:

The environment used was a set of shapes in different colours and sizes, attached to boards (cf. Figure 7), and such boards were located at several laboratories around the world. The robotic part of the agent had three components; a camera to observe the environment, a computer for perception, categorisation and lexicon lookup, and a screen, which shows the 'internal state' of the agent, i.e., what it was currently observing

[1] Strictly speaking, these are the only agents that might be considered to have grounding at all.

through the camera. The software part of the agent could travel by ‘teleportation’ on the Internet to different locations and load itself into a physical ‘body’. The features were based on an object’s horizontal and vertical positions in the context, colours and greyscale, and a scale between 0 and 1 was used, e.g. [HPOS - 0.5,1.0] to express a horizontal position to the right of the centre (which is at position 0.5). There were two components to the agent architecture: a conceptualisation module, which categorised reality or found a referent in a category (applies categories), and a verbalisation module for saying a word or interpreting a form to reconstruct its meaning. The conceptualisation module used the discrimination trees mentioned earlier, where each node divided the set into two parts: one satisfying a category and another that satisfied its opposition. If there was more than one possible solution, all the solutions were passed on to the lexicon module. The verbalisation (lexicon) module contained two-way associations between words and meanings, and each association had a score. There was no central co-ordinator, since the modules were structurally coupled.

Experimental setup:

So, let the language game begin: one agent was the speaker, another the listener (all participants took turns), and both agents collected sensory data about the objects such as colours, grey-scale, and position. The set of objects and their data was the context, a chosen object was the topic, and the rest was background. The speaker started by uttering a word (a random combination of syllables), and the hearer tried to guess the topic. The words did not have to be directly translatable into human concepts, but associated with features, so if an agent wanted to identify a red square as the topic, it ”may say ‘malewina’ to mean [UPPER EXTREME-LEFT LOW-REDNESS]”. Back to the game: the hearer pointed to the object it thought was the correct one (by looking at it). If it was the same, there was success, if not the speaker would point to the correct topic as feedback, and both agents tried to correct their internal structures in order to be more successful in the future.

Results:

The results indicate that the dynamics of the guessing game leads to incoherence, but there are tendencies towards coherence. This means that four tendencies, also found in natural language, have turned up in the experiments, namely: *synonyms*, *homonyms*

(same word, different meaning), *multi-referentiality* (a meaning has different referents for different agents in the same context, e.g. left/right), and *categorical indeterminacy* (a referent in a context can be conceptualised in more than one way; e.g., to the left *and* much higher than all others). But as the authors assert (2000): ”The goal of the game is to find the referent. It does not matter whether or not the meanings are the same”, and continues clarifying that this will never be possible, since they have no access to the brain states of other agents. The main result is that agents *can* bootstrap a successful lexicon from scratch. The success drops temporarily when new objects are added, but regains, as new words are invented or new meanings created. The authors state that they have shown for the first time evolution of an open-ended set of meanings and words by autonomous distributed agents interacting with a physical environment through sensors, which also can be scaled up to more complex environments.

To recapitulate:

The agent was a software agent, loaded into a robotic ‘body’. All agents were the same, and had neither a supplied ontology nor a lexicon. The environment was a whiteboard with shapes on it, viewed through a camera. They interacted directly, linguistically, with random combinations of syllables. The agents ‘had to’ talk and listen, it was pure labelling for its own sake. There was no noise, and they interpreted according to the discrimination tree where a meaning or a word with the highest score was used more often. The discrimination trees grew and were pruned, word-category associations were formed by guessing, and if an agent wanted to express a meaning and had no word, it could create one. There was a success or failure message that lead to association scores being adjusted. The only learning mechanism was lifetime learning; there was no ‘genetic’ but a ‘cultural’ learning process.

3.3 Billard and Dautenhahn

Billard and Dautenhahn have, separately and together, using simulations and robots, explored communication from a socially based viewpoint. According to the authors, communication is an interactive process, which is embedded in a social context. Social behaviour is claimed to be desirable for artificial agent societies, as it is a key mechanism in the cohesion and evolution of primate societies. The role of social interaction for

achieving a shared context is especially emphasised, and hence, grounding is studied alongside the use of communication as such.

3.3.1 Billard and Dautenhahn, 1997

In this paper, by Billard and Dautenhahn (1997), learning to communicate was defined as a learner achieving a similar interpretation of the environment as a teacher, on basis of its own sensory-motor interactions. They used a non-supervised teaching approach, and achieved some positive results, and also showed the limitations of imitation as a means of learning, due to spatial displacement and temporal delay.

Billard and Dautenhahn point out that the teacher and the learner did not need to have the same shape or sensori-motor characteristics, since they did not need to use the same sensory data to characterise a 'word'. When communicating about an object for instance, one agent might use its position and the other its shape.

The environment and the agents:

The environment was a hill surrounded by a level surface, with a light positioned above the hill. The agents were two heterogeneous robots; the learner, using a built-in imitation behaviour, followed the teacher around in the environment. The teacher knew the words (bit-strings, no claims about human language) for useful places in the environment, and emitted them at regular intervals at certain pre-determined control points, transmitting via a radio link. The learner associated its sensor data with the simultaneous signal from teacher.

The system architecture used (DRAMA, Dynamical Recurrent Associative Memory Architecture) was a single cognitive architecture with associative learning, selective attention and creation of mutual binding between two agents (by phototaxis). DRAMA can learn to associate words with meaning but in this experiment only the associative memory part was used. Life-long learning was combined with built-in basic behaviours: two "desires", '*mother' need* and *internal equilibrium need*. The mother need was meant to imitate a 'survival' instinct and was satisfied when the learner saw the teacher (by receiving its infra-red signal), and the internal equilibrium need, which provided the motivation to learn and move, was satisfied when the real sensory input corresponded to

the desired. The mother need was privileged over the equilibrium need, such that about 80% of the energy available was used to satisfy it.

Experimental setup:

The experiments were performed in two parts; the first concerned learning a language by following, and the second finding the teacher. In the first experiment, the teacher emitted while the learner followed, making associations along the way, as illustrated in Figure 10. In the second, the agents were separated, and the learner had to find the teacher. It did this by random search combined with using the teachers ('mother') signal to locate it.

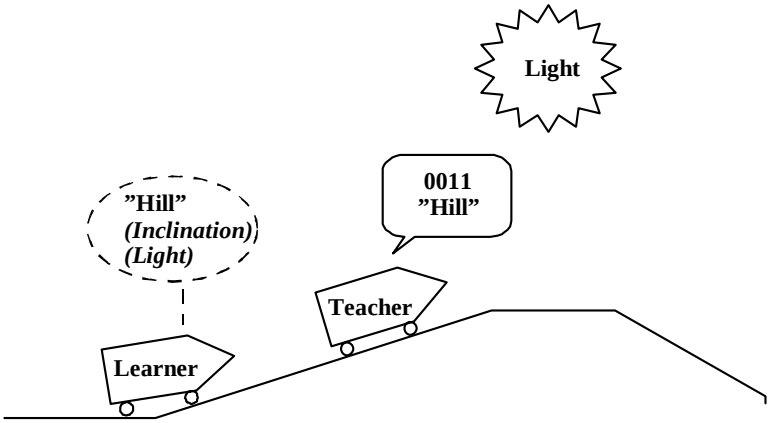

Figure 10: Robots using imitation learning, adapted from Billard and Dautenhahn (1997)

In the first experiment the learner followed the teacher, climbing up and down the hill 10 times. The teacher emitted signals, presumably continuously, which the learner heard and associated with its sensory information. In the second experiment ("mother-child") where the learner was to search for the teacher using learned associations and the teacher's signal, the setup was slightly changed. The relevant concepts here were 'hill' and 'plane', and the sensory stimuli to be associated were inclination and light (in the hill area), see Figure 11. The two agents climbed up and down the hill, until "…concepts ... and stimuli ... [were] well associated". When the learner later searches for the teacher so much energy is used for moving that it does not learn, which explains why it is not relearning even though it is hearing the signals.

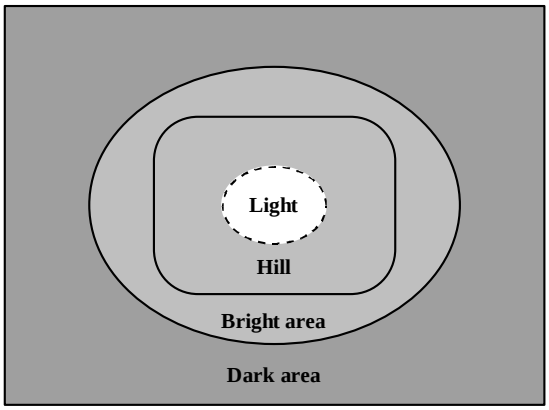

Figure 11: The environment, adapted from Billard and Dautenhahn (1997)

Results:

In the first experiment the results show that, due to spatial displacement and temporal delay, the learner most often associates a teaching signal for a position (e.g. down), with the preceding one (up). This is a consequence of the teacher being in front of the learner physically, which displaces the associated word-position by a 'robot-length'. Even if this is compensated for, the robots never obtain identical sensory input from the same position due to small shifts in their orientation, and thus the teaching also suffers from the time delay. In the second part of the experiment, 10 tests were performed, and the results show that the learner agent's reactions were correct. When it reaches the hill region, the (self-generated) desire for light is satisfied, but neither is the (self-generated) desire for inclination, nor the (built-in) desire to "find the teacher", which is why it slowly moves around the hill, which satisfies the need for light and inclination, until it finds the other agent.

Billard and Dautenhahn conclude by asserting that communication does give an agent insight into another agent's world, but since their perceptions are not identical the learning is not perfect.

3.3.2 Billard and Dautenhahn, 2000

Billard and Dautenhahn (2000) declare that the motivation is to study grounding and the use of communication in heterogeneous agents and in particular, the role of social interactions to develop a faster and better common understanding of the language, and to

speed up transmission of information. A teacher teaches word-signal pairs to several learners, using the imitation strategy.

The paper reports on three studies, all using a simulated environment, of which the first regards grounding communication among a group of robots, the second evaluates how imitation improves learning, and the third is a case study exploring how the communication system from the first simulation improves learning in another task. The first and third experiments will be reviewed here.

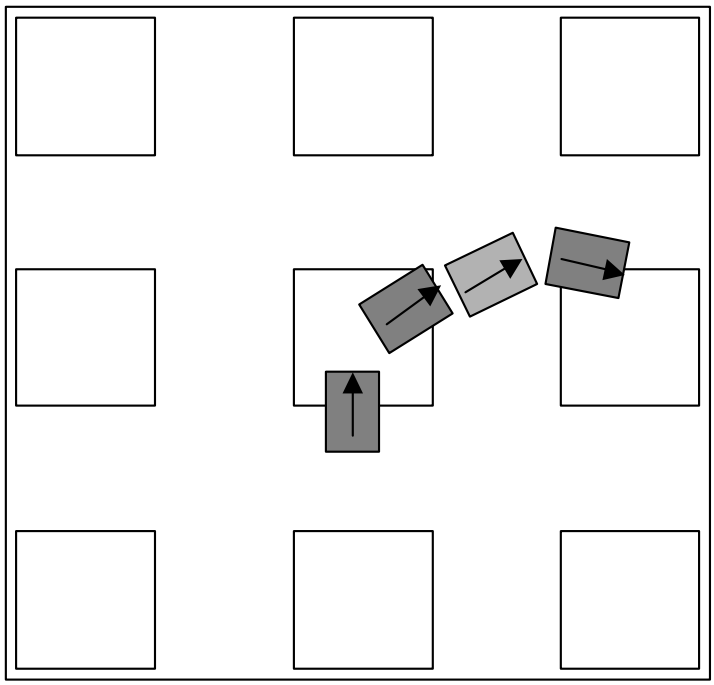

Figure 12: The simulated environment, adapted from (Billard & Dautenhahn, 2000). For the sake of clarity only four of the nine robots are shown. The arrows indicate the robots' heading.

The environment and the agents:

The environment was a 700x700 unit arena containing nine objects (coloured patches) distributed at regular intervals, see Figure 12. There was one teacher and eight learners, and the agents had common internal contexts (same movement, same distance, same ground), and external contexts (facing direction) due to the imitation behaviour. The vocabulary consisted of 31 words describing 9 colours, 14 distances, and 8 angles, in order to tell coloured patches apart and describe their locations. The DRAMA architecture, briefly described earlier, used a connectionist model for learning 'word'–

observation pairs, where associations were kept or discarded depending on how often they were used.

Experiment 1:

Only the nine colour words were used, and the vocabulary was transmitted from a teacher with full knowledge to the eight learners. Agents wandered randomly and followed each other, and the teacher sent radio signals ('words') describing the external perceptions (seeing a patch), repeating a signal 5 times during the passage of an object. A learner attached meaning to signals in terms of its own sensori-motor perceptions, and when it reached some degree of confidence, could become a teacher in turn. The teacher did not learn, but the previous learners continued learning even after they became teachers. Each robot could hear other robots at a distance of 1.5 x body size, which implies that a maximum of 4 robots could gather around a learner.

Experiment 3:

All agents knew the complete vocabulary consisting of 31 words, and wandered around, learning locations. When one agent had learnt a location it could transmit it to all others via the communication channel, which meant that an agent could learn a location without having been there. Two scenarios (inspired by honeybees) were used for information transmission: one-to-all when one agent found a patch, or one-to-one when two agents met.

Results:

The results of the first experiment show that the robots do not gather in groups, but rather form chains, so usually an agent is only taught by two other learners (one in front and one behind it). It is shown that at least 10 object passages are needed to learn a word, which corresponds to a confidence factor of 50. Below this value, learning was unsuccessful. Even so, 20% of the associations were incorrect, since there was a lot of noise due to bad learners becoming teachers to soon, using incorrect associations, and also the fact that only two agents teach another, which means that if they are bad teachers, 'bad language' will quickly spread in the population. In experiment 3, the results show that the group learn the locations faster when knowledge is transmitted, and especially so when the one-to-all strategy is used rather than the one-to-one.

In their discussion, the authors mention that the system can be scaled up to larger vocabulary (no-of-units^2), and that work is in progress, examining the use of sequences of inputs – ”sentences”, with the DRAMA architecture in a doll robot.

3.4 Other A-Life models

3.4.1 MacLennan and Burghardt, 1991, 1994, Noble and Cliff, 1996

MacLennan (1991) motivated his approach, synthetic ethology (cf. Section 2.3), by claiming that by synthesising an artificial world, it would be more amenable to scientific investigation than trying to analyse the natural world. A further claim was that communication was occurring for real, but this did not mean that the agents, the ”simorgs”, were alive. A central assumption MacLennan made is that the environment must permit some simorgs to ‘see’ things that others cannot, otherwise there would be no advantage of communicating.

The environment and the agents:

Each simorg had a local environment (Figure 13), in which the states were called *situations*. The medium for communication was a shared global environment, in which simorgs could produce, or perceive, a *symbol*. There could only be one symbol in the environment at the time. Simorgs survived if they interacted correctly with the states. An agent was a finite state machine (FSM), even though MacLennan commented that he would have preferred to use neural networks. Each machine was defined by a table that mapped symbol/situation pairs into responses.

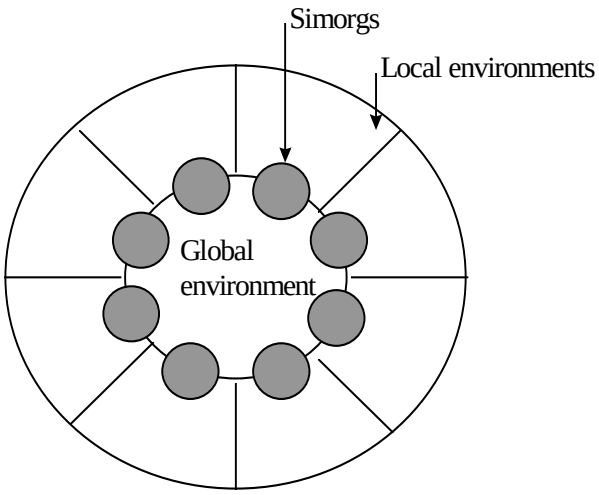

Figure 13: The agents and environment, adapted from MacLennan (1991).

Method of self-organisation:

The genotype was the transition table, and crossover (which did not break table entries) and a small amount of mutation were used. To this a simple kind of learning was added, which involved correcting the table after an ineffective act, so that it *would* have acted effectively.

Experimental setup:

When a simorg acted, the action was compared to the situation of the most recent emitter. If they matched, both were given a credit point. At the end of a cycle, the fitness value was used to select two agents to breed, and one to die, to keep the population constant. Overlapping generations were employed, so that the simorgs did not all reproduce/die at the same time, not as in a traditional genetic algorithm, which replaces entire generations, since this was said to prevent the passing of cultural information through learning. To explore the meaning of the symbols and how they were acquired, a denotation matrix was used, with an entry for each symbol/situation pair. The population size used was 100, and the number of local and global states was 8 each.

Results:

A run without communication and learning was used as a 'baseline' to denote guessing (a value of about 6), and the maximum fitness was assumed to be close to 90. A simulation with communication raised the fitness average to 11.5, which was further increased when learning was added - to about 50. When the denotation matrices were

analysed, they showed that in the absence of communication and learning, the matrix is very uniform. When communication was used, the matrix was more structured, but when learning was added, it became slightly less structured but the population had a higher fitness.

MacLennan and Burghardt (1994) examined whether or not complex signals could be evolved, with agents using memory. The agents had two bits of memory, which gave them a theoretical capability to emit coordinated pairs of symbols on successive action cycles. However, the denotation matrix shows they did not evolve to make full use of the possibilities to denote the situations, because they did not use the available memory.

Noble and Cliff (1996) replicated MacLennan and Burghardt's simulations, and the central result was replicated; that communication leads to high rates of fitness increase and a "structured language". But, on the other hand, they found that communication with learning did not lead to the highest rate of fitness increase, so communication with learning is inferior to communication alone. In analysing the denotation matrix significantly more structure was found in the 'control' condition, and significantly less in communication with learning, compared to MacLennan and Burghardt's results.

Noble and Cliff suggest that the ambiguity in the denotation matrices is due to the whole population using an inefficient language, in which a symbol can represent more than one state or a state can be represented by more than one symbol. The authors expected to find, if simorgs were evolving a 'language', a one-to-one correspondence between symbols and states, but this was not seen. Popular symbols get more popular, and infrequently used symbols drop out of use.

3.4.2 Werner and Dyer, 1991

The goal of Werner and Dyer (1991) was to explore the evolution of language in artificial organisms to evolve animal-like communication, and, with increasingly complex organisms, e.g. learning ability, evolve a primitive language in a population. The evaluation function should not judge how well they communicate but how well they solve the task, because the authors did not want direct pressure to communicate. The specific problem was to evolve a directional mating signal.

The environment and the agents:

The environment was toroidal, 200x200 squares. In this world there were 800 males and 800 females, which means 4% of the locations were occupied. Each animal had a genome, interpreted to produce the recurrent neural network used. A female could sense the location and orientation of males in a 5x5 ‘visual field’.

Self-organisation:

There was no learning, only evolution.

Experimental setup:

The closest male was detected, and the female produced an output interpreted as a ‘sound’ that was transmitted to all males within the visual field. The output of males was interpreted as moves (forward, stand still, turn right or turn left by 90°). When a male found a female, they produced two offspring, a male and a female, using crossover and mutation. The parents were moved to random locations and two animals were removed randomly. The authors claimed that it was important for the evolution of language to allow inter-generational communication, so for this reason, and for greater realism in simulations the generations were overlapping. It was also emphasised that mating was due to direct selection rather than a result of some unrelated fitness function.

In the experiment a 13-bit input was used to the males, and females had three-bit outputs (8 sounds). Since male output was interpreted as movement, female output could be interpreted as messages telling males how to move. The task was to co-evolve a population of males and females who agreed on the same interpretation of the eight signals. More than one sound could be mapped onto the same motion, nothing was a priori correct.

Results:

In about half the runs only one ‘turn’ signal evolved. This is enough since with three turns in one direction you can do what one turn in the other direction accomplishes, but it is less efficient. When comparing the experiments, at first the non-communicating agents did better, due to the ‘listeners’ listening to ‘bad directions’, but later the listeners

did better than any non-communicators. The control group never reached the maximum possible reproduction rate because of mutations away from the optimal strategy. By travelling straight 'deaf' males could have found a mate in average in 50 moves, but never did better than 100. The listeners used an average of 40 moves to find a female.

3.4.3 Hutchins and Hazlehurst, 1991, 1994

Several computer simulations were performed by Hutchins and Hazlehurst (1991), to demonstrate a form of adaptation the authors believed to be characteristic of human intelligence. Can a population discover things that are not learnable in an individual lifetime, if culture has an effect on learners? The authors explain that culture involves creating representations of the world, which is characteristic of human mental life, but as culture cannot be said to reside within an individual, it has been largely ignored by cognitive science.

An assumption of the simulations is that there exists a useful regularity in the environment, which is too complex for any individual to learn to predict in a lifetime. The 'background' story in this case is about Indians gathering shellfish on a beach, which is better during some tide states. Moon phases are a predictor (and cause) of tide states. This will be shown by letting the agents construct artefacts, and then using several kinds of learning: direct learning from the environment, mediated learning from the artefacts, and learning a language that will permit a mapping between these structures.

The environment and the agents:

The environment was a physical representation of 28 moon phase/tide state pairs that roughly correspond to the days of the lunar month. The moon phase was represented by two real numbers between 0 and 1, where the first showed how much of the left half of the moon was visible, and the second how much of the right half. Tide state was encoded by a single real number. These two representations were the events, the 'real world'. The agents were neural networks, where all architectures were the same, starting with random connections. An agent consisted of three feed-forward networks: two language nets and one task net. The language nets were auto-associating nets, see Figure 14, which when trained produced (dashed arrows in the figure) a symbolic description of an event (moon and tide language) and the event itself (the representation of having 'seen'

the event, moon phase or tide state). The task net was a six-unit, one hidden layer, XOR network.

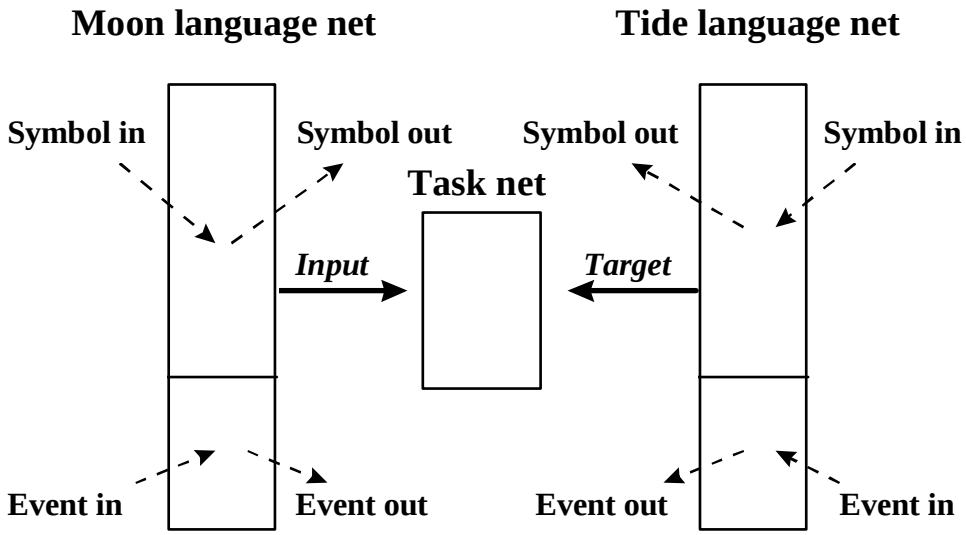

Figure 14: Networks adapted from Hutchins and Hazlehurst (1991).

Method of self-organisation:

- Learning the language by training the auto-associative nets.
- Direct learning from the environment – moon phase was the input for the task net, tide state was the output, and error was backpropagated (language nets were not used).
- Mediated learning. The two language nets produced the input and target for the task net, where moon phase was the input and tidal state is the target (unbroken arrows, Figure 14).

An artefact was a combination of four pairs of symbols. Each pair had two elements, the first a symbol for the moon phase and the second for tide state. They were created by passing information through the moon language net to create a symbol, which was the first half of the artefact, and passing an event through the moon event net, to the task net and further through the tide net, to create the second half of the artefact. In one time step, each citizen learned from the environment and the artefacts, generated one artefact, got one offspring and died. There is no genetic evolution, and the only contribution to

the next generation is an artefact. Novices choose artefacts randomly or biased by the creating agent's success.

Experimental setup and results:

Two learning scenarios were tested: direct from the environment and mediated learning from a perfect artefact. Direct learning involves all possible (28) cases, and mediated learning only four cases with perfect artefacts for the four quarters of the moon. In some trials an artefact selection bias was introduced, which reflects the MSE of the creator of the artefact. Without this bias unlucky choices of artefacts lead to slower and less dramatic learning, but even with random selection, knowledge is accumulated for later generations. Mediated learning needs only half the number of trials that the environment learning does to reach the same results. The conclusion is that a cultural process is a way to produce and maintain coordination between internal and external structures.

Hutchins and Hazlehurst (1995) continued to produce a simulation model of how shared symbols can emerge from simple interactions. A lexicon was created from scratch by agents sharing a visual experience. The authors do not claim to be treating the origins of language, but focus on developing shared symbols (or lexical distinctions) when there was no structure before. The model was based on a theory of distributed cognition (Hutchins, 1995). There are two constraints to be met: the first is that words must discriminate between objects in the environment and the second that word meanings must be shared.

The environment and the agents:

The agents were similar to the 1991 version, but consisted of one network, as shown in Figure 15. The scenes to be classified were 12 phases of the moon, shown as patterns in 6x6 arrays.

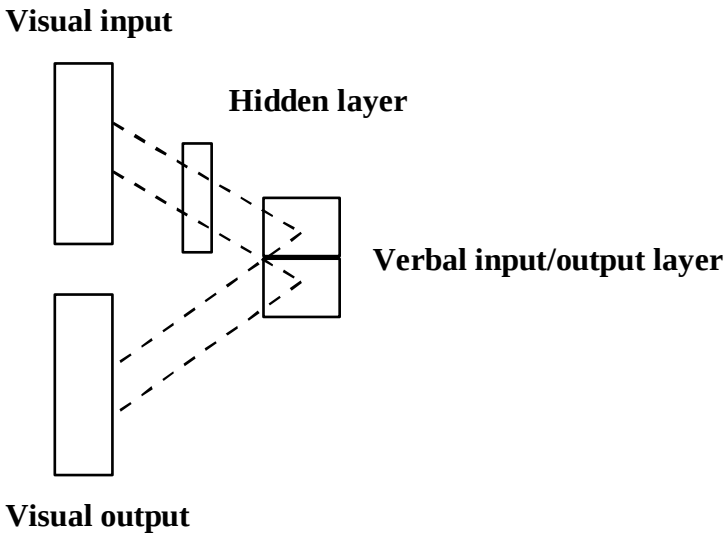

Figure 15: Network adapted from Hutchins and Hazlehurst (1995).

Experimental setup and results:

In the simulation a chosen scene was shown to two individuals, A and B, chosen from a population of six. They produced an output on their verbal layers, A's output was used as target for B's, and both got more similar, which includes the visual output. After 2000 interactions with each of the other five individuals, the agents responded differently to each of the scenes *and* agreed on how to respond to them, which means there was a consensus on a set of distinctions.

In the discussion, the authors comment that they would like to implement a critical period for language learning, which would stabilise the effect of introducing new, disorganised individuals, but even without this, new individuals exposed to well-formed lexicons did learn much faster: "individuals' abilities to learn the visual classification problem are enhanced by the existence of a coherent lexicon".

Hutchins is aware that these simulations are far from realistic, in fact they were not meant to copy 'the real world'. If we imagine a line with the real world at one end, and mathematically analysable simulations at the other, he is also aware that they are removed from the analysis side, making it difficult to completely validate them 'scientifically' (personal communication).

3.4.4 Yanco and Stein, 1993

An early attempt by Yanco and Stein (1993) describes robots having to coordinate their movements by learning to communicate. By observing a robot teacher that has access to environment information given by a human, the rest of the robots should learn to associate appropriate actions with commands given by the teacher. A robot implementation was used to test two robots with two words, and a simulator for three agents and 2-20 words. Environmental reinforcement is given by a human, but only if all agents perform the appropriate actions. The communication from human to teacher is by IR remote control, and between the robots a radio link.

The vocabulary contains signals for 'high' and 'low', the possible actions are 'go straight' or 'spin'. The teacher must learn to select one of four possible action/communication pairs, and the follower to associate each word with one of two possible actions. Neither robot is aware of actions taken by the other. The vocabulary is fixed and not interpreted, which means that two 'wrongs' can lead to a 'right'. If the environment (human) wants the teacher to do '1', it can perform action '1', give the 'wrong' signal *and* be interpreted wrongly which will lead to a correct action, and positive reinforcement. The results show that the robots do develop a shared language, but when the language size or the number of robots are scaled up in the simulator, learning time grows exponentially, which Steels (1996a) comments on in the following terse way: "Yanco and Stein get combinatorial explosions". Yanco and Stein suggest that one solution would be to let a few robots agree on a dialect, and then bootstrap the others. The authors also comment that they do want the human out of the loop, by which they mean using a more natural environment that itself can provide the feedback.

3.4.5 Balch and Arkin, 1994

Balch and Arkin (1994) explore communication in reactive multi-agent robot systems tested in simulation and later in real robots. There are three tasks: forage - find object and return it to a base, consume - wander and 'work' on found objects in place, graze - cover the whole environment. Three types of communication were examined:

- No communication - a robot can perceive 3 classes: robot, obstacle and attractor.
- State communication - a robot can detect internal state: 0 (zero) – in wander mode, 1 - in any other mode, this is not necessary intentional, but can be

observed, and the one following will be lead by the ‘leader’ who has found something.

- Goal communication: the sender must broadcast the position of the object deliberately so that the listener can go there directly.

The authors comment that implicit communication is important, and since it emerges from an interaction between agent and environment it cannot be turned off. In this case it relates to one agent seeing where another agent has grazed already. The simulation shows that the results for the forage and consume tasks improve as communication becomes more complex, but for the consume task there is a slight increase in run time for goal over state communication, and finally for the graze task there is no effect from communication, probably due to the implicit communication; seeing the graze swath. Initial results from the mobile robots were shown to support the simulation study. The authors summarise that communication improves performance if there is little environment communication, it is not essential if there is implicit, communication, and complex strategies offer little or no benefit over low-level (state vs. goal communication).

3.4.6 Moukas and Hayes, 1996

Inspired by bee-dances, this experiment concerns a robot learning a movement-based language. Moukas and Hayes (1996) believe communication is needed for cooperation, which is needed for intelligent social behaviour, and for this communication a common language is needed that all agents can use to learn about the world in their own terms. There are teacher Lego robots, that know from start how to use the language, and a learner that will learn by observing the teachers by camera. “The experiments require no human intervention; the society of agents is fully autonomous” (p. 571). The communication is direct, and one-to-many, information is sent for its own sake. The language is a set of movements. Internal representations are described in the agent’s own terms, e.g., distance is measured in the time it would take to travel it, angle by time to turn. The authors acknowledge that the language is synthetic and introduced by them. The results show that learners are able to associate symbols with relations in the world, (learn the language), and generate symbols (use the language) to communicate with others.

3.4.7 Saunders and Pollack, 1996

In a computer simulation of a multi-agent system, Saunders and Pollack (1996) evolved communication schemes. Instead of transmitting discrete communication, the agents used continuous channels, which were assumed to be able to support many communication schemes. All agents were the same; simulated ants in a grid world. In this world, there was too much food for one ant to consume in a limited time, and thus they needed to communicate to achieve a complete success. Each agent was represented by a recurrent neural network, and half the population was replaced each generation. The ants output signals that decayed over distance in the environment. Different scenarios using two or three agents were used, each with two communication channels, one of which was noisy. The results were varying and difficult to interpret, but they did find that the agents learned to rely on the clear channel and ignore the noisy one, as well as in a few occasions also "recruit" other ants, i.e. calling (or becoming quiet) when they found food attracted the other agent(s). As to the actual contribution of communication to the solving of the task, the authors comment that when they tried blocking the two channels in turn, only in one case was fitness reduced. This is explained as goal-directed communication, in that communication is the means to increase performance, but they do not know what the role of the noisy channel is. Blocking it might not be "masking a sensor'" but rather "severing an appendage", why they hesitate to draw any decisive conclusions about the role of communication.

3.4.8 Balkenius and Winter, 1997

Balkenius and Winter (1997) wished to explore pragmatic principles of communication, and examined the influence of two parameters, the cost of communication and the rate of change in the environment. In a simulation, where an agent was represented as a table, two agents played a game: there are two doors; behind one there is a prize. The prize will stay in the same place over a number of trials, and the agents take turns in guessing. They can communicate their choice by "posting a message" between the doors indicating what they chose in the previous game, as a hint to the other agent. From the outset, however, they have no common language, which means they must agree on an interpretation of the two possible messages (label A or B, for left or right door).

In the first trial agents developed common meaning by acting cooperatively, such that an agent tried to cooperate by sending the right message, and the other acted as if the first one was trying to cooperate (i.e. 'trusted' it). In later trials, it was found that when the reward is constantly in the same place, communication is of limited use, as is also the case when the world changes too often. When a cost is added to the communication, the result is that communication is used only when the agents' expectations are not met in a changing environment, by using a memory of what one said last and not repeating it if it the same the next time.

3.4.9 Di Paolo, 1997, 1998

Di Paolo (1997; 1998) was critical of two fixed ideas used in the study of communication; the first was the assumption that there has to be a selective advantage in communicating, used by many already in their definitions, and the second was the concept of information as a 'thing' being transmitted. He proposed that communication should be viewed as "coordinated activity in a consensual domain", which is related to autopoiesis. Autopoiesis is a theory of the organisation of living organisms as composite, autonomous entities (Maturana & Varela, 1980). An autopoietic system is a dynamical system whose organisation is maintained as a consequence of its own operation. Autopoiesis is a property of the organisation of the system, such that a given autopoietic system is embodied in a structure, and each state of the system is determined only by that structure and a previous state. "It implies that any state of the system that we, as observers, can relate to a particular behavior when it is situated in an environment, is a direct result of the system's own structure and of its history" (p 465). This connects to the enactive paradigm, where having a body and being in the world, with a history in the world are central (Varela, Thompson & Rosch, 1991). To take a first step in the direction pointed out by him for exploring communication, Di Paolo conducts a set of experiments, mathematical modelling as well as computational.

The environment and the agents:

In the mathematical model agents existed in a shared environment, with energy that could be received or spent. Energy was distributed to agents when they performed a correct action on a food source. An agent had two components to its actions: the effective, which decided the payoff by matching the required action of the food source

and the external component, which was the appearance of the movements implied in those behaviours to another organism, or sounds or gestures. The payoff was divided between the two agents by varying a parameter ‘c’.

Experiment and results:

In a time step one agent was selected to play first role ‘A1’, and selected another to be second role, ‘A2’. A food source was randomly selected from A1’s vicinity, A1 perceived the food type, and acted. A2 perceived the external manifestation of A1, but not the food type, A2 acted, and the payoff was distributed. The results were not very conclusive in terms of a process leading to evolution of coordinated activity.

The experiment continued with a computational model, where the environment was a toroidal grid (100x100), containing agents and food. A neighbourhood was defined as the 10x10 area around an agent. If the energy level reached zero, the agent died, if it increased, the agent might reproduce with a mate selected from the neighbourhood. As before, correct actions had to be made by A1 ‘on the food’ and a correct response made by A2, using only what he saw A1 do, and not seeing the actual food type. The results show that coordination is perfect when the payoff is divided exactly in half by the two agents. This result was unexpected, and is explained by the spatial organisation of system, how the agents are distributed, which leads to clusters being formed.

The computational model was then extended to show that coordinating *recursive* actions will lead to communication in a consensual domain, or at least to approach such a situation. In this game, the agents were stateless, i.e. had no memory, the energy was released stepwise, and both agents had access to the food type. To access the energy a sequence of actions was required by both agents. The game had a “dialogic structure” in that the actions of an agent depended on perceived food type and its partner’s previous actions, which recursively depended on the first agent’ s previous actions.

Experimental setup and results:

There were four food types, where two required action sequences such as A, B, C, D, which meant that agent 1 had to perform A, C and agent 2 B, D. The other two food types required C, D, A, B, which meant that the agents had to perform their actions in

reverse order. Payoff was distributed after the first two actions and again after the last two. The results show that dialogic activity does evolve, which means that the agents perform tasks beyond their individual abilities, they could not do this alone since they are stateless. Di Paolo claims that a description in terms of traditional information is useless, since: “If by information we mean information about features of the environment, we find that these are equally accessible to both participants, if we mean information about the state/intention of the agents, they haven’t got any”.

3.4.10 Mataric, 1998

Mataric (1998) wanted to solve two problems - hidden state, which is due to agents not being able to sense all information that is relevant to solving a task, and credit assignment for a distributed system – where it is necessary to divide credit over the multiple agents that have had different degrees of impact on a result. Both these may be solved by communication, which will have the added benefit of speeding up learning. Communication comes in two forms, according to Mataric: direct, which has the sole purpose of transmitting information aimed at a particular receiver (or many), and indirect, in biological literature referred to as stigmergic, based on observed behaviour and its effects on the environment. Cooperation is further defined to be a form of interaction based on communication, such that explicit cooperation involves the exchange of information or performing actions that benefit other agents, and implicit cooperation concerns actions that have effects in the world that help others.

The environment was an arena containing pucks (food) to be gathered (a pre-programmed ability) and the goal to improve individual gathering by learning social rules; yielding and sharing information. The behaviour-based robots used radio signals to broadcast their internal (reinforcement received) and external (their behaviour) state. A group of four robots accomplish this by learning four “rules”: to broadcast when they found food, to proceed to food when they received a food-message, to yield when close to others, and to proceed when yielded to. Mataric summarises that learning social rules can be a task that is time-consuming and difficult to program, and hence, the ‘communication’ approach is much more useful since the system learns autonomously.

3.4.11 Jung and Zelinsky, 2000

Jung and Zelinsky (2000) wanted to engineer grounded symbolic representation for communication of locations in a behaviour-based system. Two mobile robots were to perform a cooperative cleaning task. The experiments start with a location labelling procedure to attain shared grounding. One robot (1) teaches the other (2) a location by demonstration. When 2 tracks 1 and there are no labelled locations nearby, 2 signals to 1 that the current location should be labelled. They calculate the position of the location and label it. It should then be possible to represent an arbitrary location in relation to the already known, and this is provided by the authors (distance and orientation between two location indices), who comment: “As symbols systems go, ours is as impoverished as it can be”.

The results show that the addition of symbolic communication seems to improve task performance, but when the authors examine their results critically, they concede that the common process by which shared symbol grounding is developed is the design process. They conclude that their system will not scale up to complex symbol systems, and argue that it is impossible to use complex symbolic representational systems in situated agents, to be responsible for appropriate behaviour, and they propose an adaptive symbol grounding hypothesis.

3.5 Summary

This chapter has reviewed a large amount of past and ‘present’ work in communication between autonomous agents. We have mentioned purely formal models, seen examples of many software agents, and studied applications using robots of various degrees of sophistication. Some researchers combine two approaches, in using simulations first and porting their programs to robots later. When assessing the work of a few research teams more comprehensively, we found that some seem to follow a rather straight line in their research, e.g. the language game of Steels *et al*. (cf. Section 3.2), while others branch out and examine many aspects of communication, e.g. Cangelosi *et al.* (cf. Section 3.1). The next chapter will discuss the possibilities of comparing various seemingly dissimilar experiments, look closer at why agents communicate and what they are communicating, and evaluate these attempts to self-organise communication.

4 Evaluation

This chapter contains a critical evaluation and discussion of the models reviewed in chapter 3, from a number of viewpoints. In analysing to what degree self-organisation has been accomplished, as outlined by the quote from Steels and Vogt in chapter 1, "…these communication systems must be developed by the robots [agents] themselves and not designed and programmed in by an external observer", the main questions mentioned in chapter 1 are used:

- Why are the agents communicating: is their communication self-organised, and who profits from the communication?
- Is there communication in a cooperative sense?
- What is the communication about; are the symbols arbitrary labels, or are they meaningful to the agent?
- Can and should the signals be interpreted into 'human language'?
- Is there spatial and/or temporal displacement in communication?

Before going into any details, it should be pointed out that a comparison of these very diverse models is difficult and perhaps also unfair. Since the inspiration for the experiments comes from quite distinct and sometimes unrelated areas of study, a comparison on equal terms is practically impossible. Most researchers claim to at least be inspired by biological systems, even if they do not claim to simulate them. This is the case with Cangelosi (1998) – where food quality is signalled (cf. Section 3.1.1), Cangelosi (1999) – inspired by ape communication (cf. Section 3.1.3), Werner and Dyer (1991) – signalling to find a mate (cf. Section 3.4.2), and Moukas and Hayes (1996) – inspired by bee-dances (cf. Section 3.4.6). The truly biological models, such as Noble (1998b) and Noble (1999) are excluded as mentioned before. Noble (1998a) is of the opinion that many models are not inspired enough by biology, with the result that all they show is that different procedures lead to different outcomes. This implies that Noble is looking to study only biologically true models, whereas this dissertation allows for a wider outlook. There is, to be sure, no consensus on what constitutes A-Life research, as Langton (1995) explains: "the field is still in the process of defining itself".

4.1 Self-organised communication?

This section aims to investigate the question: Why are the agents communicating? This is divided into three sub-questions. Firstly, why do they start? This might be considered the essence of self-organised communication, that is, are the agents communicating as a result of a 'choice' between behavioural strategies or because there is no 'choice' – communication has been more or less hard-wired into the system? Secondly, why do they continue in learning/evolving a language? These two questions will be discussed jointly in section 4.1.1, since they are intertwined, and not easily separated. Section 4.1.2 discusses the third question, if there is communication in a cooperative sense. Is communication always cooperative, and do agents use 'the same language' or the same reason for communication?

4.1.1 Why start and continue?

This section will discuss the first two sub-questions; firstly, why do the agents communicate in the first place, is it a voluntary or forced 'choice' of behaviour, and secondly, what is the learning mechanism that makes them 'good at it'? When it comes to 'voluntary' communication, the question is really if there is self-organisation in the sense that communication is part of an integrated co-evolution of behaviours, and if there is structural coupling (cf. Sections 2.3 and 3.4.9) between the agents and their ecological niche/environment or not. The 'voluntariness' then refers to the possibility to react with communication or another behaviour. Such a structural coupling was not seen within GOFAI or connectionism, where, for example, ELIZA (cf. Section 2.1) did not answer voluntarily, but because it was all the system could do, and simply produced an output for every input. It could be argued that when animals communicate, it is voluntary since there are other possible ways in which they could have responded to a certain situation or object. A response may be any type of behaviour; it need not necessarily be communication. There are, probably much due to evolution and natural selection, a multitude of behaviours that animals from different species can use when faced with, for instance, a potentially dangerous situation (cf., e.g., Kirsh, 1996, and section 1).

Imagine our agent in such a situation, bearing in mind that it is not necessarily animal-like, but nevertheless inspired by biological systems and their behaviour. It might by all means call out to warn conspecifics, change its own appearance (as a signal), or mark the

spot as dangerous. What might it do, voluntarily, besides communicate? As Burghardt's definition of communication in section 1.1 implies, communication is about influencing someone's behaviour for the good of oneself or the group. An agent can certainly exert this influence by using other means than signals. It could e.g. drive the others off physically, move away itself, and perhaps lead the others to a different location, or it might 'choose' to ignore the situation completely, 'deciding' that the situation does not warrant (by some cost) the work. If we look at what agents as a species or population could do, in evolutionary time, instead of communication abilities better discrimination or perceptive abilities may be evolved or some other behaviour that is beneficial, e.g. running faster, or evolving some camouflage properties.

For a language to be considered "the agents' own language" (cf. the quote from Steels and Vogt, 1997, in chapter 1 and above), one of the first demands of such a system could be considered letting the agent 'decide' if, and when, it wants to say something. There is also the question of whether or not communication should be explicitly rewarded, in which case the further use, or successful evolution, of the language should not be considered voluntary either. Steels (1997a) points out that many experiments assume that successful communication has a direct benefit, but a fitness measure can be anything that impacts future survival, for instance another behaviour as argued above.

In the Cangelosi simulations, it is obvious that the agents are forced to speak, or emit signals, in every timestep. They are therefore not free to choose if, and what, to signal as commented above, because signalling is the one and only way they can (and must) react. In Cangelosi and Parisi (1998), a 'word' is the input to 3 units, which is either fed directly from the researchers or from a conspecific's output (cf. Section 3.1.1). In these cases though, behaviour is co-evolved with the language, which is not entirely true of Cangelosi (1999), in which the agents learn to differentiate between mushrooms for a number of generations before they start communicating, at which point there is no question of choosing whether they want to communicate or not (cf. Section 3.1.3). One timestep for an agent actually comprises the following steps: input is received and action decided upon, the mushroom is named with teacher feedback, and an imitation of the naming, again with teacher feedback. A 'quiet' period is also used in Cangelosi and Harnad (2000), where the 'mature' agents learn to vocalise on sight of a mushroom, by

using backpropagation (cf. Section 3.1.5), whereas in Cangelosi, Greco, and Harnad (2000), the only thing the 'agent' does is seeing symbols and learning, by backpropagation, to name (and categorise) them (cf. Section 3.1.4).

Steels (e.g. 1996a, 1996b) uses the somewhat impoverished 'seeing and naming' view of communication in all his experiments (cf. Section 3.2). The agents do communicate, and learn a language which is successful for discriminating objects, but it must be noted that the whole language game is taken for granted as a kind of ritual, and as such lends little flexibility to the learning, because so much of the basis for learning has been pre-programmed into the system. It is the language itself that self-organises and the agents are somewhat removed from the interaction with the world by being used as passive senders/receivers with appropriate innate mechanisms. One interesting difference from many other models is that the updating/learning mechanism is not a fitness function or backpropagation, but another kind of feedback, namely the agents themselves. If a game is a failure, both agents try to adapt their internal structures. This might be considered reinforcement learning, since the agents do not find out *what* they did wrong, but they do find out that *something* was wrong. This is accomplished by the hearer comparing a 'word' to its expectations and, if the game is not a success, providing feedback via radio link to the speaker (Steels & Vogt, 1997, cf. Section 3.2.4) which means both agents are aware of the failure. In the talking heads project (Steels & Kaplan 1999, 2000, cf. Section 3.2.5), the agents also 'just talk', and the feedback mechanism is even more sophisticated, since the speaker in case of failure indicates which of the objects was meant to be the topic. This contradicts in some way the introductory claim in Steels and Kaplan (2000) regarding communicative situations; that it is more realistic to get feedback on communicative success and not on meaning. In this case, though, the agents will both know that the game was a failure, *and* what the correct topic (object) was, which is not exactly feedback on meaning, but far more that just being told that there was a failure in communication.

Steels claims to apply a selectionist pressure, and not a genetic one, but since the lifetime of the agents is practically unlimited, it might be argued that the agents are subject to Lamarckian evolution or cultural transmission rather than a purely genetic one, it is merely a question of how a 'lifetime' is defined. The use and success in use of word-

meaning pairs is recorded, and naturally the most used and most successful ones are preferred, which pushes the system towards coherence, a fitness function as good as any. Genetic transmission might be seen as just another way to emit and receive signals on a different time-scale.

The experiments conducted by Billard and Dautenhahn (1997) use a teacher robot, with complete knowledge of the language, which gets no fitness for speaking (cf. Section 3.3.1). The communication amounts to the teacher calling out its position, and the learner trying to satisfy a pre-programmed need. This is a definite 'profit' situation, quite comparable to increased fitness, awarded for using the language in the right way. A confidence factor was used to ascertain when an association was learnt, which is also a kind of fitness, awarded to the association rather than to the agent. The DRAMA architecture discards incorrect associations by statistical elimination depending on their relative frequency of occurrence. Combined with the built-in basic behaviours, life-long learning is used, but when searching for the teacher so much energy is used for moving that it does not learn, which explains why it is not relearning even though it is hearing signals.

All agents in Di Paolo's (1997, 1998) experiments are forced to communicate, but the speaker and listener have equal access to information in the environment (cf. Section 3.4.9), which contradicts the belief of MacLennan (1991) that there has to be something 'private' for the agents to want to communicate (cf. Section 3.4.1). These agents were also forced to talk, and the principal goal of the selective criteria was to lead to emergence of communication without being overly 'rigged'. It turns out, however, that the fitness was a direct measure of number of times that a simorg's response to the last emitter led to effective action, but this is defended as being reasonable according to the definition of communication by Burghardt (cf. Section 1.1). Since the agents are, in fact, their language it is the language that evolves (similar to Steels' experiments), by a genetic algorithm, and 'corrective' learning, which is also similar to Steels, in that their tables are corrected after a failure, so that they would have acted correctly.

Werner and Dyer (1991) use overlapping generations to allow inter-generational communication (cf. Section 3.4.2). Mating is (said to be) due to direct selection and

males that stand still become extinct, so very few interpret a signal as meaning 'stay still'. This is the mechanism that indirectly pressures towards communication, which is supposedly better than what the authors call "some unrelated fitness function". Saunders and Pollack (1996) found some non-intuitive communication schemes since some agents were 'signalling' while searching for food, and became silent when finding it, and in other simulations the contrary effect was found (cf. Section 3.4.7). Since the agents are programmed to produce an output in every time step, however, this communication is also prompted, and it boils down to defining communication, where the absence of a signal could be considered a signal too.

In many robot experiments, the teacher is pre-programmed with the signals to emit, and emits them reflexively, i.e. when receiving a cue from the researchers (Yanco & Stein 1993, section 3.4.4), or when coming across something in the environment (Balch & Arkin, 1994, section 3.4.5; Moukas & Hayes, 1996, section 3.4.6; Mataric, 1998, section 3.4.10; Jung & Zelinsky, 2000, section 3.4.11).

Finally, Hutchins and Hazlehurst (1991, 1994) teach the neural nets the 'language' by auto-association, and when learning to create good artefacts the agents use the environment and earlier artefacts as feedback, sometimes with an added bias towards artefacts created by successful agents (cf. Section 3.4.3). This means that the agents are not actually communicating voluntarily, but on the other hand they are self-organising to some degree the artefacts, which may be considered communication since language is by many considered an artefact, e.g. Clark (1997).

To summarise, we have not found that any of the models reviewed use voluntary communication in any strong sense, but all agents are prompted to emit signals. Some models even make communicating the sole reason of the experiment. In most simulations, agents emit or listen in every timestep, by reacting to very prepared environments or chosen features of them, and in robotic models usually reflexive emitting is used, either reacting to a prepared aspect of the environment, or when meeting another robot. This shows that our first criterion of self-organised communication, the possibility and the privilege to 'speak' voluntarily, is not present in any of the reviewed models. This is not to say that all, or any, of the reviewed models were intending to show this.

4.1.2 Cooperation and asymmetric communication

This section discusses the third sub-question, is communication cooperative or perhaps competitive? There are often assumptions built into the models concerning the possible enhancement of agent interaction or cooperation. A follow-up on that issue regards 'asymmetric' communication, where sender and receiver are not communicating directly, whether that is because they are not 'aware' of each other, are not using the same 'language' or in some other way not contributing equally to the communicative process. Some agents learn to (or know how to) speak and others only learn to listen and 'interpret' signals in the way intended by the designer. There is a definite bias in such situations towards 'rote learning' i.e. learning to react in a reflexive way when perceiving signals. It is also possible that the agents are 'unaware' of their surroundings/environment to such a degree that they do not perceive their 'peers'. Should such signalling systems be considered communication at all? This question has been asked by, among others, Bullock (2000) who claims that for this to be the case, signallers and receivers should have co-evolved to achieve a common aim, as mentioned in the introduction.

Gozzi Jr. (1998) proposes that the Wittgensteinian game metaphor for language can be interpreted as being either a competitive game or a cooperative one. Language is a game that is used when we want something, which suggests that there be a goal, which fits neatly with our agents who need to have one. But is the game metaphor for language so appropriate? Are agents cooperating or competing? Most experiments do assume some kind of cooperation, as mentioned above, like the shared solving of a task for example. It might be more correct to assume that all forms of interaction, and not only cooperative situations, between agents is improved by communication, as pointed out by Steels (1997b). There is support for both alternatives, cooperative and competitive communication. Noble (1998b) has shown that there is a definite possibility that communication can evolve and survive only if there is a positive effect of fitness for all involved, in a general model where communication evolves and stabilises only when it is the interest of both signaller and receiver.

Simulations mentioned in Cangelosi and Harnad (2000) imply that the will to communicate depends on the amount of 'food' (fitness) available, where speakers stay

mute if food is scarce, and will signal to kin before signalling to all (cf. Section 3.1.5). There are also experiments that implicitly pressure towards cooperation, since reinforcement is based on group behaviour (Di Paolo, 1997, 1998, section 3.4.9; Mataric, 1998, section 3.4.10; Yanco & Stein, 1993 section 3.4.4), and some robot simulations explicitly assume that cooperation and/or coordination is the way to achieve communication, most often by some kind of teaching in a 'social' setting by a robot that has knowledge of a language in beforehand. The learning typically takes place by the learner imitating or following the teacher (Billard & Dautenhahn, 1997, 2000, section 3.3; Moukas & Hayes, 1996, section 3.4.6; Yanco & Stein, 1993, section 3.4.4). On the other hand, Bullock (1997) shows with "action-response games" that it may be valid to talk about signals with shared meanings, even if the agents share no interests, Ackley and Littman (1994) show that altruistic speakers can evolve, even if they do not receive a benefit, using the concept of kin selection, and de Bourcier and Wheeler (e.g. 1997) show that communication does arise even if the agents have adversary relations.

Another interesting issue to consider is whether or not the sender of a signal is 'aware' of the receiver, i.e. if the sender in any way will adapt its behaviour or signals due to the presence of a receiver. According to Evans and Marler (1995), it is typically assumed that animals have reflexive call production, but 'audience effects', i.e. sensitivity to the presence of a potential receiver, have been found in e.g. monkeys and birds. The possible applications for artificial communication systems are not clear, but it may be useful to consider in tasks where cooperation or coordination is necessary, and/or when there is a cost attached to communicating. Examples of this are found in Di Paolo (1997, 1998) where the concept of a game is present, in that the agents take turns in acting/communicating to jointly work towards receiving a payoff, which is distributed unevenly and periodically (cf. Section 3.4.9), in Mataric (1998), where if one robot receives and follows another's message both receive credit (cf. Section 3.4.10, and of course in Steels' language games (Steels & Kaplan, 1999, 2000, cf. Section 3.2.5). This resembles 'kin selection' where the good of the group is as important as the good of oneself, but at least in the Steels case it is probably due to the 'conversation is the name of the game'-concept used.

Up until now, we have looked at agents that apparently use the same language to communicate with each other, or seem have the same reason for communicating. In an alternative way of looking at these models, some asymmetries will be examined. In Billard and Dautenhahn (1997), the teacher is emitting signals, apparently unaware of the learner, which is only listening and not using the words (cf. Section 3.3.1). There is no apparent advantage for the teacher, compared to the satisfaction of the needs of the learner. This mismatch in the use/understanding of language is similar to the simulations by Werner and Dyer (1991), where an immobile female is calling to attract males that cannot see her (cf. Section 3.4.2), but in this case the language evolves (in the female), and the understanding evolves (in the male). The interpretation of the output is interpreted to mean 'sound' in the case of the female and 'movement' in the case of the male but it is duly noted that it could have been interpreted otherwise.

In at least one simulation there are communication *effects* even when no one is listening, which will not be considered true communication here, since it is reasonable to assume that for communication to take place there should be at least one listener or receiver, otherwise we have some kind of indirect communication. This would mean that one agent, not 'intending' to communicate, leaves some trace or signal in the environment, which may or may not be picked up by another agent. This effect is apparent when the 'silent' population in Cangelosi and Parisi (1998) still evolves a 'language' (cf. Section 3.1.1), which indicates that there is excessive pressure towards communication, or that it is slightly misguided to interpret the output of the network as a language, instead of just regarding it as a categorisation or pattern matching process. Such a process might, by all means, be a first step towards the emergence of communication, as the authors argue, but for now it is simply categorising a prepared and undemanding dataset. On the other hand, 'speaking aloud' to yourself may still be helpful in order to be more efficient but this is not the focus here (an example of this is children who speak aloud when learning, for instance, to tie their shoelaces). Many robot teachers, as mentioned above, emit reflexively, and as such cannot be said to be 'aware' of their students, which is another example of 'un-directed' communication.

In summary, adopting a view of enhanced agent interaction, rather than co-operative situations, seems more plausible, since it seems that interaction is improved by

communication and conversely. This is another argument for using a socially based approach to self-organised communication. This further implies that we should regard communication as an ability that increases fitness, for sender and receiver or for a whole group (cf. Burghardt's definition, section 1.1) but not to the extent where communication *is* fitness, which is a critique voiced by many (e.g. Ackley & Littman, 1994; Steels, 1997b).

4.2 The meaning of signals

In many of these simple communication systems reviewed, the final goal of the researchers is to evolve human language (e.g. Steels, 1997b), which may be why methods from traditional approaches still linger. Typically, when an agent is to communicate or learn/evolve a language, the notion of 'one word, one object' is used, cf. also Prem (1998) and Ziemke (1999). This entails that the designer of a model of communication must decide not only on what constitutes an object, and which objects should be present in an environment, but also choose which labels to attach to these objects (cf. Figure 2 and 3 in the introduction). The only thing left for the agent to do is to associate the prepared labels with the pre-conceptualised objects in a 'correct' way. An agent's world should not be filled with labels and preconceptions, cf. the PacMan syndrome (Peschl & Riegler, 1999) where the authors point out that predators do not run around with a label saying "I'm your enemy". Brooks (1991a) points out that this notion of 'word-object' labelling is a problem within classical AI, and Parisi (1997) voices a similar criticism against "classical connectionism" as being an abstract input/output system, but A-Life researchers declare that it can be solved by grounding representations in order for them to assume an inherent meaning for the agents. Cangelosi (2000), for example, believes that grounded symbolic associations can be simulated with evolutionary computation techniques, but not with other computational techniques. Similarly, Cliff (1991) claims that if the semantics of a network is well grounded, results are generated by observation rather than by interpretation, which leads to "hard" objective measurements rather than "soft" subjective ones, cf. also Franklin (1995). The aspect of the interpretation on the part of the researcher will be discussed later, in section 4.2.2.

Continuing with the argument that for an 'own language' to evolve, the agent should then be allowed to 'choose' what it wants to communicate about, and which features of the world are relevant to it in this particular world or situation. Who is to say, then, which objects and features are relevant? It will most probably depend on the situation, of course, and if we are dealing with a world that is at least partly understandable for humans, Prem (1995) suggests that it might be useful for an agent in unknown territory to produce labels for what it experiences in the environment or report from new territory. But this implies that it should be the agent itself that chooses relevant experiences to attach 'labels' to, and that the 'words' it uses are grounded in its own experience and environment to let agents communicate about things that help them interact with each other and their world. Is this what happens in current A-Life models, and are they still just attaching 'useless' labels to objects, acting as 'PacMen', as it were?

Before we look closer at the existing models, and how they approach this problem, a quick look at some communication or 'signalling' concepts might be helpful. According to Prem (1998), based on a Peircean division, a sign is an icon, an index (signal) or a symbol. An icon has a physical resemblance to its referent, an index is associated to the object in time or space, and a symbol represents a referent according to social conventions or mutual agreement. Animals other that humans communicate by signal use (Cangelosi, 2000) where there is an association between the signal and the object but none between the signals themselves. Humans, on the other hand, use symbolic communication, where there are relationships between symbols and object and also between symbols and all other signs ('grammar'). This dissertation focuses, as mentioned in the introduction, on examples of signal use, and not 'grammatical' symbol associations, which does not preclude the use of symbols for referring to objects. Note also that we use the concepts of signal and signalling in a broader sense than what, for instance, Prem suggests. In this dissertation they have been used as follows: whenever the agent 'says' something it is a 'signal' and the process is 'signalling', which means that it could be about icons, indices or symbols.

The following sections will first, in section 4.2.1, examine whether or not symbols in communication systems are arbitrary or have meaning with respect to the agents, and what such a meaning could be. The next section, 4.2.2, looks into what the

communication might mean to humans, and whether or not the signals can and should be interpreted. Section 4.2.3 aims to examine the possibility of immediate vs. detached use and learning of symbols, i.e. do, and, can, agents communicate about things that are not temporally or spatially present?

4.2.1 What does it mean to the agent?

This section discusses what signals can and should 'mean' to the agents. If the communication was iconic, it would be immediately interpretable, but not so with symbols. If we do not use indexes or symbols, we will not be able to represent anything abstract or non-present, and sooner or later this will be required, not to mention desired. In order to use symbols to denote something meaningful for the agents, a referential language might be needed. Hauser (1996) describes a referential system as one with functional significance, so that we can understand what a signal is about without seeing what's going on, i.e., do not need to see the evoking context. The concept of referential representations is discussed by Peschl and Riegler (1999), but their conceptualisation appears to be a more stable construct than the one intended here. The referentiality suggested here, is a functional one, and not a one-to-one correspondence. It seems that a pragmatic view should be adopted, where signals are useful to the agents by referring to something important for their survival or adaptation. This does not imply that there is a 'universal' meaning, but on the contrary a rather open meaning, and it could be argued that agents 'interpret' according to their present situation, goals, and needs. We do not want to teach, as it were, agents meaningless labels for objects we, as designers, imagine may be useful. A pragmatic view of communication is held by Winter (1998), in whose view semantics is conventionalised pragmatics, and syntax is conventionalised semantics. A meaning of a word then captures some useful generalisations from the pragmatic level. He further stresses the importance of studying how meaning is built up, from situations or objects with inherent meaning for an agent, to associating these with "previously meaningless" labels.

Continuing on this note, Prem (1998) is critical of the view that internal representations labelled with names is sufficient to explain semiosis (sign use) in embodied autonomous systems. This gives a restricted view of the purpose of sign use, and disregards that representations may be formed by system-environment interaction (Prem, 1998). An

agent should decide on its own what there is in the environment and what can be used to pursue its own goal. When doing this, Prem points out, encountered objects seem different depending on context, which means that the traditional 'word-object' labels will not work. Signs are to be seen as tools for indication purposes, and the receiver of this 'indication' could be another agent but it is not uncommon to use signs for oneself. When an agent comes across a sign in the environment it should 'react' to it rather than try to understand what it is supposed to mean. Prem (1998) suggests that appropriate reactions could be to, for example, 'give way' or 'start digging', and that a sign thus should serve to guide the agent's interaction with objects and other agents in an environment. Such an interaction consists of the sign 'showing' or 'pointing to' (indicating) something and a successful result would be the one that the sign users 'anticipated' when using the sign. This means that a signal should probably be functionally referential rather than being interpretable semantically, since the former is advantageous to the agent and the latter is more interesting to the 'observer'.

To use a more specific example, since the discussion has been rather abstract; when two agents meet it is more reasonable that they would say something like "come here" (e.g. Werner and Dyer,1991, section 3.4.2), and not find a label that says, for example, "potential mate", or, "tell me where the food is" (e.g. Cangelosi & Parisi, 1998, section 3.1.1) and not just hear "potential food item". In the spirit of the PacMan syndrome (cf. above) a designer might, for instance, designate a certain object in the world as being 'food', but this may not be true from the perspective of the agent. Is the world perceived differently with prior knowledge of a language? The strong version of the Sapir/Whorf hypothesis states that any given language will *determine* the way we perceive the world, but a more reasonable interpretation is that a language in some ways *influences* the way we think of the world (Linell, 1978). With this in mind, it would be unwise to force our world-view onto agents, who reside in worlds that may be very different from ours, and certainly exist under conditions completely unlike ours, by supplying them with a language that governs their perceptions instead of letting them 'self-organise' to manage for themselves. What is the perspective of the agent, then?

Agre and Chapman (1987) proposed the notion of indexical-functional representations. Their agent 'Pengi' used functional representations that are relative to itself. When an

agent finds itself in a situation, there are *aspects* of the current situation subjective and important only to the agent. The agent will come across entities, which can be different objects depending on the current function. An entity is embedded in an aspect, such that the function is central and the representation subjective to the agent, for instance, "the-wall-I'm-hiding-behind" and not e.g. {Wall [east, no.3]}. Indexical-functional representations will undoubtedly be more useful to the agent than abstract, arbitrary labels on objects it may not even 'have a use for'. Prem adds to this by pointing out that while the symbols may be arbitrary, the meaning should be tied to purposeful interaction and social aspects (Prem, 1998), and clarifies that this is still at the level of signs and not language. The first step should be to examine how autonomous systems create signs, and the second to study how they socially use them, hence proposing a view where communication is to be seen as motivated by a system's social interaction purposes. Many researchers are convinced that social behaviour and social context are important for communication, e.g. Billard and Dautenhahn (1997, 2000, cf. Section 3.3), and Hutchins and Hazlehurst (1991, 1994, cf. Section 3.4.3).

Returning to the current models then, are there any experiments where agents are allowed to categorise their own world, or are they all given labels? In the mushroom worlds of Cangelosi *et al.*, the world is structured so that it allows/supplies discrimination in one time step, already in the sensor input, (cf. the PacMan syndrome above). In Cangelosi & Parisi (1998), a 'word' is constrained to be one of eight possible signals, due to the design of the network, which means there is *some* freedom in choosing the label for poisonous and edible (cf. Section 3.1.1). In Cangelosi (1999), the communication units consist of 2+6 units, most probably to accommodate the fact that there are six 'categories' of mushrooms, and two appropriate actions (avoid/approach, cf. Section 3.1.3). It turns out, however, that the agents only need to distinguish four functional categories: toadstools and three edibles, which does confer more flexibility than was originally intended. Still, the world is so prepared that there is nothing that agents can do to avoid making the 'correct' associations, given time and training. In fact, in 70% of the populations the learned language is based on symbolic associations, as the authors themselves comment, most of the ANNs use a symbolic strategy when learning linguistic symbols and the syntactic rules for combining them.

The ‘word-object’ problem was tentatively explained by Parisi, Denaro and Cangelosi (1996), who suggest that perhaps we (humans) tend to think of categories and word meanings as well defined single entities because we have well defined words for them, and instead should regard them as virtual collections of activation patterns with properties that allow the organism to respond adaptively to environmental input (cf. Section 3.1.2). This is shown in their experiment, where organisms encounter mushrooms and/or hear a signal. The activation patterns vary when the agents receive ‘double input’ but are exactly the same if the agents only hear a signal, which is taken to mean that in absence of a referent, signal meanings are single mental entities, but when the input consists of perceiving a referent *and* a signal (‘word’) they are not. This could be interpreted to mean that concepts in fact are something like Prem’s concepts based on adaptive categories (above) and not the combination of signal and referent, and that the ‘words’ in the experiments result in the same internal representations is just a result of the way ANNs work, a simple input-output mapping. It also implies that a large part of the problem lies with the designer and the designer’s preconceptions.

Steels *et al.* claim to study communication that is self-organised without a designer or teacher involved (e.g. Steels & Kaplan, 1999, 2000, section 3.2.5; Steels & Vogt, 1997, section 3.2.4), and also that: ”The agents thus not only construct their own distinctions but their own language for verbalising these distinctions as well, all without the intervention of a teacher” (Steels, 1997a). The agents do create their own distinctions, and the internal representations are rather flexible, such that they may be changed in case of communication failure, but, on the other hand, the framework within which to create them is relatively inflexible, in that the categories are given, (vertical and horizontal position and colour/greyscale), and it is just a question of differentiating between the objects using whatever features necessary. These features then form the basis for the ‘language’, albeit “random combinations of syllables”, the agents are verbalising only the attribute values needed to clearly distinguish one object from another. There is, as commented above, no other reason to communicate other that the forced exchange of symbols, which makes these experiments a prime target for discussing Prem’s (1998) view that an agent should decide for itself what there is in the environment and what can be used to pursue its own goal. It is unclear to what degree the agents actually decide for themselves what there is in the environment, but at least in Steels and Vogt (cf. Section

3.2.4) and Steels (cf. Section 3.2.3), the agents scan the environment and segment it into objects, albeit objects with no actual 'meaning' or function for them. In the "talking heads" experiments, on the other hand, the environment is very limited, prepared for quite simple discrimination of objects, and constructed to be the topic of conversation and nothing else. On the issue of pursuing subjective 'goals', it is quite clear that there is no goal to speak of (no pun intended) for these agents, except 'talk for its own sake'.

Billard and Dautenhahn's experiments (1997, 2000) propose that agents can be heterogeneous, since they do not need to use same sensory data to characterise a 'word'; one agent could use a position and the other its shape (cf. Section 3.3). They further define their type of communication as the transmission of symbolic signals or symbols, and that two agents are communicating when they have achieved a similar categorisation of sensor perceptions and have successfully attached them with the same set of arbitrary signals. Their aim was stated as the study of correct transmission of a fixed pre-defined vocabulary, which in a way disqualifies them somewhat from this study of self-organised communication. Most probably, the robots cannot represent anything else than the amount of light and the inclination (1997, cf. Section 3.3.1) and the colour and position (2000, cf. Section 3.3.2), which is a pretty well structured world in the spirit of the PacMan syndrome (cf. above). Furthermore, there is no goal explicitly stated for the robots, they simply wander and cannot help but signal when they come across an object. The authors conclude that communication does give insight into another agent's world, but since their perceptions are not identical, the learning is not perfect, but on the other hand the communication is constructed in (embodied) interaction with the environment to a greater degree than, for instance, Steels' or Cangelosi's models.

In many of the remaining experiments, there is 'a best/right way' to communicate, i.e., associate the 'word' and the object as close as possible to a solution wanted by the researcher, and rewarded for in a variety of ways. This happens in Di Paolo (1997, 1998), who uses sequences of actions to be 'guessed' by the agents (cf. Section 3.4.9), Hutchins and Hazlehurst (1991, 1994) where there does exist a 'perfect artefact' (cf. Section 3.4.3), Yanco and Stein (1993) in which robots must associate a signal with the correct action (even though two wrongs may give a right, cf. Section 3.4.4) and Jung and Zelinsky (2000), whose labels are simply serial numbers attached to 'new' places (cf.

Section 3.4.11). Most computer simulations also use pre-structured input to such a degree that the agents cannot but 'do the right thing', whereas robot simulations often use more 'realistic' (noisy) worlds. On the other hand, the language in robot models is more often completely pre-structured, which is a disadvantage. Moukas and Hayes (1996) comment that by *introducing* communication in the form of *a common language* in a multi-agent system they enable each agent to conceive and learn about the world *in its own terms* (cf. Section 3.4.6), which in the light of the previous discussion must be considered a somewhat questionable conclusion.

To summarise this section, it has been shown that many of the reviewed models do employ situated agents in one way or another, which means that they have more 'context' than GOFAI and connectionist models. This is not always enough, since we want the signals (words) to have some 'innate meaning' for the agents; otherwise we are back to simple labelling, and the use of the designer-as-context. There is also a potential problem in that way agents and environments are prepared separately from each other and only later combined. For communication to self-organise, it is not appropriate to set down an agent in an unknown, but pre-structured or pre-conceptualised, territory and force it to communicate about it with a more or less prepared language.

4.2.2 What does it mean to us?

It turns out that one of the original criticisms against traditional AI and connectionism has been dealt with, since none of the agents discussed are learning human language. On the other hand there are often many pre-conceptions as to what the agent may need to, or has to, communicate, such as the use of categories that are pre-determined or heavily constrained by an experiment designer. This section discusses two aspects of interpretation on the designers' behalf: the pre-interpretations that are made when preparing an experiment or a model, and the 'post-interpretation' when we observe the behaviour and analyse the results. How can an experiment be prepared with as little 'designer bias' as possible?

As soon as we even start thinking about an experiment, we start conceptualising, which implies that it may be very difficult to remove the designer 'completely' from the process and achieve 'total self-organisation'. Jung and Zelinsky (2000) point out that since

elements of design are labelled by a designer, the anthropocentric groundings used to interpret these labels will effect the design, e.g. if naming a behaviour component “Wall-Following”, human assumptions of walls might be included, whereas an agent most probably has no concept of a wall as we do. Saunders and Pollack (1996) describe one of the problems they experienced when trying to analyse their results, using two communication channels, of which one was noisy. They assumed that communication had occurred if there was a drop in performance if one channel was blocked. However, on closer reflection, it was not clear if the agent used the noise as some kind of regularity in the environment, in which case blocking a noisy channel might resemble “severing an appendage rather than masking a sensor”. This shows that researchers must be careful not to let their preconceptions get the better of them when designing and interpreting experiments.

How should we interpret what the agents are ‘saying’? Should we interpret at all? An important distinction to make here is the one between distal and proximal descriptions (se e.g. Nolfi, 1998). A distal description is a high-level description from an observers’ point of view of some behaviour. Proximal regards the actual mechanism at work, which is not necessarily intentional, but could just be reflexive. There is a definite risk that we over-interpret or anthropomorphise what we observe, as will be clear to anyone who has seen people observe a robot that displays some behaviour. There will always be someone who exclaims something like: “Oh look, it’s hungry/angry/happy...”, definitely a distal description. Researchers must be ever vigilant not to be caught up in over-interpretations, and might be helped by Cliff (cf. above), who claims that if the semantics of a network is well grounded, results are generated by observation rather than by interpretation, which leads to “hard” objective measurements rather than “soft” subjective ones (Cliff, 1991), cf. also Franklin (1995).

A similar view is held by Prem (1998), when discussing how to verify that a sign has been understood correctly, claims that we will need an outer criterion such as behaviour, since behaviour can be observed, since the system will be “conceptually opaque” to an observer. As the sign user is assumed to care about the effect a sign has on behaviour this means that if the reaction is the desired one, the understanding may be interpreted as being correct. From the point of the agent, then, we must have a communication system

that is grounded by the agent in an environment and from the point of the researcher, results should be observed and not interpreted, also pointed out by Sharkey and Ziemke (1998) warning us not to make "wishful attributions" from what we see happening in a system by using inappropriate terminology. Such attributions are easily made when observing apparently 'intelligent' systems. Consider, for instance, the following quotation from Billard and Dautenhahn (1997) "…association words and sensor data leads to [the robot] understanding the radio words" which cannot be strictly true, and most probably is just a misuse of the word 'understand' but nonetheless, the point is made. It should also be noted that to understand signs used in a system, Prem suggests that we observe behaviour, but if there is no behaviour (except the communication) to observe, what then? This strongly implies that it is necessary to co-evolve several behaviours, which will be discussed in chapter 5.

Dawkins (1982) describes a useful analogy in the context of existing organisms, which naturally will not always be totally relevant, but serves well as an inspiration. When watching a computer program being run, it is impossible to know which programming language, if any (it could have been machine code, or hard-wired in), it was originally programmed in, in effect, a "lost program":

> "A biologist looking at an animal is in somewhat the same position as an engineer looking at a computer running a lost program. The animal is behaving in what appears to be an organized, purposeful way, as if it was obeying a program, an orderly sequence of imperative instructions. The animal's program has not actually been lost, for it was never written out" (p 119).

Rather, natural selection has, over generations, favoured solutions that let agents behave in appropriate ways, which means (here) appropriate for the survival and propagation of the genes concerned. So it is convenient to think of it as a 'program', to imagine alternative programs or subroutines that could compete with the present one and treat an agent as a temporary executor of these. However, Dawkins warns us to treat the analogy carefully. In this we find, besides a reminder that behaviour may look purposeful but not

necessarily is, the concept of “competing subroutines” which relates well to the notion of co-evolving behaviours.

As an added emphasis on the careful handling that is needed of the program analogy, Hendriks-Jansen (1996) warns us that computational models can be used to model behaviour but not minds, but the programs must not be used as explanations in themselves. Instead we must use the interrelationship between the agent’s activities and its environment as the explanation. There is a delightfully simple way of summarising these views, as put by Gozzi Jr. (1998), inspired by Wittgenstein: How do we know what something means? Watch how it is used, and find out.

This might be done by observing the agents in, for instance, Werner and Dyer (1991) who search for mates (cf. Section 3.4.2). The authors point out that they consider interpreting the meaning of a message to be problematic, reading them as e.g. “turn right”, since they could be only ‘hot’ and ‘cold’ messages. Since the signals are described in terms of the motions (a behaviour, discussed further in section 5.1) taken by the males, we are in a position to watch how the signals are used, and find out what they mean, in the context of these very agents in this very situation. There will be no absolutely universal or exact meaning forthcoming, but it could be argued that this is something we, on closer reflection, do not even have in human language. MacLennan (1991) draws the analogy to human language where it is possible for an utterance to have different pragmatic significance when it is spoken as when it is heard, and says that it would be wrong to claim that a symbol only can mean one specific situation. Noble and Cliff (1996) do not agree, and suggest that a language where a symbol can represent more than one state or a state can be represented by more than one symbol is inefficient. For the present purposes, a language that does not have a one-to-one correspondence with states is probably the most realistic and useful, and, as Dorffner, Prem, and Trost (1993) suggest, we do not need to look at semantic interpretability since symbols only exist through an individual’s interpretation. The observers and the interpreters are in this case the researchers.

Another, similar, way of interpreting the meaning of a signal used by an agent is to watch what it accomplishes, as Hutchins and Hazlehurst (1995) comment that the meaning of a

'good' lexicon can only be determined by the functional properties entailed in agents using the lexicon, and the authors further cite Clancey (1989): "Representations do not get to be symbols by virtue of *us* creating them or calling them such, but rather, by our reading of what they do for the members of a community of cognitive agents". The agents in Hutchins and Hazlehurst (cf. Section 3.4.3) do just that when they choose artefacts made by agents that have been successful in the past, and in turn are more successful. Since the artefacts have helped a successful agent to 'find more food when there is a favourable tide', the learning agent has done what we should do, watches what the artefact/signal/symbol does for other agents.

When discussing the possibilities of interpreting what we see, a discrepancy between the views of Cliff (1991) and Harnad (1994) is that Cliff claims that the semantics are grounded in computational mode, which Harnad would state as an interpretation not intrinsic to the system, but projected onto it by an interpreter (with a mind). This is interpreted as being due to different levels of analysis, and perhaps also different ways of defining grounding. Cliff may be looking for a justified way of interpreting simulation results, whereas Harnad is out to define artificial life.

To summarise, it seems that the researcher should be doubly careful, firstly when designing an experiment so as not to introduce intuitions or pre-conceptions and secondly, when observing and interpreting the results, in order to keep "wishful attributions" out of the objective results. If this is possible, it is reasonable to suppose that the system will sufficiently free from designer influence to be interpreted as a self-organised 'community'.

4.2.3 Spatial/temporal detachment

For agent communities to make full use of a communication system, it would be advantageous if they could communicate about objects they have perceived in a location that is the not the present one, e.g. "I found food over there", or perhaps "Beware of the north, there is a predator there". Are there agents that communicate about anything that is not 'right in front of them'? This is a question about spatial displacement, closely related to temporal displacement, which means the ability to talk about the past, which spatially translates to "where I was when I found the food" and the future, which is

"where you will be when you have found it". However, Evans and Marler (1995) claim that there is no evidence of temporal displacement in non-human animals (except honeybees), and hence it seems to be an aspect of language that is uniquely human.

Gärdenfors (1995) describes this as "detached representations", representations that concern an object or event that is neither present now, nor triggered by something that happened recently, i.e. something that can be evoked independently of the context in which the "memory" was created. The alternative is cued representations, which are either present or triggered by a recent situation. But there seems to be degrees of detachment; an example by Gärdenfors is object permanence: an agent knows that something "hid behind the curtain", which is at least independent of information provided by the senses. This reminds us of Brooks' agents (cf. Section 2.2), which would not 'survive' without sensory information, since they had no representations. It looks as if there are advantages to using detached representations, as an agent can use or 'learn' information about its environment that is not immediately perceived. To avoid completely reflexive agents, we obviously need some detachment, and Gärdenfors clarifies that call systems are described as automatic reactions, whereas language is a voluntary and detached system. Even if he in this case is discussing human language, it is clear that a call system, or an 'iconic' language will not suffice, and as an added bonus we find that, with detachment, we may be getting closer to an voluntary communication system.

The fact that practically no non-human animals have a detached language should not exclude the possibility of spatial or temporal displacement in A-Life models, although in many models the experiments are set up so that a prepared 'topic' is given directly as input or placed in such a way that the agents have to talk about it at the moment (or location) of perception. This happens in Steels' talking heads models (cf. Section 3.2.5), where the topic is forced upon the agents, and there is no real choice of what to speak of and where since the agents only have to agree on which of the objects in front of them is the topic. But in the robot experiments they segment the world into parts of their own 'choice'.

Furthermore, it could be argued that the 'words' constructed will not in their present form work for a detached language, since they usually consist of only the features necessary to discriminate objects in the present context. In the models studied by Cangelosi and co-workers (cf. Section 3.1), the mushroom features are directly input to the network, and the topic is always the closest mushroom, regardless of its distance. No agent discusses mushrooms in distant locations. The agents in Hutchins and Hazlehurst (1991, 1994), are not actually situated, and are shown only chosen aspects of the world (the moon phases and the artefacts, cf. Section 3.4.3), but are making an inference about the state of the world (tide) using what they perceive (moon). This is a detached representation, since the agents do not need to 'go to the beach' to find out the state of the tide.

In Billard and Dautenhahn (1997) there is some freedom in the topic choice, as the robots wander around, finding hills and plains, and there is in fact spatial displacement between the learner and the topic, but this turns out to be a problem that is hard to handle since the learning, due to this, is not perfect (cf. Section 3.3.1). The more recent experiment (2000, cf. Section 3.3.2) examines transmission of information regarding patches that have been encountered since two robots last met (this is slower than the one-to-many strategy, cf. below). The robots exchange signals regarding positions and colours of unknown (to the receiver) patches when they meet, which is similar to Moukas and Hayes (1996), where teachers return to a home base to communicate where they have found 'food' by performing a dance (cf. Section 3.4.6). What is characteristic of these two models is that they both claim to be inspired by bee dances, which, as commented above, has long been the only known animal communication system that touches on displacement of the topic. There are, however, results that point to that ants have a much more impressive communication system than has been known before (Reznikova & Ryabko, 2000). Is there no other way, then, to communicate about objects that are not immediately perceivable? Perhaps the researchers are too inspired by 'insect' communication systems to find another solution?

Several models work around this by allowing signalling/receiving distance to be more or less independent of the distance between sender and receiver/s. There are two strategies; the one-to-all 'infinite' distance (to the boundaries of the present environment), used by,

for instance, Billard and Dautenhahn (2000) in which a robot calls all other robots when a patch is found (cf. Section 3.3.2), and the one-to-many limited emission, shown by Werner and Dyer (1991), since the female emits a signal that is heard by all males within her "visual field" (cf. Section 3.4.2), and Mataric (1998): the sender transmits to all robots within the receptive field of the radio (cf. Section 3.4.10). Saunders and Pollack (1996) use a more realistic approach; the signal decays with distance (cf. Section 3.4.7). However, the receiver has no possibility of choosing whom to listen to, since the input is a summation of all other agents' signals.

In summary, this seems to be by far the most problematic aspect of self-organisation. In order to scale up the complexity of agents and worlds, it is reasonable to assume that this aspect must be explored further. It is also clear that only a few models approach anything like detachment, and this only because they seek inspiration from the honeybee, the only non-human animal that is known to have any detachment in its communication (apart from possibly ants). The implication of this is not clear, but it might be that inspiration should be sought from humans and other animals, including insects, and not just one of these areas.

5 Discussion and conclusions

The first section of this chapter discusses the contribution of A-Life approaches towards self-organised communication, and briefly compares this with results from traditional AI and connectionism. An emphasis is placed on the importance of co-evolution of behaviours, as opposed to looking at communication alone. The final section sums up the progress made towards self-organised communication in A-Life models.

5.1 Discussion

A central ambition of A-Life is, and has been, to study behaviour/s in a bottom-up synthetic fashion. In the very word synthesis lies an important hint, in that the goal is to integrate and not to separate behaviours. This chapter discusses this issue in the context of the reviewed models, where the focus has been on self-organised communication. Communication, it has been shown, is more often than not an interaction between agents, and as such one behaviour among many possible ones. For instance, Burghardt (1970) comments that natural selection produces and maintains end-directed behaviour, and that the prime example is communication, which not only tells us that communication *is* a behaviour, but that it might be end-directed, i.e. a goal is needed, which is discussed below. It could further be argued that A-Life researchers should co-evolve communication, as a behaviour, together with other behaviours to a greater degree than is done now. This should take place in an environment that among other things contain con-specifics, effectively leading to the joint self-organisation of several behaviours, including communication. When examining different levels of self-organisation, some subquestions have been asked, and this section summarises the answers found.

If we compare GOFAI, connectionism, and A-Life, we find that communication, in a strong sense, is not self-organised in any of the approaches. The first two have been discussed in sections 2.1 and 2.2, respectively. In A-Life different degrees of feedback effectively force agents to communicate, and the only difference lies in the kind of feedback given to the agents. In the animal world, communication has of course evolved due to environmental pressure, but animals or species practically always have alternative ways of reacting to some environmental influence or situation (cf. Section 4.1.1), and are

hardly ‘rewarded’ directly for using communication ‘in itself’ over other behaviour. One possible way of achieving self-organised communication in A-Life models, would be to co-evolve behaviours, such that agents might ‘choose’ one of many possible alternative reactions, one of which then would be communication. This approach was tried (briefly) and abandoned by Cangelosi and Parisi (1998) and Cangelosi (1999) respectively (cf. the previous section).

The customary way to rank feedback ranges from instructive (supervised) via evaluative (reinforcement) to absent (unsupervised)[2]. However, the feedback considered in this dissertation is the one given by the designer explicitly in the form of a fitness function or backpropagation of error. Feedback that the agent itself generates, e.g., by evaluating the effect of an action will not be considered designer feedback, which differs from de Jong (2000) who considers this reinforcement learning. On the other hand, it agrees with Dorffner (1997) who regards self-organisation as unsupervised learning, at least in the context of radical connectionism (cf. Section 2.2). An extension of this argument is that the problem is not actually the kind of feedback given, but why and how it is given. In nature, fitness is related to the survival of the fittest, and fitness is not ‘smart’, meaning that fitness does not ‘look ahead’. There is no direct fitness increase for, for instance, communication acts that in turn lead to finding food or mates, which then in turn increases the chances of survival for the animal/s involved. In A-Life, the designer typically decides in beforehand which behaviour will lead to positive feedback (increased fitness) for one or several agents. This feedback may be explicitly given by the designer, and most certainly has been explicitly chosen.

None of the agents in the models reviewed are ‘allowed’ to ‘choose’ their behaviour, but are forced (by a ‘fitness function’) to communicate, since all researchers use different degrees of instructive or evaluative feedback, rewarding communication, and actually not giving any options as to how to behave. In Cangelosi’s experiments, various backpropagation algorithms and imitation tasks are used as feedback, and at first (Cangelosi & Parisi, 1998) foraging behaviour is co-evolved with communication (cf. Section 3.1.1), but later (1999) the author comments that the simultaneous evolution

[2] These phrases are used in de Jong (2000), whereas the parenthesised expressions are the standard terms in the field.

proved difficult, and therefore is divided into two stages (cf. Section 3.1.3). In, for instance, Steels' talking heads project (Steels & Kaplan, 1999, 2000) it could be argued that agents are generating their own feedback, but on the other hand this is all they do, and they are well prepared for it (cf. Section 3.2.5). In fact, in all language games there is no co-evolution or integrated self-organisation of behaviour, as for instance the robots in Steels and Vogt (1997) use pre-programmed behaviour modes, and in all simulations the agents are immobile (cf. Section 3.2.4). In the robot models by Billard and Dautenhahn (1997, 2000) the robots use a combination of pre-programmed random wandering and following behaviours, and as feedback built-in "needs" assumed to be 'instinct-like' (cf. Section 3.3.1) and a confidence factor in the later experiments (cf. Section 3.3.2). This means that there is no co-evolution of behaviours, and that evaluative feedback is given. In the 1997 model, life-long learning is said to be combined with built-in basic behaviours, but after the search stage begins the learning is not used, why it could be argued that lifetime learning in this case is not relevant, since it neither used, nor really sought after (cf. Section 3.3.1).

When we look at what it is the agents are 'talking' about, and compare to earlier approaches, it seems that not much has changed. In GOFAI abstract symbols were syntactically manipulated, while connectionism used self-organised patterns. In both GOFAI and connectionism input and output was prepared and interpreted by a designer, and A-Life was thought a likely candidate to remedy this. The symbols or signals that are used now are grounded in agent-environment interaction, but carry no inherent meaning since the objects were chosen to have certain meanings for the agents, as in the PacMan syndrome: "this is food" (cf. Section 4.2 and 4.2.1).

If behaviours were co-evolved, i.e., several behaviours were somehow equally likely for the agent to evolve and/or choose as a reaction to a situation (where both the situation could evolve and the self-organisation could take place in a short or long time span), this would probably assist in giving objects and situations in the agents' world meaning for them. For example, foraging (alone or in a group), finding food, feeding and letting others know are activities that are interrelated and not easily separated without loss of coherence in 'the life of an agent'. Being situated is a step in the right direction, but not enough; an agent must be 'interactive' as well. The 'classic' way of labelling abstract

objects therefore still remains, leaving A-life not quite as far along as one might have expected in the quest for self-organised communication. The designers of experiments are thus still wielding 'abstraction as a dangerous weapon' (Brooks, 1991a, cf. Section 2.1) and the agents have but one task, to ascertain which label goes with which object.

The question of detachment is not really applicable to GOFAI or connectionism, since situatedness most probably is necessary for speaking of detached objects, definitely if they are spatially detached, but also temporally. This seems to be a question to ask of A-Life alone, and is not one that has been studied explicitly by anyone (to our knowledge) but this will be necessary soon in order to expand the agent's worlds. Many experiments assume that some information about the environment is available to the sender but not perceivable by the receiver/s. Communication is often about something that is spatially (or temporally) displaced, in the sense of being 'not here', whether it is not directly perceivable by the receiver or it is not actually at the present position. There are of course studies, especially in robotics, where for instance positions of objects are taught to 'learners', but they are about static worlds and contain no predictions. An example of such a prediction would be something like this: "the box is at X, Y now and will be at X1, Y1 in timestep+1". Prediction within one agent has been simulated, but as far as is known not communicated to another agent. If we want agents to cope in dynamic worlds this issue must be seriously considered. A special case might be the work of Hutchins and Hazlehurst (1991, 1994), in which agents cooperate to create artefacts with which the state of the world can be predicted, even though the state is cyclic and easily predictable (moon phase vs. state of the tide, cf. Section 3.4.3). This implies that social and cultural aspects may be important in 'detached' communication.

5.2 Conclusions

The criticism of the A-Life approach as exemplified by the models reviewed here, could be condensed to that there is little or no integrated co-evolution of behaviours. In the enactive approach, the coupling of individual and environment is central, and MacLennan (1991) defines synthetic ethology, very closely related to A-Life as mentioned in section 2.3, as the study of synthetic life forms in a synthetic world to which they have become coupled through evolution. Furthermore, it could be argued that coupling an agent to its

environment encompasses all its behaviours, and not just communication, as in many of the cases reviewed. Steels (1995) claims that:

> "To make selectionism work for robots...we should not concentrate on the acquisition of a single behavior system...there should be many diverse behavior systems that are complementary but still in competition. Paradoxically, it might be easier to develop many behavior systems at once than to concentrate on one behavior system in isolation. Each behavior system, except for the basic reflexes, should remain adaptive, just as each individual organism remains adaptive" (p 99).

Returning to the issues discussed in the previous chapter, how might co-evolution of behaviours assist in self-organised communication and the integration of communication and other behaviour?

Regarding the 'start' of communication, it could be assumed that communication will 'appear' when the dynamics of a group of agents with several, possibly composite, behaviours in some way makes it profitable to choose communication over other behaviour (cf. Kirsh, 1996, section 1). The 'choosing' of communication as the adaptive behaviour may be due to purely social reasons or perhaps cooperative needs. The problem of rewarding communication in order to make it continuously more successful disappears, since agents presumably will evolve and/or learn what they need to communicate about and which signals to use, and evolution will take care of solutions that are not useful, at least not by themselves. Whether communication turns out to be cooperative or asymmetric will also be shown in time, and different strategies will most probably surface in community use.

What will the signs mean to the agents? This is something that, as above, will evolve, and most probably consist of functional and dynamic meanings, since this is the way the world is, not static and well prepared for an agent's use. How, then, will *we* understand what it all means, since it looks like it could be a closed autonomous system? We will get a pretty good idea of what it is all about by observing, and also seeing what signs 'do'

for the agents functionally. We might have to do as we do when we find ourselves in unknown situations, perhaps abroad, 'talk to the natives' and gradually build a subjective understanding of what is going on.

Regarding the question of spatial and/or temporal detachment, this is the least explored issue, and there is little evidence of explicit studies. This is an aspect that most probably will be important when agents and environments are scaled up to higher complexity, not to mention when environments are dynamic. It may also be important to consider when adding social and cultural aspects to the models, for without the possibility to communicate about abstract, 'non-present' topics, how will it be possible for an agent to communicate 'thoughts' or 'feelings' to another? This issue is one that will have to be left more or less open, and left to researchers to pursue further.

More importantly, how should this system be 'designed' with little designer influence? There is, as yet, no possible way that a system can come into existence without being 'constructed', and if we want to examine communication in addition to other behaviours, there must be a reasonable possibility for an agent to evolve it. We might use the basic ideas from radical connectionism (Dorffner, 1997; Dorffner, Prem & Trost, 1993), i.e. using self-organisation – by combining evolution and learning, only using sensori-motor interfaces to the environment, and naturally using autonomous, situated agents. Furthermore, there is pre-design of some components; e.g. concept formation and referential links (Dorffner, Prem & Trost, 1993). If this were complemented by incremental evolution, as advocated by Nolfi (1998) (working on evolutionary robotics, but not explicitly on communication) to gradually increase the level complexity of the environment and the agents' behaviours, there would not be much need for human pre-design or supervision in the ways we have seen used in current models. On the other hand, the pre-design is shifted towards the breaking down of tasks or competencies into incremental steps or parts (e.g. Nolfi, 1998). It might also be possible to use fitness functions or 'rewards' consisting of several levels, considering an individual lifetime, the evolutionary time of a species or population. Further, it is conceivable to use some sort of fitness value assigned to the environment or to some meta-level of the system, since we are interested in the system as a whole and not just the separate parts. Finally, if this is combined with Cliff's ideas (cf. Section 4.2), generating 'hard' objective measurements

by observing well grounded agents, we may be well on the way to realising our goal of self-organised communication between autonomous agents.

References

Ackley, D. H., & Littman, M. L. (1994). Altruism in the Evolution of Communication. In R. A. Brooks & P. Maes (Eds.), *Artificial Life IV: Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems.* pp. 40-48. Cambridge, MA: The MIT Press.

Agre, P. E., & Chapman, D. (1987). Pengi: An Implementation of a Theory of Activity. *Proceedings of AAAI-87.* pp. 268-272. Menlo Park, CA: AAAI.

Aitchison, J. (1998). On discontinuing the continuity-discontinuity debate. In J. R. Hurford, M. Studdert-Kennedy, & C. Knight (Eds.), *Approaches to the Evolution of Language: Social and Cognitive Bases.* Cambridge: Cambridge University Press.

Balch, T., & Arkin, R. C. (1994). Communication in Reactive Multiagent Robotic Systems. *Autonomous Robots*, 1, pp. 1-25.

Balkenius, C. & Winter, S. (1999). Explorations in Synthetic Pragmatics. In A. Riegler, M. F. Peschl, & A. von Stein (Eds.), *Understanding Representation in the Cognitive Sciences: Does Representation Need Reality?* New York: Kluwer Academic / Plenum Publishers.

Billard, A., & Dautenhahn, K. (1997). The social aspect of communication: a case study in the use and usefulness of communication for embodied agents. In P. Husbands & I. Harvey (Eds.), *Proceedings of the Fourth European Conference on Artificial Life.* Cambridge, MA: The MIT Press.

Billard, A., & Dautenhahn, K. (2000). Experiments in social robotics: Grounding and Use of Communication in Autonomous Agents. *Adaptive Behavior*, 7 (3-4).

Brooks, R. A. (1990). Elephants don't Play Chess. *Robotics and Autonomous Systems,* 6 (1-2), pp. 1-16.

Brooks, R. A. (1991a). Intelligence without Representation. *Artificial Intelligence*, 47, pp. 139-159.

Brooks, R. A. (1991b). Intelligence without reason. *Proceedings of IJCAI '91,* Sydney, Australia.

Bullock, S. (1997). An exploration of signalling behaviour by both analytic and simulation means for both discrete and continuous models. In P. Husbands & I. Harvey (Eds.), *Proceedings of the Fourth European Conference on Artificial Life*. pp. 454-463. Cambridge, MA: The MIT Press.

Bullock, S. (2000). *Something to talk about: Conflict and coincidence of interest in the evolution of shared meaning*. Paper presented at the Proceedings of the 3rd Conference on The Evolution of Language, Paris, France.

Burghardt, G. M. (1970). Defining "Communication". In J. W. J. Johnston, D. G. Moulton, & T. Amos (Eds.), *Communication by Chemical Signals.* pp. 5-18. New York: Appleton-Century-Crofts.

Cangelosi, A. (1999). Modeling the Evolution of Communication: From Stimulus Associations to Grounded Symbolic Associations. In D. Floreano, J-D. Nicoud, & F. Mondada (Eds.), *Advances in Artificial Life: Proceedings of the 5th European Conference, ECAL'99.* Berlin: Springer-Verlag.

Cangelosi, A. (2000) (submitted). Evolution of Communication and Language using Signals, Symbols, and Words. *IEEE Transactions in Evolutionary Computing*.

Cangelosi, A., Greco, A., & Harnad, S. (2000). From Robotic Toil to Symbolic Theft: Grounding Transfer from Entry-Level to Higher-Level Categories. *Connection Science*, 12(2), pp. 143-162.

Cangelosi, A., & Harnad, S. (2000) (accepted for publication). The Adaptive Advantage of Symbolic Theft over Sensorimotor Toil: Grounding Language in Perceptual Categories. *Evolution of Communication*.

Cangelosi, A., & Parisi, D. (1998). The Emergence of a 'Language' in an Evolving Population of Neural Networks. *Connection Science*, 10(2), pp. 83-97.

Clancey, W. (1989). *Ontological commitments and cognitive models*. Paper presented at the Proceedings of The eleventh annual conference of the cognitive science society, Ann Arbor, Michigan.

Clancey, W. J. (1995). A Boy Scout, Toto, and a Bird: How situated cognition is different from situated robotics. In L. Steels & R. Brooks (Eds.), *The "Artificial Life" Route to "Artificial Intelligence": Building Situated Embodied Agents.* pp. 227-236. Hillsdale, NJ: Lawrence Erlbaum Associates.

Clark, A. (1997). *Being There*. Cambridge, MA: The MIT Press.

Cliff, D. (1991). Computational Neuroethology; A Provisional Manifesto. In J-A. Meyer & S. W. Wilson (Eds.), *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*. pp. 29-39. Cambridge, MA: The MIT Press.

Dawkins, R. (1982). *The Extended Phenotype*. Oxford: Oxford University Press.

de Bourcier, P., & Wheeler, M. (1997). The Truth Is Out There: the Evolution of Reliability in Aggressive Communication Systems. In P. Husbands & I. Harvey (Eds.), *Proceedings of the Fourth European Conference on Artificial Life.* pp. 444-453. Cambridge, MA: The MIT Press.

de Jong, E. D. (2000). *Autonomous Formation of Concepts and Communication*. Ph.D. thesis, Vrije Universiteit, Brussels.

Di Paolo, E. A. (1997). Social coordination and spatial organization: Steps towards the evolution of communication. In P. Husbands & I. Harvey (Eds.), *Proceedings of the Fourth European Conference on Artificial Life.* pp. 464-473. Cambridge, MA: The MIT Press.

Di Paolo, E. A. (1998). An Investigation into the Evolution of Communication. *Adaptive Behavior*, 6 (2), pp. 285-325.

Dorffner, G. (1997). Radical connectionism - a neural bottom-up approach to AI. In G. Dorffner (Ed.), *Neural Networks and a New Artificial Intelligence*. London: International Thomson Computer Press.

Dorffner, G., Prem, E., & Trost, H. (1993). *Words, Symbols, and Symbol Grounding* (TR-93-30). Austrian Research Institute for Artificial Intelligence.

Dreyfus, H. L. (1979) *What Computer's Can't Do, Revised Edition: The Limits of Artificial Intelligence*. New York: Harper & Row.

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science*, 14, pp. 179-211.

Elman, J. L. (1998). Connectionism, artificial life, and dynamical systems. In W. Bechtel & G. Graham (Eds.), *A Companion to Cognitive Science*. Malden, MA: Blackwell.

Evans, C. S., & Marler, P. (1995). Language and Animal Communication: Parallels and Contrasts. In H. L. Roitblat & J-A. Meyer (Eds.), *Comparative Approaches to Cognitive Science.* pp. 341-382. Cambridge, MA: MIT Press.

Franklin, S. (1995). *Artificial Minds*. Cambridge, MA: The MIT Press.

Gozzi Jr., R. (1998). Is language a game? *ETC: A Review of General Semantics, Summer98*, 55 (2), pp. 189-195.

Gärdenfors, P. (1995). *Language and the Evolution of Cognition* (LUCS 41). Lund University Cognitive Studies.

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

Harnad, S. (1994). Artificial Life; Synthetic vs. Virtual. In C.G Langton (Ed.), *Artificial Life III: Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. XVII*. Reading, MA: Addison-Wesley.

Harnad, S. (1996). The Origin of Words: A Psychophysical Hypothesis. In W. Durham & B. Velichkovsky (Eds.), *Communicating Meaning: Evolution and Development of Language*. Hillsdale, NJ: Erlbaum.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. Cambridge, MA: The MIT Press.

Hauser, M. D. (1996). *The Evolution of Communication*. Cambridge, MA: MIT Press.

Hendriks-Jansen, H. (1996). *Catching Ourselves in the Act*. Cambridge, MA: The MIT Press.

Hurford, J. R. (2000) The roles of communication and representation in language evolution. *Proceedings of the 3rd Conference on The Evolution of Language.* Paris, France.

Hutchins, E. (1995). *Cognition in the Wild*. Cambridge, MA: The MIT Press.

Hutchins, E., & Hazlehurst, B. (1991). Learning in the Cultural Process. In C. G. Langton, C. Taylor, J. D. Farmer, & S. Rasmussen (Eds.), *Artificial Life II. Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. X*. pp. 689-706. Redwood, CA: Addison-Wesley.

Hutchins, E., & 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*. London: UCL Press.

Jung, D., & Zelinsky, A. (2000). Grounded Symbolic Communication between Heterogeneous Cooperating Robots. *Autonomous Robots*, 8 (3). pp. 269-292.

Kirby, S. (1999). Syntax out of Learning: The Cultural Evolution of Structured Communication in a Population of Induction Algorithms. In D. Floreano, J-D. Nicoud, & F. Mondada (Eds.), *Advances in Artificial Life: Proceedings of the 5th European Conference, ECAL'99.* Berlin: Springer-Verlag.

Kirby, S., & Hurford, J. (1997). Learning, Culture and Evolution in the Origin of Linguistic Constraints. In P. Husbands & I. Harvey (Eds.), *Proceedings of the Fourth European Conference on Artificial Life.* Cambridge, MA: The MIT Press.

Kirsh, D. (1996). Adapting the Environment Instead of Oneself. *Adaptive Behavior*, 4 (3/4), pp. 415-452.

Langton, C. G. (1995). Editor's Introduction. In C. G. Langton (Ed.), *Artificial Life: An Overview.* Cambridge,MA: The MIT Press.

Linell, P. (1978). *Människans språk*. Malmö, Sweden: Gleerups.

MacLennan, B. (1991). Synthetic Ethology: An Approach to the Study of Communication. In C. G. Langton, C. Taylor, J. D. Farmer, & S. Rasmussen (Eds.), *Artificial Life II. Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. X*. pp. 631-657. Redwood, CA: Addison-Wesley.

MacLennan, B. J., & Burghardt, G. M. (1994). Synthetic Ethology and the Evolution of Cooperative Communication. *Adaptive Behavior*, 2 (2), pp. 161-188.

Mataric, M. J. (1998). Using communication to reduce locality in distributed multi-agent learning. *Journal of Experimental and Theoretical Artificial Intelligence*, 10, pp. 357-369.

Maturana, H., & Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht, Holland: D. Reidel Publishing.

Moukas, A., & Hayes, G. (1996). Synthetic Robotic Language Acquisition by Observation. In P. Maes, M. J. Mataric, J-A. Meyer, J. Pollack, & S. W. Wilson (Eds.), *From Animals to Animats 4, Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*. pp. 568-579. Cambridge, MA: The MIT Press.

Noble, J. (1998a). *The Evolution of Animal Communication Systems: Questions of Function Examined through Simulation*. D.Phil. thesis, University of Sussex, Brighton, UK.

Noble, J. (1998b). Evolved Signals: Expensive Hype vs. Conspiratorial Whispers. In C. Adami, R. K. Belew, H. Kitano, & C. E. Taylor (Eds.), *Proceedings of the Sixth International Conference on Artificial Life, ALife 6*. Cambridge, MA: The MIT Press.

Noble, J. (1999). Sexual Signalling in an Artificial Population: When Does the Handicap Principle Work? In D. Floreano, J-D. Nicoud, & F. Mondada (Eds.), *Advances in Artificial Life: Proceedings of the 5th European Conference, ECAL '99*. New York: Springer.

Noble, J. (2000). Defining animal communication, and why it matters for understanding the evolution of language. *Proceedings of the 3rd Conference on The Evolution of Language*. Paris, France.

Noble, J., & Cliff, D. (1996). On Simulating the Evolution of Communication. In P. Maes, M. J. Mataric, J-A. Meyer, J. Pollack, & S. W. Wilson (Eds.), *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*. pp. 608-617. Cambridge, MA: MIT Press.

Nolfi, S. (1998). Evolutionary Robotics: Exploiting the Full Power of Self-organisation. *Connection Science*, 10 (3&4), pp. 167-184.

Parisi, D. (1997). An Artificial Life Approach to Language. *Brain and Language*, 59, pp. 121-146.

Parisi, D., Denaro, D., & Cangelosi, A. (1996). *Categories and Word Meanings in Adaptive Organisms* (Technical Report No. NSAL-96-004). Institute of Psychology, National Research Council, Rome.

Peschl, M. F., & Riegler, A. (1999). Does Representation Need Reality? Rethinking Epistemological Issues in the Light of Recent Developments and Concepts in Cognitive Science. In A. Riegler, M. F. Peschl, & A. von Stein (Eds.), *Understanding Representation in the Cognitive Sciences: Does Representation Need Reality?* New York: Kluwer Academic / Plenum Publishers.

Pfeifer, R. (1996). Building "Fungus Eaters": Design Principles of Autonomous Agents. In P. Maes, M. J. Mataric, J-A. Meyer, J. Pollack, & S. W. Wilson (Eds.), *From Animals to Animats 4, Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*. pp. 3-12. Cambridge, MA: The MIT Press.

Pinker, S. (1994). *The Language Instinct*. London: Penguin Books.

Prem, E. (1995). Dynamic Symbol Grounding, State Construction, and the Problem of Teleology. In J. Mira & F. Sandoval (Eds.), *From Natural to Artificial Neural Computation, Proceedings of the International Workshop on Artificial Neural Networks*. pp. 619-626. New York: Springer.

Prem, E. (1998). *Semiosis in embodied autonomous systems.* Paper presented at the Proc. of the ISIC/CIRA/ISAS98, Madison, WI.

Reznikova, Z., & Ryabko, B. (2000) Using Information Theory Approach to study the Communication System and Numerical Competence in Ants. In J-A. Meyer, A. Berthoz, D. Floreano, H. Roitblat, & S.W. Wilson (Eds.) *From Animals to Animats 6, Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior.* pp. 501-506.

Ronald, E. M. A., Sipper, M., & Capcarrère, M. S. (1999) Design, Observation, Surprise! A Test of Emergence. *Artificial Life*, 5, pp. 225-239.

Rumelhart, D. E., & McClelland, J. L. (1986). On Learning the Past Tenses of English Verbs. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* pp. 216-271. Cambridge, MA: The MIT Press.

Russell, S. J., & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. London: Prentice-Hall.

Saunders, G. M., & Pollack, J. B. (1996). The Evolution of Communication Schemes Over Continuous Channels. In P. Maes, M. J. Mataric, J-A. Meyer, J. Pollack, & S. W. Wilson (Eds.), *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior.* Cambridge, MA: MIT Press.

Savage-Rumbaugh, S., & Rumbaugh, D. M. (1978). Symbolization, language, and Chimpanzees: A theoretical reevaluation on Initial language acquisition processes in four Young *Pan troglodytes. Brain and Language*, 6, pp. 265-300.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals and understanding*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Searle, J. R. (1980) Minds, Brains, and Programs. *Behavioral and Brain Sciences,* 1, pp. 417-424.

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel Networks That Learn to Pronounce English Text. *Complex Systems*, 1, pp. 145-168.

Shannon, C. E. (1948) A mathematical theory of communication. *Bell System Technical Journal*, vol. 27, pp. 379-423 and 623-656.

Sharkey, N. E., & Ziemke, T. (1998). A Consideration of the Biological and Psychological Foundations of Autonomous Robotics. *Connection Science*, 10 (3-4), pp. 361-391.

Sjölander, S. (1995). Some Cognitive Breakthroughs in the Evolution of Cognition and Consciousness and their Impact on the Biology of Language. *Evolution and Cognition*, 1 (1), pp. 3-11.

Steels, L. (1995). The Artificial Life Roots of Artificial Intelligence. In C. G. Langton (Ed.), *Artificial Life: An Overview.* pp. 75-110. Cambridge, MA: The MIT Press.

Steels, L. (1996a). Emergent Adaptive Lexicons. In P. Maes, M. J. Mataric, J-A. Meyer, J. Pollack, & S. W. Wilson (Eds.), *From Animals to Animats 4, Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior.* pp. 562-567. Cambridge, MA: The MIT Press.

Steels, L. (1996b). A Self-Organizing Spatial Vocabulary. *Artificial Life*, 2 (3), pp. 319-332.

Steels, L. (1997a). Constructing and Sharing Perceptual Distinctions. In M. van Someren & G. Widmer (Eds.), *Proceedings of the European Conference on Machine Learning*. Berlin: Springer-Verlag.

Steels, L. (1997b). The Synthetic Modeling of Language Origins. *Evolution of Communication*, 1 (1), pp. 1-34.

Steels, L., & Kaplan, F. (1999). Situated Grounded Word Semantics. In T. Dean (Ed.), *Proceedings of IJCAI '99.* pp. 862-867: Morgan Kaufmann Publishers.

Steels, L., & Kaplan, P. (2000). Bootstrapping Grounded Word Semantics. In T. Briscoe (Ed.), *Linguistic Evolution through Language Acquisition: Formal and Computational Models.* Cambridge: Cambridge University Press.

Steels, L., & Vogt, P. (1997). Grounding adaptive language games in robotic agents. In P. Husbands & I. Harvey (Eds.), *Proceedings of the Fourth European Conference on Artificial Life.* pp. 474-482. Cambridge, MA: The MIT Press.

Studdert-Kennedy, M., Knight, C., & Hurford, J. R. (Eds.). (1998). *Approaches to the Evolution of Language*. Cambridge, UK: Cambridge University Press.

Thelen, E., & Smith, L. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, Ma: The MIT Press.

Theraulaz, G., & Bonabeau, E. (1999). A Brief History of Stigmergy. *Artificial Life*, 5, pp. 97-116.

Turing, A. M. (1950) Computing machinery and intelligence. *Mind,* 59, pp. 433-460.

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The Embodied Mind*. Cambridge, MA: The MIT Press.

Weizenbaum, J. (1965). ELIZA - A Computer Program For the Study of Natural Language Communication Between Man and Machine. *Communications of the Association for Computing Machinery*, 9 (1), pp. 36-45.

Weizenbaum, J. (1976). *Computer Power and Human Reason: from judgment to calculation*. New York: W.H. Freeman and Co.

Werner, G. M., & Dyer, M. G. (1991). Evolution of Communication in Artificial Organisms. In C. G. Langton, C. Taylor, J. D. Farmer, & S. Rasmussen (Eds.), *Artificial Life II. Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. X*. pp. 659-687. Redwood, CA: Addison-Wesley.

Winter, S. (1998). *Expectations and Linguistic Meaning.* Doctoral Dissertation, LUCS 71, University of Lund, Lund.

Yanco, H., & Stein, L. A. (1993). An Adaptive Communication Protocol for Cooperating Mobile Robots. In S. W. Wilson, J-A. Meyer, & H. L. Roitblat (Eds.), *From Animals to Animats II: Proceedings of the Second International Conference on Simulation of Adaptive Behavior.* pp. 478-485. Cambridge, MA: The MIT Press.

Ziemke, T. (1997). Embodiment and Context. In Proceedings 'Workshop on Context', European Conference on Cognitive Science, Manchester, UK.

Ziemke, T. (1999). Rethinking Grounding. In A. Riegler, M. F. Peschl, & A. von Stein (Eds.), *Understanding Representation in the Cognitive Sciences: Does Representation Need Reality?* New York: Kluwer Academic / Plenum Publishers.

Ziemke, T., & Sharkey, N. E. (in press). A stroll through the worlds of robots and animals: Applying Jakob von Uexküll's theory of meaning to adaptive robots and artificial life. Semiotica.