Language evolution and robotics: Issues on symbol grounding and language acquisition

Paul Vogt[1,2]

[1]University of Edinburgh
School of Philosophy, Psychology and Language Sciences
Language Evolution and Computation Research Unit
40 George Square, Edinburgh EH8 9LL, UK
[2]Tilburg University
Computational Linguistics
Induction of Linguistic Knowledge
P.O. Box 90153, 5000 LE Tilburg, The Netherlands
email: paulv@ling.ed.ac.uk
URL: http://www.ling.ed.ac.uk/~paulv

DRAFT VERSION. To appear in:

Artificial Cognition Systems
Angelo Loula, Ricardo Gudwin, João Queiroz (Editors)

1. Introduction

One of the key aspects that distinguishes humans from other species is that humans use a complex communication system that is - among other things - symbolic, learnt, compositional and recursive, whereas all other species' communication systems typically lack these properties. It is often thought that this unique human feature is the key to understanding the nature (and nurture!) of human cognition. In order to understand the foundations of this distinction between humans and other species, scientists study the *origins and evolution of language*.

Traditionally the origins and evolution of language has been studied by biologists, anthropologists, psychologists, palaeontologists, philosophers and linguists - although the Linguistic Society of Paris had strangely enough banned any studies on this issue between 1866 and 1974, because too many theories were proposed that were hard to verify at the time. With the recent advancements in computational resources an increasing number of simulations studying various aspects of language origins and evolution have emerged (see, e.g., Steels, 1997; Cangelosi and Parisi, 2002; Briscoe, 2002; Kirby, 2002 for overviews).

Mostly, these computational studies incorporate a multi-agent system that can learn, or evolve, a communication system of varying complexity that allows the system to communicate about a predefined set of meanings. However, as human communication is about the real world, understanding the underlying principles of language requires an understanding of the mechanisms with which the languages' meanings are rooted in reality. Models based on predefined meanings therefore face what is often referred to as the *symbol grounding problem* (Harnad, 1990). Few studies have tried to tackle this problem using robotic models of language origins and evolution, most notably (Marocco et al., 2003; Vogt, 2000a; Steels and Vogt, 1997; Steels et al., 2002).

In this chapter, I will present an overview of robotic (and other related) studies on the evolution of language. The aim is to present why robotics is a fruitful approach to study language origins and evolution, identify the main topics, report the major achievements and problems and provide a roadmap to future studies. Although I will cover most robotic studies on the evolution of language, the overview is not exhaustive and will, for instance, not cover studies on language learning robots, such as (Oates et al., 2000; Steels and Kaplan, 2000; Sugita and Tani, 2005; Roy, 2000), since these deal with human-robot interaction rather than with multi-robot communication.

In the next section, I will provide some theoretical background on language evolution, discuss an alternative view on the symbol grounding problem and present some foundations toward studying language evolution using robots. In Section 3, I will present an overview of topics that have been studied in language evolution robotics. These topics will be illustrated with case studies and a critical review of the approaches taken. An outlook to future endeavours is presented in Section 4. Section 5 concludes this chapter.

2. Background

The question of why humans have evolved the ability to use natural language is one of the most intriguing in contemporary cognitive science, and possibly one of the hardest problems in science (Christiansen and Kirby, 2003). Looking at recent collections on the evolution of language (e.g., Christiansen and Kirby, 2003), we can find that most prominent questions include: For what purpose has human languages evolved? How have human sound systems evolved? How have we become symbolic species? How have we established a shared signalling system of symbols? How has syntax emerged? How has linguistic diversity emerged? How do languages change through time? Among these questions, the emergence of syntax is considered by many to be the most important question.

One of the most prominent debates regarding language evolution concerns the nature versus nurture paradigm. On the one side, many scholars adhere to the nativist approach, which aims at explaining language universals in terms of biological adaptations (Chomsky, 1980; Pinker and Bloom, 1990). Only a few modellers take up this approach by developing models that try to evolve an innate Universal Grammar (e.g., Briscoe, 2000; Nowak et al., 2000; Yamauchi, 2004). On the other side of the debate are those who believe language is an empirically learnt system (Elman et al., 1996, MacWhinney, 1999) or a culturally evolved system (Tomasello, 1999). Most computer modellers follow the cultural evolution paradigm and assume that language is a *complex adaptive dynamical system* (Steels, 1997). In this paradigm, language is assumed to have evolved through self-organisation resulting from cultural interactions and individual learning. Following Kirby and Hurford (2002), this paradigm regards language evolution as an interplay "between three complex adaptive systems:

Learning.
> During *ontogeny* children adapt their knowledge of language in response to the environment in such a way that they optimise their ability to comprehend others and to produce comprehensible utterances.

Cultural evolution.
> On a historic (or *glossogenetic*) time scale, languages change. Words enter and leave the language, meanings shift, and phonological and syntactic rules adjust.

Biological evolution.
> The learning (and processing) mechanisms with which our species has been equipped for language, adapt in response to selection pressures from the environment, for survival and reproduction." (Kirby and Hurford, 2002, p. 122)

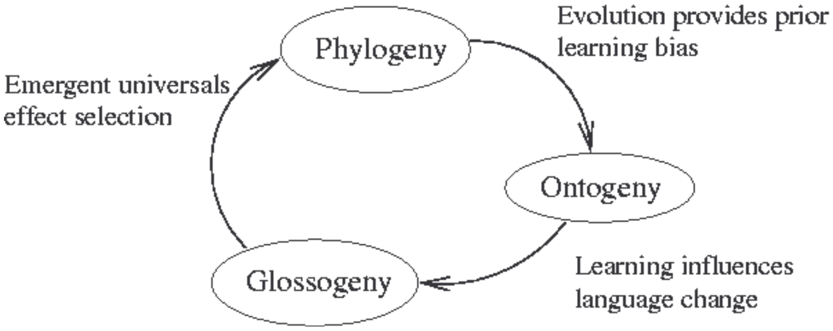
Phylogeny
Evolution provides prior learning bias
Emergent universals effect selection
Ontogeny
Glossogeny
Learning influences language change

Figure 1: This figure illustrates how the three adaptive systems interact to explain the emergence of language. (Adapted from Kirby and Hurford (2002).)

Figure 1 illustrates the interaction between these three complex adaptive systems. The remainder of this chapter will primarily focus on studies involving the complex adaptive system, as this is - up to now - the most studied paradigm in robotic studies.

Although, as mentioned, questions relating to the emergence of syntax are generally considered the most important questions, it has been argued that the first important questions should relate to the emergence of *symbolic* communication (Deacon, 1997; Jackendoff, 1999). Traditional cognitivist approaches in cognitive science have assumed that human cognition can be seen as a *physical symbol system* (Newell and Simon, 1976). Physical symbol systems are systems that can store, manipulate and interpret symbolic structures according to some specified rules. Assuming this is a correct characterisation, we need to define symbols. Cognitivists have treated symbols as internal structures that - following De Saussure (1974) - relate representations of meanings with arbitrary signals or labels. In these approaches, it is left unspecified how the meaning relates to reality, which has caused famous fundamental problems such as the *frame problem* (McCarthy and Hayes, 1969), the *Chinese Room problem* (Searle, 1980), and the *symbol grounding problem* (Harnad, 1990). The main issue with these problems is that in the cognitivist approach, symbols are neither situated –

i.e. they are not acquired in interaction with an environment, nor embodied – i.e. they are not based on bodily experiences (Clancey, 1997; Pfeifer and Scheier, 1999).

To deal with these problems of cognitivism, Brooks (1990) proposed the *physical grounding hypothesis*, which states that intelligence should be grounded in the interaction between a physical agent and its environment. In the physical grounding hypothesis, Brooks has argued that symbols are no longer necessary; intelligent behaviour can be established by parallel operating sensorimotor couplings. Although physically grounded systems are both situated and embodied, from the point of linguistics, Brooks' hypothesis is problematic, since human language is indeed considered to be symbolic.

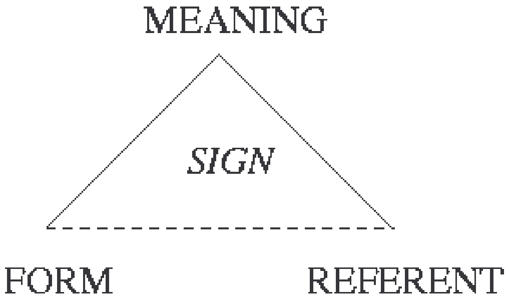

Figure 2: The semiotic triangle illustrates the relations that constitute a sign. When the form is either arbitrary or conventionalized, the sign can be interpreted as a symbol. (Adapted from Ogden & Richards (1923).)

It has been argued that the true problem of the cognitivist approach lies in the definition of symbols. If we were to accept there must be symbols, they should be defined as structural couplings connecting objects to their categories based on their sensorimotor projections (Clancey, 1997; Maturana and Varela, 1992). I have argued (Vogt, 2002b), however, that such a definition is already present from the semiotic theorist Charles Sanders Peirce (1931-1958). Peirce has defined symbols as a triadic relation between a referent, a meaning and a form[1] as illustrated in the semiotic triangle (Figure 2), where the relation between meaning and form is either arbitrary or conventionalised such that the relation must be learnt. This definition is very common among linguists and other cognitive scientists, e.g., (Lakoff, 1987; Deacon, 1997; Barsalou, 1999). To distinguish between the cognitivist and Peircean definitions of symbols, I have coined Peirce's definition *semiotic symbols* (Vogt, 2002b).

In this definition, the term *meaning* requires extra care. According to Peirce, the meaning of a symbol arises from the process of *semiosis*, which is the interaction between form, meaning and referent. This means that the meaning depends on how the symbol is constructed and with what function. As such, the meaning of a symbol is a functional relation between its form and referent based on an agent's bodily experience and interaction with the referent. The experience is based on the agent's history of interactions with the referent and/or form. The ways these bodily experiences are represented and memorized form the internal representation of the

[1] Peirce actually used the terms *object*, *interpretant* and *representamen* to denote what I call *reference*, *meaning* and *form* respectively. Throughout the text I will also use the terms label, word and signal interchangebly to denote a form.

meaning. The actual interaction between an agent and a referent 'defines' the functional relation.

I have argued that when the semiotic definition of symbols is adopted, the symbol grounding problem no longer exists as a fundamental problem (Vogt, 2002b). This is primarily because semiotic symbols are *per definition* meaningful and grounded. The problem, however shifts into a hard technical problem, which I have called the *physical symbol grounding problem*, and which relates to the *construction* of the triadic relation between referent, meaning and form (Vogt, 2002b).[2] (In the remainder of this paper, I will use the term *symbol* to denote a *semiotic symbol*. When I refer to the cognitivist sense of a symbol, this will be made explicit.)

So, when studying the origins of symbolic communication, the symbols should arise from an interaction of an agent with its environment; this justifies the use of robotic models. Preferably physical robots are used, but simulated robots can offer a suitable platform too, since experiments with real robots may be very time consuming and costly. Ideally, the experiments have an ecological validity in that the robots have a "life task" to solve (Ziemke and Sharkey, 2001), but - as will become clear - most robotic models so far have little or no ecological validity.

During the course of language evolution symbols have become culturally shared in a population, i.e. the members of a language society have learnt more or less similar meanings and references of signals. Learning the meanings of words is - in principle - a notoriously hard problem. In a seminal work, Quine (1960) has shown that when you hear a novel word, this word can - logically - have an infinite number of meanings. He illustrates this point by considering a linguist studying a language that he or she does not know. The linguist observes a native speaker exclaiming "gavagai!" when a rabbit scurries by. It would be natural for the linguist to note that gavagai means `rabbit`, but logically, gavagai could mean an infinite number of different things, such as `undetached rabbit parts`, `a running rabbit` or even `it is going to rain today`. Deciding which meaning to associate with a signal is an extremely hard problem for robotic models, but humans solve this task seemingly very easily. Researchers in child language acquisition have proposed a number of constraints and means to reduce the number of possible inferences as to the meanings and references of words. Examples include *representational constraints*, such as a *whole object bias* (Macnamara, 1982) or a *shape bias* (Landau et al., 1988); *interpretational constraints*, such as *mutual exclusivity* (Markman, 1989) and the *principle of contrast* (Clark, 1993); and *social constraints* such as *joint attention* (Tomasello, 1999), *corrective feedback* (Chouinard and Clark, 2003) and *Theory of Mind* (Premack and Woodruff, 1978) that allows individuals to understand that others have intentions similar to themselves. For an overview, consult, e.g., (Smith, 2005b).

Once a shared symbolic communication system was in place, humans are thought to have developed a *protolanguage* (Jackendoff, 1999). Some have argued that the protolanguage was formed from unstructured expressions of multiple words

[2] This problem is similar to the *anchoring problem* (Coradeschi and Saffiotti, 2000), which deals with the technical problem of connecting traditionally defined symbols to the real world, see Vogt (2003a) for a discussion.

(Bickerton, 1984), others have argued that expressions of the protolanguage were mainly single holistic utterances (Wray, 1998). It is widely assumed that from the protolanguage, grammatical structures have emerged that resemble modern languages (Jackendoff, 1999). Some scientists believe this transition was due to a biological adaptation in the brain (Bickerton, 1984; Pinker and Bloom, 1990); others think that the language itself has adapted to become learnable (Deacon, 1997). Although there have emerged many ungrounded computational models that simulate this transition (Batali, 2002; Brighton, 2002; Kirby, 2001), this area is still largely unexplored in robotics (but see Steels 2004; Vogt 2005a). One idea on which robotics can contribute to the study of grammar evolution is that robots can exploit structures that occur in the interaction between robots and their environment. Both the world and our interaction with the world contain combinatorial structures that could serve as the basis of the semantic structures, which - in turn - form the basis of the syntactic structures in languages. For instance, objects, such as apples, can have different colours, sizes or other properties, which could serve as the basis of what could be called *adjective noun phrases*, such as “the red apple”. Another structure that could be exploited is that actions are typically performed by a subject on an object, which could have resulted in the universal tendency of languages to have expressions combining subjects, objects and verbs.

Philosophically, the physical symbol grounding problem has provided sufficient ground to favour robotic models over what I call *ungrounded* models (i.e., models that have predefined meanings or no meanings at all), but what about more pragmatic reasons? Here I mention two reasons: First, ungrounded models may be built on false assumptions. For instance, in most ungrounded models, all agents have the same meanings. This is clearly not a realistic assumption, because in real life meanings arise from interactions of an individual with its environment and with other individuals, and therefore each individual will have different experiences. In addition, when the model uses a population turnover, in realistic models, the older experienced agents should have a matured set of meanings, while the new agents have not developed any meanings at all. Ungrounded models are completely ignorant about this. Ungrounded models also tend to assume that hearers can observe both the communicated signal and its meaning. Clearly, humans do not observe the meaning of a word that resides in a speaker’s brain, because that would make the signal redundant (Smith, 2003). In an idealised world, humans can observe a word’s reference, though - as mentioned - logically each word can have an infinite number of meanings (Quine, 1960) and a learner has to infer the word’s meanings. In robotic models, the meanings typically develop during an agent’s lifetime. As a result, meanings are private and may differ substantially from agent to agent. Moreover, as a matter of principle, robots cannot observe other agents’ internal representations, unless the experimenter ‘cheats’, which can be useful if, for instance, joint attention is hard to achieve as in (Vogt, 2000a).

Second, robotic models can actually exploit the nature of the interaction with the environment. I have already mentioned exploiting structures for constructing semantic structures. In addition, interactions and feedback mechanisms could be exploited to reduce the number of possible meanings of an expression. If the response of an action induced by a communication act is positive, the agents participating in the communication could use the positive reward to reinforce the used association between expression and meaning, thus allowing the agent to learn a word's meaning more easily.

3. Topics and case studies

This section presents an overview of some robotic studies on the evolution of language. The overview is not exhaustive; for instance, I will not discuss the very interesting study on the evolution of communication channels by Quinn (2001), nor will I discuss language learning models based on human-robot interaction, such as (Oates et al., 2000; Sugita and Tani, 2005; Roy, 2000). Instead, the focus will be on language development in multi-robot systems, providing a clear review of the topics studied showing the state-of-the-art. The topics in this section are – more or less – increasingly complex. First, I present how semiotic symbols can be constructed. The learning of conventions is the subject of Section 3.2. At first, meaning formation is only treated as forming an internal representation of meaning (or *category*), rather than in a functional manner. The functional development of meaning is discussed in Section 3.4.[3]

3.1 Constructing semiotic symbols

The first problem that needs to be solved is the physical symbol grounding problem (i.e. creating the semiotic triangle). The problem can be decomposed into three parts: (1) Sensing and pre-processing of raw sensorimotor images, (2) categorisation or meaning construction, and (3) labelling. The labelling problem is either trivial (in case of using arbitrary forms) or it is based on learning conventions through language. In this subsection, I will assume the trivial solution and focus on the sensing, pre-processing and meaning formation. Learning conventions will be discussed later.

The discussion of how semiotic symbols can be constructed is presented for individual robots. Constructing a semiotic symbol usually starts with a sensori(motor) stimulation based on the robot's interaction with the real world (when embedded in communication, construction can also start upon 'hearing' an expression). Sensorimotor stimulation can be based on a scene acquired by a camera, which may be static as in the Talking Heads experiment (Steels et al., 2002), or dynamic as in more recent experiments of Luc Steels (Steels, 2004; Steels and Baillie, 2003); the activation of infrared, sonar or simple light sensors (Vogt, 2003a); the flow of sensorimotor activity (Billard and Dautenhahn, 1999; Vogt, 2000b); or the activation of a sensorimotor coupling (Vogt, 2002a). Often, the raw data is pre-processed to reduce the huge amount of data. Typically regions of interest are identified and some feature extraction algorithm is used to describe such regions in terms of feature vectors. How this is done can be quite complex and is not discussed further in this chapter, for details consult the individual papers. When the sensing is based on the activation of sensorimotor couplings, pre-processing may not be required (Vogt, 2002a). Furthermore, in simulations, the image is often more abstract, such as a bitstring representing mushrooms (Cangelosi et al., 2000) or just random vectors (Smith, 2003), which do not require any more pre-processing.

[3] Although this is not necessarily more complex than the evolution grammar, it is treated at the end of this section because it is helpful to have the background provided in the first three sections.

At the heart of creating semiotic symbols lies - technically - an agent's ability to categorise the (pre-processed) perceptual data. Once these categories are in place, the agent can simply associate a label (or form) to this category, thus constructing the symbol. (Whether this symbol is useful or functional is another question, which will be dealt with in Section 3.4.) A number of techniques have been developed that allow a robot to construct categories from scratch with which it is able to recognise or discriminate one experience from another. These techniques usually rely on techniques that have been present in AI for quite some time, such as pattern recognition and neural networks. Some researchers use neural networks to associate (pre-processed) sensorimotor images with forms, e.g., (Marocco et al., 2003; Cangelosi et al., 2000; Billard and Dautenhahn, 1999), which - although they work well - makes it hard to analyse how the meanings are represented. Moreover, these techniques are often inflexible with respect to the *openness* of the system (see Section 3.2), because typically, the number of nodes in a neural network are fixed. Another technique that is frequently used in grounded models of language evolution is the *discrimination game* (Steels, 1996b).

The aim of the discrimination game is to categorise a sensorimotor experience such that this category distinguishes this experience from other experiences. If such a *distinctive category* (or meaning) is found, the game is considered a success. If it fails, a new category is formed based on the experience that is categorised, such that discrimination can succeed in a future situation. This allows the agent to construct a repertoire of categories from scratch, as illustrated in Figure 3.

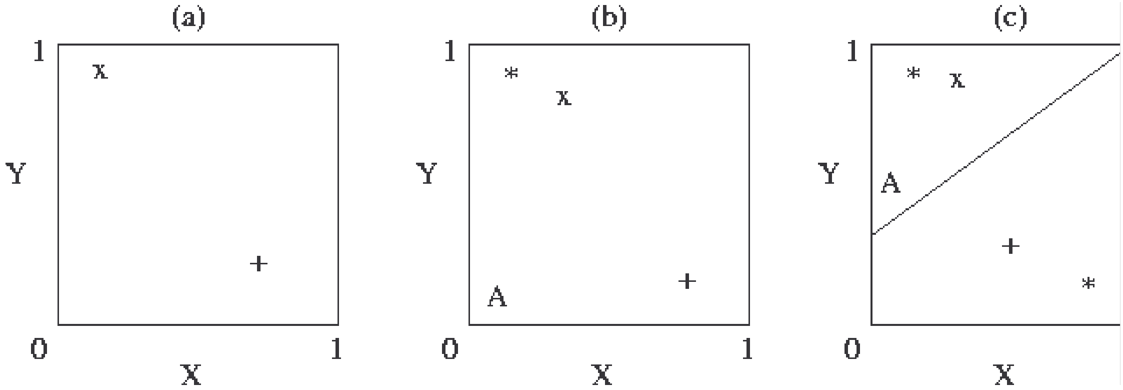

Figure 3: An illustration of three subsequent discrimination games using a prototype representation. The figure shows three instances of a combined feature space and a conceptual space. The x and + are feature vectors of observed objects (e.g., the location of an object in a 2D plane), the * denotes a prototype, while A and B are categories. Each individual plot represents one discrimination game. In game (a), the robot observed 2 objects (x and +), but has not yet formed any categories. Consequently, the game fails and a new category (A) is added to the conceptual space of which the feature vector of the target object serves as an exemplar for its prototype - the * in figure (b). In the second situation (b), again two objects are observed, which are now both categorised with category A. In this case, no distinction can be made. Suppose the + was the target (a.k.a. the topic) of the discrimination game, then a new category is formed by adding the feature vector + to the conceptual space as the new prototype of category B in figure (c). Note that this alters the initial category A. In the third game (c), both objects can be categorised distinctively. Irrespective of which one is the topic, the discrimination game succeeds. Typically, when a discrimination game

succeeds, the prototype is moved slightly in the direction of the topic's feature vector.

The discrimination game illustrated in Fig. 3 has successfully been implemented in the Talking Heads simulation THSim (Vogt, 2003c).[4] Experiments have shown that the discrimination game is typically a fast learning mechanism and is very robust in using different representations for categories. The original implementation used *binary trees* (Steels, 1996b), which was used in various robotic experiments (Steels and Vogt, 1997; Steels et al., 2002) and simulations (Smith, 2003). Other representations that were used include *binary subspaces* (de Jong, 2000), *radial basis function networks* (Steels and Belpaeme, 2005), *neural networks* (Berthouze and Tijsseling, 2002), *predicate logic* (De Beule, 2004; Sierra-Santibáñez, 2001) and different variants of the prototype representation (Vogt, 2005a, 2003a, 2004).

The discrimination game is context dependent; the robot always contrasts the topic with respect to other objects in the context. This has the consequence that the game may succeed, even if the observed feature vector has a relatively large distance to the category's prototype, leading to an overgeneralisation of symbols. However, after a while, the categories become finer grained, thus allowing the agents to resolve overgeneralisation. It is a well known fact, however, that young children also tend to overgeneralise during early word-learning (Bloom, 2000).

When the sensing typically yields only one region of interest (i.e. there is only one object or action), the discrimination game can only be applied in contrast to some sensorimotor images that are in the robot's memory. In such cases different models can be used as well. The *classification game* was used in experiments with Sony's AIBO where whole (segmented) images of the camera were stored as exemplars (Steels and Kaplan, 2000). The *identification game* was used to categorise the motor flow of robots following each other (Vogt, 2000b). The latter used a pre-processing of the raw sensorimotor flow based on constructing delay vectors from time series (Rosenstein and Cohen, 1998). The identification game is very similar to the discrimination game in that the delay vector (or feature vector) is categorised with the nearest prototype, provided its distance was within a certain threshold. If not, the delay vector is added to the ontology as an exemplar.

As mentioned, once a category, which is a representation of the meaning in a semiotic symbol, is in place, the category can be associated with a form. This form may be arbitrary, but in language they need to be conventionalised. In language evolution models, this is often modelled by interactions called *language games* (Steels, 1996a) which will be explained hereafter.

3.2 Sharing semiotic symbols

Among the biggest problems in modelling language evolution using robots is the development of a shared communication system, which is related to Quine's problem of the indeterminacy of meaning. The models will have to include some mechanism to established shared or joint attention to some object or event. Since human children face the same problems when they grow, it is important that robotic models are based on what is known about how children learn language.

[4] THSim is freely downloadable from http://www.ling.ed.ac.uk/~paulv/thsim.html.

Another important aspect relates to the *openness* of human languages. Unlike the communication systems of other animals, human languages are open systems (i.e. new words, meanings, objects, agents and grammatical structures appear, disappear and change rapidly). Many models assume that language is a closed system, for instance, by fixing the number of signals and meanings. Although this occurs in some grounded models (Billard and Dautenhahn, 1999; Marocco et al., 2003; Cangelosi and Parisi, 1998), it most frequently occurs in ungrounded models (e.g., Smith, 2002; Oliphant, 1999). Furthermore, in order to maintain efficiency in an open language system, humans must learn these changes; they cannot be innate, as is for instance the case with Vervet monkeys (Seyfarth and Cheney, 1986). Most models of language evolution (both grounded and ungrounded) acknowledge this principle and allow agents to learn language (Smith, 2002; Cangelosi et al., 2000; Oliphant, 1999). However, there are some models that violate the idea of language acquisition by using genetic algorithms to evolve the communication system (e.g., Marocco et al., 2003; Cangelosi and Parisi, 1998). One widely used open system is the language game model, which has successfully been implemented on physical robots to study the emergence of lexicons (Steels and Vogt, 1997; Steels et al., 2002).

3.2.1 The language game

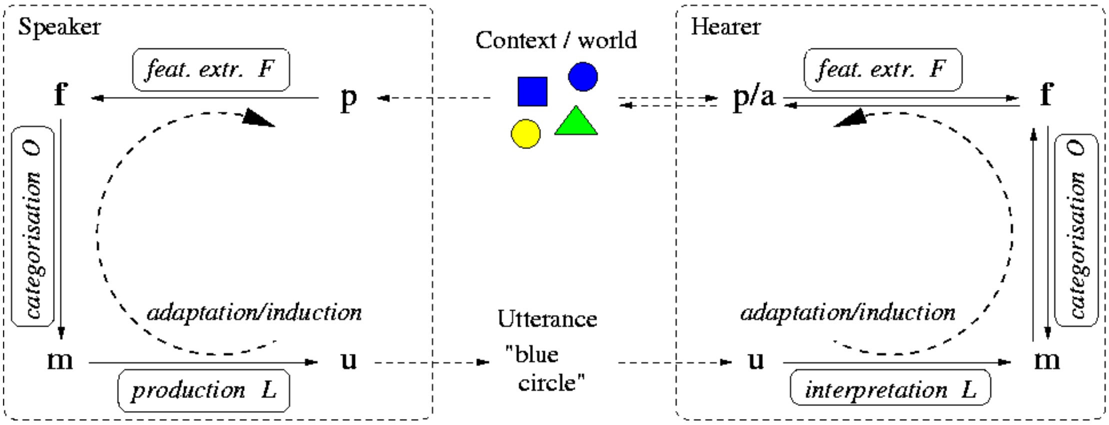

Figure 4: The semiotic square illustrates the working of a language game. See the text for details.

The language game, which is illustrated in Fig. 4, typically involves two agents - one speaker and one hearer - and a context of objects and/or actions. Both agents perceive (p) the context, extract feature vectors (f) and categorise these with meanings (m), e.g., by using the discrimination game. The speaker selects one object as the topic and tries to produce an utterance (u) based on its lexicon. The lexicon is typically an associative memory between meaning representations (or meanings for short) and forms, see Fig. 5 (a). Each association has a score σ_{ij} that indicates the effectiveness (or occurrence frequency) of the association based on past interactions. These lexicons, like the ontologies, are private and thus can differ from one agent to another. The speaker searches its lexicon for an association that corresponds to the meaning of the topic and that has the highest score. If such an association is found, the corresponding form is uttered. When hearing a form, the hearer searches, for this form, the association that has the highest score ($\sigma_{ij}>0$) for those meanings that are in the context or that relate to the topic, if this is known.

	m_1	m_2	…	m_M
w_1	0.5	0.2	…	σ_{1M}
w_2	0.6	0.8	…	σ_{2M}
…	…	…	…	…
w_N	σ_{N1}	σ_{N2}	…	σ_{NM}

(a)

	m_1	m_2	…	m_M
w_1		-		
w_2	-	+	…	-
…		…		
w_N		-		

(b)

Figure 5: Figure (a) shows a typical open association matrix where forms w_i are associated with meanings m_j with some strength (or score) σ_{ij}. Due to the openness of the system, *N* and *M* may grow - in principle - indefinitely and need not have the same values. However, since the memories are limited in size, *N* and *M* are bounded. When the speaker of a language game tries to produce an utterance about meaning m_j, it searches the corresponding column for the highest score σ_{ij} and finds its utterance w_i in that row. When the hearer tries to interpret a word w_i, it searches the corresponding row for the meaning m_j with the highest score σ_{ij}, provided this meaning is pragmatically possible (i.e. it should fit the context). Figure (b) illustrates the update of scores in case of a successful language game. Suppose that the association between word w_2 and m_2 was used successfully, the strength of σ_{22} is increased, while all competing associations σ_{i2} and σ_{2j} are laterally inhibited (i.e. those associations that are in the same row or column as the successful association: i=1,3,4,…,*N* and j=1,3,4,…,*M*).

Typically, the success of the game is evaluated and if the game succeeds, the used association are reinforced, while competing associations are laterally inhibited as illustrated in Fig. 5 (b). If the game fails, the scores of the used associations are decreased. These adaptations ensure that successfully used elements tend to be reused again and again, while unsuccessful ones tend to get weaker. This serves a self-organisation of the lexicon shared at the global population. At the start of an agent’s lifetime, its lexicon is empty, so initially, most language games fail. When they do, the lexicon needs to be expanded. When the speaker encounters a meaning that has no association in its lexicon, the speaker can invent a new form. When the hearer receives a form that has no association in its lexicon, it will adopt the form associated this with the meaning of the topic, or with the meanings of all objects in the context if the topic is unknown. In this way, the language is an open system.

3.2.2 Lexicon grounding on mobile robots

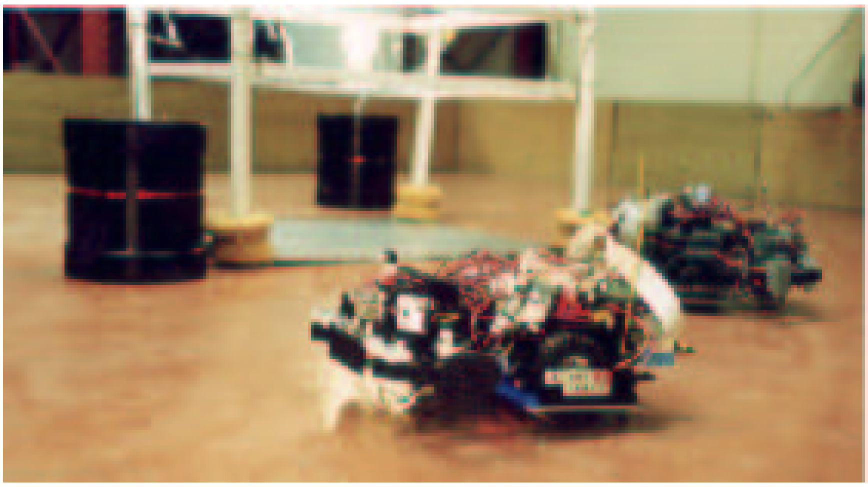

Figure 6: The LEGO vehicles used in the first physical implementation of language games.

The first implementation of the language game on a physical system was done using LEGO robots, such as shown in Figure 6 (Steels and Vogt, 1997). In this experiment, the robots evolved a lexicon to name the different types of light sources in their environment, which they could detect using very simple light sensors mounted on the front of the robots. In order to acquire a sensory image of their environment, the two robots participating in the game first approached each other to stand facing each other at a close distance, after which each robot rotated 360 degrees. The raw image was then pre-processed to identify the different light sources, which were then described in feature vectors. The speaker selected a topic and ‘pointed’ at this object, so that the hearer could identify the topic as well. Then both robots played a discrimination game to categorise the topic’s feature vector. When the discrimination game succeeded, the remainder of the language game was played as explained above. Note that the type of language game in which the speaker points at the topic, thus establishing joint attention, has become known as the *observational game* (Vogt, 2000c, 2002b).

Although the experiments were very successful, many problems have arisen during the development of the model (for more up-to-date details consult Vogt, 2000a,2000c,2002b,2003a). Most problems had to do with the inconsistencies between what the two robots had seen during a language game, thus leading to different contexts, and with the difficulty in achieving joint attention by means of pointing. Importantly, these problems have helped in realising the effect that the false assumptions in ungrounded models have on the soundness and realism of their results.

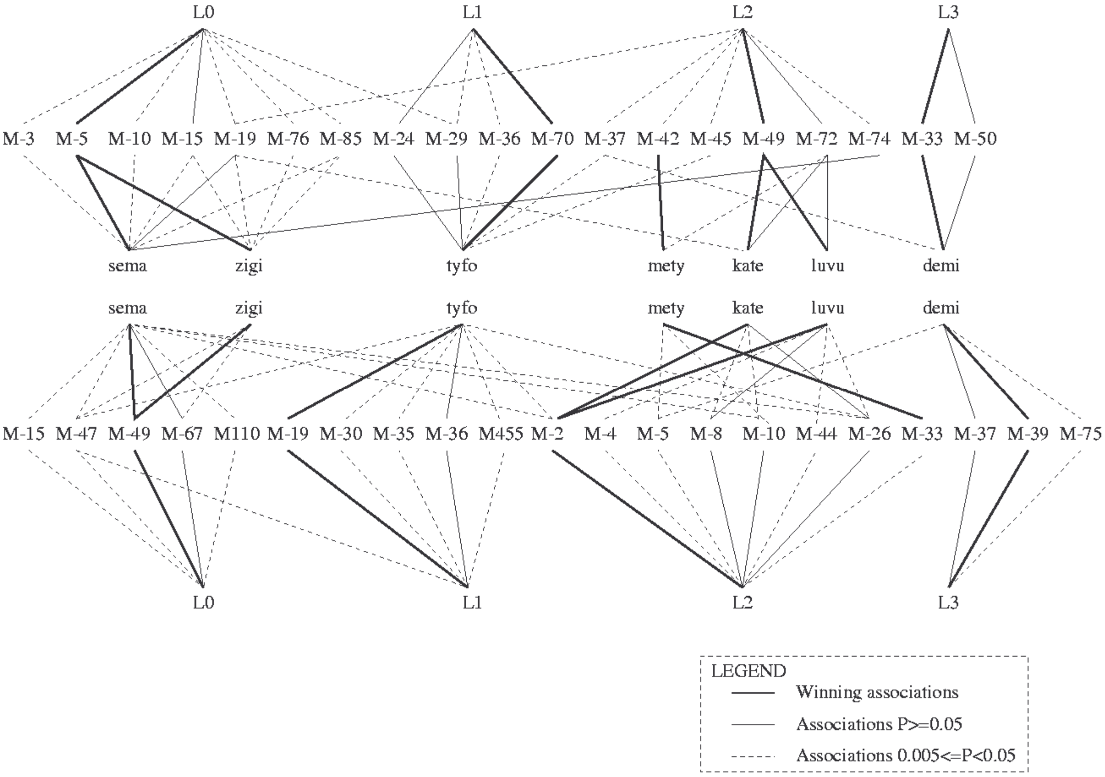

Figure 7: This semiotic landscape shows the associations between referents (L_i), meanings (m_i) and word-forms evolved for two robots in one of the robotic experiments on lexicon grounding. As can be seen, each referent was categorised using different meanings in different situations, but only a few forms were used to name the referents. The thickness and type of the line indicate the frequency with which associations were used. See the text for more details.

The inconsistencies between the robots' sensory images were due to the fact that when two physical bodies were standing opposite of each other, then each individual sensed something different, for instance, because one robot obscured the visibility of an object, or the distance from a light source was too large for one of the robots to detect it. Also differences in the sensitivity to noise of the sensors and different lighting conditions played a significant role. So, although the robots were designed to talk about the 'here and now' (something what young children also tend to do), the hearer may not have seen what the speaker was talking about. Moreover, even if the hearer saw the speaker's topic, it could have detected a completely different sensory image of this object. Interestingly, the self-organisation and adaptiveness of the language game model partly solved this problem. By allowing an agent to acquire many categories, which they could employ in different situations for the same referent, while maintaining only one or two forms associated with a referent, the robots could communicate about the referents consistently and reliably. So, the robots acquired (near) one-to-one mappings between referent and form, and one-to-many mappings between referent and meaning and between form and meaning (Fig. 7). One consequence of this result is that the model thus might help us explaining how we deal with notions such as *family resemblance* (Wittgenstein, 1958) and *object constancy*. It is beyond the scope of this paper to repeat this argumentation, for details consult (Vogt, 2000c,2003a).

Problems with respect to the unreliability of pointing made us look at other ways to solve the mapping problem. While trying to model how the robots could estimate the reliability of pointing, based on Steels and Kaplan (1998), it was found that the speaker's pointing was not necessary, provided the agents could verify the success of the game in order to provide rewards to the learning mechanism. This has lead to the development of what has become known as the *guessing game*, which has also been applied in the Talking Heads experiment (Steels et al., 2002). In the guessing game, the hearer guesses the speaker's reference and then success of the game is evaluated by the speaker who can - in case of failure - provide corrective feedback on the word's meaning. In the LEGO robots (and the Talking Heads) this was done by the hearer 'pointing' (unreliably) at the object it guessed the speaker referred to, thus allowing the speaker to verify success, which is then signalled back to the hearer. In case of failure, the speaker could then point at its topic to provide the hearer with feedback in order to the hearer to acquire the proper meaning. But since this was adding another source of noise and the world only contained 4 light sources, associating the form with a randomly chosen meaning appeared just as effective (Vogt, 2000c).

Experiments comparing the guessing game with the observational game (the game where the speaker points at the topic prior to the verbal communication) have shown that both methods can achieve high levels of communicative success. Although the high levels are reached faster by playing observational games, the lexicons that emerge from the guessing games contain more information, meaning that a form is used more specificly and consistently in naming a referent. This latter result is explained by realising that when hearers have to guess the reference of an utterance, the words have to be informative. When the hearer already knows the topic, this is not required. It has been argued that in both games a strong competition between associations exists, and that the guessing game provides more pressure to disambiguate the language (Vogt, 2000a; Vogt and Coumans, 2003); an effect that has recently been confirmed in a simple ungrounded model (Wedel, personal communication).

Although many researchers in child language acquisition believe there is ample evidence of caregivers providing corrective feedback with respect to the meaning of words (Chouinard and Clark, 2003; Brown and Hanlon, 1970), its availability is controversial (Bloom, 2000). Since corrective feedback may be unrealistic and joint attention - in principle - cannot be assumed to be precise, a third game was developed. This *selfish game*[5] is based on the principle of cross-situational learning (Hurford, 1999; Siskind, 1996), or more precisely *cross-situational statistical learning* (Vogt and Smith, 2005). The idea is that the robots learn the meaning of a word solely based on co-variances that occur across different situations. Unfortunately, cross-situational statistical learning (CSSL) did not work on the in the LEGO robot experiments, because the environment was very minimal and few variations could be detected across situations - even when variation was imposed by the experimenter (Vogt, 2000a, 2000c). Work by Andrew Smith (2003) and recent simulations of the Talking Heads (Vogt, 2004,2003b) have proved that CSSL may become a viable learning

[5] The - unfortunately chosen - term selfish game refers to the selfishness of the robots' not caring about the effectiveness of the game.

strategy. Although the learning mechanism is much slower than the observational and guessing game models and coherence between the agents in the population is hard to achieve (Vogt and Coumans, 2003), results can improve if additional constraints on acquiring the meaning of new words, such as *mutual exclusivity* (Markman, 1989) are added (Smith, 2005a).

3.2.3 Talking Heads and other related work

Probably the best known robotic experiment regarding language evolution is the Talking Heads experiment (Belpaeme et al., 1998; Steels et al., 2002). This large scale experiment consisted of several installations distributed across the world and connected with each other through the Internet.[6] Each installation contained two physical robots embodied as pan-tilt cameras connected to a computational unit. Each camera (or Talking Head) was oriented towards a white board on which geometrical coloured figures were pasted. The population contained a large number agents that could migrate from one site to another through the Internet; at each site, each agent then played a given number of guessing games. If an agent participated in a game, it first embodied itself inside a Talking Head. The speaker would select an arbitrary object from a randomly selected region of interest (a subset of the white board) as topic. The speaker then indicated (or ‘pointed’) to the hearer what the region of interest was, thus establishing the context, and the guessing game (explained above) started. Feedback was evaluated by the hearer ‘pointing’ at its guessed topic.

Human users could interact with the experiment by launching agents, which they could send around the different installations, and by changing the words agents had acquired with words given by the human user. So, new agents entered the population and others left regularly. Furthermore, at the physical sites, people were allowed to alter the geometrical world, thus introducing new objects and removing others. Although this made the experiment largely uncontrolled, it added to the openness of the system. Two experiments were launched; the second and longest experiment had lasted for about 7 months in which a total of approximately 6,000 agents had played roughly 400,000 guessing games. The average communicative success was around 60% during this period, see (Van Looveren, 2001; Steels et al., 2002) for detailed results. Although many agents had participated, one must realise that not all agents were present during the entire experiment; most were probably only present for short periods. Furthermore, although there was no central control of the language, some agents were present almost the entire experiment and were thus likely to have a large impact on the lexicon that evolved. Nevertheless, the Talking Heads experiment was a significant contribution in showing that a large open system of robotic agents was able to evolve a stable, though dynamic set of shared semiotic symbols in a world that had many different and varying conditions, especially with respect to the illumination.

Another related set of experiments on mobile LEGO robots and in simulations was carried out by Billard and colleagues (Billard and Dautenhahn, 1999; Billard and Hayes, 1999). In these experiments, a (group of) learner robot(s) learnt a lexicon through interacting with a teacher robot that had its lexicon predefined. Although these experiments did not explicitly study the origins and evolution of language, the

[6] See http://talking-heads.csl.sony.fr.

experiments are related, since the experiments involved autonomous robot-robot communication and imitation learning. The robots were designed using a dynamical recurrent associative neural network architecture (DRAMA) that fully connected three sensorimotor modules: a communication input/output module, a sensor input module and an actuator output module. In essence, the learner robots were to follow the teacher in an environment that contained different patches on the surface (objects or places of interest) that the robots could communicate about. In addition, the robots could communicate about proprioceptive states and events, such as orientation, inclination and action. Interestingly, although the design was set up such that the robots avoided problems involved with pointing in, e.g, Steels and Vogt (1997), they were faced with other problems concerning the establishment of joint attention. These were mainly caused by the delay with which learners observed the objects or events the teacher talked about. The model was a closed system in the sense that the lexicon for the teacher was predefined with a fixed number of words. This disallowed the introduction of completely new objects without altering the teacher’s lexicon. Although different numbers of learners were allowed in the system, it might have proved difficult to allow a fully open system in terms of population dynamics, where agents continuously enter and leave the society.

An alternative approach of evolving a shared vocabulary in a robotic model was issued by Marocco et al. (2003), in which a genetic algorithm (GA) was used to evolve a lexicon to coordinate interactions of a robot arm with two different objects: a sphere and a cube. The arm was a configuration of three segments with a total of 6 degrees of freedom. The controller was an artificial neural network, of which the weights were evolved using a GA. Fitness was not calculated based on communication, but was assessed by counting the number of appropriate interactions of the robots with the objects: the arm had to touch the sphere and avoid the cube. It was shown that a reasonably well shared lexicon evolved that improved the fitness. The problem with this approach is that Marocco et al. used a GA as a model of evolution, but the system lacked a learning episode of each individual. This is thus not a realistic model of language evolution; rather these simulations use a GA as a machine learning technique to optimise the robot controller, which makes use of the evolved lexicon. A more realistic approach using a GA is to let the GA evolve the connections of the controller and/or the initial values of the weights in the neural network and then use a learning mechanism to adjust the weights of individuals while they interact with each. An example of such an approach is in Cangelosi et al. (2000), which will be discussed in Section 3.4.

3.3 The emergence of grammar

As mentioned, one of the most distinctive features of human languages is the high degree of compositionality they contain. This means that the utterances of human languages are highly structured in that parts of the utterances map onto parts of the whole meaning of these utterances. For instance, in the phrase “orange square”, the word “orange” refers to the colour `orange` and the word “square” to a `square`. In contrast, in a holistic phrase such as “kick the bucket” (referring to `dying`), no part of the utterance refers to a part of its meaning. One influential hypothesis suggests that during the course of evolution, human languages have changed into compositional languages from initially holistic *protolanguages* (Wray, 1998). Many ungrounded models have been developed, which provide support to this idea (Brighton, 2002; Smith et al., 2003; Kirby, 2001).

What Brighton, Kirby, K. Smith and others (BKS for short) have shown is that when learners learn the language of an adult population while observing only a part of the language (i.e. there is a transmission *bottleneck*), holistic languages are not sufficient to allow for stable learnable communication systems. This can be understood by realising that when the learners become adults and start communicating to the next generation of learners, they have no means to produce expressions about objects/meanings they have not encountered before. Compositional languages, however, could allow a learner to produce utterances for previously unseen meanings when the learnt structures can be combined. For instance, if an agent has learnt the proper structures from the phrases “orange square”, “orange triangle” and “red square”, it would be able to produce the phrase “red triangle”, even though it would never have encountered a red triangle before. BKS have shown that - given a predefined structured semantics and a learning mechanism that can discover such compositional structures - a compositional language can emerge from an initially holistic language, *provided* the language is transmitted through a bottleneck. In a way, the language changes to become more learnable for future generations.

This approach has recently been implemented in a simulation of the Talking Heads experiment, in which the semantics was not predefined, but co-developed with the language (Vogt, 2005a,b,c,d). The agents could detect four perceptual features of objects: the three components of the RGB colour space and one feature indicating the shape of an object. The semantic structures developed from a combination of the discrimination game to construct categorical features (elements in one dimension) and an inducer to discover conceptual spaces[7] of one or more dimensions that could serve to represent linguistic categories, such as colours or shapes (note that there was no restriction on which dimensions would constitute a conceptual space - all possible combinations were allowed). One the other hand, syntactic structures could be discovered by looking for coinciding substrings at the utterance level, in a very similar way to the approach taken in Kirby (2001). The model thus investigated the following twofold hypothesis:

1. The emergence of compositional linguistic structures is based on exploiting regularities in (possibly random and holistic) expressions, though constrained by semantic structures.
2. The emergence of combinatorial semantic structures is based on exploiting regularities found in the (interaction with the) world, though constrained by compositional linguistic structures.

The model combines the two most familiar approaches taken in modelling language evolution: the *iterated learning model* (ILM) of BKS and the language game model. The iterated learning model typically implements a vertical transmission of language, in which the population contains adults and learners, the learners learn from utterances produced by adults. At some given moment the adults are replaced by the

[7] The term *conceptual spaces* (Gärdenfors, 2000) is used to denote an *n*-dimensional space in which categories are represented by prototypes. The conceptual space is spanned by *n* quality dimensions that relate to some (preprocessed) sensorimotor quality. Gärdenfors (2000) has argued that conceptual spaces can form the semantic basis for linguistic categories.

learners and new learners enter the population and the process repeats, thus providing a generational turnover. Typically (a part of) the language is transmitted from one generation to the next in one pass; without competition, but see Kirby (2000) for a model with competition; and in a population of size 2, i.e. with 1 adult and 1 learner. The integration of the ILM with the language game allows for competition between different rules and structures, but it requires more passes through the language in order for the language to be learnt sufficiently well.

The experiments reported in Vogt (2005a) have revealed that learners from the first generation already develop a compositional structure, even in the absence of a transmission bottleneck. The reason for this rapid development of compositionality is to be sought in the statistically high level of reoccurring structures in both the feature spaces (thus speeding up the development of semantic structures) and in the signal space (thus increasing the likelihood of finding structures at the syntactic level), see Vogt (2005b) for a detailed analysis. In the case that the population was of size 2, this compositionality was rather stable, but when the population increased to a size of 6, a transmission bottleneck was required to provide stability in compositionality. (Instability of compositionality means that the compositional languages collapse and holistic ones take over.) This difference can be understood by realising that when a learner learns from only one adult, the input received by the learner is consistent, allowing them to adopt the compositional structures reliably. When multiple learners learn from multiple adults, who do not speak to each other, then the input to each learner is highly inconsistent, making it harder to learn the language and to converge on the language.

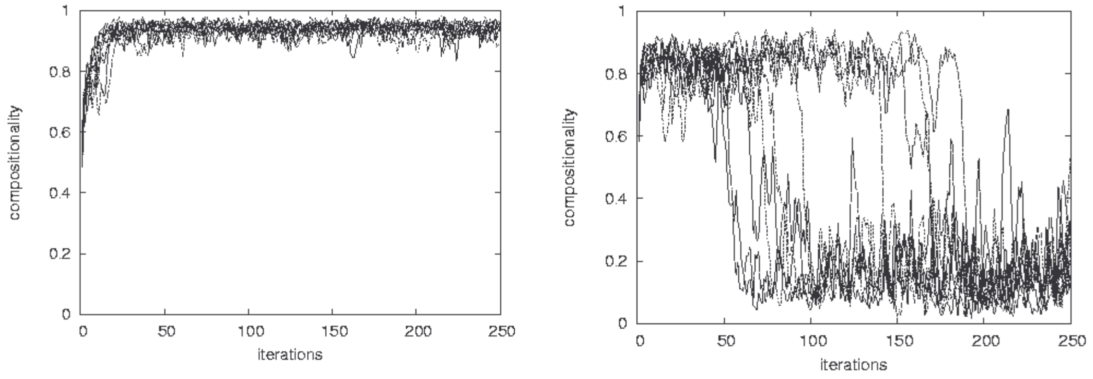

Figure 8: The results of an experiment comparing the guessing game (left) with the observational game (right). In the experiment, the language is transmitted from one generation to another during an iteration (x-axis). The graphs show the evolution of compositionality, which measures the degree with which produced or interpreted expression have a compositional structure. Each line shows the evolution in one run of the experiment. The experiment was done with a population of size 6 and with a transmission bottleneck of 50%.

With the ‘larger’ population size of 6 agents, compositionality was only stable over multiple generations when the guessing game model was used. When the observational game was used, the transmission bottleneck caused compositional structures to remain longer in the population, but they eventually died out (Fig. 8). Like for the experiments on lexicon development, the differences between the guessing game and observational game could be explained by the pressure to disambiguate competing structures. Where in the guessing game this pressure is high, because the hearer has to guess the speaker's topic based on the received expression, this pressure is absent in the observational game, because the hearer already knows the topic and the information of the expression is redundant. It appears that the lack of this pressure allows the meanings to drift through the conceptual spaces. If the meanings are part of a whole meaning, i.e. they are part of a compositional structure, a meaning shift affects a larger part of the language. However, when the meaning is associated holistically, a shift has little effect on the rest of the language. Consequently, the shift of meanings makes the language less stable and thus harder to learn than holistic structures.

Another interesting result that was found is that when learners are allowed to speak as well, which is typically not the case in the BKS models, then no experimentally imposed transmission bottleneck is required for populations of 6 agents (Vogt, 2005c). Instead of starting to speak when they are adults, the learners now speak during their development. In this way, the learners face the bottleneck earlier in life, because they have to produce utterance about previously unseen objects. This may be an important results, because it may help to explain why children are so good at learning grammar early in life. Moreover, this property may even explain why children are thought to be the driving force for the development of grammar in Nicaraguan sign language (Senghas et al., 2004).

A recent study in which the population size was varied from 2 to 100 revealed that this model can lead to compositional structures in larger populations (Vogt, 2005d). However, the larger the population becomes, the less frequently compositionality is stable.

A more complex model implemented in an extended physical version of the Talking Heads experiment is being developed by Steels and his co-workers (Steels, 2004; Steels and Baillie, 2003). In this experiment the cameras do not look at a static scene pasted on the white board, but the cameras observe a dynamic scene played in front of them, such as ‘pick up red ball’. The events are processed through a visual processing system, which - although advanced - is still very limited. Only slow movements can be captured and only a few objects can be recognised, but only *after* training the visual module. The pre-processed events are then matched with top down generated world knowledge, which is represented in the form of predicate calculus of which the basic building blocks are predefined (Steels and Baillie, 2003).

Using these event descriptions, the guessing game (or *description game*) proceeds. Where possible, the agents use the knowledge (lexicon, syntax and semantics) they already acquired, but when events or semantics cannot be described with the given knowledge, new parts of the language is invented, abducted or induced. New words, semantic categories, syntactic categories and hierarchical structures can be constructed using some complex techniques, which are largely based on existing techniques from computational linguistics. This way *grounded construction grammars* (Lakoff, 1987) can develop, as some preliminary experiments have shown (Steels, 2004).

The experiments reported so far were carried out with two robots of one generation, which took turns in taking the role of speaker. Although using such small populations without a population turnover is not uncommon in robotic models (see, e.g., Steels and Vogt, 1997; Vogt, 2003a), the results achieved may not be stable in larger populations with a generational turnover as shown in Vogt (2005a,d). However, given that the model also incorporates the guessing game, which is in favour for a strong selective pressure on the competition between different structures in the language, the model is likely to scale up in terms of population size and population dynamics. Apart from the higher complexity in experimental set up and learning mechanisms, one of the main differences between Steels' model and my own is that in Steels' model the speaker can invent new grammatical structures, whereas in my model, this can only be achieved by the hearer. Although it is unclear what the implications are for this distinction, experiments with my model have confirmed that the productive power of speakers - even though they do not invent new compositional structures - has a positive effect on the stability of the evolved grammars (Vogt, 2005c).

3.4 Ecological models

All models discussed so far have completely ignored the ecological value of language and thus fail to investigate the functional meanings of semiotic symbols. In human societies language is clearly used to exchange information that can be used to enhance some aspect of behaviour. For instance, language can be used to indicate the whereabouts of food sources, the presence of predators or other dangers. So far, no physical robot study has been carried in which the evolved language is used to achieve something. There exist, however, a few grounded simulations in which the language is used to coordinate activity that improves the viability of the artificial organism.

Well known are the studies on the emergence of symbolic communication in a world of edible and non-edible mushrooms (Cangelosi and Harnad, 2000; Cangelosi and Parisi, 1998). In these studies, the task of the agents was to approach the edible mushrooms and avoid the poisonous ones. The mushrooms were perceptually distinguishable through the encoding of a bitstring. The controller of the agents was implemented as a multilayered feedforward network. In Cangelosi and Parisi (1998) the weights of the network were trained using a GA, where the fitness was based on the organisms' energy levels (the profit of eating edible mushrooms was smaller than the cost of eating non-edible ones). As mentioned before, training a neural network using only a GA is far from realistic for modelling language evolution, since it lacks a model of language acquisition.

In Cangelosi and Harnad (2000) the agents do have a learning cycle in which the neural network is adapted using backpropagation and the initial weights evolve using the GA. Although this model is more realistic, backpropagation uses the output vector of the target behaviour to update the weights. Cangelosi and Harnad (2000) investigated two conditions: (I) one in which no communication was used to classify edible and non-edible mushrooms, and (II) one in which communication was used. They have shown that using communication achieved higher performance on categorisation than in the condition without communication, *provided* the population in condition II first evolved the ability to categorise the mushroom world using the method of condition I. This is in contrast with the language game model, in which the ability to categorise the world co-develops with the language. Given the rapid changing nature of natural language, co-development of meaning and forms seems more realistic.

Interestingly, the analysis of the evolved neural networks has shown how the networks had emerged different internal representations in both conditions. Condition II yielded a more structured representation of the categories, allowing Cangelosi and Harnad (2000) to conclude that language influences the way in which the individuals observe the world. A similar Whorfian effect (Whorf, 1956) has also been observed with the language games (Steels and Belpaeme, 2005). A nice property of Cangelosi et al.'s experiments is that they show how language can emerge to improve the population's fitness, which was not defined in terms of communicative accuracy.

A related ecological experiment, where the categories co-developed with the language was carried out by de Jong (2000). This simulation, which was inspired by the alarm call system of Vervet monkeys, contained a number of agents, which were placed on a 2 dimensional grid of size *N*x3.[8] At given times, a predator was present in one of the 3 rows; the row in which it was present resembled the type of predator. In order to avoid the predator, the agents that were in that row had to move to another row. The agents had a vision system with which they could see the predator, but this system was subject to noise so the presence was not always detected. The agents were also equipped with a language game model, in which they could develop categories that represented their own position, location of the predator and the appropriate action to take. The categorisation was modelled using the discrimination game with an adaptive subspace representation that allowed the emergence of *situation concepts* (de Jong 2000). Situation concepts relate – based on past experiences – perceptual categories with actions that need to be performed in order to remain viable. De Jong showed that the agents could successfully evolve a lexicon to avoid predators. Moreover, de Jong successfully showed that the lexicon development can be classified as attractor dynamics, thus providing support for considering the language game as a complex dynamical adaptive system.

[8] See Loula et al. (2003) for another grounded study on the emergence of alarm calls among vervet monkeys.

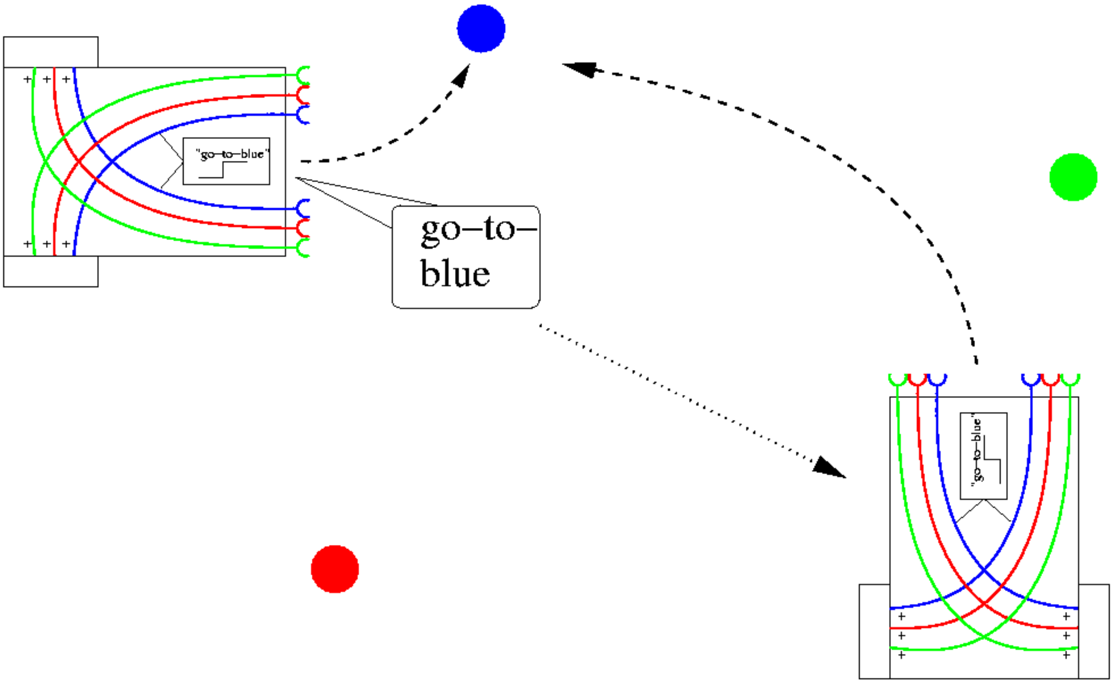

Figure 9: This figure illustrates the principles of the Braitenberg vehicles used in the experiments of Vogt (2002a). Different sensorimotor modules connect the sensors that are sensitive to a particular colour directly to the motors of the two wheels of the robot. These reactive systems make sure that when one module is activated, the robot will move towards the light source of the corresponding colour. Activation of a coupling could be regulated by the need of some energy source, or by the interpretation of a received signal. When a coupling was activated due to the need of some energy source, this robot would try to produce an utterance that is associated with this coupling (e.g., the "go-to-blue" signal in the figure), using the standard guessing game. In turn, the hearer would interpret the signal, which could activate its associated sensorimotor coupling. If both robots then ended up visiting the same energy source, they receive energy which then served as a positive feedback mechanism for updating the association scores in the lexicon.

Inspired by earlier work of Steels (1994) on ecological robot models, Vogt (2002a) reported an experiment in which two simulated mobile robots developed a lexicon to improve on their ability to remain viable over longer periods of time. The robots operated in an environment that contained four different charging stations at which they could increase their energy levels. The only way to get access to the energy was to arrive at a charging station simultaneously. Each charging station had a different colour and at some randomly selected moment[9], one robot initiated a game and selected a charging station that was in its visual field. The robots played a guessing game and the hearer would guess where the initiator was going to and activate the corresponding sensorimotor coupling. When both robots arrived at the same charging station, their energy levels would be refilled, thus providing a positive feedback loop for updating the scores in their lexicons. The robots' control mechanisms were based on a simple Braitenberg vehicle (Braitenberg, 1984) and had four excitatory

[9] Currently, research is carried out to regulate the selection based on the need for energy and use the evolved language to pass on knowledge how to survive in this environment.

sensorimotor connections sensitive to the four different light colours, which they used to navigate towards the corresponding light source (Fig. 9).

Each connection served as a category and when activated, they could activate an associated form. This way, the semiotic symbols could be viewed as symbols that had as referent the action of moving towards a charging station in order to refill their energy supplies; the form is obviously the signal, which they conventionalised using the guessing game; and the meaning was represented by the activation level (0 or 1) of the sensorimotor connection that functioned as the “life-task” of the robots to maintain their energy levels. This, then could be an interesting step toward the development of a robotic model in which the meanings of the symbols are truly meaningful according to (Ziemke and Sharkey, 2001). A nice aspect with respect to the implementation is that the meaning of the actions are not represented in terms of the temporal (sensori)motor flow, as was the case in (Vogt, 2000b), but more directly in the activation of a reactive mechanism. This is much in line with Brooks’ (1990) remark that “once the essence of being and reacting are available” the development of higher order functions, such as using language, would be rather simple.

4. Future trends

Up to date a lot has been achieved by using robots to study the origins and evolution of language. However, it is clear from the overview in this chapter, that we are still far from understanding the entire picture. Can we ever design a group of robots that can evolve languages similar to human languages? Personally, I think that the human brain and body is so complex that we may never be able to unravel all its secrets - like we may never unravel the complete working of our universe, which is similarly complex. Nevertheless, I think that we can use robotics profitably to answer some of the questions that are posed in our effort to understand language evolution. Much future research will need to focus on ecological models, models of grammar, categorisation of more human-like concepts, and on models of the theory of mind. Furthermore, the models will have to be scaled up at many levels, such as population size, sensorimotor complexity, and complexity of the world in which the robots operate.

A recently started European project called New Ties[10] aims at developing a simulation in which a large community of robots (over 1,000 agents) evolve a cultural society, including language. The society ‘lives’ in an environment where they have to cooperate in order to survive (Gilbert et al., 2005). The agents will evolve and learn techniques to deal with the constraints set by the environment in order to improve their viability (Griffioen et al., 2005). In addition, the agents will be designed to evolve language as the motor for evolving the cultural society (Vogt and Divina, 2005). The techniques these agents will use are heavily based on the language game techniques that have been developed so far. One of the major innovations - apart from its complexity and ecological setting - will be a design on the Theory of Mind, which at a later stage is intended to become subject of the evolution.

[10] New Emerging World models Through Individual, Evolutionary and Social learning (http://www.new-ties.org).

Given that many problems in modelling language evolution on robots relate to the difficulties in establishing joint or shared attention to the reference of the communication, studies into the nature of this ability is extremely important. Only few studies are known that investigate how joint attention can emerge, e.g., (Kaplan and Hafner, 2004). One of the key aspects with respect to joint attention and related issues is that agents need to infer the intentions of other agents. This ability can loosely be characterised by the Theory of Mind (Premack and Woordruff, 1978). It may well be that the ToM is one of the major innovations of the human species with respect to language origins, and therefore its origins deserves more attention in models of language evolution.

On the physical side of the implementation, research is starting to focus on the development of more humanoid like platforms. Such implementations have more complex sensorimotor systems, which inevitably will provide more complex data from which a more complex language - in terms of grammar and vocabulary size - can develop.

With respect to the origins and evolution of grammar and compositionality, research needs to done to study how the learning mechanisms have evolved that allow individuals to construct grammatical and compositional structures. Up to now, all studies on grammaticalisation have assumed that such learning mechanisms exist and therefore only investigate how grammar can emerge given these mechanisms.

Another important direction that needs to be tackled is in relation to the grounding of more abstract and higher level symbols, such as - for instance - number systems, arithmetic, planning and ‘feelings’ (internal states). Up to now, all research has focused on the emergence of language about events or objects that are directly observable to the robots. We humans often use language to communicate about events that happened in the past or that may happen in the future. Some work on the development of time concepts is done (De Beule, 2004), but it would be good if a robot could communicate, for instance, the presence of an interesting object at a distant location, which the robot has visited before.

Most techniques used so far are based on simple language games where some aspect of a visible entity is communicated in one direction to investigate learning techniques. However, human language use is much more based on dialogues. Future robotic models should investigate how dialogues can aid in evolving language. This could be particularly interesting for applications where robots develop their own language in order to cooperate in environments we don’t know, such as planets, or using sensors that we find difficult to read, such as infrared, sonar or other exotic sensors.

5. Conclusions

In this chapter, an overview of robotic (and other grounded) models of language evolution is presented. There are many reasons for using robotic models to study the evolution of language. The most fundamental one is that robots - necessarily - have to solve the *symbol grounding problem* (Harnad, 1990). It is argued that by using the Peircean definition of symbols (or *semiotic symbols*), the symbol grounding problem is solved, because these symbols are *per definition* grounded in the real world. This, however, shifts the problem in the *physical symbol grounding problem* (Vogt, 2002b), which deals with constructing these semiotic symbols. More pragmatic reasons for using robots to study language origins and evolution are:

1. Robotic models can reveal false assumptions that are typically incorporated in ungrounded models, e.g., the assumption that all agents have the same meanings and are given the meaning of a signal during communication, and
2. Robots can actually exploit the nature of the interaction with the environment to develop certain structures in language.

The overview shows that in many areas, robots can successfully be used to study certain aspects of language evolution, though the state of the art is still extremely limited, especially when compared to human languages. The design of robotic models is extremely hard and time consuming for many difficult problems need to be solved. One of the most difficult problems that was identified in most models deals with the robots' ability to infer the reference of a heard utterance. If we assume that this ability is the key aspect of a Theory of Mind (Premack and Woodruff, 1978; Bloom, 2000), then the studies indicate that the evolution of a ToM is perhaps the most important transition in human evolution with respect to language evolution.

Robotics has successfully been applied to study the emergence of small lexicons and simple grammatical structures, and to study how language evolution can aid in operating cooperatively in an ecological environment. Most of the models presented started from the assumption that language is a complex dynamical adaptive system in which language evolves through self-organisation as a result of cultural interactions and individual learning. The studies reported have shown that the language game in general provides a robust model in which robots can develop a culturally shared symbol system despite (1) the difficulties they face in establishing joint attention and (2) the differences in their perception and meaning development. The studies have also revealed how agents can exploit structures they find in their interaction with their environment to construct simple grammars that resemble this structure. Furthermore, the studies reveal some of the effects that the nature of social interactions and the co-development of language and meaning have on the emergence of language. In addition, ecological models on the evolution of language indicate how the functional use of language can provide feedback on the effectiveness of communication, which individuals can use to learn the language. Concluding, robotics provides a fruitful platform to study the origins and evolution of language, thus allowing us to gain more insights about the nature and nurture of human language.

Bibliography

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences 22*, 577-609.

Batali, J. (2002). The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In T. Briscoe (Ed.), *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge University Press.

Belpaeme, T., L. Steels, and J. van Looveren (1998). The construction and acquisition of visual categories. In A. Birk and J. Demiris (Eds.), *Learning Robots, Proceedings of the EWLR-6, Lecture Notes on Artificial Intelligence 1545*. Springer.

Berthouze, L. and A. Tijsseling (2002). Acquiring ontological categories through interaction. *The Journal of Three Dimensional Images 16*(4), 141-147.

Bickerton, D. (1984). The language bioprogram hypothesis. *Behavioral and Brain Sciences 7*, 173-212.

Billard, A. and K. Dautenhahn (1999). Experiments in social robotics: grounding and use of communication in autonomous agents. *Adaptive Behavior 7(3-4)*, 415-438.

Billard, A. and G. Hayes (1999). Drama, a connectionist architecture for control and learning in autonomous robots. *Adaptive Behaviour 7(1)*, 35-64.

Bloom, P. (2000). *How Children Learn the Meanings of Words*. Cambridge, MA. and London, UK.: The MIT Press.

Braitenberg, V. (1984). *Vehicles, Experiments in Synthetic Psychology*. Cambridge MA.: The MIT Press.

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial Life 8(1)*, 25-54.

Briscoe, E. (2000). Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device. *Language 76*(2), 245-296.

Briscoe, E. J. (Ed.) (2002). *Linguistic evolution through language acquisition: formal and computational models*. Cambridge: Cambridge University Press.

Brooks, R. A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems 6*, 3-15.

Brown, R. and C. Hanlon (1970). Derivational complexity and order of acquisition in child speech. In *Cognition and the Development of Language*. New York: Wiley.

Cangelosi, A. and S. Harnad (2000). The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication* 4(1), 117-142

Cangelosi, A. and D. Parisi (1998). The emergence of "language" in an evolving population of neural networks. *Connection Science 10*, 83-93.

Cangelosi, A. and D. Parisi (Eds.) (2002). *Simulating the Evolution of Language*. London: Springer.

Chomsky, N. (1980). Rules and representations. *The behavioral and brain sciences 3*, 1-61.

Chouinard, M. M. and E. V. Clark (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language 30(3)*, 637-669.

Christiansen, M. H. and S. Kirby (Eds.) (2003). *Language Evolution*. Oxford: Oxford University Press.

Clancey, W. J. (1997). *Situated Cognition*. Cambridge University Press.

Clark, E. V. (1993). *The lexicon in acquisition*. Cambridge University Press.

Coradeschi, S. and A. Saffiotti (2000). Anchoring symbols to sensor data: preliminary report. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, Austin, pp. 129-135.

De Beule, J. (2004). Creating temporal categories for an ontology of time. In *Proceedings of the 16th Belgian-Dutch Conference on Artificial Intelligence (BNAIC04)*.

de Jong, E. D. (2000). *The Development of Communication*. Ph. D. thesis, Vrije Universiteit Brussel.

De Saussure, F. (1974). *Course in general linguistics*. New York: Fontana.

Deacon, T. (1997). *The Symbolic Species*. New York, NY.: W. Norton and Co.

Elman, J. L., E. A. Bates, M. H. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett (1996). *Rethinking innateness: A connectionist perspective on development*.

Gärdenfors, P. (2000). *Conceptual Spaces*. Bradford Books, MIT Press.

Gilbert, N., S. Schuster, M. den Besten, and L. Yang (2005). Environment design for emerging artificial societies. In *Proceedings of AISB 2005: Socially inspired computing joint symposium*. In press.

Griffioen, A., M. Schut, A. Eiben, A. Bontovics, G. Hévízi, and A. Lõrincz (2005). New Ties agent. In *Proceedings of AISB 2005: Socially inspired computing joint symposium*. In press.

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

Hurford, J. R. (1999). Language learning from fragmentary input. In K. Dautenhahn and C. Nehaniv (Eds.), *Proceedings of the AISB'99 Symposium on Imitation in Animals and Artifacts*, pp. 121-129. Society for the Study of Artificial Intelligence and the Simulation of Behaviour.

Jackendoff, R. (1999). Possible stages in the evolution of the language capacity. *Trends in Cognitive Science 3(7)*, 272-279.

Kaplan, F. and V. Hafner (2004). The challenges of joint attention. In L. Berthouze, H. Kozima, C. Prince, G. Sandini, G. Stojanov, G. Metta, and C. Balkenius (Eds.), *Proceedings of the 4th International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic System*, pp. 67-74. Lund University Cognitive Studies 117.

Kirby, S. (2000). Syntax without natural selection: how compositionality emerges from vocabulary in a population of learners. In C. Knight, M. Studdert-Kennedy, and J. R. Hurford (Eds.), *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, pp. 303-323. Cambridge: Cambridge University Press.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation 5(2)*, 102-110.

Kirby, S. (2002). Natural language from artificial life. *Artificial Life 8(3)*.

Kirby, S. and J. R. Hurford (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi and D. Parisi (Eds.), *Simulating the Evolution of Language*, London, pp. 121-148. Springer.

Lakoff, G. (1987). *Women, Fire and Dangerous Things*. The University of Chicago Press.

Landau, B., L. B. Smith, and S. S. Jones (1988). The importance of shape in early lexical learning. *Cognitive Development 3*, 299-321.

Loula, A., R. Gudwin, and J. Queiroz (2003). Synthetic approach of symbolic creatures. *S.E.E.D. Journal -- Semiotics, Evolution, Energy, and Development*, 3(3), p.125-133

Macnamara, J. (1982). *Names for things: a study of human learning*. Cambridge, MA: MIT Press.

MacWhinney, B. (1999) *Emergence of Language*. Lawrence Earlbaum Associates.

Markman, E. (1989). *Categorization and naming in children*. Cambridge, Ma.: MIT Press.

Marocco, D., A. Cangelosi, and S. Nolfi (2003). The emergence of communication in evolutionary robots. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences 361*(1811), 2397-2421.

Maturana, H. R. and F. R. Varela (1992). *The tree of knowledge: the biological roots of human understanding*. Boston: Shambhala.

McCarthy, J. and P. J. Hayes (1969). Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence 4*, 463-502.

Newell, A. and H. A. Simon (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM 19*, 113-126.

Nowak, M. A., J. B. Plotkin, and V. A. A. Jansen (2000) The evolution of syntactic communication. *Nature 404*, 495-498.

Oates, T., Z. Eyler-Walker, and P. R. Cohen (2000). Toward natural language interfaces for robotic agents: Grounding linguistic meaning in sensors. In *Proceedings of the Fourth International Conference on Autonomous Agents*.

Ogden, C. K. and I. A. Richards (1923). *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. London: Routledge & Kegan Paul Ltd.

Oliphant, M. (1999). The learning barrier: Moving from innate to learned systems of communication. *Adaptive Behavior 7 (3-4)*, 371-384.

Peirce, C. S. (1931). *Collected Papers*, Volume I-VIII. Cambridge Ma.: Harvard University Press. (The volumes were published from 1931 to 1958).

Pfeifer, R. and C. Scheier (1999). *Understanding Intelligence*. MIT Press.

Pinker, S. and P. Bloom (1990). Natural language and natural selecion. *Behavioral and brain sciences 13*, 707-789.

Premack, D. and G. Woodruff (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences 1(4)*, 515-526.

Quine, W. V. O. (1960). *Word and object*. Cambridge University Press.

Quinn, M. (2001). Evolving communication without dedicated communication channels. In J. Kelemen and P. Sosík (Eds.), *Proceeding of the 6th European Conference on Artificial Life, ECAL 2001*, LNAI 2159, Berlin Heidelberg, pp. 357-366. Springer-Verlag.

Rosenstein, M. and P. R. Cohen (1998). Symbol grounding with delay coordinates. In *Working notes of the AAAI-98 workshop on: The Grounding of Word Meaning*, Menlo Park Ca. AAAI Press.

Roy, D. (2000). A computational model of word learning from multimodal sensory input. In *International conference of cognitive modeling*, Groningen, The Netherlands.

Searle, J. R. (1980). Minds, brains and programs. *Behavioral and Brain Sciences 3*, 417-457.

Senghas, A., S. Kita, and A. Özyürek (2004). Children creating core properties of language: Evidence from an emerging sign language in Nicaragua. *Science 305(5691)*, 1779-1782.

Seyfarth, R. and D. Cheney (1986). Vocal development in vervet monkeys. *Animal Behavior 34*, 1640-1658.

Sierra-Santibáñez, J. (2001). Grounded models as a basis for intuitive reasoning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 401-406.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition 61*, 39-91.

Smith, A. D. M. (2003). Intelligent meaning creation in a clumpy world helps communication. *Artificial Life 9(2)*, 559-574.

Smith, A. D. M. (2005a). Mutual exclusivity: Communicative success despite conceptual divergence. In M. Tallerman (Ed.), *Language Origins: perspectives on evolution*, pp. 372-388. Oxford: Oxford University Press.

Smith, A. D. M. (2005b). The inferential transmission of language. *Adaptive Behavior*, In press.

Smith, K. (2002). The cultural evolution of communication in a population of neural networks. *Connection Science 14*(1), 65-84.

Smith, K., H. Brighton, and S. Kirby (2003). Complex systems in language evolution: the cultural emergence of compositional structure. *Advances in Complex Systems 6*(4), 537-558.

Steels, L. (1994). A case study in the behavior-oriented design of autonomous agents. In: D. Cliff, P. Husbands, J.-A. Meyer, and S.W. Wilson (Eds.) *From Animals to Animats 3. Proceedings of the Third International Conference on Simulation of Adaptive Behavior, SAB'94*, Complex Adaptive Systems, pp. 445-452, Cambridge, MA: The MIT Press.

Steels, L. (1996a). Emergent adaptive lexicons. In P. Maes (Ed.), *From Animals to Animats 4: Proceedings of the Fourth International Conference On Simulating Adaptive Behavior*, Cambridge Ma. The MIT Press.

Steels, L. (1996b). Perceptually grounded meaning creation. In M. Tokoro (Ed.), *Proceedings of the International Conference on Multi-Agent Systems*, Menlo Park Ca. AAAI Press.

Steels, L. (1997). Language learning and language contact. In W. Daelemans, A. Van den Bosch, and A. Weijters (Eds.), *Workshop Notes of the ECML/MLnet Familiarization Workshop on Empirical Learning of Natural Language Processing Tasks*, Prague, pp. 11-24.

Steels, L. (2004). Constructivist development of grounded construction grammars. In W. Daelemans (Ed.), *Proceedings Annual Meeting of Association for Computational Linguistics*.

Steels, L. and J.-C. Baillie (2003). Shared grounding of event descriptions by autonomous robots. *Robotics and Autonomous Systems 43*(2-3), 163-173.

Steels, L. and T. Belpaeme (2005). Coordinating perceptually grounded categories through language. a case study for colour. *Behavioral and Brain Sciences*. In press.

Steels, L. and F. Kaplan (1998). Stochasticity as a source of innovation in language games. In *Proceedings of Alive VI*.

Steels, L. and F. Kaplan (2000). Aibo's first words. the social learning of language and meaning. *Evolution of Communication 4(1)*.

Steels, L., F. Kaplan, A. McIntyre, and J. Van Looveren (2002). Crucial factors in the origins of word-meaning. In A. Wray (Ed.), *The Transition to Language*, Oxford, UK. Oxford University Press.

Steels, L. and P. Vogt (1997). Grounding adaptive language games in robotic agents. In C. Husbands and I. Harvey (Eds.), *Proceedings of the Fourth European Conference on Artificial Life*, Cambridge Ma. and London. MIT Press.

Sugita, Y. and J. Tani (2005). Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. *Adaptive Behavior in press*.

Tomasello, M. (1999). *The cultural origins of human cognition*. Harvard University Press.

Van Looveren, J. (2001). Robotic experiments on the emergence of a lexicon. In B. Kröse, M. de Rijke, G. Schreiber, and M. van Someren (Eds.), *Proceedings of the 13th Belgian/Netherlands Artificial Intelligence Conference, BNAIC'01*.

Vogt, P. (2000a). Bootstrapping grounded symbols by minimal autonomous robots. *Evolution of Communication 4(1)*, 89-118.

Vogt, P. (2000b). Grounding language about actions: Mobile robots playing follow me games. In Meyer, Bertholz, Floreano, Roitblat, and Wilson (Eds.), *SAB2000*

Proceedings Supplement Book, Honolulu. International Society for Adaptive Behavior.

Vogt, P. (2000c). *Lexicon Grounding on Mobile Robots*. Ph. D. thesis, Vrije Universiteit Brussel.

Vogt, P. (2002a). Anchoring symbols to sensorimotor control. In *Proceedings of the 14th Belgian/Netherlands Artificial Intelligence Conference, BNAIC'02*.

Vogt, P. (2002b). The physical symbol grounding problem. *Cognitive Systems Research 3(3)*, 429-457.

Vogt, P. (2003a). Anchoring of semiotic symbols. *Robotics and Autonomous Systems 43(2)*, 109-120.

Vogt, P. (2003b). Grounded lexicon formation without explicit meaning transfer: who's talking to who? In W. Banzhaf, T. Christaller, P. Dittrich, J. T. Kim, and J. Ziegler (Eds.), *Advances in Artificial Life - Proceedings of the 7th European Conference on Artificial Life (ECAL)*. Springer Verlag Berlin, Heidelberg.

Vogt, P. (2003c). THSim v3.2: The Talking Heads simulation tool. In W. Banzhaf, T. Christaller, P. Dittrich, J. T. Kim, and J. Ziegler (Eds.), *Advances in Artificial Life - Proceedings of the 7th European Conference on Artificial Life (ECAL)*. Springer Verlag Berlin, Heidelberg.

Vogt, P. (2004). Minimum cost and the emergence of the Zipf-Mandelbrot law. In J. Pollack, M. Bedau, P. Husbands, T. Ikegami, and R. A. Watson (Eds.), *Artificial Life IX Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems*, pp. 214-219. The MIT Press.

Vogt, P. (2005a). The emergence of compositional structures in perceptually grounded language games. Submitted for publication.

Vogt, P. (2005b). Meaning development versus predefined meanings in language evolution models. In *Proceedings of IJCAI-05*. In press.

Vogt, P. (2005c). On the acquisition and evolution of compositional languages. *Adaptive Behavior*. In press.

Vogt, P. (2005d). Stability conditions in the evolution of compositional languages: issues in scaling population sizes. In preparation.

Vogt, P. and H. Coumans (2003). Investigating social interaction strategies for bootstrapping lexicon development. *Journal for Artificial Societies and Social Simulation 6(1)*. http://jasss.soc.surrey.ac.uk.

Vogt, P. and F. Divina (2005). Language evolution in large populations of autonomous agents: issues in scaling. In *Proceedings of AISB 2005: Socially inspired computing joint symposium*. In press.

Vogt, P. and A. D. M. Smith (2005). Learning colour words is slow: a cross-situational learning account. *Behavioral and Brain Sciences*. In press.

Whorf, B. L. (1956). *Language, Thought, and Reality*. Cambridge Ma.: MIT Press.

Wittgenstein, L. (1958). *Philosophical Investigations*. Oxford, UK: Basil Blackwell.

Wray, A. (1998). Protolanguage as a holistic system for social interaction. *Language and Communication 18*, 47-67.

Yamauchi, H. (2004). *Baldwinian Accounts of Language Evolution*. Ph.D. thesis, University of Edinburgh.

Ziemke, T. and N. E. Sharkey (2001). A stroll through the worlds of robots and animals: Applying Jakob von Uexküll's theory of meaning to adaptive robots and artificial life. *Semiotica 134(1-4)*, 701-746.