The Origins of Syntax in Visually Grounded Robotic Agents

Luc Steels
SONY Computer Science Laboratory
6 Rue Amyot, 75005 Paris
VUB AI Laboratory
Pleinlaan 2, 1050 Brussels
steels@arti.vub.ac.be
Submission to Artificial Intelligence Journal

Abstract

The paper proposes a set of principles and a general architecture that may explain how language and meaning may originate and complexify in a group of physically grounded distributed agents. An experimental setup is introduced for concretising and validating specific mechanisms based on these principles. The setup consists of two robotic heads that watch static or dynamic scenes and engage in language games where one robot describes to the other what they see. The first results from experiments showing the emergence of distinctions, of a lexicon, and of primitive syntactic structures are reported.

1 Introduction

Artificial Intelligence research has made remarkable progress the last decades by showing how operations over symbolic models may explain various aspects of intelligent behavior, such as planning, problem solving, or natural language processing. However, the problem of the origin of these symbolic models has so far not been adequately addressed. Most of the time it is the programmer who designs formalisms and datastructures, who provides the ontology of

objects, concepts and their relations, and who interprets the world and feeds examples to the AI system. Even most learning systems start from a prior ontology, carefully designed formalisms or networks, and carefully prepared example sets.

The research discussed in this paper attempts to address the lack of grounding and the lack of self-construction in present-day AI systems. It focuses on how representations could originate and become more complex, without the intervention of human designers. We are interested to understand both the origin of the *form* of representations (the origin of syntactic structure) and the *content* (e.g. the origin of concepts of space, time, objecthood, etc.).

We hypothesize that communication through language can be a driving force to bootstrap the representational capacities of intelligent agents and that it is the way through which agents manage to share ontologies and world views, even though one agent cannot inspect directly the internal states of another agent. Language and meaning co-evolve: Language becomes more complex because more complex meanings need to be expressed, and meanings become more complex because a more complex language enables its expression. Sufficiently complex meaning then becomes the basis for other cognitive activities like planning, cooperation, or problem solving.

We have already reported how agents may autonomously develop distinctions [19], and how they may develop autonomously a lexicon for expressing these distinctions [20]. A first experiment in physical grounding, in which these capabilities were instantiated on robotic agents playing adaptive language games, has been presented in [23]. The present paper goes beyond this earlier work by using vision as source of sensory data and by showing the very beginnings of syntax.

The research reported here is related to a lot of work currently being done in machine learning as well as recent work on the origins of language, as discussed in [14], [9], [1], [7] and [10]. This related research is extensively surveyed in [22].

The rest of the paper is in four sections. The next section (section 2) introduces the experimental setup used to validate mechanisms for the origins of language and meaning and study their performance. Then the main principles underlying our approach are briefly presented. Section 4 moves into concrete technical details. It discusses processes for segmenting raw images and collecting data about image segments, for constructing symbolic descrip-

tions about each segment, and for coding and decoding symbolic descriptions into words. Section 5 then turns to the problem of the origins of syntax. It examines under which conditions syntax may emerge and what additional structure is needed in the agents. Some conclusions end the paper.

2 The Talking Heads Experiment

It is an important tradition in AI to design and implement challenging experimental settings in which various issues can be addressed in an integrated fashion. We have therefore designed and implemented a setup to be able to focus on the problem of the origins of language and meaning.

The setup features two 'robotic heads'. Each head consists of a black and white camera mounted on a pan-tilt unit, connected to electronics for low-level signal processing and actuator control (figure 1), and a main computer for running the symbolic processes described further in the paper. Each head is autonomous and can only influence another head through language. At present, communication goes through a local area network connecting the computers. In a later phase, the language communication is planned to be through sound. Although we have only two physical heads, we can simulate multiple agents by 'loading' the state of different agents into each head.

The heads are either holding still, observing a static scene before them, or they move, trying to track dynamically moving objects. Figure 2 shows a typical example of the dynamical environments we use in our experiments. It consists of one or more robots moving about in an ecosystem which contains a charging station and other objects. As heads turn while tracking a robot, other objects come occasionally into view: the charging station, obstacles, other robots, etc. These objects are distinguished against the background by standard low-level visual processing.

The robotic heads engage in *language games* in which they describe to each other what they see. Observation starts after the previous conversation has terminated and goes until the beginning of the next conversation. During this time period various objects (or more precisely coherent image-fragments) will have been in view. These image-fragments constitute the *context* of a conversation. One element from the context and its dynamical behavior is chosen by the speaking agent as the *topic*. Distinctive descriptions characterising the topic are conceptualised by the speaker and encoded in language.

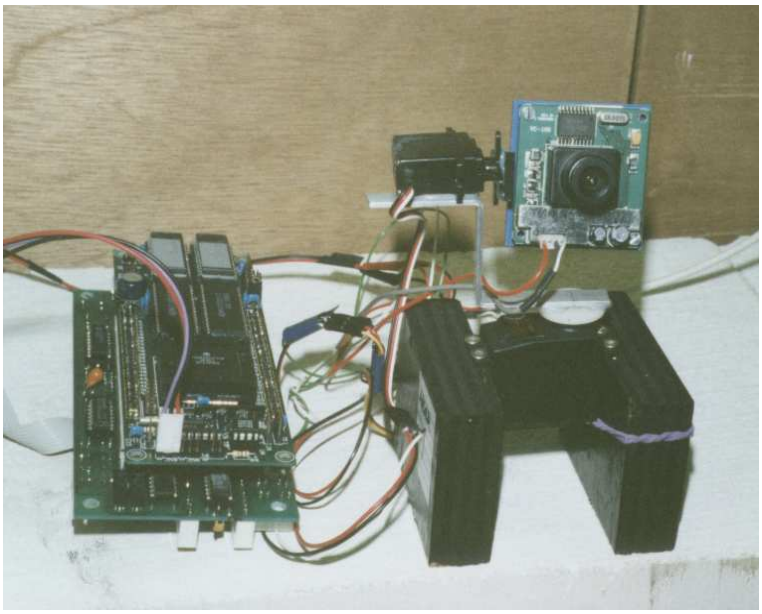

Figure 1: Language and meaning creation is performed by robotic heads which have a camera and 2 degrees of freedom movement. The black-and-white camera unit is shown with on the left dedicated hardware for low level signal processing and actuator control. The heads track moving objects and engage in language games expressing what happened most recently before them.

Figure 2: Typical example of dynamical scenes used in the talking heads experiment. They consist of autonomous robots roaming in an ecosystem with a charging station, competitors, and obstacles. These robots have been used in other language grounding experiments, as reported in [23].

They are then decoded by the hearer. A language game suceeds if the meaning decoded by the hearer fits with his observations and conceptualisations of the same scene. Otherwise the game fails and various repair actions are undertaken by each agent. The observation time is initially short so that typically only two objects are involved (the topic and one or sometimes zero objects). It becomes progressively longer as the heads develop more concepts and more language. The agents either take turns, so that both construct and acquire language, or one plays the role of language creator and the other one as language acquirer. The examples in the remainder of the paper are taken from the second type of game.

In order to have a successful language game, many conditions must be satisfied:

1. There must be low level sensory processes that extract sufficiently rich data from the raw images in real-time.

2. There must be a repertoire of concepts for categorising these data. This repertoire must be sufficiently rich to distinguish the topic from the other elements making up the context.

3. There must be a set of shared words lexicalising the concepts. This set must cover all the distinctions that need to be expressed in this environment.

4. If grammar has become necessary or useful, there must be a set of shared grammatical conventions.

The experimental (and theoretical) challenge is to show how all this may emerge *without* a specific language or ontology being programmed in and *without* human intervention during development. It is in addition required that the system is open, i.e. new unseen objects may enter into the environment at any time, possibly requiring extensions of the set of low level sensory routines, the conceptual repertoire, the lexicon and the syntax. The total system must also be open from the viewpoint of the agents: A new agent should be allowed to enter the community and this agent should be able to acquire the conceptual distinctions and language already present in the community. It might also happen that an agent leaves the group. This should not cause a total collapse of the linguistic and conceptual capabilities of the other agents.

In a way, the 'talking heads' experimental setup is restricted because we want to be able to do controlled and repeatable experiments. But it is at the same time rich enough to address the issues raised in this paper for now and future work: The ontology potentially present in this environment includes objects, invariant properties of objects, time, space, dynamic state changes and actions, and situations involving multiple objects (the robot pushing against another object, an object disappearing behind another one, etc.). As more and more complex meanings require expression, the arsenal of linguistic means must steadily expand to include expression of roles of objects in situations or actions, temporal expression (tense, mood, aspect), etc. Although we have been able to bootstrap the whole system, only a very small fraction of the rich potential of the experimental setup has been tapped so far.

3 Major Hypotheses

Before embarking on a more detailed description of the various components and processes implemented at this point, it is useful to state briefly the main principles underlying our approach:

1. *Progressive Increase in Complexity.* We hypothesise that agents construct and acquire concepts and language in a stepwise fashion, starting from very simple and basic constructions and gradually leading up to more complex ones. The total system is never in a steady state but keeps evolving as new challenges arise. This progressive increase must have happened at the species level during the time language originated and can still be observed in the formation and evolution of language. For example, new sounds emerge in languages and there are continuous shifts and changes to established sound systems [11], lexicons keep evolving to cope with new meanings, various grammaticalisation processes give rise to novel syntactic constructions and shifts in basic grammatical patterns [26]. All of these phenomena are heavily at work in the case of creole formation [24] but happen on a smaller scale in stable languages. The progressive origins and complexification of language and meaning can also be seen at the level of each individual. For example it is only around the age of two, when a stable initial lexicon has been constructed/acquired, that a child starts constructing and using the first simple grammatical devices similar to the ones to be discussed later in this paper

[25].

2. *Adaptive (language) games.* The second basic principle is that the overall system relating perception and language can be decomposed into a series of adaptive games. A game is a particular kind of interaction between agents or between an agent and the environment. The nature of the game is determined by the activity concerned: Imitation games are used to develop a common sound repertoire, discrimination games are used to develop distinctions, naming games lead to the formation of a lexicon, and memory games give rise to syntax. A game is adaptive when the participants in the game change their internal structure after a game in such a way that they are more successful in future games. In the present case, the change may take various forms:

- An agent may infer new information about the language or about concepts held by the other agent, for example he may acquire the use of a new word.
- An agent may construct new concepts or new linguistic conventions - possibly by analogy with existing ones. This constructive aspect is crucial because it is the way in which the system is bootstrapped from scratch.
- An agent may adapt already existing structures. For example, extend or restrict the scope of use of a grammatical construction.

Note that adaptive games imply a cultural transmission and evolution of concepts and language. Our approach therefore contrasts sharply with the proposal that language and meaning have originated in a genetic fashion [17] or that language or meaning acquisition is a matter of instantiating and setting parameters by an innate language acquisition device [2].

3. *Selectionism.* Although we do not assume genetic evolution to be the main driving force in the evolution of language or meaning, our approach is nevertheless selectionist: Structures are being created or adopted by an agent based on only local information and imperfect knowledge. These structures are then subjected to various selectionist constraints in subsequent games. For example, sounds which are too close to be distinctive will progressively disappear. Distinctions that were created but turn out to be irrelevant, will be forgotten. Words that an agent invented to refer to certain descriptions

but which are not picked up by other agents will be abandoned. Syntactic constructions that are confusing or too difficult to parse will give way to clearer and simpler structures. A key question for the future is to identify precisely the selectionist pressures that drive language and meaning to adopt the universal tendencies observed in natural languages (see [13] for a similar approach to phonetics).

4. *Co-evolution* The different games are not played in isolation but are coupled in two ways: The result of one game provides building blocks for the next game. For example, distinctions produced by discrimination games are the basis for the descriptions lexicalised in naming games. Conversely, selectionist constraints flow in the opposite direction. For example, those distinctions are prefered that are lexicalised and whose lexicalisations have been adopted by the rest of the agent population. These two-way flows not only cause a progressive coordination but they also drive the increases in complexity at each level.

5. *Self-organisation* A group of agents engaging in language games and interactions with the world and others form an open distributed system. No agent is in full control, and agents have only limited knowledge of the behavior or internals of other agents. This raises the issue how there might ever arise coherence. Here we rely on a principle which has first been proposed and discovered in physico-chemical and biological systems, namely the principle of self-organisation [16]. Given a system in which there is natural variation through local fluctuations, global coherence in the form of a so-called dissipative structures may emerge provided certain kinds of positive feedback loops are in place. The system gets locked into a particular metastable state due to a process of symmetry breaking. This state is not predictable in advance, nor is it necessarily the most "efficient" state.

Self-organisation takes place here because each agent keeps track of the use and success of a certain linguistic or conceptual device. Because an agent wants to maximise success in future games, it prefers to use those structures that have had most success. This causes a positive feedback in the total multi-agent system. The more something has success the more it is used, and the more it is used the more success it has. The resulting coherence is not only self-organised but also keeps dynamically evolving and adapting itself.

It is of interest that all these principles have been used in other fields to explain complexity, particularly for the explanation of biological complexity

[15]. It is therefore reasonable to assume that they are at work in the origins of cognitive complexity and language. In the remainder of the paper the general principles are instantiated so that concrete experiments are possible. Of course, this is only one possible instantiation and many more factors enter in the evolution of human natural languages. Our goal is to understand the principles, not mimick human language genesis *in toto*.

4 Inventing and lexicalising distinctions

This section focuses on mechanisms for sensory processing, meaning creation and lexicon formation. They have already been described in other papers [19], [20], [21] which should be consulted for a more extensive and formal discussion. The grammatical component is discussed in the next section. Ignoring for the time being grammar, the general architecture of the system built so far is as in figure 3. Each of these components is now discussed in some detail.

4.1 Sensory Processing

The tracking and image processing algorithms identify coherent *image- fragments*. Fragments are either formed because an object moves against the background, detected based on subtracting consecutive images, or because they form a region whose intensity differs significantly from the average intensity of the complete image. For example, the robot moving around yields one continuous image-fragment, as long as it does not disappear out of side. Other objects yield other image-fragments possibly coming into view for a brief time period and disappearing again. During the observation period, various image-fragments are created and monitored as long as they stay in view. Note that there is no notion of object permanence yet.

For each image-fragment, low level sensory routines collect a variety of data: the size of the bounding box of the image-fragment, the average intensity, the ratio between the significant area and the total area of the bounding box, the orientation of the head with respect to the central point of the image-fragment, the maximum and minimum value, the time the image-fragment was seen, the sharpness or visibility (i.e. how much the image-fragment is in focus), etc. Some of these data will be relatively constant during the time

Figure 3: There are four major components leading up to syntax: A sensory processing component, a categorisation component, a discrimination component and a lexical component. One component provides input for the next one and conversely a higher level component supplies selectionist constraints for a lower one. When a component fails, adaptation takes place.

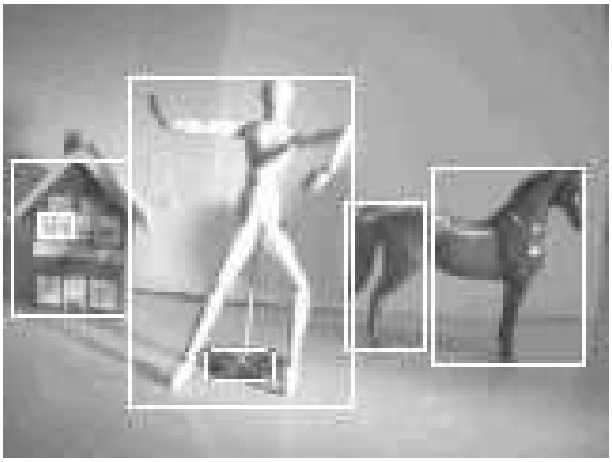

Figure 4: Typical example of the static scenes used in the talking heads experiment. The identified fragments are surrounded by a bounding box.

the image-fragment is seen. Others will be changing. In that case, an *image-segment* is created as part of an image-fragment. For example, when the angle towards the head steadily increases (caused by the robot moving to the left) and then steadily decreases (caused by the robot moving to the right) two segments are created as part of the same image-fragment for the robot. At the moment a conversation is about to start, all the image-fragments that are still ongoing are closed, their global properties computed, and the total set of recent image-fragments is passed on to the next component.

A typical example of a (static) scene is shown in figure 4. It contains a horse, a house and a wooden puppet. The bounding boxes in figure 4 indicate the fragments. A sample of data collected about each of these fragments is displayed in the table below. The data channels are:

- *hor*: The horizontal angle between the head and the center point of the fragment. This is a function of the pan angle and the x-distance between the center of the image and the center of the fragment.

- *ver*: The vertical angle between the head and the center point of the total image. This is a function of the tilt angle and the y-distance

between the center of the image and the center of the fragment.

- *area*: The total area of the fragment.
- *vis*: The distance between the center of the fragment and the center of the total image, which indicates how much the fragment is into the center of attention.
- *int*: The average intensity inside the bounding box of the fragment.
- *ratio*: The ratio between the area filled by the region identified as significant and the total bounding box around the region.

x1	y1	x2	y2	hor	ver	area	vis	int	ratio
33	19	92	106	0.465	0.504	0.515	0.852	0.449	0.389
9	55	18	61	0.400	0.494	0.553	0.047	0.867	0.530
2	41	33	82	0.407	0.497	0.548	0.189	0.284	0.333
113	43	153	95	0.571	0.523	0.453	0.340	0.146	0.269
90	52	111	91	0.532	0.528	0.470	0.139	0.343	0.409
53	91	71	99	0.473	0.593	0.467	0.026	0.271	0.543

Obviously many more kinds of data could be extracted from the image. So far we have used only the most rudimentary image processing techniques. Our goal is not to perform sophisticated vision processing but to study the way in which a symbolic process could use whatever is provided by a low level visual processor.

4.2 Categorisation and Discrimination

The sensory processing component yields a set of image-fragments I and data for each image-fragment or its segments. One image-fragment $t \in I$ is chosen by the speaker to be the topic of the language game. The others make up the context $C = I - \{t\}$. The next component translates these data in a symbolic description in the form of attribute-value pairs. There are many possible ways to do this. We have been experimenting with (binary) discrimination trees which segment the continuous domain of each data channel into finer and finer regions. Each data source corresponds to one attribute and each region to a value of a description. For example for the channel associated with average intensity (INT), the tree first splits into two branches. One

delineates data between [0.0,0.5] which in English could be lexicalised as 'light', and another one between [0.5 and 1.0] which could be lexicalised as 'dark'. By going through the different discrimination trees associated with the various channels for which data are available, a description-set can be constructed for each image fragment.

In order to engage in a conversation, it necessary to find a description-set which distinguishes the topic from the other elements in the context. This proceeds in three steps:

1. The existing discrimination trees are used to derive the first (and therefore most abstract) descriptions for each data channel associated with an image-fragment. Each image-fragment i has thus an associated description-set F_i.

2. Distinctive description-sets are computed. Let F_t be the description-set of the topic, then a distinctive description-set $S_t \subset F_t$ is such that there is no $c \in C$ such that $S_t \subset F_c$.

3. There are now three cases:

   (a) There are no distinctive description-sets $F_t = \emptyset$ but it is possible to refine existing descriptions because the discrimination trees contain more refinements. In that case, new description-sets with these refinements are computed and step 2 above is reconsidered.

   (b) There may be no distinctive description-sets and the discrimination trees were exhaustively explored. In this case, a new distinction is created by randomly selecting one of the active endpoints of the tree and dividing its associated region into two subregions. There is no guarantee that this is the right solution - this will become clear in subsequent discrimination games.

   (c) There are distinctive description-sets. When there is more than one possibility, the distinctive description-sets are ordered based on a number of selectionist criteria: A smaller set, a set with more successful descriptions, and a set where the descriptions are lexicalised, is prefered. The best distinctive description-set according to these criteria is used in the remainder of the game.

Here are some examples of this process operating on scenes as shown in figure 4. First a discrimination game is shown where head-104 is the speaker. There are six image fragments and the fifth is chosen as topic. A new distinction is created (as there are no trees available yet). The new description divides the average intensity of an image fragment into two subregions, thus creating the values v-339 and v-440 for the attribute INT.

```
1 ++> Speaker: head-104 Topic: 5 Context: 6 4 3 2 1 0
Failure: INSUFFICIENT-descriptionS
=> New distinction: int[-1.0 1.0]: v-339 v-340
Failure
```

The next game also fails, a new distinction is created by dividing the VER channel.

```
2 ++> Speaker: head-104 Topic: 4 Context: 6 5 3 2 1 0
Failure: INSUFFICIENT-descriptionS
=> New distinction: VER [-1.0 1.0]: v-341 v-342
Failure
```

In the next game, a first success is seen. For each channel on which discriminations can be made (namely for average intensity (INT) and vertical angle (VER) the data and the categorised data is printed out. Then the distinctive description-set is computed.

```
3 ++> Speaker: head-104 Topic: 6 Context: 5 4 3 2 1 0
Categorised data:
```

INT			VER		
Obj	data	value	Obj	data	value
o6	0.319	v-339	o6	0.507	v-342
o5	0.350	v-339	o5	0.496	v-341
o4	0.370	v-339	o4	0.415	v-341
o3	0.747	v-340	o3	0.486	v-341
o2	0.908	v-340	o2	0.482	v-341
o1	0.949	v-340	o1	0.482	v-341
o0	0.877	v-340	o0	0.525	v-342

```
Distinctive combinations:
   ((ver v-342) (int-339))
```

Here is an example after about 30 games. Discrimination trees are much more developed and for some channels refinements 4 levels deep are made before a distinctive description-set is found. (NIL means no further refinement could be made.)

```
38 ++> Speaker: head-104 Topic: 6 Context: 5 4 3 2 1 0
Categorised data:
```

RATIO						INT				
Obj	data	value	value	value	value	Obj	data	value	value	value
o6	0.437	v-345	v-364	v-376	v-379	o6	0.321	v-339	NIL	NIL
o5	0.457	v-345	v-364	v-376	v-380	o5	0.320	v-339	NIL	NIL
o4	0.608	v-346	v-357	NIL	NIL	o4	0.333	v-339	NIL	NIL
o3	0.926	v-346	v-358	v-362	NIL	o3	0.368	v-339	NIL	NIL
o2	0.641	v-346	v-357	NIL	NIL	o2	0.567	v-340	v-359	NIL
o1	0.009	v-345	v-363	NIL	NIL	o1	0.904	v-340	v-360	NIL
o0	0.438	v-345	v-364	v-376	v-380	o0	0.954	v-340	v-360	NIL

HOR				VER				DIST					
Obj	data	value	value	Obj	data	value	value	Obj	data	value	value	value	v
o6	0.459	v-343	NIL	o6	0.507	v-342	NIL	o6	0.573	v-352	v-367	v-369	
o5	0.462	v-343	NIL	o5	0.506	v-342	NIL	o5	0.567	v-352	v-367	v-369	
o4	0.486	v-343	NIL	o4	0.500	v-341	NIL	o4	0.527	v-352	v-367	v-369	
o3	0.538	v-344	NIL	o3	0.459	v-341	NIL	o3	0.463	v-351	NIL	NIL	
o2	0.524	v-344	NIL	o2	0.423	v-341	NIL	o2	0.527	v-352	v-367	v-369	
o1	0.473	v-343	NIL	o1	0.486	v-341	NIL	o1	0.566	v-352	v-367	v-369	
o0	0.506	v-344	NIL	o0	0.487	v-341	NIL	o0	0.500	v-351	NIL	NIL	

```
Distinctive combinations:
   ((RATIO v-379))
```

The evolution of the conceptual repertoire is illustrated in figure 5. There is a steady increase in the number of distinctions, in order to cope with situations that have not been seen before. But as more and more distinctions become available, success in discrimination is steady and the discrimination trees stabilise.

Figure 5: Increase in discrimination success and in number of distinctions for a single agent in a series of 500 discrimination games. The agent becomes progressively capable to deal with the situations he is confronted with. 45 distinctions have been created, implying 90 possible descriptions.

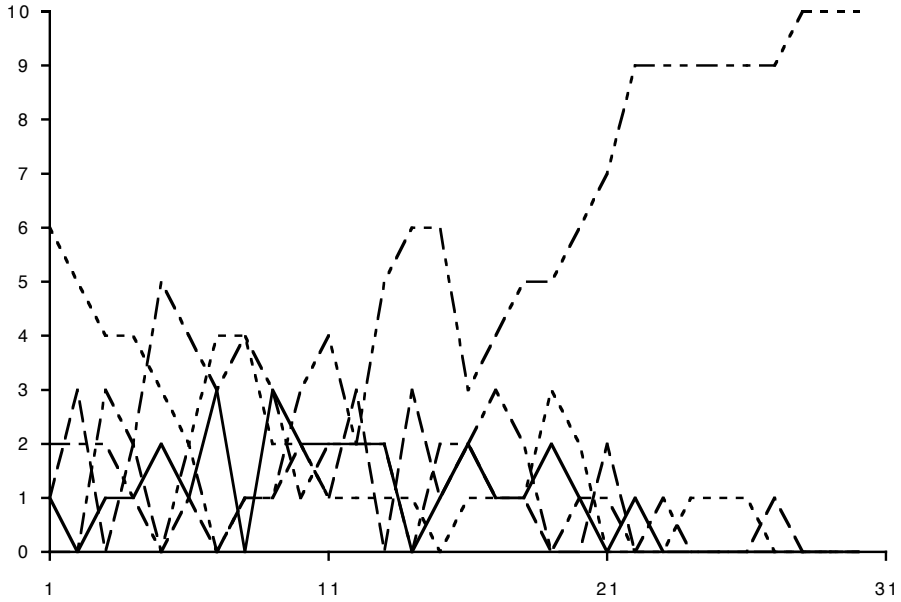

Figure 6: This figure shows the results of an experiment where 10 agents progressively agree on the word to be used for expressing a particular meaning. The y-axis shows the average communicative success of a word and the x-axis a series of language games. Different associations compete until one gains complete dominance.

4.3 Lexicon formation

A lexicon consists of a set of word-meaning associations, where a word is a description-set describing a word form, which at present consists of a sequence of letters drawn from a finite shared alphabet, and a meaning is a description-set describing aspects of reality, as produced by the discrimination games discussed in the previous section. One word may be associated with many meanings and one meaning may be associated with many words. Each agent has his own lexicon and an agent cannot directly inspect the lexicon of another one. Each agent monitors how often a word-meaning association has been used and how successful it has been in its use. While encoding, a speaker will prefer word-meaning associations that have been used more often and were more succesful in use. This establishes a positive feedback loop pushing the group towards self-organised coherence (figure 6).

Let L_a be the lexicon of a single agent $a \in A$. It is initially empty. The possible meanings of a word in L are denoted as $F_{w,L}$. The following functions can be defined:

- $cover(F, L)$ defines a set of expressions, where each expression U is such that $\forall f \in F, \exists w \; such \; that \; D \in F_{w,L}, \; f \in D \; and \; w \in U$
- $uncover(U, L)$ defines a set of description-sets, where each description-set K is such that $\forall w \in U, \exists f \; such \; that \; D \in F_{w,L} \; and \; D \subset K$

As part of the language game, the speaker selects one distinctive description-set output by the categorisation component and translates it to words using the cover function. The hearer interprets this expression using the uncover function and compares it with his expectations, i.e. the description-set uncovered from the expression should be a subset of the distinctive description-sets extracted using categorisation by the hearer.

As a side effect of a language game, various language formation steps take place:

1. *The speaker does not have a word*: In this case at least one distinctive description-set S is detected but the speaker s has no word(s) yet to express it. The language game fails. However the speaker may create a new word (with a probability typically $w_c = 0.05$) and associate it in his lexicon with S.

2. *The hearer does not have a word*: At least one distinctive description set S is detected and the speaker s can construct an expression to express it, i.e. $\exists u \in cover(S, L_s)$. However, the hearer does not know the word. Because the hearer has a hypothesis about possible description-sets that might be used, he is able to extend his lexicon to create associations between the word used and each possible description-set. If there is more than one possibility, the hearer cannot disambiguate the word and the ambiguity is retained in the lexicon.

3. *The speaker and the hearer know the word*: In this case there are two possible outcomes:

   (a) *The meanings are compatible with the situation*: The dialog is a success and both speaker and hearer achieve communicative success. Note that it is possible that the speaker and the hearer use

different description-sets, but because the communication is a success there is no way to know this. Semantic incoherences persist until new distinctions become important and disambiguate.

(b) *The meanings are not compatible with the situation*: The same situation as before may arise, except that the description-set uncovered by the hearer is not one of the description sets expected to be distinctive. In this case, there is no communicative success, neither for the speaker or the hearer.

Here are some examples of this process. The first example shows a language game where the speaker does not have a word to cover for the distinctive feature set that has been found. A new word is created but the game ends in failure

```
5 ++> Speaker: head-40 Topic: 0 Context: 6 5 4 3 2 1
Categorial Perception: ((AREA v-364)(RATIO v-367))
Failure: MISSING-WORD-SPEAKER
Failure: MISSING-WORD-FORM
=> Extend word repertoire: (R U)
=> Extend Lexicon:
 Association-193
  Function:([AREA v-364] [RATIO v-367])
  Form:([WORD (R U)])
Failure
```

The new word is now used but unknown to the hearer. The hearer has still insufficient features to make the distinction between the topic and the other objects in the context.

```
6 ++> Speaker: head-40 Topic: 0 Context: 6 5 4 3 2 1
Categorial Perception: ((AREA v-364)(RATIO v-367))
Word-schemas: (Association-193)
Meaning:([RATIO v-367] [AREA v-364])
Expression: ((R U))
++> Hearer head-41
Expression: ((R U))
Failure: INSUFFICIENT-FEATURES
=> New distinction: INTENS [0.0 1.0]: v-369 v-370
Failure
```

A few games later, the hearer has made the necessary distinctions and is ready to construct a lexical association:

```
9 ++> Speaker: head-40 Topic: 0 Context: 6 5 4 3 2 1
Categorial Perception: ((AREA v-364)(RATIO v-367))
Word-schemas: (Association-193)
Meaning:([RATIO v-367] [AREA v-364])
Expression: ((R U))
++> Hearer head-41 Topic: 0 Context: 6 5 4 3 2 1
Failure: MISSING-WORD-FORM
=> Extend word repertoire: (R U)
Categorial Perception: ((INTENS v-370)(HOR v-371))
Expression: ((R U))
Failure: MISSING-LEMMA-HEARER
Extend Lexicon:
  Association-194
    Function:([HOR v-371] [INTENS v-370])
    Form:([WORD (R U)])
Failure
```

Still later, both speaker and hearer have a word and the game ends in success.

```
14 ++> Speaker: head-40 Topic: 0 Context: 6 5 4 3 2 1
Categorial Perception: ((AREA v-364)(RATIO v-367))
Word-schemas: (Association-193)
Meaning:([RATIO v-367] [AREA v-364])
Expression: ((R U))
++> Hearer head-41 Topic: 0 Context: 6 5 4 3 2 1
Categorial Perception: ((INTENS v-370)(HOR v-371))
Word-schemas: (Association-194)
Meaning:([INTENS v-370] [HOR v-371])
Expression: ((R U))
Success
```

Overall, we can see a steady co-evolution of the discrimination trees, which grow as more distinctions need to be made, and the lexicon, which lexicalises these distinctions. After a few hundreds of games, 25 words have been created. Part of the lexicons of the two agents is as follows:

Form	Speaker	Hearer
([WORD (R U)])	([AREA v-364][RATIO v-367])	([HOR v-371][INTENS v-370])
		([RATIO v-377][HOR v-371])
		([RATIO v-377][AREA v-382])
([WORD (B O)])	([RATIO v-368][AREA v-363])	([HOR v-372][RATIO v-378])
		([VER v-376][RATIO v-378])
([WORD (M A)])	([VER v-379][AREA v-363])	([AREA v-381][VER v-375])
		([AREA v-381][INTENS v-370])
([WORD (L U)])	([AREA v-364][HOR v-384])	([HOR v-372][AREA v-382])
		([INTENS v-370][AREA v-382])
		([RATIO v-377][AREA v-382])
([WORD (N U)])	([VER v-397])	([VISIB v-404] [HOR v-371])
		([VISIB v-404][VER v-375])
		([VISIB v-404][AREA v-382])

The meaning of a word for one agent is always different from the meaning understood by the other because the names of the values are proper to each agent, nevertheless they may refer to the same region. Additional differences and ambiguities come in because often more than one distinctive description-set is compatible with the situation occurring in a particular game. The lexicons progressively stabilise and reach coherence due to the positive feedback loop based on use and success. The evolution of the lexicon of a single agent is illustrated in figure 8. We see clearly that there is a progressive buildup of the lexicon as needed to cover the distinctive description-sets that distinguish the chosen topic from the other objects. We see also that progressively the other agent picks up the vocabulary.

5 The emergence of syntax

Linguists in the structuralist tradition view grammar as a formal device that has no functional or cognitive motivation [2]. At the same time, there is an opposing long tradition in linguistics, which views grammar in functional terms and grammatical processing or grammar formation as an integral part and special case of general cognitive processing [4], [12], [6]. Our approach follows this second direction. This implies that in order to understand how grammar may emerge, we must understand why grammar is useful and necessary, i.e. what it is for. By grammar, we mean any kind of linguistic device

Figure 7: Evolution of the lexicon of a single agent. Both the increase in the success in covering the distinctive description-set of a game (*produced*) and the success in understanding and agreeing with the lexicalisation by the hearer (*understood*) is shown.

that goes beyond the use of individual words in isolation. This includes word order, function words (such as the auxiliary "do" in English to form negation), morphological variation (affixes, suffixes), agreement phenomena such as number concord between subject and verb, intonation contours, etc. These devices are used for a variety of purposes:

- *To express additional aspects of meaning.* For example, subject and verb are inversed to express questions (as in Dutch), word order is used to express case roles (as in "John gave Mary a book"), etc.

- *To aid in conveying the grammatical (and hence semantic) functions of a word.* For example, the distinction between adjectives and nouns or the distinction between topic and comment. This raises the predictive characteristics of the language. If we do not know a word but can guess its grammatical function we can more easily guess its meaning.

- *To aid in managing the complexity of parsing and producing.* Very quickly combinatorial explosions arise when multiple words which each form different groups are combined. Grammatical devices help by embodying conventions that establish what belongs to what. For example, the verb in English affirmative sentences signals that the noun group identifying the subject has terminated.

In this paper, we focus only on the very simplest form of grammar with hierarchical structures and word order, i.e. pure syntax. We also focus only on the very beginnings of grammar.

Our hypothesis is that grammar arises when a *cognitive memory system* intervenes in the case of multiple word expressions. Such expressions are already generated by the processes described in the previous sections. The cognitive memory system is capable of (1) recording a situation when it arises, (2) recognising a recorded situation, and (3) re-enacting a previous situation. Situations are recorded in the form of schemas and associations between schemas. It is not assumed to be specific to language but underlies memory and reuse of action sequences in planning, recurrent problems in expert problem solving, or scene recognition in complex visual processing. A schema consists of a set of slots, restrictions on the fillers of a slot, and constraints on the total. The memory system attempts to compact its internal structures by generalising or specialising schemas. It also attempts

to re-use existing structure by allowing partial matches between a new situation and a previous situation and re-enactment based on analogy. Many of the techniques necessary to build such as cognitive memory system have been explored earlier in research on frames and schemas and case-based reasoning. The technical details of the cognitive memory system that we use goes beyond the scope of the present paper. We focus instead on how such a memory device could give rise to syntax.

5.1 Form of the grammar

The grammar is seen as a natural continuation of the lexicon, in the sense that it consists also of associations between forms and meanings. Use and success are monitored for each association so that the same type of self-organised coherence arises in the group, as seen in the lexicon (see figure 6). The form is now a more complex structure, defined as a syntactic schema. The meaning is a semantic schema. The schemas circumscribe a description-set in terms of a set of slots, restrictions on the fillers of each slot, and constraints on the combination of the fillers to form the total covered by the schema.

Syntactic schemas describe word groups. They have an associated category which corresponds in linguistic terms to group categories like noun-group, verb-group, sentence. The slots in syntactic schemas correspond to syntactic functions (also called grammatical relations) such as subject, object, modifier, complement. They name the roles that certain words or word groups play in the group. The categories used to restrict possible slot-fillers correspond in linguistic terms to syntactic categories like noun, verb, adjective, etc. An example of a syntactic schema generated by the cognitive memory system is the following:

```
Schema-541
  SLOTS: (syn-slot-51 syn-slot-50)
  DESCRIPTION-SET:
    ([syn-slot-50 syn-cat-75]
     [syn-slot-51 syn-cat-76])
  CONSTRAINTS: ((PRECEEDS (>> syn-slot-50) (>> syn-slot-51)))
  CATEGORY: syn-cat-77
  USE: 10
  SUCCESS: 3
```

The constraints on the schema are represented in a constraint system. Each constraint has a dual procedural encoding: to enforce the constraint when re-enacting the situation described by a schema or to test the constraint when recognising the schema. In the present case only a precedence relation is recorded. Agreement, intonation patterns, morphological variations, are some other possible constraints on syntactic schemas.

The categories restricting slot-fillers are either themselves defined in terms of schemas (for example, syn-cat-77 could be the restriction on a slot-filler in another schema), or they are defined as rules that are applied in a forward-chaining fashion during matching. Two examples of rules related to the above schema are:

```
rule 101: ([WORD (W U)]) => ([MEMBER syn-cat-76])
rule 99: ([WORD (W O)]) => ([MEMBER syn-cat-75])
```

Semantic schemas describe the language-specific semantic structures underlying the meanings of complete word groups. The closest linguistic correspondent to a semantic schema is the notion of a case-frame. The constraints indicate how the total meaning is constructed/decomposed into the meaning of the parts. During interpretation such constraints therefore perform the same role as Montague style semantic interpretation functions. The slots correspond to cases such as agent, patient, time, distance, or arguments of semantic functions. The categories used to constrain what can fill a slot correspond in linguistic terms to selection restrictions like animate, human, edible, future, etc. The schema has also an associated category for the whole so that hierarchical combination is possible. An example of a semantic schema is:

```
Schema-542
  SLOTS: (sem-slot-51 sem-slot-50)
  DESCRIPTION-SET:
   ([sem-slot-50 sem-cat-75]
    [sem-slot-51 sem-cat-76])
  CONSTRAINTS: ((CONJUNCTION (>> sem-slot-50) (>> sem-slot-51)))
  CATEGORY: sem-cat-77
  USE: 10
  SUCCESS: 3
```

Inference rules such as the following define the selection restrictions:

```
rule 102: ([VISIB v-411]) => ([MEMBER sem-cat-76])
rule 100: ([VER v-431]) => ([MEMBER sem-cat-75])
```

Each association in the grammar associates a syntactic schema with a semantic schema. The association can be used in two directions. If a syntactic schema is recognised, i.e. can be mapped onto a description of a group of words or word groups, the semantic schema is used to reconstruct its meaning. If a semantic schema is recognised, the syntactic schema is used to reconstruct the form. The association contains a mapping of the slots in order to enable this reconstruction. The association combining the above two schemas is as follows:

```
Association-271
  FUNCTION: Schema-542
  FORM: Schema-541
  MAPPING: ((syn-slot-51 sem-slot-51) (syn-slot-50 sem-slot-50))
  USE: 10
  SUCCESS: 3
```

5.2 Operation of the grammar

As discussed in the previous sections, a speaker chooses a distinctive description-set resulting from perception and discrimination, and then performs lexicon lookup. The lexicon may yield a group of words which cover the chosen distinctive description-set (see figure 8). Rather than simply transmitting these words to the hearer, the cognitive memory system comes in action and attempts to find an association (or set of associations) for the total. The inference rules operate on both the word forms and the meanings to see what syntactic categories and semantic categories apply. Schemas match when all slots are filled by elements which belong to the appropriate categories (figure 8) and when the constraints on the semantic schema apply. When there is a match, the constraints on the syntactic schema are enacted and added to the description of the form. For example, the precedence relation between the words is added. Only then the form is rendered and transmitted to the hearer. A language game in which all this is happening with the above example schemas and rules is given below:

```
127 ++> Speaker: head-40 Topic: 2 Context: 6 5 4 3 1 0
```

```
Categorial Perception:
     ([VISIB v-411] [VER v-431])([VER v-431] [AREA v-406])
Lexicon lookup: (Association-259 Association-234)
Syntactic structure:
   (syn-cat-77
      (syn-slot-50 (syn-cat-75 |(W O)|))
      (syn-slot-51 (syn-cat-76 |(W U)|)))
Semantic structure:
   (sem-cat-78 (sem-slot-50 (sem-cat-75 (VER v-431)))
               (sem-slot-51 (sem-cat-76 (VISIB v-411))))
Meaning:
   ([VISIB v-411] [VER v-431])
Expression: ((W O) (W U))
```

The hearer engages in similar operations. They are more complex due to the ambiguity of words and uncertainty about the topic. When more than one word is transmitted, several associations will come out of the lexicon and syntactic schemas need to be found that match with the form transmitted by the speaker. When a syntactic schema could be constructed, the associated semantic schema is used to reconstruct the meaning of the total expression, and this meaning is compared with the distinctive description-sets resulting from perception and discrimination, as before. The game succeeds if one of the meanings is compatible with one of the distinctive description-sets.

5.3 Build up of the grammar

As in the case of the lexicon, the build up of the grammar happens when there is a failure, i.e. when no schemas can be found that match with the present situation. The speaker can do this because schemas, associations, categories and inference rules can be constructed once the word forms and the meanings are available (which they are as given by the lexicon). The syntactic constraints are in a first phase partly arbitrary. For example, word-1 may have been coming out of the lexicon process before word-2 leading to the constraint that word-1 (or more precisely the filler of slot-1 in the syntactic schema) preceeds word-2 (i.e. the filler of slot-2). The cognitive memory system acts in a first instant purely as a device that records a particular way in which language has been produced so that it can later be re-produced in

Figure 8: Top: Two lexicon associations have become active to cover the distinctive description-set resulting from discrimination and perception. Bottom: The group of words and the meanings match with schemas forming part of a grammatical association. The matching implies some inferencing to see whether the categories associated with the schemas apply.

the same way.

Also the hearer can perform this recording operation. He is presented with a specific set of word form from which he can abstract a syntactic schema. He derives meaning from the definition of the words in the lexicon and the distinctive description-set coming out of perception. From this a semantic schema can be extracted.

An example of an initial grammar, reached after about 500 language games, and using the lexicon briefly illustrated earlier, is as follows:

```
== Categories ==
102: ([VISIB v-411]) => ([MEMBER sem-cat-76])
101: ([WORD (W U)]) => ([MEMBER syn-cat-76])
100: ([VER v-431]) => ([MEMBER sem-cat-75])
99: ([WORD (W O)]) => ([MEMBER syn-cat-75])
90: ([VISIB v-411]) => ([MEMBER sem-cat-67])
89: ([WORD (W U)]) => ([MEMBER syn-cat-67])
88: ([INTENS v-430] [VER v-398]) => ([MEMBER sem-cat-66])
87: ([WORD (N I)]) => ([MEMBER syn-cat-66])
82: ([VISIB v-411]) => ([MEMBER sem-cat-61])
81: ([WORD (W U)]) => ([MEMBER syn-cat-61])
80: ([VER v-380]) => ([MEMBER sem-cat-60])
79: ([WORD (D I)]) => ([MEMBER syn-cat-60])
54: ([VER v-379] [AREA v-363]) => ([MEMBER sem-cat-40])
53: ([WORD (M A)]) => ([MEMBER syn-cat-40])
52: ([RATIO v-367]) => ([MEMBER sem-cat-39])
51: ([WORD (S O)]) => ([MEMBER syn-cat-39])

== Syntax ==
271: ([sem-slot-50 sem-cat-75] [sem-slot-51 sem-cat-76])
      <=> ([syn-slot-50 syn-cat-75] [syn-slot-51 syn-cat-76])
     (PRECEEDS (>> syn-slot-50) (>> syn-slot-51))
263: ([sem-slot-44 sem-cat-66] [sem-slot-45 sem-cat-67])
      <=> ([syn-slot-44 syn-cat-66] [syn-slot-45 syn-cat-67])
     (PRECEEDS (>> syn-slot-44) (>> syn-slot-45))
237: ([sem-slot-40 sem-cat-60] [sem-slot-41 sem-cat-61])
      <=> ([syn-slot-40 syn-cat-60] [syn-slot-41 syn-cat-61])
     (PRECEEDS (>> syn-slot-40) (>> syn-slot-41))
```

```
222: ([sem-slot-26 sem-cat-39] [sem-slot-27 sem-cat-40])
      <=> ([syn-slot-26 syn-cat-39] [syn-slot-27 syn-cat-40])
     (PRECEEDS (>> syn-slot-26) (>> syn-slot-27))
```

Although the pure recording and replaying of syntactic structures yields a shared set of conventions for multi-word sentences which features hierarchy, exploitation of word order, and the steady build up of a set of linguistic categories, more interesting grammars only emerge when the memory system exercises its powers of generalisation: When no schema is found that matches precisely with an existing schema, weaker matches are tolerated. The best weaker match is chosen and the relevant schemas are adopted. For example, if only one of the two slots in a syntactic schemas are filled, then the schema can still be accepted and extended by adding inference rules that will make the non-filling word form a member of the syntactic category of the non-filled slot. The semantic categories are then also adopted to make the grammatical association as a whole adapted to the expanded use. The elasticity in matching is one of the major factors that determines the nature of the grammar, and universal tendencies observed in grammars of natural languages can help to find out what elasticity should be tolerated. More interesting grammars also emerge when additional operations to restructure the set of grammatical associations are used. For example, it is possible that two partially overlapping schemas become active, leading to the construction of a new schema that integrates both, and thus to a more encompassing syntactic and semantic structure.

6 Conclusions

The paper discussed an agent architecture for the autonomous build up of a repertoire of distinctions, a lexicon for verbalising these distinctions, and a set of syntactic conventions for structuring multiple word sentences. The architecture consists of a set of coupled adaptive games. Each game consists of a particular kind of interaction between two agents or between an agent and the environment. The game is adaptive in the sense that agents change their internal structure to be more successful in future games. The games are coupled because one game delivers building blocks for the next one and selectionist constraints flow from the user to the provider.

The paper proposed also an experimental testbed for testing this architecture on streams of experiences by two robotic heads that are watching real world scenes. Some experimental results which explore the proposed architecture and its underlying principles were presented.

There is obviously a large amount of work left to do, both theoretically and experimentally. Particularly in the area of syntax, we have just reached the very first steps and the further progression towards more complexity will require several additional processes in the memory system. Nevertheless, the progress already achieved raises exciting prospects for understanding the autonomous progressive self-construction of cognitive capacity by a physically embodied agent in an emergent, bottom-up fashion.

7 Acknowledgement

The lexicon, grammar, and meaning formation programs were designed and implemented by Luc Steels. This research was conducted and financed by the Sony Computer Science Laboratory in Paris. The implementation and maintenance of the robotic environments used in the present paper is a group effort at the VUB AI laboratory in Brussels, in which Tony Belpaeme, Andreas Birk, Luc Steels, Peter Stuer, Dany Vereertbrugghen and Paul Vogt have made major contributions. The robot research was financed (until december 1996) by an IUAP project of the Belgian government. The head, the tracking mechanism, and the low level sensory processing were implemented by Tony Belpaeme.

References

[1] Batali, J. (1997) Computational Simulations of the Emergence of Grammar. To appear in Hurford, J., et.al. (1997).

[2] Chomsky, N. (1975) *Reflections on Language.* Pantheon books, New York.

[3] De Boer, B. (1997) Emergent Vowel Systems in a Population of Agents. In Harvey, I. et.al. (eds.) *Proceedings of ECAL 97*, Brighton UK, July 1997. The MIT Press, Cambridge Ma.

[4] Dik, S. (1980) *Studies in Functional Grammar.* Academic Press, London.

[5] G.M. Edelman. *Neural Darwinism: The Theory of Neuronal Group Selection.* Basic Books, New York.

[6] Greenberg, J.H. (1966) *Universals of Language* The MIT Press, Cambridge Ma.

[7] Hurford, J. (1989) Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua*, 77:187-222, 1989.

[8] Hurford, J., C. Knight and M. Studdert-Kennedy (eds.) (1997) *Evolution of Human Language.* Edinburgh Univ. Press. Edinburgh.

[9] Hutchins, E. and B. Hazelhurst (1995) How to invent a lexicon. The development of shared symbols in interaction. In: Gilbert, N. and R. Conte (eds.) *Artificial societies: The computer simulation of social life.* UCL Press, London.

[10] Kirby, S. (1996) Function, Selection, and Innateness: The Emergence of Language Universals. Ph.D. Thesis. University of Edinburgh.

[11] Labov, W. (1994) *Principles of Linguistic Change. Volume 1: Internal Factors.* Blackwell, Oxford.

[12] Langacker, R. (1986) *Foundations of Cognitive Grammar.* Stanford University Press, Stanford.

[13] Lindblom, B. (1986) Phonetic universals in vowel systems. In: Ohala, J. and J. Jaeger (eds.) Experimental Phonology. Academic Press, London.

[14] MacLennan, B. (1991) Synthetic Ethology: An approach to the study of communication. In: Langton, C., et.al. (1991) *Artificial Life II*, Addison-Wesley Pub. Cy, Redwood City Ca.

[15] Maynard-Smith, J. and E. Szathmary (1994) *The major transitions in evolution.* Freeman Spektrum, Oxford.

[16] Nicolis, G. and I. Prigogine (1993) *Exploring Complexity.* Piper, Berlin.

[17] Pinker, S. (1994) *The language instinct.* Penguin Books, London, 1994.

[18] Steels, L. (1994) A case study in the behavior-oriented design of autonomous agents. In Brooks, R. et.al. (eds.) *Proceedings of the third Simulation of Adaptive Behavior Conference.* The MIT Press, Cambridge Ma, 1994.

[19] Steels, L. (1996a) Perceptually grounded meaning creation. In: Tokoro, M. (ed.) *Proceedings of the International Conference on Multi-Agent Systems.* pages 338-344. AAAI Press, Menlo Park Ca, 1996.

[20] Steels, L. (1996b) Self-organising vocabularies. In Langton, C. (ed.) *Proceedings of Artificial Life V.* Nara, 1996.

[21] Steels, L. (1997a) Constructing and Sharing Perceptual Distinctions. In van Someren, M. and G. Widmer (eds.) *Proceedings of the European Conference on Machine Learning*, Prague, April 1997. Springer-Verlag, Berlin, 1997.

[22] Steels, L. (1997b) The synthetic modeling of language origins. *Evolution of Communication*, 1(1):1-35.

[23] Steels, L. and P. Vogt (1997) Grounding adaptive language games in robotic agents. In Harvey, I. et.al. (eds.) *Proceedings of ECAL 97*, Brighton UK, July 1997. The MIT Press, Cambridge Ma., 1997.

[24] Thomason, S. and T. Kaufman (1988) *Language Contact, Creolization, and Genetic Linguistics.* University of California Press, Berkeley, 1988.

[25] Tomasello, M. (1992) *First verbs. A case study of early grammatical development* Cambridge University Press, Cambridge, 1992.

[26] Traugott, E. and Heine, B. (1991) *Approaches to Grammaticalization. Volume I and II.* John Benjamins Publishing Company, Amsterdam, 1991.