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

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 a scene in which a robot moves around in its ecosystem. 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, natural language processing, etc. 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 (including most neural network experiments) start from a prior ontology, carefully designed formalisms or networks, and carefully prepared example sets. This gap in current AI has been severely criticised, for example by Searle through his Chinese Room metaphor.

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 (including the origin of syntactic structure) and its *content* (e.g. the origin of space, time, objecthood, etc.). This research is related to a lot of work currently being done in machine learning but most specifically to recent work on the origins of language, such as by [MacLennan, 1991], [Hutchins and Hazelhurst, 1995], [Batali, 1997], [Hurford, 1989] [Kirby, 1996], and others, as has been extensively surveyed in [Steels, 1997b].

One of the key hypotheses underlying our approach is that communication through language is the main driving force in bootstrapping the representational capacities of intelligent agents. It is also the way through which agents which are part of the same community, 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, problem solving, etc.

This paper reports on concrete progress towards the goals expressed above. It builds on our earlier work showing how a shared language medium in the form of a shared phonology may arise in a group of distributed agents [De Boer, 1997], how agents may autonomously develop distinctions [Steels, 1996a], and how they may develop autonomously a lexicon for expressing these distinctions [Steels, 1996b]. A first experiment in physical grounding, in which these components were instantiated on robotic agents playing adaptive language games, has been reported in [Steels and Vogt, 1997]. The present paper goes beyond this earlier work by showing the very beginnings of syntax.

The rest of the paper is in four sections. The next section (section 2) introduces the experimental setup used to validate the various proposed mechanisms and study their performance. Then the main hypotheses underlying our approach are briefly presented. Section 4 discusses the component responsible for producing sen-

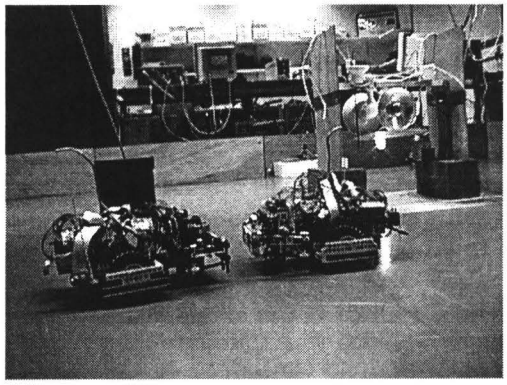

Figure 1: The source of visual experiences consists of a robotic ecosystem in which robots can survive by recharging and doing work. Two robots are shown with a black cylindrical box behind them and the charging station to the right.

sory data points from raw images, the component responsible for turning data points into feature structures, and the component responsible for coding and decoding feature structures 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 has two parts.

First there is a robotic ecosystem consisting of an arena in which one or more robots can survive by recharging and doing work to get enough energy in the charging station (see figure 1) [Steels, 1994]. The moving robots are autonomous Lego-vehicles (size: 30 x 20 x 15 cm) with various types of sensors (infrared, visible light, sound, touch) and two actuators, a left and right motor. The overall processing capacity resides in a Motorala MC86332 micro controller with 128 kB ROM and 256 kB RAM located on a Vesta board. The Vesta board is extended with a second board dedicated to low level sensory-motor processing and buffering. The details of the behavior of these robots and their implementation fall outside the scope of the present paper.

Second there are two 'robotic heads' located on the border of the ecosystem but relatively close to each other. Each head has a black and white camera and can rotate around its axis (figure 2). A head has the same hardware and software architecture as the moving robots, but is augmented with an additional computer that runs the higher level activities discussed in this paper. At present communication between the heads goes through a network, although in a later phase the language communication is planned to be through sound. Although we have only two physical heads, we simulate multiple agents by 'loading' the state of different agents into each head.

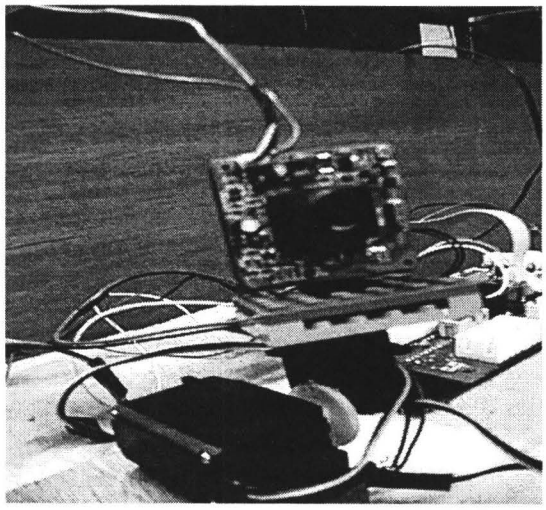

Figure 2: Language and meaning creation is performed by robotic heads which have a camera and rotate around their axis. The heads track moving objects and engage in language games expressing what happened most recently in the ecosystem.

The robotic heads track moving objects in real-time. Consecutive bitmaps are compared for differences, so that moving objects stand out against the background. One object is the focus of attention and the tracker tries to remain focused, even though the object may occasionally stand still or the difference matching may fail. In the experiments reported later, there is only one moving object, which is a single robot performing its normal activities (pushing boxes, obstacle avoidance, recharging). As heads turn while tracking this robot, other objects come occasionally into view: the charging station, obstacles, other robots, etc. These other (static) objects are distinguished against the background by standard low-level visual processing. The robotic heads are thus watching a dynamically evolving scene.

In addition, the heads engage in language games in which they describe to each other what they see. Observation starts after a conversation has terminated and

goes until the beginning of the next conversation. During a specific observational time period various objects (or more precisely image-elements) will have been in view. These image-elements 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 features characterising the topic are conceptualised by the speaker and encoded in language. They are then decoded by the hearer. A language game suceeds if the meaning decoded by the hearer fits with his observations and conceptualisations. Otherwise the game fails and various repair actions to be discussed later are undertaken by each agent. The observation time is initially short so that typically only two objects are involved (the topic and one element or even none in the remaining context), but it becomes progressively longer as the heads develop more concepts and more language, and so they can identify the topic partly through linguistic means.

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

- There must be low level sensory routines that extract sufficiently rich data streams from the raw images.
- 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.
- 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.
- If syntax has become necessary or useful, there must be a set of shared syntactic conventions.

The experimental (and theoretical) challenge is to show how all this may emerge *without* 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 ecosystem 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.

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. Only a small fraction of this potential has been realised in our experiments so far.

3 Major Hypotheses

Before embarking on a more detailed description of the various components and processes implemented so far, it is useful to state briefly the main hypotheses underlying our approach. Basically, there are five guiding principles.

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 [Labov, 1994], lexicons keep evolving to cope with new meanings, various grammaticalisation processes give rise to novel syntactic constructions and shifts in basic grammatical patterns [Traugott and Heine, 1991] All of these phenomena are heavily at work in the case of creole formation [Thomason and Kaufman, 1988] but happen even 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 to be discussed later in this paper [Tomasello, 1992]. It should be possible to develop and validate a very precise scenario for the gradual origin of ontological and linguistic complexity, similar to scenarios that have been proposed for the evolution of complexity in biology [Maynard-Smith and Szathmary, 1994]).

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 more complex language 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 induce from the material available new information about the language or about concepts held by the other agent.
- 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, to succeed better in future imitation games, an agent may slightly change its articulation of a certain vowel.

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 [Pinker 1994] and that language or meaning acquisition is a matter of instantiating and setting parameters determined by a basically innate language acquisition device [Chomsky 1975].

3. *Selectionism.* Although we do not assume genetic evolution to be the main driving force in language or ontological development, our approach is nevertheless selectionist: Structures are being created or adopted by an agent based on only local information and in imperfect ways. These structures are 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 in the present context, will be forgotten. Words that an agent invented to refer to certain features 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.

4. *Level Formation* The different games are not played in isolation but are coupled in two ways: The result of one game provides building blocks for the game at the next level. For example, distinctions produced by discrimination games are the basis for the features 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 [Nicolis and Prigogine, 1994]. 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 particular kinds of positive feedback loops are in place. More concretely, each agent keeps track for each structure at whatever level what the use and success has been. 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 a structure 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 will keep dynamically evolving.

It is of interest that these various principles have been identified in other efforts to explain complexity, particularly for the explanation of biological complexity, and it is therefore reasonable to assume that they are at work also for the origins of cognitive complexity and language.

4 Inventing and lexicalising distinctions

The general architecture of the system built so far is as in figure 3. This section examines briefly the components for sensory processing, meaning creation and lexicon formation. They have been described already in other papers [Steels, 1996a], [Steels, 1996b] which should be consulted for a more extensive and formal discussion of these mechanisms. The syntactic component is discussed in the next section.

4.1 Sensory Processing

The tracking and image processing algorithms identify coherent *image-elements.* For example, the robot moving around yields one continuous image-element, as long as it does not disappear out of side. Other objects yield other image-elements which come into view for a brief time period and disappear again. During the observation period, various image-elements are thus created and monitored as long as they stay in view. Note that there is no notion of object permanence yet.

For each image-element, low level sensory routines collect a variety of data: the size of the bounding box of the image-element, the average grey level, the orientation of the head with respect to the central point of the image-element, the maximum and minimum value, the

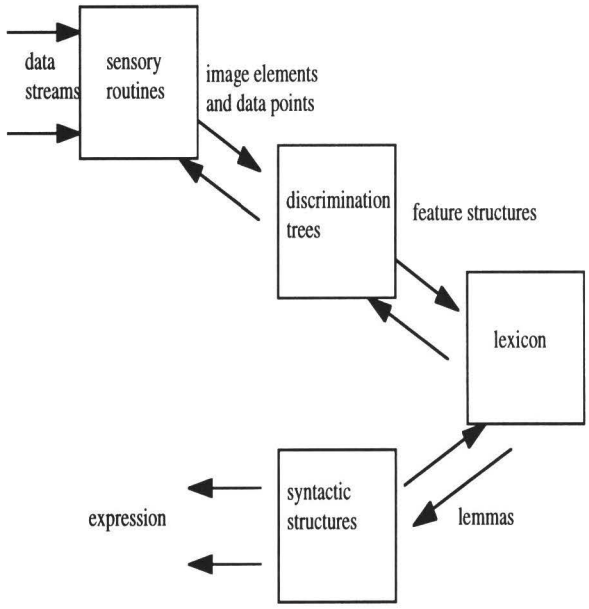

Figure 3: There are four major components in the system. One component provides input for the next one and conversely a higher level component supplies selectionist constraints for a lower one. When a component fails, adaptation takes place.

time the image-element was seen, the sharpness or visibility (i.e. how much the image-element is in focus), etc. (see figure 4). Some of these data will be relatively constant during the time the image-element is seen. Others will be changing. A direction of change might be constant during the whole period an image-element is seen, or it might also change. In that case, an *image-segment* is created for that *image-element*. For example, when the angle towards the head steadily increases (caused by the object moving to the left) and then steadily decreases (caused by the object moving to the right) two segments are created as part of the same image-element. An image-element may have different segmentations because the direction of change of each data source may change.

At the moment a conversation is about to start, all the image-elements that are still ongoing are closed, their global properties computed, and the total set of recent image-elements is passed on to the next component The various sensory routines that perform data collections are currently hard coded and fixed. However, they should also be a dynamically expanding set which is subjected to selectionist pressure from subsequent usage of the data. Concrete proposals in this direction and a possible neurological implementation has been introduced by [Edelman, 1987].

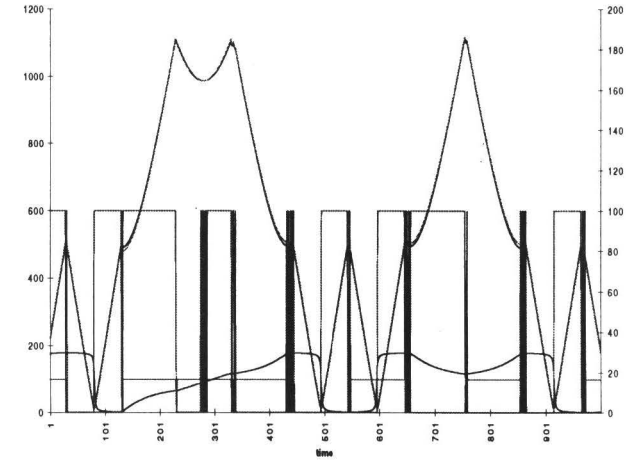

Figure 4: Examples of continuously collected data for various image-elements. These datastreams are segmented based on changes in the direction of change.

4.2 Forming Distinctions

The data-extraction component yields a set of image-elements I and data for each image-element or its segments. One image-element $t \in I$ is chosen by the speaker to be the topic of the language game. The others make up the context $C = I - \{t\}$. The next step is to categorise the data in terms of a feature structure consisting of a set of attribute-value pairs. The categorisation is achieved through (binary) discrimination trees which segment the continuous domain of each data point into finer and finer regions. Each data source corresponds to one attribute and each region to a value. For a given set of image elements a distinctive feature structure is found in three steps:

1. The existing discrimination trees are used to derive the first (and therefore most abstract) features for each data point associated with an image-element. Each image-element i has thus an associated feature set F_i.
2. Distinctive feature sets are computed. Let F_t be the feature set of the topic, then a distinctive feature 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 may be no distinctive feature sets $F_t = \emptyset$ but it is possible to refine existing features because the discrimination tree contains more refinements. In that case, new feature sets with these refinements are computed and step 2 above is reconsidered.
   (b) There may be no distinctive feature sets and the discrimination trees were exhaustively explored. In this case, a new distinction is cre-

ated 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 feature sets. When there is more than one possibility, the distinctive feature sets are ordered based on a number of selectionist criteria: A smaller set, a set with more successful features, and a set where the features are lexicalised, is prefered. A feature is more successful if it has been used more and had more success in usage. The best distinctive feature set according to these criteria is used in the remainder of the game.

Here are some examples of this process at work. First a discrimination game is shown in which there are not enough distinctive features and hence a new one is created. The new feature divides the range of possible slopes for the angle of the image element with respect to the head into two subregions, thus creating the values v-3 and v-4 for the attribute "slope of angle".

```
>> Speaker: a-4 Topic: i-6 Context: i-1
Failure: INSUFFICIENT-FEATURES-SPEAKER
New feature: ANGLE SLOPE [-1.0 1.0]: v-3 v-4
```

In a subsequent game this feature will be used to successfully differentiate $i-6$ from $i-1$:

```
>> Speaker: a-4 Topic: i-6 Context: i-1
Distinctive:
  ((#<FEATURE #x39515F6> v-4))
```

Here is a more complex example after about 100 discrimination games showing for various features the successive refinements with the topic values listed first. There are two possible distinctive feature sets

```
#<FEATURE #x37F2686> (IMAGE ANGLE CONSTANT)
(v-10) (v-10) nil (v-10)
#<FEATURE #x37F2716> (IMAGE DISTANCE CONSTANT)
(v-6) (v-6) NIL (v-6)
#<FEATURE #x37F27A6> (IMAGE GREYLEVEL CONSTANT)
(v-8 v-12 v-18) (v-8 v-12 v-17) (v-8 v-12 v-18)
     (v-8 v-11 nil)
Distinctive feature sets:
((#<FEATURE #x37F27A6> v-18)
     (#<FEATURE #x37F2686> v-10))
((#<FEATURE #x37F27A6> v-18)
     (#<FEATURE #x37F2716> v-6))
```

4.3 Lexicon formation

A lexicon consists of a set of word-meaning pairs, where the meaning consists of a feature set. One word may have many meanings and one meaning may be expressed by many words. Each agent has his own lexicon and an agent cannot directly inspect the lexicon of another one. Each agent maintains how often a word-meaning pair has been used and how successful it has been in its use. While encoding, a speaker will prefer word-meaning pairs that have been used more often and were more succesful in use.

A discrimination game results in a series of possible distinctive feature sets of which one is chosen by the speaker as the basis of a naming game. This feature set is encoded by the speaker and then decoded by the hearer. Several things can go wrong in this process and each failure results in appropriate actions:

1. *The speaker does not have a word for a certain feature set.* In this case, the speaker is allowed to construct a new word (formed by a random combination drawn from an alphabet) and associate that in his lexicon with the feature set. This happens with a low probability because a word may already exist in the population for this feature set.
2. *The hearer may lack a word used by the speaker.* In this case, the hearer can infer possible feature sets that might be meant by that word, based on the distinctive feature sets that he is expecting. In the simplest situation, there is only one feature necessary to distinguish the topic from the objects, so that the meaning is unequivocally known. It could also be that some words are known but not others. The meaning of the missing words must then be reconstructed from the remaining unknowns. Because there may be more than one distinctive feature set, it is inevitable that ambiguity creeps into the lexicon of the hearer. These ambiguities are weeded out by future use and success in use, which determine what word-meaning pairs will become most common.
3. *Some of the feature sets decoded by the hearer do not match with the expected distinctive feature sets.* This means that there are some word-meaning pairs which are not shared by some of the agents. For the successful word-meaning pairs, both success and use is incremented, whereas for the others only the use is incremented, so that their future use diminishes.
4. *The feature set decoded by the hearer does not match with any of the expected distinctive feature sets.* In that case, the hearer extends the lexicon, using the same procedure as for situation 2 above.

Note that this mechanism is again selectionist. Agents create or infer word-meaning pairs. Which pairs 'survive' depends on use and success in use, and this is determined by how many agents have adopted the same word-meaning pairs. Typically we see a phase transition when one word starts to dominate for the expression of a particular meaning. This phase transition is due to the

positive feedback loop inherent in the system. Software simulations reported in [Steels, 1996b] have shown that a group of agents indeed converges towards a common lexicon after a sufficient number of adaptive naming games. Moreover new agents may enter at any time, and due to the adaptive nature of the discrimination games, new features may enter the repertoire of possible meanings.

Here are some examples of this process within the context of the present experimental setup. The following is an example of a naming game that succeeded because the meaning hearer decoded by the hearer fits with i-1 being the topic in this context.

```
>> Speaker: a-4 Topic: i-1 Context: i-6
Distinctive feature sets:
((#<FEATURE #x37F2716> v-6))
((#<FEATURE #x37F2686> v-10))
((#<FEATURE #x37F2636> v-1))
Expression:
  (A F) (#<FEATURE #x37F2716> v-6)
     (IMAGE DISTANCE CONSTANT) [0.000 1.000]
<< Hearer: a-5
Features: ((#<FEATURE #x37F2716> v-6))
Accepted
```

The following is a naming game that did not succeed because the speaker had no word to express the feature that he wanted to express. A new word is created.

```
>> Speaker: a-4 Topic: i-2 Context: i-3 i-6
Distinctive feature sets:
  ((#<FEATURE #x39516D6> v-17))
Failure: MISSING-LEMMA-SPEAKER
New word: (F A)
```

Next an example where the hearer has no word and adopts the word used by the speaker. The example is more complex also because the speaker has used more than one word. The hearer already knows one of these and hypotheses a meaning for the other.

```
>> Speaker: a-4 Topic: i-3 Context: i-6 i-1 i-2
Distinctive feature sets:
  ((#<FEATURE #x39515B6> v-10)
        (#<FEATURE #x39516D6> v-18))
  ((#<FEATURE #x3951646> v-6)
        (#<FEATURE #x39516D6> v-18))
Expression:
  (E K) (#<FEATURE #x39516D6> v-18)
     (IMAGE GREYLEVEL CONSTANT) [0.750 1.000]
  (A F) (#<FEATURE #x3951646> v-6)
     (IMAGE DISTANCE CONSTANT) [0.000 1.000]
<< Hearer: a-5
Failure: MISSING-LEMMA-HEARER
Acquire word: (E K)
    ((#<FEATURE #x39516D6> v-18))
```

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 in order to engage in language games. After a few hundreds of games the lexicon of one agent is as follows:

```
(F A) ((#<FEATURE #x37FB82E> v-17))
    (IMAGE GREYLEVEL CONSTANT) [0.500 0.750]
(E K) ((#<FEATURE #x37FB82E> v-18))
    (IMAGE GREYLEVEL CONSTANT) [0.750 1.000]
(A F) ((#<FEATURE #x37FB79E> v-6))
    (IMAGE DISTANCE CONSTANT) [0.000 1.000]
(K I) ((#<FEATURE #x37FB74E> v-3))
    (IMAGE ANGLE SLOPE) [-1.000 0.000]
(K D) ((#<FEATURE #x37FB74E> v-4))
    (IMAGE ANGLE SLOPE) [0.000 1.000]
```

`(F A)` and `(E K)` capture color distinctions of objects. `(K I)` and `(K D)` mean moving to the left and moving to the right respectively. `(A F)` expresses that the image element is at a constant distance from the agent.

5 The emergence of syntax

Linguists in the Chomskian tradition view syntax as a formal device that has no functional or cognitive motivation. But 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 [Dik, 1980] [Langacker, 1986]. Our approach follows this second direction. This implies that in order to understand how syntax may emerge, we must understand why syntax is useful and necessary, i.e. what syntax is for. By syntax, we mean any kind of linguistic device 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 syntactic 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), case roles are expressed using word order, case markings or prepositions, 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.
- *To aid in managing the complexity of parsing and producing.* Very quickly combinatorial explosions will arise when multiple words which each form different groups are combined. Syntactic devices help by embodying conventions that help to establish what belongs to what. For example, the verb in

English affirmative sentences signals that the noun group identifying the subject has terminated.

We hypothesise that syntax relies on a general cognitive capability to recognise and re-instantiate frame-like structures or Piagetian schemata. A frame groups a set of elements which play particular roles. The frame gets its coherence based on constraints of each element and on the role between the elements. In the case of syntax, the elements are individual words or word groups and the constraints are the various syntactic devices mentioned earlier, such as word order constraints.

The emergence of grammar starts as soon as there are multiple word sentences (which we saw earlier to arise naturally from the naming games discussed in the previous section). The different words are grouped in a frame which is initially completely word/situation specific. When one of the words re-occurs again, it triggers the same frame so that expectations are set up about the other elements and their constraints. From this humble basis, the formation of grammatical complexity takes off through a variety of operations:

- Frames may be fused to give larger frames. For example, when w1 and w2 were grouped in f1 and w2 and w3 in f2, a new frame can be created by fusing f1 and f2.
- Semantic and syntactic classes form on what can fill a slot in an existing frame, i.e. the constraints on a frame may generalise. The generalisation could be based on semantic criteria (features of the objects or situations concerned) but could also be based on tagging words as belonging to certain syntactic classes.
- Hierarchical structure emerges because the elements in a frame become themselves frames. This hierarchical structure co-evolves with the emergence of the hierarchical structure implicit in more complex language games.
- Constraints in the form of additional syntactic devices are added to distinguish one frame from another one, to enable more rapid recognition, etc. The addition takes place by a variety of operators. One of them is overinterpretation. Observed properties of a frame, such as word order, are taken to be obligatory as opposed to a side effect of conceptual processing.

It should also be emphasised that a large part of the growth in complexity of language is a side effect of the language population dynamics as a whole. New syntactic categories form, lexical words become purely functional, lexemes shift in meaning (for example from spatial to temporal expression), etc. [Traugott and Heine, 1991].

Here is an example of these mechanisms at work in the context of the present experimental setup. A language game takes place in which the speaker uses two words. The syntactic component triggers and constructs a new group ($group - 11$).

```
>> Speaker: a-4 Topic: a-2 Context: a-3 a-6 a-3
Distinctive feature sets:
  ((#<FEATURE #x39516D6> v-17))
  ((#<FEATURE #x39515B6> v-10)
        (#<FEATURE #x39516D6> v-17))
  ((#<FEATURE #x3951646> v-6)
        (#<FEATURE #x39516D6> v-17))
  ((#<FEATURE #x3951566> v-15)
        (#<FEATURE #x39516D6> v-17))
Expression:
  (A F) (#<FEATURE #x3951646> v-6)
     (IMAGE DISTANCE CONSTANT) [0.000 1.000]
  (F A) (#<FEATURE #x39516D6> v-17)
     (IMAGE GREYLEVEL CONSTANT) [0.500 0.750]
Failure: MISSING-SYNTACTIC-FRAME-SPEAKER
New syntactic frame:
GROUP: #:|group-11|
STRUCTURE: ((#:|role-7| . #:|class-7|)
            (#:|role-8| . #:|class-8|))
CONSTRAINTS: ((ORDER (#:|role-7| #:|role-8|)))
```

The group includes two slots for elements (called `role-7` and `role-8`). There is an order constraint between the fillers and the dictionary entries for `(A F)` and `(F A)` is expanded to indicate that they belong to the class of possible fillers of these roles: `class-7` and `class-8`.

Later on the following game takes place:

```
>> Speaker: a-4 Topic: i-3 Context: i-2 i-6 i-1
Distinctive feature sets:
((#<FEATURE #x37F27A6> v-18)
     (#<FEATURE #x37F2686> v-10))
((#<FEATURE #x37F27A6> v-18)
     (#<FEATURE #x37F2716> v-6))
Expression:
  (E K) #<FEATURE #x37F27A6> v-18)
     (IMAGE GREYLEVEL CONSTANT) [0.750 1.000]
  (A F) (#<FEATURE #x3951646> v-6)
     (IMAGE DISTANCE CONSTANT) [0.000 1.000]
Failure: MISSING-SYNTACTIC-FRAME-SPEAKER
```

Although there is no frame that matches completely, `(A F)` fills `role-7` in `group-11` and `(E K)` can be integrated by assuming that it fills the same role:

```
Integration into #:|group-11|:
((#:|role-8| . #<LEMMA #x37F8946>)
 (#:|role-7| . #<LEMMA #x37F449E>))
```

When the syntactic frame is subsequently used, the ordering constraint of `group-11` would put `(A F)` before `(E K)` and not vice versa as would be based purely on the conceptualisation processes. In other words, the normal flow of producing words is interrupted and the syntactic

constraints associated with the group in which the different words fit are enacted. Conversely, during parsing the group structure sets up expectations and all syntactic constraints are tested. The dictionary entries are now as follows:

```
(F A) ((#<FEATURE #x37F27A6> v-17)) (class-8)
    (IMAGE GREYLEVEL CONSTANT) [0.500 0.750]
(E K) ((#<FEATURE #x37F27A6> v-18)) (class-8)
    (IMAGE GREYLEVEL CONSTANT) [0.750 1.000]
(A F) ((#<FEATURE #x37F2716> v-6)) (class-7)
    (IMAGE DISTANCE CONSTANT) [0.000 1.000]
```

Two 'syntactic' classes have been formed: `class-8` and `class-7`. `Class-7` contains so far only one word `(A F)` expressing that the topic is at a constant distance from the head (and therefore would be a static object). `Class-8` contains two words: `(F A)` and `(E K)` which qualify this description in terms of the color of both objects.

6 Conclusions

The paper proposed a general 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 dynamic scenes involving a robot moving around in its ecosystem. Some experimental results which partly validate 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. Another key problem which has not been addressed yet is how the games themselves may come into existence. 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 implementation and maintenance of the robotic ecosystem is a group effort at the VUB AI laboratory in which Tony Belpaeme, Andreas Birk, Luc Steels, Peter Stuer, Dany Vereertbrugghen and Paul Vogt have made major contributions. This 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 was implemented by Tony Belpaeme. The language and meaning formation programs were designed and implemented by Luc Steels. His research was conducted and financed by the Sony Computer Science Laboratory in Paris. The author is strongly indebted to Toshi Doi and Mario Tokoro for being able to work in this superb research environment.

8 References

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

[Chomsky, 1975] Noam Chomsky. *Reflections on Language.* Pantheon books, New York, 1975.

[De Boer, 1997] Bart De Boer. 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.

[Dik, 1980] Simon Dik. *Studies in Functional Grammar.* Academic Press, London, 1980.

[Edelman, 1987] G.M. Edelman. *Neural Darwinism: The Theory of Neuronal Group Selection.* New York: Basic Books, 1987.

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

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

[Hutchins and Hazelhurst, 1995] E. Hutchins and B. Hazelhurst. 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, 1995.

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

[Labov, 1994] William Labov. *Principles of Linguistic Change. Volume 1: Internal Factors.* Blackwell, Oxford, 1994.

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

[MacLennan, 1991] Bruce MacLennan. 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., 1991.

[Maynard-Smith and Szathmary, 1994] John Maynard-Smith and Eors Szathmary. *The major transitions in evolution.* Freeman Spektrum, Oxford, 1994.

[Nicolis and Prigogine, 1993] Gregoire Nicolis and Ilya Prigogine. *Exploring Complexity.* Piper, Berlin, 1993.

[Pinker, 1994] Steven Pinker. *The language instinct.* Penguin Books, London, 1994.

[Steels, 1994] Luc Steels. 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.

[Steels, 1996a] Luc Steels. 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.

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

[Steels, 1997a] Luc Steels. 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.

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

[Steels and 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.

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

[Tomasello, 1992] Michael Tomasello. *First verbs. A case study of early grammatical development* Cambridge University Press, Cambridge, 1992.

[Traugott and Heine, 1991] Elizabeth Traugott and Bernd Heine. *Approaches to Grammaticalization. Volume I and II.* John Benjamins Publishing Company, Amsterdam, 1991.