

Grounding Symbols through Evolutionary Language Games

Luc Steels

VUB Artificial Intelligence Lab - Brussels

Sony Computer Science Lab - Paris

steels@arti.vub.ac.be

Abstract

There is a broad consensus today that the representations used by a cognitive agent must be grounded in external reality through a sensori-motor apparatus. These representations must also be sufficiently similar to those used by other agents in the group to enable coordinated action and communication, and they must be acquired autonomously by each agent. This paper argues that language plays a crucial role in the learning of grounded representations because it is a source of feedback and constrains the degrees of freedom of the representations used in the group. Evolutionary language games are introduced as framework for concretising the structural coupling between concept formation and symbol acquisition and some example experiments with physical robots are discussed.

1 Defining the problem

Given the terminological confusion in the cognitive sciences it is worthwhile to define more precisely the problems we try to approach.

1. In computer science, a *representation* is a physical state of a machine (computer memory for example) which acts as a "stand in" for something else. The physical state becomes thus a means to store information and physical processes operating over the state can implement whatever representational transformations we wish to enact. Thus a representation of a

number in a computer is a configuration of digital states. Calculation takes place by changing these states. Objects, concepts, actions, etc. can likewise be represented in an artificial agent by postulating internal states for each of these. Cognitive processes like decision-making, language parsing, object recognition, etc. can then be conceived as physical operations over these internal states.

A representation (both the physical medium chosen and the convention used to map information onto this medium) is arbitrary with respect to what one wants to represent. The only requirement is that the mapping is systematic and that processes operating over a representation are consistent with respect to the mapping that has been adopted. Thus we can not only use binary representations of numbers but also hexadecimal representations, and make use of marks on a laser disk as well as electromagnetic states of an electronic circuit as the state medium. Note that once we are at the level of physical states and physical processes, no additional homunculus is involved to "interpret" representations.

In neuroscience parlance, the equivalent of the computer science notion of a representation is the notion of a neural correlate. This is a biological state (for example the activation of a neuron or set of neurons) which stands for something else, like a control signal for an arm, the recognition of a concept, experience of the color red, etc. Neural processes operating over these physical states, usually thought to take the form of the selective propagation of signals through a network, are the neural correspondence of the physical operations carried out over computational states.

So even if computational implementations are very different from biological implementations, the notion of representation is similar in AI and (cognitive) neuroscience. Using representations and operations over representations to explain cognitive functions seem now so natural and obvious that it is difficult to follow philosophers who claim that cognition does not involve representations. Perhaps they simply have not understood what representations are or are using the notion of representation in another way.

2. We say that a representation is *grounded* when there is an autonomous process that transforms sensations (i.e. data flowing from sensors or motors into internal states) into internal representations and transforms internal representations into motor activations. Through these grounding processes, the agent can coordinate his activities with the world and other agents. The representation need not be an exact, full, veridical representation of the world,

and it can be analog or categorical, but it needs to be sufficiently detailed and faithful to support the agent's interaction with the world and others.

Grounding is trivially achieved for devices like a calculator. The user pushes buttons which directly activate internal representations. It is obviously much more complex for representations about the world. Sensors reflect physical properties of the environment which are not necessarily those that the agent needs to focus on. The information is hidden in the sensory-motor data and requires complex processing to get out. Often there is not enough information in the sensory data and so representations have to be hypothesised in a top down fashion and mapped onto the sensory-motor data.

In a lot of (pre 1990) AI work the problem of grounding was (temporarily) abstracted out by supplying the representations directly to the computer and only focusing on the processing aspect. This was a useful strategy for a while but it has been rightfully criticised because not all the representations assumed by early AI programs can be grounded on a physically embodied robot [5], [34]. For example, it is far from obvious that abstract geometric representations about the world, as envisioned by David Marr [22], can be extracted from real world images given the available resources. This has led to a healthy move away to simpler representations and better exploitation of bodily interaction [26]. We should nevertheless keep in mind that many experiments which involve complex representations - even from the very beginning of AI - have considered the problem of grounding these representations. A typical example is the SRI Shakey robot [24] which was a model of many subsequent robotics efforts. So there is nothing in the notion of representation that makes them inherently not groundable. It is only that the grounding of representations is a very non-trivial and difficult technical problem involving a whole arsenal of statistical and pattern recognition techniques and that not every abstract representation can be grounded.

3. Representations are *symbolised* when there are external tokens (speech sounds, gestures, scratches on a piece of paper, configurations on a display) that are associated with the representation and used for external communication with another agent. The relationship is entirely conventional. Sender and receiver must agree, but there are in principle endless possibilities. The process of relating a representation to its symbolisation and vice-versa must be carried out autonomously by each agent. We say that a symbol is grounded iff its representation is grounded.

The relations considered so far are summarised in the semiotic square

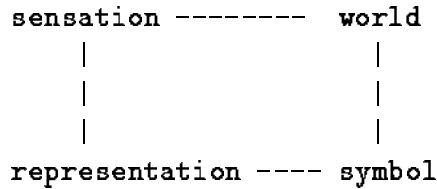


Figure 1: The semiotic square summarises the relation between world, sensation, representation, and symbol.

depicted in figure 1. By sensation, I mean the perceptual or motor data streams that directly connect the agent to the world. By representation I mean a conceptual representation useful for decision making, language or other cognitive tasks. The semiotic square is reminiscent of the semiotic triangle familiar from the semiotic literature [11] which relates world, concept, and symbol. The relation between a symbol and the world is the *reference relation*. The relation between a symbol and a representation is the *meaning relation*. In the philosophy of language literature, the reference relation and the meaning relation are studied theoretically, independently of how this relation is established by a cognitive agent. This kind of research in formal semantics is of interest when one wants to investigate how a symbol system can in principle be related to the world, but is a very different topic from the one considered in this paper.

The problem of symbolisation is trivially solved by a calculator which transforms the internal representation of a number into an external representation on a display and which displays on the buttons the conventional representations of numbers so that users know which button to push. It is extraordinarily difficult in the case of natural language, because the conventions are not universal, they are open-ended, and involve a non-trivial multi-level mapping from representations to symbols.

Just for clarification, it is perhaps important to sharpen in this context the notion of symbolic processing, as it has been used in AI. Symbolic processing means that a set of symbols, possibly without any relation to the world, such as postulated in a logical calculus, is mapped onto internal representations and the internal representations are processed conform to given symbol manipulation rules, for example the rules of natural deduction of the

predicate calculus. The outcome of processing is then translated back into symbols. A logic programming systems such as PROLOG or a functional programming language like LISP support this kind of operation. They can handle millions of symbols and their compilers optimise for fast symbolic processing. This technology is of enormous value for building non-trivial cognitive agents but does not address in itself grounding nor learning.

4. *Learning grounded representations* means that the agent allocates states for certain representations but also - and more importantly - that the agent learns to use the representation appropriately, more specifically (1) the ability to relate the representation to the world through the sensorimotor apparatus, and (2) the use of the representation for some purpose, such as making a decision about the next action to take.

Learning a symbolisation means to acquire the relation between representations and symbols as required for communication in a specific community. The agent must acquire the ability to activate the intended internal representations given a set of symbols or to select a set of symbols to externalise a particular representation.

Learning a semiotic system means to acquire both grounded representations and their symbolisations, i.e. the relation between world, sensation, representation, and symbol. This is the problem that the child faces when growing up in a language community and the problem that concerns us further in this paper. We take this problem to be equivalent to the *symbol grounding problem*.

2 Two approaches

The first approach to the symbol grounding problem is to follow a divide-and-conquer strategy. It assumes that there is on the one hand a process that learns grounded representations. Once the representations are in place, it then postulates a second independent process that associates symbols with the already acquired representations. It has been suggested that this is the way symbol grounding happens in humans [14] and various experiments have been done following this approach [10] [7].

However a second approach is possible, as first suggested in Steels [30], [32], in which there is a strong *structural coupling* between the two. This means that learning representations and learning their symbolisation go hand

in hand and influence each other. A representation that has been learned can be the subject of symbolisation but communication through symbols provides important feedback to representational learning.

In my opinion the structural coupling approach is the only viable way to explain the massive build up of representations and symbols that humans use and it can be profitably used in artificial systems. This approach seems paradoxical at first because instead of solving two difficult problems one by one, we try to solve both of them at the same time, which intuitively seems to be even more difficult. Here are the reasons why I nevertheless believe that a structural coupling approach is better.

There are many ways to learn grounded representations, but broadly speaking mechanisms fall into two classes: unsupervised or supervised learning. In the case of unsupervised learning, clustering techniques (possibly implemented as neural networks such as the Kohonen network) extract from a series of data invariances that are then associated with internal representations. However we note two things: (1) not all representations of interest to a cognitive agent are reflected as invariants in sensori-motor data, and (2) often there is more than one possible way to cluster the data depending on the dimensions that are considered or the parameter settings of the clustering algorithm.

This generates the problem that if different agents each independently develop representations about the world, there is no guarantee that they arrive at mutually compatible representations. Today the experimenter carefully designs the features that are input to the learning system, carefully selects appropriate example sets, and then tweaks parameters until an appropriate clustering comes out. This is not quite the autonomous learning that we would hope for. but using sensory data in their raw form, i.e. bitmaps captured by a camera, motor states, the audio signal directly coming from a microphone, etc. does not leave any other choice.

In the case of supervised learning, the agent is given a series of cases as well as feedback whether the representations being developed are appropriate with respect to some task. Thus if the task is classification, the agent would be given examples and counterexamples, if the task is action in the world, the agent gets a feedback signal whether the action was successful (as in reinforcement learning algorithms). Because the task can incorporate some form of coordination with other agents, it is in principle possible to steer the acquisition of representations in such a way that they are compatible

with those used by others, by incorporating in the feedback some element that is related to representation sharing. But the critical question here is: Where does the feedback come from? In real world circumstances, feedback is never direct and obvious, specifically not concerning internal representations. Feedback comes only through the *use* of a representation. If the designer has to carefully determine feedback and prepare the example sets, then we are missing something fundamental. By letting the use of a representation generate feedback, the learning setup becomes self-supervised.

There are many users of representations. For example for planning actions, particularly at a microlevel (like for grasping an object), the agent needs adequate categorisations of reality dedicated to that task. So action execution can be a possible source of feedback. Language is another big user of representations because before anything can be said the world must be conceptualised in the way that has been lexicalised and grammaticalised in the language (and this can differ substantially from one language to another [27]). But language is not only useful because it provides representational feedback, it also helps a community of agents to settle on similar representations.

The next thing we need is a good framework to study these issues concretely. I claim that evolutionary language games are such framework. We started to work on this around 1995 [?] and have since shown in an increasing number of papers and experiments that the framework is a rich foundation for studying both language formation and concept acquisition. The remainder of this paper reports these developments in some more detail.

3 Evolutionary Language Games

Evolutionary games are now widely used to study issues in evolution [?] or economics []. They form part of the larger framework of game theory. A game is an interaction between two agents according to certain rules and having a certain outcome. Games can be adversary (like the Prisoner's Dilemma [?]) or cooperative. A game is evolutionary if the players change their internal states in order to be more successful in the future [?].

In the past few years, we have done several experiments exploring the co-evolution of language and meaning. Although we have also studied multi-word expressions and the emergence of syntax within the same experimental

context [?], this paper only discusses single word utterances so that we can focus completely on the issue of grounding meaning.

4 The Talking Heads experiment

The main experiment that will be briefly described here is known as the Talking Heads experiment. The robotic setup consists of a set of ‘Talking Heads’ connected through the Internet. Each Talking Head features a Sony EVI-D31 camera with controllable pan/tilt motors for horizontal and vertical movement (figure 2), a computer for cognitive processing (perception, categorisation, lexicon lookup, etc.), a screen on which the internal states of the agent currently loaded in the body are shown, a TV-monitor showing the scene as seen through the camera, and devices for audio in- and output. Agents can load themselves in a physical Talking Head and teleport themselves to another Head by travelling through the Internet. By design, an agent can only interact with another one when it is physically instantiated in a body located in a shared physical environment. The experimental infrastructure also features a commentator which reports and comments on dialogs, displays measures of the ontologies and languages of the agents and game statistics, such as average communicative success, lexical coherence, average ontology and lexicon size, etc.

For the experiments reported in this paper, the shared environment consists of a magnetic white board on which various shapes are pasted: colored triangles, circles, rectangles, etc. Although this may seem a strong restriction, we have learned that the environment should be simple enough to be able to follow and experimentally investigate the complex dynamics taken place in the agent population.

4.1 Components

The guessing game

The interaction between the agents consists of a language game, called the guessing game. The guessing game is played between two visually grounded agents. One agent plays the role of *speaker* and the other one then plays the role of *hearer*. Agents take turns playing games so all of them develop the



Figure 2: Two Talking Head cameras and associated monitors showing what each camera perceives.

capacity to be speaker or hearer. Agents are capable of segmenting the image perceived through the camera into objects and of collecting various sensory data about each object, such as the color (decomposed in RGB channels), average gray-scale or position. The set of objects and their data constitute a *context*. The speaker chooses one object from this context, further called the *topic*. The other objects form the *background*. The speaker then gives a linguistic hint to the hearer.

The linguistic hint is an utterance that identifies the topic with respect to the objects in the background. For example, if the context contains [1] a red square, [2] a blue triangle, and [3] a green circle, then the speaker may say something like "the red one" to communicate that [1] is the topic. If the context contains also a red triangle, he has to be more precise and say something like "the red square". Of course, the Talking Heads do not say "the red square" but use their own language and concepts which are never going to be the same as those used in English. For example, they may say "malewina" to mean [UPPER EXTREME-LEFT LOW-REDNESS].

Based on the linguistic hint, the hearer tries to guess what topic the speaker has chosen, and he communicates his choice to the speaker by pointing to the object. A robot points by transmitting in which direction he is looking in his own agent-centered coordinates. The other robot is calibrated

in the beginning of the experiment to be able to convert these coordinates into his own agent-centered coordinates. The game succeeds if the topic guessed by the hearer is equal to the topic chosen by the speaker. The game fails if the guess was wrong or if the speaker or the hearer failed at some earlier point in the game. In case of a failure, the speaker gives an extra-linguistic hint by pointing to the topic he had in mind, and both agents try to repair their internal structures to be more successful in future games.

The architecture of the agents has two components: a conceptualisation module responsible for categorising reality or for applying categories to find back the referent in the perceptual image, and a verbalisation module responsible for verbalising a conceptualisation or for interpreting a form to reconstruct its meaning. Agents start with no prior designer-supplied ontology nor lexicon. A shared ontology and lexicon must emerge from scratch in a self-organised process. The agents therefore not only play the game but also expand or adapt their ontology or lexicon to be more successful in future games.

The Conceptualisation Module

Meanings are categories that distinguish the topic from the other objects in the context. The categories are organised in discrimination trees (figure 3) where each node contains a discriminator able to filter the set of objects into a subset that satisfies a category and another one that satisfies its opposition. For example, there might be a discriminator based on the horizontal position (HPOS) of the center of an object (scaled between 0.0 and 1.0) sorting the objects in the context in a bin for the category ‘left’ when $HPOS < 0.5$, (further labeled as [HPOS-0.0,0.5]) and one for ‘right’ when $HPOS > 0.5$ (labeled as [HPOS-0.5,1.0]). Further subcategories are created by restricting the region of each category. For example, the category ‘very left’ (or [HPOS-0.0,0.25]) applies when an object’s HPOS value is in the region [0.0,0.25]. For the experiments in this paper, the agents have only channels for horizontal position (HPOS), vertical position (VPOS), color (RGB indicated as RED, GREEN, BLUE), and grayscale (GRAY). The system is open to exploit any channel with additional raw data, such as audio, or results from more complex image processing.

A distinctive category set is found by filtering the objects in the context

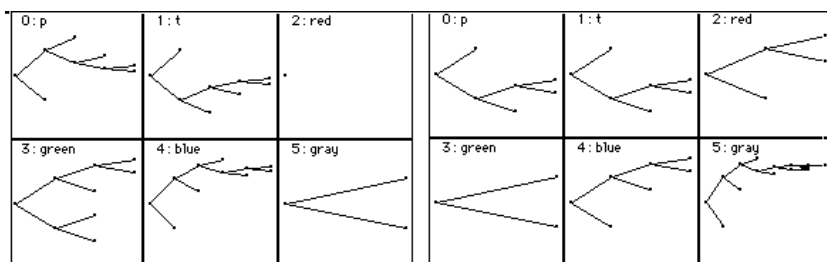


Figure 3: The discrimination trees of two agents.

from the top in each discrimination tree until there is a bin which only contains the topic. This means that only the topic falls within the category associated with that bin, and so this category uniquely filters out the topic from all the other objects in the scene. Often more than one solution is possible, but all solutions are passed on to the lexicon module.

The discrimination trees of each agent are formed using a growth and pruning dynamics coupled to the environment, which creates an ecology of distinctions. Discrimination trees grow randomly by the addition of new categorisers splitting the region of existing categories. Categorisers compete in each guessing game. The use and success of a categoriser is monitored and categorisers that are irrelevant for the environments encountered by the agent are pruned. More details about the discrimination game can be found in [?].

Verbalisation module

The lexicon of each agents consists of a two-way association between forms (which are individual words) and meanings (which are single categories). Each association has a score. Words are random combinations of syllables, although any set of distinct word symbols could be used. When a speaker needs to verbalise a category, he looks up all possible words associated with that category, orders them and picks the one with the best score for transmission to the hearer. When a hearer needs to interpret a word, he looks up all possible meanings, tests which meanings are applicable in the present context, i.e. which ones yield a possible single referent, and uses the remaining meaning with the highest score as the winner. The topic guessed by the

hearer is the referent of this meaning.

Based on feedback on the outcome of the guessing game, the speaker and the hearer update the scores. When the game has succeeded, they increase the score of the winning association and decrease the competitors, thus implementing lateral inhibition. When the game has failed, they each decrease the score of the association they used. Occasionally new associations are stored. A speaker creates a new word when he does not have a word yet for a meaning he wants to express. A hearer may encounter a new word he has never heard before and then store a new association between this word and the best guess of the possible meaning. This guess is based on first guessing the topic using the extra-linguistic hint provided by the speaker, and on performing categorisation using his own discrimination trees as developed thus far. These lexicon bootstrapping mechanisms have been explained and validated extensively in earlier papers [35] and are basically the same as those reported by Oliphant [25].

The conceptualisation module proposes several solutions to the verbalisation module which prefers those that have already been lexicalised. Agents monitor success of categories in the total game and use this to target growth and pruning. The language therefore strongly influences the ontologies agents retain. The two modules are structurally coupled and thus get coordinated without a central coordinator.

Examples

Here is the simplest possible case of a language game. The speaker, **a1**, has picked a triangular object at the bottom of the scene as the topic. There is only one other rectangular object in the scene, nearer to the top. Consequently, the category $[VPOS-0.0,0.5]_{a1}$, which is valid when the vertical position $VPOS < 0.5$, is applicable because it is valid for the triangle but not for the rectangle. Assuming that **a1** has an association in his lexicon relating $[VPOS-0.0,0.5]_{a1}$ with the word "lu", then **a1** will retrieve this association and transmits the word "lu" to the hearer, which is agent **a2**.

Now suppose that **a2** has stored in his lexicon an association between "lu" and $[RED-0.0,0.5]_{a2}$. He therefore hypothesises that $[RED-0.0,0.5]_{a2}$ must be the meaning of "lu". When he applies this category to the present scene, in other words when he filters out the objects whose value for the

redness channel (RED) do not fall in the region [0.0,0.5], he obtains only one remaining object, the triangle. Hence **a2** concludes that this must be the topic and points to it. The speaker recognises that the hearer has pointed to the right object and so the game succeeds.

The complete dialog is reported by the commentator as follows:

Game 125.

a1 is the speaker. a2 is the hearer.
a1 segments the context into 2 objects
a1 categorises the topic as [VPOS-0.0,0.5]
a1 says: "lu"
a2 interprets "lu" as [RED-0.0,0.5]
a2 points to the topic
a1 says: "OK"

This game illustrates a situation where the speaker and the hearer picks out the same referent even though they use a different meaning. The speaker uses vertical position and the hearer the degree of redness in RGB space.

Here is a second example, The speaker is again **a1** and he uses the same category and the same word "lu". But the hearer, **a3**, interprets "lu" in terms of horizontal position [HPOS-0.0,0.5]_{a3} (left of the scene). Because there is more than one object satisfying this category in the scene the agents look at, the hearer is confused. The speaker then points to the topic and the hearer acquires a new association between "lu" and [VPOS-0.0,0.5]_{a3}, which starts to compete with the one he already had. The commentator reports this kind of interaction as follows:

Game 137.

a1 is the speaker. a3 is the hearer.
a1 segments the context into 2 objects
a1 categorises the topic as [VPOS-0.0,0.5]
a1 says: "lu"
a3 interprets "lu" as [HPOS-0.0,0.5]
There is more than one such object
a3 says: "lu?"
a1 points to the topic
a3 categorises the topic as [VPOS-0.0,0.5]
a3 stores "lu" as [VPOS-0.0,0.5]

The table below shows part of a vocabulary of a single agent after 3,000 language games. The table shows also the score.

| <i>Form</i> | <i>Meaning</i> | <i>Score</i> | <i>Form</i> | <i>Meaning</i> | <i>Score</i> |
|-------------|-----------------|--------------|-------------|-----------------|--------------|
| wovota | [RED-0.0,0.125] | 1.0 | sogavo | [GREEN-0.5,1.0] | 0.0 |
| tu | [GRAY-0.25,0.5] | 0.0 | naxesi | [GREEN-0.5,1.0] | 0.0 |
| gorepe | [VPOS-0.0,0.5] | 0.3 | ko | [GREEN-0.5,1.0] | 0.0 |
| zuga | [VPOS-0.0,0.5] | 0.1 | ve | [GREEN-0.5,1.0] | 0.0 |
| lora | [VPOS-0.25,0.5] | 0.1 | migine | [GREEN-0.5,1.0] | 0.0 |
| wovota | [VPOS-0.25,0.5] | 0.2 | zota | [GREEN-0.5,1.0] | 0.9 |
| di | [VPOS-0.25,0.5] | 0.0 | zafe | [GREEN-0.5,1.0] | 0.1 |
| zafe | [VPOS-0.0,0.25] | 0.2 | zulebo | [HPOS-0.0,1.0] | 0.0 |
| wowore | [VPOS-0.0,0.25] | 0.9 | xí | [HPOS-0.0,1.0] | 0.0 |
| mifo | [HPOS-0.0,1.0] | 1.0 | | | |

We see in this table that for some meanings (such as [RED-0.0,0.125]) a single form "wovota" has firmly established itself. For other meanings, like [GRAY-0.25,0.5], a word was known at some point but is now no longer in use. For other meanings, like [VPOS-0.0,0.5], two words are still competing: "gorepe" and "zuga". There are words, like "zafe", which have two possible meanings [VPOS-0.0,0.25] and [GREEN-0.5,1.0].

5 Discussion

Natural languages are clearly not totally coherent even in the same language community, and languages developed autonomously by physically embodied agents will not be fully coherent either.

1. Different agents may prefer a different word for the same meaning. These words are said to be *synonyms* of each other. An example is "pavement" versus "sidewalk". The situation arises because an agent may construct a new word not knowing that one is already in existence. Synonymy is often an intermediate stage for new meanings whose lexicalisation has not stabilised yet. Natural languages show a clear tendency for the elimination of synonyms. Accidental synonyms tend to specialise, incorporating different shades of meaning from the context or reflecting socio-linguistic and dialectal differences of speaker and hearer.

2. The same word may have different preferred meanings in the population. These words thus become *ambiguous*. This situation may arise completely accidentally, as in the case of "bank" which can mean river bank and financial institution. These words are then called *homonyms*. The situation may also arise whenever there is more than one possible meaning compatible with the same situation. An agent on hearing an unknown word may therefore incorrectly guess its meaning. Ambiguity also arises because most words are *polysemous*: The original source meaning has become extended by metaphor and metonymy to cover a family of meanings [38]. Real ambiguity tends to survive in natural languages only when the contexts of each meaning is sufficiently different, otherwise the hearer would be unable to derive the correct meaning.
3. The same meaning may denote different referents for different agents *in the same context*. This is the case when the application of a category is strongly situated, for example 'left' for the speaker may be 'right' for the hearer. Deictic terms like "this" and "that" are even clearer examples from natural language. In natural languages, this *multi-referentiality* is counter-acted by verbalising more information about the context or by avoiding words with multi-referential meanings when they may cause confusion.
4. It is possible and very common with a richer categorial repertoire, that a particular referent in a particular context can be conceptualised in more than one way. For example, an object may be to the left of all the others, *and* much higher positioned than all the others. In the same situation different agents may therefore use different meanings. Agents only get feedback about whether they guessed the object the speaker had in mind, not whether they used the same meaning as the speaker. This *indeterminacy* of categories is a cause ambiguity. A speaker may mean 'left' by "bovubo", but a hearer may have inferred that "bovubo" meant 'upper'.

So, although circumstances cause agents to introduce incoherence in the language system, there are at the same time opposing tendencies, attempting to restore coherence. Synonyms tend to disappear and ambiguity is avoided. In the remainder of this paper, we want to show that the dynamics of the guessing game, particularly when it is played by situated embodied robotic

agents, leads unavoidably to incoherence, but that there are tendencies towards coherence as well. Both tendencies are emergent properties of the dynamics. There is no central controlling agency that weeds out synonyms or eliminates ambiguity, rather they get pushed out as a side effect of the collective dynamics of the game. Let us examine some of the dynamics that came out of the experiments.

First it can be shown that agents indeed construct and acquire conventions in a group given the behaviors outlined above (see [31]). Moreover if agents take turn being speaker and hearer and a speaker is allowed to invent a new symbol occasionally when he does not have an association yet to symbolise a particular representation, a set of conventions can establish itself from scratch in the population (figure 5). This is due to a self-organising positive feedback loop. Associations that are successful become even more so because their score goes up, so that they become used even more frequently and hence propagate in the rest of the population. The dynamics is similar to that of ant societies self-organising a path or to increasing returns as studied in non-equilibrium economics.

The tight integration between meaning and language is illustrated by an experiment with the guessing game where a categorial repertoire for color was evolved first with and then without coupling to language [3]. The results are shown in figure ??.

6 Conclusions

This paper advocates a tight structural coupling between processes for learning grounded representations and learning symbolisations of them. Both constrain each others' degrees of freedom and enable the learner to get feedback about the adequacy of a representation. Our experiments in simulation and on real robots have sufficiently demonstrated that applications can be constructed in a straightforward way using these principles. At the moment we are specifically targeting research on language games for humanoid robots.

This work raises many additional interesting issues. For example, there has been a longstanding debate between nativists who claim that language learning amounts to learning labels for existing categories and relativists such as Whorf who claim that each language implies a different categorisation of reality. The structural coupling of concept formation and language acqui-

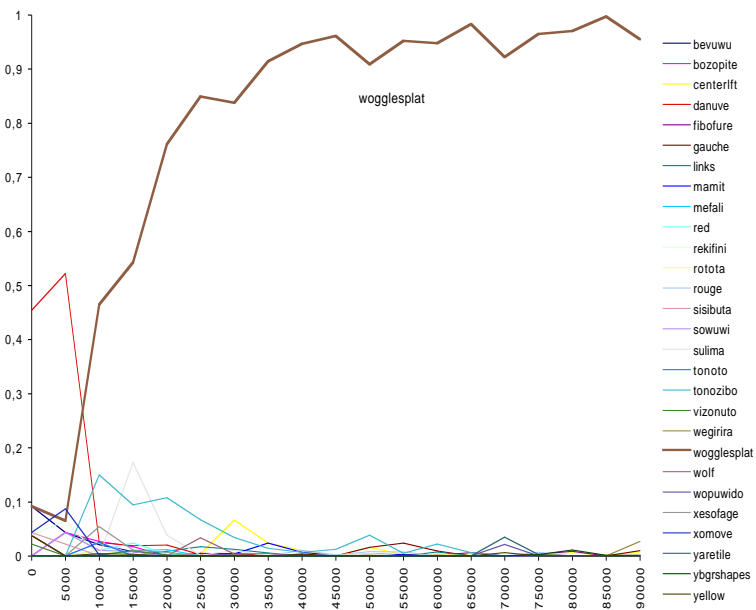


Figure 4: A meaning-form diagram which graphs for a specific meaning all the possible forms and their score. A winner-take-all situation is clearly observed. X-axis shows language games and y-axis the score of forms. There is a steadily growing population reaching 1500 agents towards the end

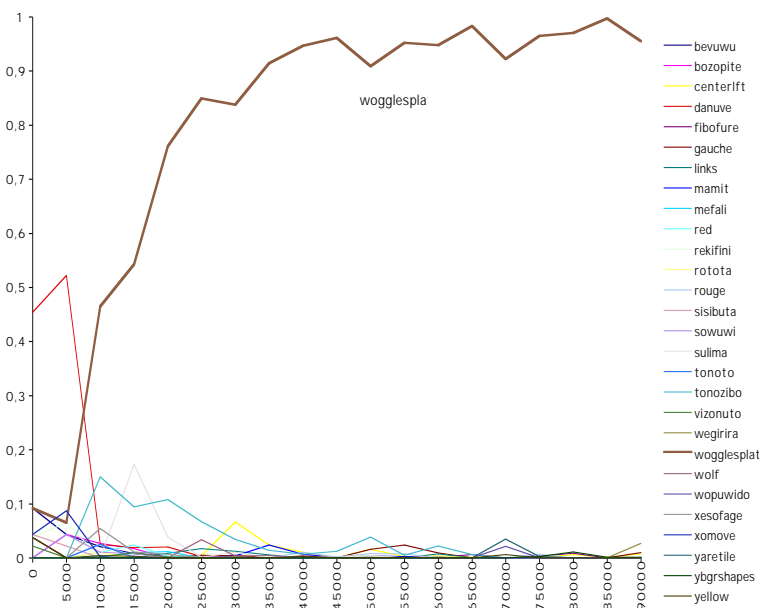


Figure 5:

sition advocated in this paper explains how a relativistic view is not only possible but unavoidable. If language enables and influences the learning of representations then it is easy to see how representations can become language specific. Of course representations are still strongly constrained by the world and tasks carried out in the world as well - they are not completely conventional or arbitrary. But they need not be innate to explain how they can become shared.

Acknowledgement

This paper was triggered by discussions with Steve Harnad as part of the committee concerning the Ph.D. thesis by Paul Vogt and by discussions with Erik Myin on the Wittgensteinian view of language. It is based on a large amount of technical work conducted in collaboration with Edwin De Jong, Frederic Kaplan, Angus McIntyre, Joris Van Looveren, Jelle Zuidema and Paul Vogt.

References

- [1] Batali, J. (1998) Computational Simulations of the Emergence of Grammar. In: Hurford, J. et.al. (1998).
- [2] Bekey, G. and A. Knoll (2000) Proceedings of the First IEEE Workshop on Humanoids. Cambridge Ma.
- [3] Belpaeme, T. (2001) Simulating the Formation of Color Categories. Submitted to ECAL-2001.
- [4] Bishop, C.M. (1995) Neural Networks for Pattern Recognition. Oxford Univ Press, Oxford.
- [5] Brooks, R. (1999) Cambrian Intelligence : The Early History of the New AI. The MIT Press, Cambridge Ma.
- [6] Byrne, A. and D.R. Hilbert (1997) Readings on Color. (Volume I: The Philosophy of Color. Volume 2: The Science of Color) The MIT Press, Cambridge Ma.
- [7] Cangelosi A., Greco A. and Harnad S. (2000). From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories. *Connection Science*, 12(2), 143-162
- [8] Clancey, W. (1997) Situated Cognition : On Human Knowledge and Computer Representations (Learning in Doing - Social, Cognitive and Computational Perspectives). Cambridge Univ. Press, Cambridge UK
- [9] Daelemans W. (1999) Memory-based language processing: introduction. - In: *Journal of experimental and theoretical artificial intelligence*, 11:3(1999), p. 287-292
- [10] De Jong, E.D. (1999). Analyzing the Evolution of Communication from a Dynamical Systems Perspective. In *Proceedings of the European Conference on Artificial Life ECAL'99*, 689-693. Springer-Verlag LNCS, Berlin.
- [11] Eco, Umberto (1968) *La struttura assente*. Milano 1968.
- [12] Edelman, S. (1999) Representation and recognition in Vision. The MIT Press, Cambridge.

- [13] Hardcastle, W. and N. Hewlett (1999) *Coarticulation. Theory, Data and Techniques*. Cambridge University Press, Cambridge.
- [14] Harnad, S. (1990) The symbol grounding problem. *Physica D* 42: 335-346.
- [15] Horswill, I. (1993) Polly: A Vision-Based Artificial Agent. In: *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI-93)*, Washington DC, MIT Press.
- [16] Hurford, J, C. Knight and M. Studdert-Kennedy (eds.) (1999) *Approaches to the Evolution of Human Language*. Cambridge Univ. Press. Cambridge.
- [17] Kirby, S. (1999) *Function, Selection and Innateness: the Emergence of Language Universals*. Oxford University Press, Oxford.
- [18] Ladefoged, P. (2000) *Vowels and Consonants: The sounds of the world's languages*. Blackwell Pub. Oxford.
- [19] Labov, W. (1994) *Principles of Linguistic Change: Internal Factors*. Blackwell Pub. Oxford.
- [20] Lindblom, B., P. MacNeilage, and M. Studdert-Kennedy (1984) Self-organizing processes and the explanation of phonological universals. *Linguistics*, 21 (1).
- [21] Marco, R. P. Sebastiani, and P. R. Cohen (2000) Bayesian Analysis of Sensory Inputs of a Mobile Robot. *Proceedings of the Fifth International Workshop on Case Studies in Bayesian Statistics. Lecture Notes in Statistics*, Springer, New York, NY, 2000.
- [22] Marr, D. (1982) *Vision*. Freeman, San Francisco.
- [23] Mitchell, T. (1997) *Machine Learning*. McGraw-Hill, New York.
- [24] Nilson, N.J. (1984). Shakey the robot. SRI A.I. Center Technical Note 323.
- [25] Oliphant, M. (1996) The dilemma of Saussurean communication. *Biosystems*, 37 (1-2), pp. 31-38.

- [26] Pfeifer, R. and C. Scheier (1999) *Understanding Intelligence*. The MIT Press, Cambridge Ma.
- [27] Talmy, L. (2000) *Toward a Cognitive Semantics: Concept Structuring Systems (Language, Speech, and Communication)* The MIT Press, Cambridge Ma.
- [28] Pylyshyn, Z.W. (ed.) (1987) *The Robot's Dilemma*. Norwood NJ: Ablex Publishing Co.
- [29] Pylyshyn, Z. (2000) Situating vision in the world. *Trends in Cognitive Science*, 4(5), May 2000, pp 197-207.
- [30] Steels, L. (1997) *Constructing and Sharing Perceptual Distinctions*. In: van Someren, M. and G. Widmer (eds.) (1997) *Proceedings of the European Conference on Machine Learning*. Springer-Verlag, Berlin.
- [31] Steels, L. (1997) The synthetic modeling of language origins. *Evolution of Communication*, 1(1):1-35.
- [32] Steels, L. (1999) *How Language Bootstraps Cognition*. In: Wachsmutt, I. and B. Jung (eds.) *KogWis99. Proceedings der 4. Fachtagung der Gesellschaft fuer Kognitionswissenschaft*. Infix, Sankt Augustin. p.1-3.
- [33] Steels, L.: *The Emergence of Grammar in Communicating Autonomous Robotic Agents*. In: Horn, W. (ed.) *Proceedings of ECAI 2000*. IOS Publishing, Amsterdam. (2000)
- [34] Steels, L. and R. Brooks (eds.) (1995) *The Artificial Life Route to Artificial Intelligence : Building Embodied, Situated Agents*. Lawrence Erlbaum Assoc, New Haven.
- [35] Steels, L. and Kaplan, F. (1998) *Situated Grounded Word Semantics*. In *Proceedings of IJCAI-99*, Stockholm. Morgan Kaufman Publishing, Los Angeles. p. 862-867.
- [36] 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.

- [37] Ullman, S. (1996) *High-level Vision. Object Recognition and Visual Cognition*. The MIT Press, Cambridge Ma.
- [38] Victorri, B. and C. Fuchs. (1996) *La polysemie. Construction dynamique du sens*. Hermes, Paris.

5