Social symbol grounding and language evolution

Paul Vogt[1] and Federico Divina[1,2]
[1]ILK / Communication and Information Science, Tilburg University,
The Netherlands / [2]School of Engineering, Pablo de Olavide University,
Seville, Spain

This paper illustrates how external (or *social*) symbol grounding can be studied in simulations with large populations. We discuss how we can simulate language evolution in a relatively complex environment which has been developed in the context of the New Ties project. This project has the objective of evolving a cultural society and, in doing so, the agents have to evolve a communication system that is grounded in their interactions with their virtual environment and with other individuals. A preliminary experiment is presented in which we investigate the effect of a number of learning mechanisms. The results show that the social symbol grounding problem is a particularly hard one; however, we provide an ideal platform to study this problem.

Keywords: agent based modelling, language evolution, referential indeterminacy, joint attention, principle of contrast, cross-situational learning

Introduction

Human language is thought to have evolved from an interaction among three adaptive systems: biological evolution, individual learning and cultural evolution (Kirby & Hurford, 2002). The New Ties project[1] aims to merge these systems in a large scale simulation to evolve a cultural society of simulated agents that are situated in a complex environment and that need to acquire behaviours to remain viable over extended periods of time. One important aspect of this simulation is to evolve language that allows social learning, while being grounded in a virtual world.

The symbol grounding problem (Harnad, 1990) has been studied in relation to both language acquisition and language evolution using various robotic models (e.g., Roy, 2005; Steels, Kaplan, McIntyre, & Van Looveren, 2002; Vogt, 2002) and related simulations (e.g., Steels & Belpaeme, 2005; Cangelosi, 2001; A. D. M. Smith,

Interaction Studies 8:1 (2007), **31–52**.
ISSN 1572–0373 / E-ISSN 1572–0381

2005; Vogt, 2005). For overviews, see, for example, Cangelosi and Parisi (2002) or Vogt (2006). Most of these studies have focused on the ability of (simulated) robots to construct a shared symbolic communication system that has no 'survival' function to the society (but see, e.g., Cangelosi, 2001, for an exception). Such a survival function, however, is a crucial aspect of symbol grounding (Ziemke & Sharkey, 2001). The New Ties project will focus on how a language can evolve in a way that is relevant to the society's survival. To this end we need to deal with what Cangelosi (2006) has called *social symbol grounding*, that is, symbol grounding in (potentially large) populations.

To arrive at a shared set of symbolic conventions, the agents have to learn language from each other. In doing that, they face a problem that is closely related to the *referential indeterminacy* problem illustrated by Quine (1960). Quine showed that when learning a new word, the word can have — logically speaking — an infinite number of meanings. He used the example of an anthropologist who is studying a native speaker of a — to him — unknown language. When a rabbit suddenly scurries by, the native exclaims "gavagai!" and the anthropologist notes that gavagai means `rabbit`. Although this may be a valid inference, gavagai could also have meant `undetached rabbit parts`, `dinner`, `running animal` or even `it's going to rain`. To reduce the number of possible meanings, the anthropologist has to acquire more information regarding the meaning of gavagai. People — especially children — are extremely good at this, but for robots this has proven to be a very hard problem (Vogt, 2006).

Inspired by the literature on children's language acquisition, several learning mechanisms have been studied using computational models (see, e.g., A. D. M. Smith, 2005; Vogt & Coumans, 2003). Based on such studies, we present a new hybrid model that combines these learning mechanisms, which involve joint attention, feedback, cross-situational learning and the principle of contrast, in one model. We investigate the effect of these learning mechanisms on the ability to evolve a shared language in a large population. The next section provides more background on the symbol grounding problem. After that we present an overview of the New Ties project, followed by a more detailed description of the hybrid model that allows the population to evolve language. This description is followed by the presentation of some experiments investigating the learning mechanisms. The experiments are then discussed in relation to social symbol grounding before the paper concludes.

Symbol grounding

Physical symbol grounding

When agents communicate about things that are relevant in the world, they have to solve the symbol grounding problem (Harnad, 1990). Vogt (2002) has argued that, to achieve this, agents need to construct a semiotic relation between a *referent* (being something concrete or abstract), a *meaning* (being a representation inside an agent's brain that has some function to the agent) and a *form* (being the signal conveyed). This triadic relation (see Figure 1) is what Peirce (1931–1958) has called a symbol, provided the relation between form and meaning is either arbitrary or conventionalised.[2] The hardest part of solving this *physical symbol grounding problem* (i.e., creating the semiotic triangle, Vogt, 2002) is the construction of the relation between referent and meaning, because this relation is often dynamic and complex. As the physical symbol grounding problem may relate only to individual agents, the form could — in principle — be any arbitrary signal or label associated with this relation (Vogt, 2006). However, in language, the forms have to be conventionalised through cultural interactions and communicating forms have to be functional (e.g., it has to invoke some response from the recipient). Hence the population has to deal with *external symbol grounding* (Cowley, 2006), which we interpret as *social symbol grounding* (Cangelosi, 2006), i.e., symbol grounding in populations.

Social symbol grounding

Various studies have shown how a shared system of symbols can evolve from scratch through (local) cultural interactions between agents and (individual) learning mechanisms (Cangelosi, 2001; Steels & Belpaeme, 2005; Vogt, 2002). In this approach, which assumes that language is a *complex adaptive dynamical*

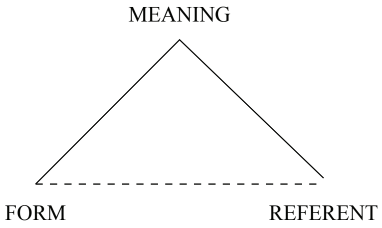

Figure 1. The semiotic triangle indicates the triadic relation among referent, meaning and form. Note that the relation between referent and form is indirect (hence the dotted line). When the relation between meaning and form is either arbitrary or conventionalised, the triangle represents a symbol. (Figure adapted from Ogden & Richards, 1923.)

system (Steels, 1997), a population of agents interact through a long series of *language games*. During these language games, agents can adapt their language, such that a shared structure (the external language) evolves through self-organisation. We believe that crucial aspects in the success of social symbol grounding are cognitive learning mechanisms, non-verbal interactions, physical properties of the environment and population dynamics.

Quine's referential indeterminacy problem can make the social symbol grounding problem more complicated than the physical symbol grounding problem. Many learning mechanisms in developmental linguistics have been proposed to deal with this problem (see, e.g. Bloom, 2000, for an overview). For instance, Tomasello (1999) has proposed that *joint attention* is a crucial mechanism by which two interlocutors focus their shared attention on a third object, allowing a child to associate utterances to the same situation the adult is attending to. This way, referential indeterminacy is reduced substantially, though the gavagai example shows this is not sufficient. The anthropologist can infer that gavagai relates to the rabbit, but is it the whole rabbit, its movement, its function, or something else? Additional mechanisms therefore have to be part of the cognitive learning mechanism.

Researchers have proposed a number of additional mechanisms that might further reduce referential indeterminacy, such as, e.g., the *principle of contrast* (Clark, 1993). In addition there is ample, though controversial, evidence that children receive feedback from their caregivers regarding their language use — especially regarding word-meaning mappings (Chouinard & Clark, 2003). Finally, evidence suggests that children may learn some words more straightforwardly by taking the intersection of their possible meanings across situations (Akhtar & Montague, 1999). This process, known as *cross-situational learning*, seems to take place from very early in the development of infants (Houston-Price et al., 2005).

Previous computational studies have investigated some of these mechanisms, but in isolation. Joint attention is most often modelled through *explicit meaning transfer*, where the hearer gets access to the exact intended meaning (e.g., Oliphant & Batali, 1997). More realistically, a number of robotic studies have used pointing as an unreliable joint attention mechanism, so hearers could not exactly determine the intended meaning, but could only estimate the intended referent (Vogt, 2000). In many studies corrective feedback on the referent or meaning has been the prime ingredient of the so-called *guessing games*, which allows agents to acquire the right association and to disambiguate competing word-meaning mappings through lateral inhibition (Steels & Belpaeme, 2005; Steels et al., 2002; Vogt, 2002). Finally, cross-situational learning, which is similar to the guessing game but without feedback, has been investigated extensively in computer models (Siskind, 1996; A. D. M. Smith, 2005; Vogt, 2000) and mathematical models (K. Smith et al., 2006).

In a study that has compared joint attention, feedback (through explicit meaning transfer) and cross-situational learning, Vogt and Coumans (2003) have found that cross-situational learning is hard to scale up for larger populations. This is because in the early stages of evolution different agents invent different words to convey the same meaning, which then have to be disambiguated during further development in order for effective communication to take place. If there is joint attention or feedback, disambiguation can be performed quite efficiently. However, cross-situational learning is based on the assumption that a word and its meaning are consistently co-occurring in different contexts. If there are many different words for a meaning, more ambiguities can enter the language and this condition may no longer hold. It has been shown, however, that cross-situational learning improves if there are additional biases such as mutual exclusivity (A. D. M. Smith, 2005) or some other synonymy damping mechanism (De Beule et al., 2006).

In the hybrid model that we introduce later, we combine the following mechanisms:

Joint attention is modelled by a pointing mechanism which allows a hearer to identify the target object reliably. This mechanism does not resolve uncertainty about the meaning of an utterance, because this relates to a feature of the object, such as colour or shape.

The principle of contrast allows agents to acquire the meaning of words such that they tend to favour meanings that have not yet been associated with other words.

Feedback is used as a non-verbal signal to indicate whether the hearer 'thinks' it has understood the speaker. Thus, the feedback may be prone to errors. Although negative feedback does not necessarily lead to correction, it increases the chance that the speaker repeats itself while using joint attention.

Cross-situational learning allows the refined learning of correct word-meaning mappings, regardless of whether joint attention is present.

In the simulations reported in this paper, we investigate the effect of each of these mechanisms on the ability to develop a shared lexicon.

New Ties

The objective of this project is to set up a simulation in which a large population of agents (i.e., more than 1,000) can evolve a cultural society using evolutionary, individual and social learning. Sub-objectives include investigating the interaction among these three adaptive systems and evolving a communication system that facilitates social learning. It is the latter aspect which is relevant to social symbol grounding and to this paper. Below is a brief description of the project; for more details consult Gilbert et al. (2006).[3]

The New Ties world — inspired by Epstein and Axtell's (1996) sugar-scape world — is a virtual world with places of varying roughness that contain objects such as tokens, edible plants and agents. Agents are provided with sensors and actuators that allow them to see and act. The sensors are configured such that an agent can see a number of perceptual features (e.g., colour, shape, direction, distance) of the objects in their visual field. The actuators allow the agents to, among others, move forward, turn left or right, eat, mate and talk. Each action costs energy, the amount of which depends, for instance, on the weight carried by the agents. When an agent's energy falls to zero, it dies, but it can also die of old age. Eating plants increases the agent's energy level, which depends on the 'ripeness' of the plant.

Agents develop their own control system using evolutionary, individual and social learning. This control system is a *decision Q-tree* (DQT), which is a stochastic decision tree that may change using reinforcement learning. The details of this DQT are beyond the scope of this paper, and the interested reader is referred to Gilbert et al. (2006). Suffice it to say that the DQT takes categories of perceived objects and interpreted messages as input and outputs an action based on some decision process. The structure of the DQT can change based on cross-over and mutation during reproduction, reinforcement learning and social learning. As we will discuss, social learning allows agents to develop shared behavioural skills using socially grounded symbols.

The genome carries, apart from the initial structure of the controller, a number of biases influencing the behaviours of agents regarding aspects such as the tendency to be social. The social bias is particularly important for language evolution, as it regulates, for instance, the frequency with which agents communicate or assist each other with learning language. Interaction is achieved by the predefined production and interpretation mechanisms, as explained in the next section.

Language evolution in New Ties

Figure 2 shows the architecture of the agents. The architecture consists of four modules, which are processed in sequential order from top to bottom. In addition, each agent has a short term memory (STM) and a long term memory (LTM). The input to an agent includes perceptual input regarding all objects an agent can see in its visual field and all messages sent within its audible range. There is no noise in perception, so it is assumed to be perfect. (Note that agents do not exactly share their visual or audible fields as they cannot be at the same location simultaneously.)

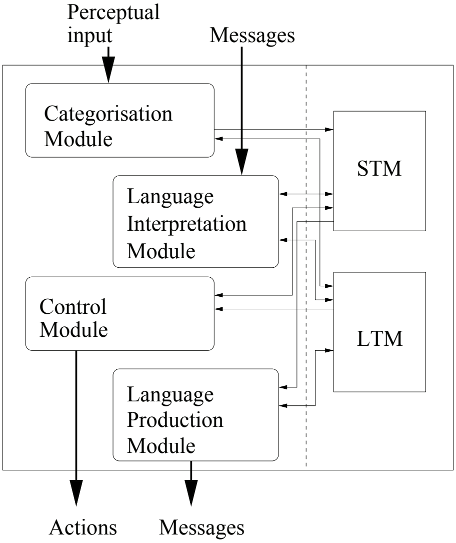

Figure 2. The basic architecture of the agents.

The perceptual input is first sent to the categorisation module, where each feature of each object is categorised with its nearest category. Categories are represented as predefined prototypes stored in the LTM in one-dimensional *conceptual spaces* (Gärdenfors, 2000). Each conceptual space relates to some quality dimension, such as colour, shape, direction, etc. It is possible to specify which categories are predefined and allow other categories to be acquired during development by playing discrimination games similar to the ones described in (Vogt, 2005). In the simulations reported here, the prototypes have a one-to-one relation among all possible perceptual features of all objects in the visual field, and the discrimination game has been switched off.

Since objects are perceived with different features (colour, shape, etc.), categorising an object results in a category set stored in the STM for further processing by other modules. The categories of all objects in the visual field constitute the *context* (*Cxt*). The context and interpreted messages are used by the control module to decide on an action to take. These actions can fail if they are in conflict with the 'physical' laws of the environment (e.g., two agents cannot be at the same location simultaneously).

When communicating, the agents construct physically grounded symbols (i.e., the semiotic triangle). With respect to the current setup of the model, the referent is a single perceptual feature of an object (e.g., a colour), the meaning is represented by its category and the form is a single word. For the moment we treat the meaning as a representation that has no real function regarding the agents' life task, though interpreting a message can influence the next action of an agent as determined by the controller.

We now explain the interpretation and production modules in more detail. Since we assume that objects' perceptual features are categorised with given prototypes, we focus this explanation on how a shared lexicon can arise as part of the social symbol grounding process.

Interpretation

The language interpretation module (LIM) processes all messages that an agent receives. A message can consist of multiple words. For each word an agent receives, the LIM searches the lexicon (stored in the LTM) for entries that match the word. The lexicon is represented by two association matrices (Figure 3), one that maintains association scores $\sigma_{ij} \in \langle 0,1 \rangle$ and one that maintains a posteriori probabilities $P_{ij} = P(w_i|m_j)$ of finding word w_i, given meaning m_j. The association scores contain information about the association's effectiveness as evaluated through feedback. However, since we assume that feedback is not always provided, nor always accurate, the agents also maintain the co-occurrence probabilities allowing for cross-situational learning. The reason for using two types of scores is that earlier studies have revealed that using the probability type score is less efficient (read slower) if feedback is present, whereas using the association scores σ_{ij} does not work well for cross-situational learning (Vogt & Coumans, 2003).

When a hearer searches its lexicon, it selects the association matching the heard word and of which the association strength $strL_{ij}$ is highest. (If the word is not in the lexicon, the interaction fails and the word is adopted as explained later.) The association strength is a coupling between the two scores σ_{ij} and P_{ij}:

$$strL_{ij} = \sigma_{ij} + (1 - \sigma_{ij})P_{ij}. \tag{1}$$

This coupling assures that the association strength relies more on the association score σ_{ij} if it is high (i.e., it has been effective in previous interactions); otherwise $strL_{ij}$ relies more on the co-occurrence probability P_{ij}.

$$\begin{array}{c|ccc} & m_1 & \cdots & m_N \\ \hline w_1 & \sigma_{11} & \cdots & \sigma_{1N} \\ \vdots & \vdots & \vdots & \vdots \\ w_M & \sigma_{M1} & \cdots & \sigma_{MN} \end{array} \qquad \begin{array}{c|ccc} & m_1 & \cdots & m_N \\ \hline w_1 & P_{11} & \cdots & P_{1N} \\ \vdots & \vdots & \vdots & \vdots \\ w_M & P_{M1} & \cdots & P_{MN} \end{array}$$

Figure 3. A simplified illustration of the lexicon. The lexicon consists of two matrices associating meanings m_j with words w_i. The left matrix stores association scores σ_{ij}, and the right matrix stores co-occurrence probabilities P_{ij}.

To establish joint attention, the speaker may have pointed to a target object o_t that relates to the message's meaning M_t, which is constructed by the speaker (s) as a subset of the target's category set CS_t^s (i.e., $M_t \subseteq CS_t^s$). Suppose m_j is the interpretation of one word (w_j) from the message. To estimate the outcome of this interpretation, the following steps are taken:

1. If an object is pointed to, the context *Cxt* is reduced to the hearer's category set CS_t^h of that target object. Now,
   (a) if $m_i \in CS_t^h$, the interpretation of word w_i is considered a *success*.
   (b) if $m_i \notin CS_t^h$, there are two possibilities:
      i. The association score $\sigma_{ij} > \Theta$ (where $\Theta = 0.8$ is a threshold), in which case it is assumed that the interpretation is correct, but the speaker got it wrong.
      ii. If $\sigma_{ij} \leq \Theta$, the interpretation is assumed to be the hearer's failure.
2. If no object is pointed to, and
   (a) $\sigma_{ij} > \Theta$, then the interpretation is considered a success. Otherwise,
   (b) the agent will — with some probability — either assume success or assume a hearer failure.

If the interpretation was considered a success for all words in a message, the controller adds it to the STM for further processing. Note that for step 2, the target relating to the interpreted meaning need not be in the hearer's context, so the agents can ground some knowledge about the environment without actually seeing it.

When the interpretation of all words has finished, the hearer may send a hardwired feedback signal to the speaker. This signal is sent with a probability proportional to the socialness gene and inversely proportional to their social bond.[4] This way, if the agent is social and does not know the speaker well, it is inclined to provide feedback, which should allow further learning. The feedback signals that can be sent are:

1. *Success* if the hearer considers the interpretation to be correct for all words.
2. *Speaker-error* if the hearer assumes the speaker to be wrong for at least one word.
3. *Interpretation-error* if there is a hearer failure for at least one word or when the hearer hears an unknown word.

Note that this feedback depends on the hearer's *estimation* of the game's outcome, but the hearer may be wrong as it has no means of verifying exactly what the speaker's meaning of a word is. This is different from the guessing games used in, for example, Steels et al. (2002), where the agents verify whether they refer to the same target.

	m_1	m_2
w_1	0.8	0.2
w_2	0.1	0.7

	m_1	m_2	m_3
w_1	0.8	0.2	0.0
w_2	0.1	0.7	0.0
w_3	0.02	0.03	0.1

Figure 4. An illustration of the principle of contrast. Suppose the leftmost table shows the lexicon containing the association scores before acquiring the word w_3, which is heard in the context $Cxt=\{m_1,m_2,m_3\}$. Both w_3 and m_3 are added to the lexicon, where the association scores of w_3 with meanings m_1 and m_2 are inversely proportional to the highest association scores already existing for these meanings, and the association with m_3 is highest, since this meaning had no existing association.

If a success feedback signal was sent, the used association scores σ_{ij} for both agents are increased, while all competing association scores σ_{kl} ($k=i$ or $l=j$, but not both) are laterally inhibited. If an interpretation error signal was sent, both agents also lower the association scores of the interpretation. In case of a speaker error, only the speaker lowers the association score. In addition, in all cases the co-occurrence probabilities P_{ij} of each word with all meanings in the context Cxt (or $CS_t^{s,h}$) are adapted accordingly. (For more details on these adaptations, consult Divina & Vogt, 2006).

When interpretation is assumed to have failed or when a word is not in the lexicon, the LIM adds this word w_n to the lexicon in association with all categories in the Cxt (or CS_t^h in case the target was pointed to), provided the association does not already exists. The frequency counters of these associations are set to 1 and — to simulate the *principle of contrast* — the association scores σ_{nj} are initialised with:

$$\sigma_{nj}=(1-\max_i(\sigma_{ij}))\sigma_0, \tag{2}$$

where $\max_i(\sigma_{ij})$ is the maximum association score that meaning m_j has with other words w_i, $i\neq n$ and $\sigma_0=0.1$ is a constant. This way, if the agent has already associated the meaning (or category) m_j with another word w_i, the agent is biased to prefer another meaning with this word (see Figure 4 for an example). It is important to note that since the agents do not share their visual fields, the hearer may not have seen the object relating to the word's meaning, so the new acquisitions may be wrong.

Production

When the LIM has finished processing, the control module will decide upon an action to take using all categories resulting from the categorisation and language interpretation as input. So, the information acquired through interpreting a message may influence this decision process.

Regardless of the action, the language production module (LPM) is started, because even if the action is not to talk, the LPM may nevertheless decide to communicate about something. This happens with a probability proportional to its socialness gene, provided the agent sees another agent. If the agent received an 'interpretation error' message, the LPM always decides to communicate about the object it communicated about before, provided the object is still in the context, but now the probability that the message is accompanied by a pointing gesture is increased. If no interpretation error was received and the agent has decided to communicate, a meaning is selected as follows.

First, a task complexity C_t is chosen. The task complexity is a value that indicates how many words the message will contain. C_t is a value between 1 and 5 such that the agent will tend to speak shorter sentences to younger agents and longer sentences to older agents. The rationale is that shorter sentences are easier to interpret by less skilled language users than longer sentences. Second, one target object is selected randomly from the objects in the speaker's visual field and the message's meaning M_t is formed from selecting C_t arbitrary categories from the category set CS_t^s relating to this target.

For each category, the LPM searches its lexicon for associations whose meaning matches the category and for which the association strength $strL_{ij}$ (Eq. 1) is highest. The associated word is then appended to the message. If a category has no entry in the lexicon yet, a new word is created as a random string and the new association is added to the lexicon. It is important to realise that agents are 'born' with an empty lexicon.

Once a message is thus constructed, the LPM decides, with a probability proportional to its social bias, whether the agent will point to an object that directly relates to the message's meaning. So, the more social the agent is, the more likely the speaker is to provide its audience with hints as to what it is referring to. This can, thus, be seen as a form of affective interaction as part of external symbol grounding (Cowley, 2006).

Experimental results

A number of experiments were done with the above model to investigate the effects of particular aspects of the learning mechanisms such as feedback, the principle of contrast, pointing and cross-situational learning. To keep our focus on these aspects, we switched off all evolutionary learning and individual learning mechanisms.

All experiments were run for 36,500 time steps, which is slightly longer than an agent's maximum lifespan. Agents were able to reproduce after they lived for

3,650 time steps, so the population size remained constant at the initial size of 100 during the first 3,650 time steps. Thereafter the population size increased slightly in all simulations, but kept fluctuating on average around 110 agents. This is because from that moment, many agents tended to die from their loss of energy expended during reproduction (offspring receives 50% of their parents' energy). Throughout all simulations, about 20% of the population initiated a language game each time step, so during one simulation — assuming an average population size of 110 agents — about 739,200 messages were sent. The agents tend to talk only in small groups because of their spatial distribution. On average, 44% of all language games were accompanied by a pointing gesture, in 12% of all games a feedback singal was sent, and in 48% of all games neither pointing nor feedback was used.

The effectiveness of the language evolution was monitored with *communicative accuracy* (or *accuracy* for short). Accuracy was measured at every 30 time steps by dividing the total number of *successful* language games by the total number of language games played during that period. A language game is determined successful if the hearer interpreted the speaker's expression with the exact intended meaning (not the intended referent). We prefer to use the term communicative accuracy rather than, for example, communicative success, because the interpretations need to yield intended meanings. Communicative success would be used if success was evaluated based on identifying the intended referent irrespective of the meaning representation. Since there is a one-to-one relation between meaning and referent in the current setting, communicative accuracy implies communicative success. For statistical purposes, all results we present are averages of 10 different trials of each setting with different random seeds.

In all simulations, the agents were given 6 feature channels (out of a maximum of 10) with which to detect 5 different types of objects having a total of 26 different perceptual features.[5] This means that each object (except agents) was perceived with 5 different perceptual features (colour, shape, characteristics, direction and distance); agents were perceived with the additional feature of sex (either male or female). So, if during a language game an object was pointed to, the a priori chance of communicative accuracy is 1/5 (or 1/6). If no object was pointed to, the a priori chance of accuracy was between 1/5 and 1/24 (or between 1/6 and 1/26), depending on the number of objects in the context.

A first experiment

A first experiment was carried out with the model as described in the previous section. Figure 5 shows the evolution of accuracy during this experiment. Accuracy increased rapidly during the first 2,000 time steps to a value around 0.6. From then on, accuracy remained more or less stable. Although accuracy did not reach a high

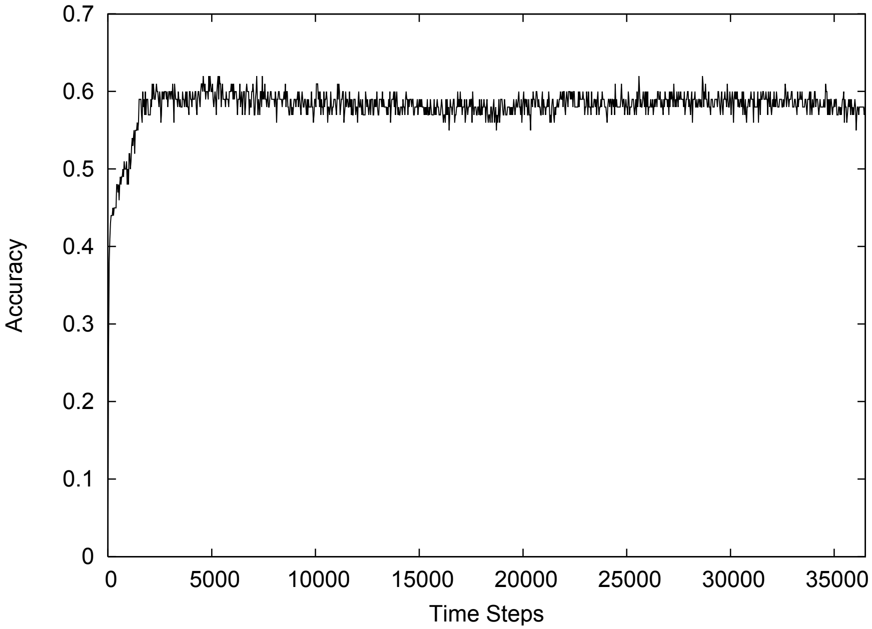

Figure 5. This figure shows the evolution of accuracy over time in the first experiment.

level (we discuss the reason why in the Discussion section), the system performed better than expected.

Although the lexicons of individual agents increased to up to an average of 250 words, they tended to use far fewer words as some were heard perhaps only once from an occasional contact with another agent. On average, the number of words used by the whole population during the final 1,000 time steps was 50. This means that — on average — only two different words were used by the population to express a meaning. This is surprising, since during their lifetime, each agent does not meet every other agent, nor are they likely to communicate with half of the population frequently enough to align their lexicons through direct communication. Hence, the language seems to have diffused over the population. However, given that individual lexicons contain an average of 250 words, there also may be considerable language change, though many of these words could come from sporadic inventions by young agents and their use across the populations.

Excluding learning mechanisms

In the second series of simulations, we varied the use of particular learning mechanisms by running experiments in which we switched off one of the following learning mechanisms:

No feedback. In these experiments, the agents did not provide feedback signals. As a result, the association scores σ_{ij} were never adapted, though their initial scores were still initialised following Eq. (2).

No principle of contrast. In these simulations, the principle of contrast was switched off (i.e., Eq. (2) was not used) and each novel association was initialised with the same association score σ_0.

No pointing. In this setting, no message was accompanied by a pointing gesture, so each context size was somewhere between 6 and 26, depending on the number of objects in the hearer's visual field.

No cross-situational learning. In these simulations, the co-occurrence probabilities P_{ij} were never updated. So $strL_{ij} = \sigma_{ij}$, which is only updated through feedback and the principle of contrast.

Figure 6 shows the results of his experiment as measured at the end of each set of simulations. For comparison, the results of the standard model used in the previous section are included. The graph shows that feedback and the principle of contrast had little influence on the level of accuracy. A Wilcoxon rank sum test has

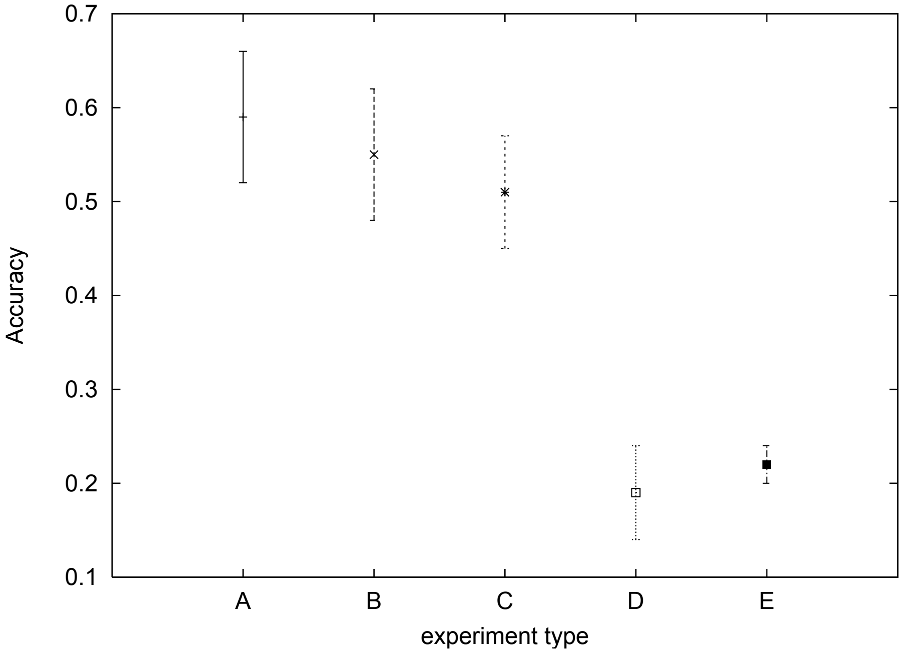

Figure 6. This figure shows accuracy measured at the end of each set of simulations for (A) the standard model and those that exclude either (B) feedback, (C) principle of contrast, (D) pointing or (E) cross-situational learning. Error bars indicate the standard deviations across different runs.

shown that the effect of removing the principle of contrast was more significant ($p = 0.0177$) than that of removing feedback ($p = 0.0526$). We introduced feedback, because we believed that this would improve learning enough to allow for damping of synonymy when adapting the association scores; a mechanism that was shown to be important in cross-situational learning in large populations (De Beule et al., 2006). We introduced the principle of contrast as an extra bias against synonymy, and although the effect is small, it appears significant.

Pointing and cross-situational learning, however, had a large impact. Pointing, of course, is used to reduce referential indeterminacy to the number of perceptual features. Cross-situational learning then refines the learning of word-meanings. Note that removing cross-situational learning does not reduce the model to the guessing game, because the infrequent updates of the association scores are based on unreliable feedback.

Discussion

In this paper we study social symbol grounding in a multi-agent simulation of a relatively complex world. This study focuses on how a shared system of grounded symbols can evolve based on a model that combines learning mechanisms such as feedback, the principle of contrast, joint attention (through pointing) and cross-situational learning. In this section, we discuss the experimental results, their implications and some future work.

We have argued that the social symbol grounding problem is probably harder than the physical symbol grounding problem (which is symbol grounding for an individual), because of the referential indeterminacy problem (Quine, 1960) that arises with the need for making conventions. The experiments have shown that to acquire and understand the verbal communication of other individuals, participants of language games benefit from engaging in triadic behaviours such as joint attention, because it reduces referential indeterminacy from all possible meanings in the context to all possible meanings relating to the attended object.

This means that the social activity of engaging in joint attention is crucial for this model. It is hard to generalise this finding to human societies, but since the point at which infants start to use joint attention activities coincides with the start of language use (Tomasello, 1999), joint attention skills seem crucial for social symbol grounding in general.

As mentioned before, pointing is not sufficient to eliminate all referential indeterminacy, because an object has a number of perceptual features and — in this model — a word refers to a single perceptual feature. (As the gavagai example illustrates, this holds in general.) To deal with the remaining uncertainty of a word's meaning, cross-situational learning could be a crucial mechanism, as this, too, has an significant impact on the success of evolving a shared lexicon in the model.

Although it has been shown mathematically that cross-situational learning can work if the context size is relatively large (K. Smith et al., 2006), for large populations it has been shown that evolving a coherent lexicon is quite hard, even for small context sizes (Vogt & Coumans, 2003). It must therefore be beneficial for cross-situational learning if agents engage in joint attention activities as this reduces the effective context size.

Feedback and the principle of contrast have little influence in this model, but it is even harder to validate these findings in general, because that would require more comparative experiments using more controlled environments. Regarding feedback, the small effect is most likely due to the fact that feedback does not make use of a mechanism to evaluate the success of a language game reliably. Agents only assume success if the association strength reaches a certain threshold. Feedback based on the evaluation of success, using explicit meaning transfer or verifying whether both agents have identified the same referent, has proven to be very effective (Steels et al., 2002; Vogt & Coumans, 2003). Future extensions of this model should therefore consider a means to evaluate successful interpretation more reliably.

It is likely that the contrast implemented (i.e., the differences between initial association scores) is too small, so that it only has an effect for a brief period. To prolong that period, we need to enlarge the initial differences by using a larger initial score σ_0 in Eq. (2).

Although much better than chance, the level of accuracy reached in the experiments (±60%) is far from optimal. It is hard to assess exactly why this is the case, but we can identify at least two reasons why grounding a shared lexicon is hard. First, a word does not always co-occur in a context containing its meaning, because the hearer may not have seen the target object intended by the speaker. The reason for this is that the speaker and hearer cannot be at the same location at the same time (see Figure 7). Furthermore, the speaker does not check the hearer's orientation, so both agents may have completely different visual fields. This is no problem if the hearer already acquired the word-meaning mapping reliably, but it is harmful for learning. Previous tests have shown that when agents use explicit meaning transfer (i.e., the exact meaning is provided to the hearer), accuracy increases to 97% (Divina & Vogt, 2006).

Second, because agents develop their own lexicon independently and from scratch, different agents may create different words to express the same meanings, so the maximum number of words created in a lexicon increases with the population. The task for the population is then to reduce the number of words being used. Such a reduction works well if the agents have a strong form of explicit meaning

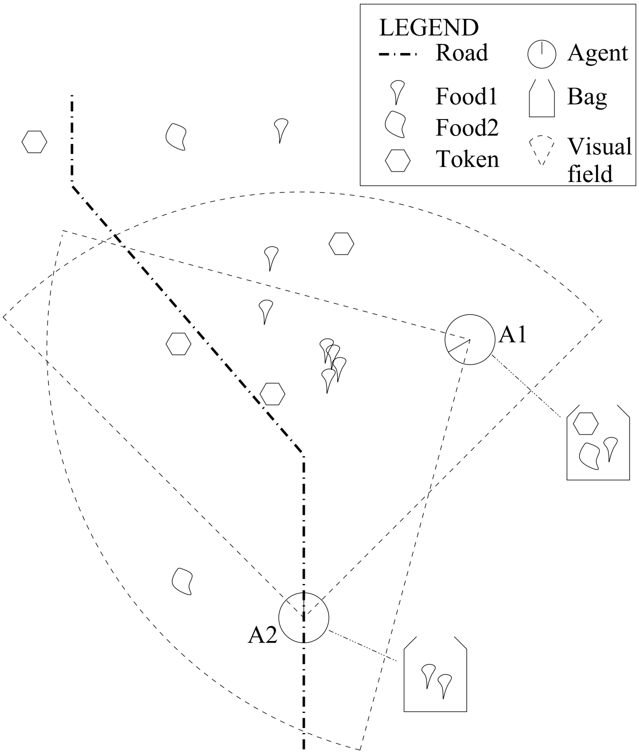

Figure 7. A possible situation for a language game. The two agents, A1 and A2 are inside each other's visual field (dashed arcs), but they do not share a context (i.e., visual field). As a result, the hearer may not have seen the intended target. Note that this is a near ideal situation; many situations will have more dissimilar contexts.

transfer, but it becomes harder if there is no strict one-to-one bias between words and meanings. We hope to improve accuracy in the future by incorporating the one-to-one bias model introduced by De Beule et al. (2006).

The extent to which accuracy can be improved remains to be seen. We are currently trying to improve the model by investigating the effect of changing parameter settings and modifying methods. We have described a different evaluation mechanism for the feedback process, changing the initial association score for improving the principle of contrast and the model of De Beule et al. However, the effectiveness of the social symbol grounding process is not only realised by the learning mechanisms and the quality of non-verbal interactions, but also by environmental constraints and population dynamics.

Regarding the former, we described the problems in sharing the context. More sophisticated methods for checking the 'focus' of attention might be developed to improve setting a common ground. This way, an agent might take into consideration what the other agent sees (and perhaps even knows). Hence, they might need something like a Theory of Mind (Premack & Woodruff, 1978). In addition, to allow better generalisations toward human societies, we need to perform comparative experiments using other platforms, such as the Talking Heads simulator (Vogt, 2005).

Regarding the population dynamic, the world is a spatial environment and agents are distributed over the entire world, though clusters (of agents and language) may emerge in the population.[6] Since agents move around, they will encounter new agents who speak another language, which has a negative effect on accuracy. The rate of population flow, their distribution in the world and the speed with which they can move will influence accuracy. Interestingly, though, language contact also seems to allow language diffusion (unless the language changes rapidly), which could explain how large language communities form. Further studies should investigate more thoroughly what exactly is happening.

In future experiments, we plan to extend the simulations, such that they are integrated with evolutionary, individual and social learning. Regarding the latter, we intend to extend the model with the social learning of skills where skills are transmitted using language. To accomplish this, the agents will communicate parts of their decision process as evaluated by their controller. As mentioned, the controller is a decision Q-tree that can be adapted using reinforcement learning (Gilbert et al., 2006). The idea is that agents adapt their DQTs by inserting new nodes based on the heard decision process of other agents, thus allowing them to align parts of their DQTs with those of others. This way, communicated meanings become more meaningful regarding the agents' survival, therefore surmounting true (social) symbol grounding (Ziemke & Sharkey, 2001)

Conclusions

In this paper we explore how the social symbol grounding problem can be investigated using large scale multi-agent systems to evolve social and other behavioural skills to survive in a complex environment over extended periods of time. In particular, we investigate a novel hybrid model of language learning that involves joint attention, feedback, cross-situational learning and the principle of contrast.

The experiments show that — although the system does not work optimally — levels of communicative accuracy better than chance evolve quite rapidly in this system. In addition, they show that accuracy is mainly achieved by the joint attention and cross-situational learning mechanisms and that feedback and the principle of contrast do not contribute much. However, further experiments using different parameter settings, platforms and learning mechanisms are required to generalise these findings.

The research to be carried out with the New Ties platform has only just started and, to increase the number of related studies, the New Ties platform has been made publicly available.[7] We encourage other researchers to use this platform — which we think allows the study of symbol grounding in a social context — and challenges will be published to set out benchmark experiments. One way to extend the current model is to allow populations using language to learn behavioural skills from each other. This would take social symbol grounding to a higher level.

Acknowledgements

This research and the New Ties project is supported by an EC FET grant under contract 003752. We thank all members the New Ties project for their invaluable contributions. Opinions and errors in this manuscript are the authors' responsibility, they do not necessarily reflect those of the EC or other New Ties members. Paul Vogt is supported by a VENI grant awarded by the Netherlands Organisation for Scientific Research (NWO). Special thanks to Andrew Smith and three anonymous reviewers for their invaluable comments on earlier versions of this manuscript.

Notes

[1] New Ties stands for New Emerging World models Through Individual, Evolutionary and Social learning. See http://www.new-ties.org.

[2] Note that Peirce used a different terminology than that adopted here. He used *object, interpretant* and *representamen* to denote what we call *referent, meaning* and *form*, respectively. The adopted terminology is more common in modern cognitive science.

[3] The current state of the project is that most parts have been implemented and tested, and preliminary experiments are being carried out.

[4] The social bond is based on the frequency with which two agents have interacted with each other.

[5] In Divina and Vogt (2006) we have investigated the effect of the number of feature channels on the level of accuracy. The results have shown that, if the agents perceive up to 6 features, accuracy evolves to a lower level and then more or less stabilises, because additional features, such as age, are not always observable for some objects.

[6] Methods are being developed to discover clusters in the population regarding similarities in language and controllers.

[7] http://www.new-ties.org.

References

Akhtar, N., & Montague, L. (1999). Early lexical acquisition: The role of cross-situational learning. *First Language*, *19*, 347–358.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA. and London, UK.: The MIT Press.

Cangelosi, A. (2001). Evolution of communication and language using signals, symbols and words. *IEEE Transactions of Evolutionary Computation*, *5*, 93–101.

Cangelosi, A. (2006). The grounding and sharing of symbols. *Pragmatics and Cognition*, *14*, 275–285.

Cangelosi, A., & Parisi, D. (Eds.). (2002). *Simulating the evolution of language*. London: Springer.

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

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

Cowley, S. J. (2006). Distributed language: biomechanics, functions and the origins of talk. In C. Lyon, C. Nehaniv, & A. Cangelosi (Eds.), *Emergence and evolution of linguistic communication.* Springer.

De Beule, J., De Vylder, B., & Belpaeme, T. (2006). A cross-situational learning algorithm for damping homonymy in the guessing game. In L. Rocha, L. Yaeger, M. Bedau, D. Floreano, R. Goldstone, & A. Vespignani (Eds.), *ALIFE X. Tenth international conference on the simulation and synthesis of living systems.* Cambridge, MA: MIT Press.

Divina, F., & Vogt, P. (2006). A hybrid model for learning word-meaning mappings. In P. Vogt, Y. Sugita, E. Tuci, & C. Nehaniv (Eds.), *Symbol grounding and beyond.* Springer.

Epstein, J. M., & Axtell, R. (1996). *Growing artificial societies: social science from the bottom up*. Cambridge, MA.: MIT Press.

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

Gilbert, N., Besten, M. den, Bontovics, A., Craenen, B., Divina, F., Eiben, A., et al. (2006). Emerging artificial societies through learning. *Journal of Artificial Societies and Social Simulation*, *9(2)*.

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

Houston-Price, C., Plunkett, K., & Harris, P. (2005). 'Word-learning wizardry' at 1;6. *Journal of Child Language*, *32*, 175–190.

Kirby, S., & Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (p. 121–148). London: Springer.

Ogden, C. K., & Richards, I. A. (1923). *The meaning of meaning: A study of the influence of language upon thought and of the science of symbolism*. London: Routledge & Kegan Paul Ltd.

Oliphant, M., & Batali, J. (1997). Learning and the emergence of coordinated communication. *Center for Research on Language Newsletter*, *11(1)*.

Peirce, C. S. (1931–1958). *Collected papers* (Vol. I-VIII). Cambridge Ma.: Harvard University Press.

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

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

Roy, D. (2005). Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, *167*, 170–205.

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

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

Smith, K., Smith, A., Blythe, R., & Vogt, P. (2006). Cross-situational learning: a mathematical approach. In P. Vogt, Y. Sugita, E. Tuci, & C. Nehaniv (Eds.), *Symbol grounding and beyond.* Springer.

Steels, L. (1997). The synthetic modeling of language origins. *Evolution of Communication*, *1(1)*, 1–34.

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, *28*, 469–529.

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

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

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

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

Vogt, P. (2005). The emergence of compositional structures in perceptually grounded language games. *Artificial Intelligence*, *167(1–2)*, 206–242.

Vogt, P. (2006). Language evolution and robotics: Issues in symbol grounding and language acquisition. In A. Loula, R. Gudwin, & J. Queiroz (Eds.), *Artificial cognition systems.* Idea Group.

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

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

Authors' addresses

Paul Vogt
Communication and Information Science, Tilburg University, P.O. Box 90153, 5000 LE
Tilburg, The Netherlands

p.a.vogt@uvt.nl

Federico Divina
School of Engineering, Pablo de Olavide University, Seville, Spain

fdiv@upo.es

About the authors

Paul Vogt received his M.Sc. ('doctoraal') in Cognitive Science and Engineering (currently Artificial Intelligence) from the University of Groningen (NL). For this degree he wrote his dissertation at the AI Lab of the Vrije Universiteit Brussel in Belgium, where he also received his Ph.D. for the thesis *Lexicon Grounding in Mobile Robots.* He was a post-doctoral researcher at the Universiteit Maastricht (NL) and at the Language Evolution and Computation unit of the University of Edinburgh (UK). Currently, he is a research fellow at the Communication and Information Science department of Tilburg University (NL). His research focuses on language evolution and acquisition, particularly on those aspects related to symbol grounding.

Federico Divina is currently an assistent professor at the School of Engineering of the Pablo de Olavide University of Seville, Spain. Previously he was a post-doctoral fellow at the Department of Communication and Information Science of the Tilburg University in the Netherlands. He received his Ph.D. in 2004 from the Computational Intelligence group of the Vrije Universiteit of Amsterdam with a dissertation on the use of hybrid evolutionary computation applied to inductive learning, under the supervision of Prof. Gusz Eiben and Elena Marchiori. He visited the Machine Learning group of the University of Seville, Spain, where he applied Evolutionary Computation to bioinformatics problems. He studied Computer Science at the Ca' Foscari University of Venice, Italy. His research in the New Ties project focuses on the emergence and evolution of language.