Perceptually grounded lexicon formation using inconsistent knowledge

Federico Divina[1], Paul Vogt[1,2]

[1] Computational Linguistics and AI Section
Tilburg University, Tilburg, The Netherlands
F.Divina@uvt.nl
[2] Language Evolution and Computation (LEC) Unit
University of Edinburgh - UK
paulv@ling.ed.ac.uk

Abstract. Typically, multi-agent models for studying the evolution of perceptually grounded lexicons assume that agents perceive the same set of objects, and that there is either joint attention, corrective feedback or cross-situational learning. In this paper we address these two assumptions, by introducing a new multi-agent model for the evolution of perceptually grounded lexicons, where agents do not perceive the same set of objects, and where agents receive a cue to focus their attention to objects, thus simulating a Theory of Mind. In addition, we vary the amount of corrective feedback provided to guide learning word-meanings. Results of simulations show that the proposed model is quite robust to the strength of these cues and the amount of feedback received.

1 Introduction

In the past decade, a number of studies have investigated the emerge of perceptually grounded lexicons in multi-robot systems [14, 13]. The aim of such studies is to investigate under what conditions a population of (possibly simulated) robots can evolve a shared vocabulary (or *lexicon*) that allow them to communicate about objects that are in their environment. Typically, these studies assume that the lexicons evolve culturally through inter-agent communication, individual adaptation and self-organisation [12]. These *perceptually grounded* studies extend other *ungrounded* models in which the meanings of the lexicons are predefined as in, e.g., [6, 8], by allowing the agents to develop their meanings from scratch based on their sensing of the environment. Typically, the lexicons and meanings are constructed through *language games* [12] in which two agents communicate about an object they detect in a certain context. This way, the grounded models add more realism to the ungrounded models, as they do not assume that agents have an innate set of meanings.

The grounded models, however, still build upon many assumptions. One such assumption is that agents perceive the same set of objects. Especially in simulations – both grounded [10, 17] and ungrounded [8, 5, 18] – this is taken for granted. In studies using real robots, this is assumed, though not necessarily

achieved [16]. When two agents communicate, one cannot simply assume that the context of one agent is similar to the other agent's context. Agents located in different places see different things – even if they are located close to each other and looking in a similar direction.

The problem agents face when learning the meaning of words is that when hearing a novel word, logically, this word can have an infinite number of meanings or references – even if the target is pointed at [9]. So, in order to learn the meaning of a word, an agent has to reduce the number of possible meanings. Humans are exceptionally good at this, and it is generally assumed that humans have some innate or acquired means to infer what a speaker's intention is, see, e.g., [2] for an overview. Among these means are *joint attention* [15], *Theory of Mind* (ToM) [2], and receiving *corrective feedback* on the meaning of words [3]. In joint attention, the participants of a communication act focus their attention to an object or event while actively checking that they share their attention. In a way, this requires that the participants understand each other as having similar intentions [15]. Loosely speaking, this is a part of the Theory of Mind. However, ToM is more: it allows someone to form theories regarding the intentions of speakers by simulating that he or she is the speaker (him)herself. For instance, a child may know that its caregiver is hungry and can therefore infer the caregiver is more interested in food than in a doll. The problem with joint attention and ToM is that it is not always precise. Suppose a rabbit passes by and someone shouts 'gavagai', even if you establish joint attention, you cannot be sure 'gavagai' refers to the rabbit; it may also refer to undetached rabbit parts or that it is going to rain [9]. To further reduce the number of possible meanings for a word, caregivers sometimes provide corrective feedback on the meaning of childrens' utterances. Although the availability of corrective feedback is highly disputed [2], recent analysis has shown that it is actually abundant, especially with respect to the meaning of words [3]. However, corrective feedback is not always present. In those cases, we assume that children can learn the meaning of words across situations by focusing on the covariance between word meaning pairs. There is some recent evidence that children indeed use such a learning strategy [1].

The current study addresses the two issues discussed and introduces a model in which a perceptually grounded lexicon is developed by a population of simulated robots based on the Talking Heads experiment [13]. In this model, the contexts that agents perceive while playing a language game differ from each other. In addition, as suggested in [16, 18], the three models are integrated together with a naive implementation of ToM. The ToM is implemented by introducing attention cues that focus the (possibly joint) attention on objects in a context. In this study, these cues are assigned randomly, but in future work we intend to implement this by having the agents estimate these cues autonomously based on some knowledge about intentions [19]. Using these cues and a verbal hint provided by the speaker, the hearer will guess the reference of the uttered word, and in some cases the agents evaluate corrective feedback. In addition, co-occurrence frequencies are maintained to allow for cross-situational statistical learning [20]. The experiments reported in this paper investigate the effect

of using these attention cues and to what extent the model is still robust when we vary the amount of corrective feedback available to the learners.

The next section introduces the new model. In Section 3 we experimentally assess the validity of the proposed model, and finally Section 4 concludes.

2 The Model

The model we propose here is implemented in the Talking Heads simulation toolkit THSim[3] [17] and extends the iterated learning model (ILM) [7] in combination with the language game model [13]. The ILM implements a generational turnover in which a population of adults transmits its acquired lexicon to the next generation of learners culturally by engaging in a series of language games. A language game is played by two agents, which are randomly selected from the population at the start of each game: a speaker taken from the adult population and a hearer taken from the learner population. In each generation, a fixed number of language games are played and after each generation, the adults are removed, the learners become adults and new agents enter the population.

Each time a language game is played, both agents a individually observe a context C_a that contains a number of objects o_i. The objects are geometrical coloured shapes, such as red triangles and blue squares. The contexts of the agents share a number of objects, while the rest are distinct. If the contexts contain n objects, the agents share $1 \leq k \leq n$ objects, where k is assigned randomly in each game. If $k = n$, then the contexts are equal.

In order to provide each agent a with some naive form of Theory of Mind, we simulate the use of *attention cues* $strA(o_i)$ assigned to each object $o_i \in C_a$. In future models [19], we intend to base these attention cues on a more sophisticated ToM, which the speaker uses to select the topic of the language game and which the hearer uses to estimate the speaker's intention (see Section 2.2). For the moment we assume that these attention cues are assigned with random values, where shared objects have a high attention cue, while objects that are not shared possess a low attention cue. In short:

$$strA(o_i) = X_i = \begin{cases} \beta_s \leq X_i \leq 1 & \text{if } o_i \text{ is shared,} \\ 0 \leq X_i \leq \beta_u & \text{if } o_i \text{ is not shared.} \end{cases} \quad (1)$$

where β_s and β_u are user supplied parameters, with default values of $\beta_s = 0.5$ and $\beta_u = 0.1$. These values were experimentally determined to yield good results; in general it was found that the results were good when $\beta_s > \beta_u$. Shared objects are assigned the same attention cue value.

2.1 Categorising Objects

Categorisation of objects is based on the *discrimination game* model [11] and implemented using a form of 1-nearest neighbourhood classification [4]. The aim of the discrimination game is to categorise an object (called the *topic* o_t) in an

[3] THSim is available from http://www.ling.ed.ac.uk/~paulv/thsim.html.

agent's context such that it is distinctive from the other objects in the context. Note that this does not require that the category is a perfect exemplar of the object. By playing a number of discrimination games, each agent a constructs its own ontology O_a, which consists of a set of categories: $O_a = \{c_0, \ldots, c_p\}$. The categories c_i are represented as prototypes $\mathbf{c}_i$ which are points in a n-dimensional conceptual space. Each agent starts its life with an empty ontology.

Each object is perceived by six perceptual features f_q: colour (expressed by Red, Green and Blue components of the RGB space), shape (S) and location (expressed by X and Y coordinates). Each agent a extracts for each object $o_i \in C_a$ a feature vector $\mathbf{f}_i = (f_1 \ldots, f_n)$, where n is equal to the number of perceptual features.

Each object $o_i \in C_a$ is categorised by searching a category $c_j \in O_a$, such that the Euclidean distance $||\mathbf{f}_i - \mathbf{c}_j||$ is smallest. It is then verified that the category found for the topic o_t is distinctive from the categories of the other objects $o_k \in C_a \backslash \{o_t\}$. If no such category exists, the discrimination game fails, and the ontology of the agent is expanded with a new category for which the feature vector $\mathbf{f}_t$ of the topic is used as an exemplar. Otherwise the discrimination game succeeds and the found category is forwarded as the topic's *meaning* m to the production or the interpretation phase.

2.2 The Language Game

Table 1. The outline of a language game, see the text for details.

speaker	hearer
-perceive context -categorisation/DG -produce utterance -update $memory_1$ -send message	
	-receive message -perceive context -categorisation/DG -interpret utterance -update $memory_1$
-corrective feedback	-corrective feedback
-update $memory_2$	-update $memory_2$

The language game we propose, outlined in Table 1, combines the guessing game, e.g., [13] with a *cross-situational statistical learner* [10, 16, 20]. In both models, the hearer h guesses the topic based on the utterance produced by the speaker, where the topic $o_t \in C_h$. In the guessing game, the agents evaluate whether or not the hearer guessed the right topic (i.e. the object referred to by the speaker). In cross-situational statistical learning (CSSL) such feedback is not evaluated, instead the agent keeps track of co-occurring word-meaning pairs. In order to provide a naive ToM, the model is adapted to include the attentional cues $strA(o_i)$.

In a nutshell, the agents start by perceiving the context of the game and categorise the objects they see using the discrimination game (DG) as explained above. Then the speaker s selects an object from its context as the topic $o_t \in$

C_s of the language game. In order to so, a roulette wheel mechanism is used, where the sectors of the roulette wheel are proportional to the attention cues $strA(o_i)$ assigned to the objects. Thus, typically, objects that are shared with the context of the hearer have more probability of being selected, since generally their attention cues are higher. As the hearer uses these attentional cues as a bias in guessing the speaker's topic, it is virtually simulating the speaker's selection process. This, we believe, is a naive form of ToM, which in future work we intend to work out more realistically, based on agents' more sophisticated selection criteria.

Each agent maintains an internal lexicon, represented by two associative memories, as illustrated in Figure 1. One of the associative memories (referred to as $memory_1$ in Figure 1) keeps an *a posteriori probability* P_{ij}, which is based on the occurrence frequencies of associations. The other matrix ($memory_2$) maintains an *association score* σ_{ij}, which indicates the effectiveness of an association based on past experiences. The reason for this twofold maintenance is that studies have revealed that when strong attentional cues (such as the corrective feedback used in the guessing game) guide learning, lexicon acquisition is much faster with the association score σ_{ij} than with the a posteriori probabilities [18]. The reverse is true when such strong attentional cues are absent as in CSSL. This is mainly because the update mechanism reinforces the score σ_{ij} more strongly than the update of usage based probabilities P_{ij}. This works well when the cues are precise, but the fluctuations of σ_{ij} would be too strong to allow statistical learning in CSSL.

	m_1	...	m_N
w_1	P_{11}	...	P_{1N}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
w_M	P_{M1}	...	P_{MN}

memory₁

	m_1	...	m_N
w_1	σ_{11}	...	σ_{1N}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
w_M	σ_{M1}	...	σ_{MN}

memory₂

Fig. 1. Two associative memories constructed and maintained as part of an agent's lexicon. The left memory ($memory_1$) associates meaning m_j with word w_i using conditional a posteriori probabilities P_{ij}. The right memory ($memory_2$) associates meanings m_j with words w_i using an association score σ_{ij}.

The probabilities are conditional probabilities, i.e.,

$$P_{ij} = P(m_j|w_i) = \frac{u_{ij}}{\sum_j u_{ij}} \tag{2}$$

where u_{ij} is the co-occurrence frequency of meaning m_j and word w_i. This usage frequency is incremented each time word w_i co-occurs with meaning m_j that is either the topic's meaning (in case of the speaker) or the meaning of an object in the context (in case of the hearer). The update is referred to in Table 1 as 'update $memory_1$'. If this principle is the only mechanism, the learning is achieved according to the CSSL principle, i.e., across different situations based on the covariance in word-meaning pairs [20].

When corrective feedback is evaluated, the association score σ_{ij} is updated according to the following formula:

$$\sigma_{ij} = \eta\sigma_{ij} + (1 - \eta)X \tag{3}$$

where η is a learning parameter (typically $\eta = 0.9$), $X = 1$ if the association is used successfully in the language game, and $X = 0$ if the association is used wrongly, or – in case of a successful language game – if the association is competing with the used association (i.e., same word, different meaning; or same meaning, different word). The latter implements lateral inhibition. If Eq. (3) is the only update, the game reduces to the guessing game. The update of association scores is referred to in Table 1 as 'update $memory_2$' and is only carried out if corrective feedback is evaluated. The rate with which corrective feedback is evaluated is subject of the second experiment.

Given these two matrices, the speaker, when trying to produce an utterance, calculates an association strength $strL(\alpha_{it})$ for each association α_{it} of a word w_i with the topic's meaning m_t. This is done using Eq. (4):

$$strL(\alpha_{it}) = \sigma_{it} + (1 - \sigma_{it})P_{it} \tag{4}$$

This formula neatly couples the two variables. When σ_{it} is high, the influence of P_{it} is low, and when σ_{it} is low, P_{it} will have more influence. This implements a bias toward basing a choice on known effectiveness vs. estimated probabilities. In Eq. (4), σ_{it} and P_{it} might be weighted, in order to rely more on the association scores or on the a posteriori probabilities. The speaker will select the association that has the highest strength $strL(\alpha_{it})$ and utters its word. If no association can be found, e.g., because the lexicon is still empty, the speaker invents a new word and adds the association to its lexicon with an initial association score $\alpha_{it} = 0.01$ and $u_{it} = 0$.

When the hearer h receives an utterance, it looks in its memories for associations with the current signal and whose meanings match the meanings for each object $o_j \in C_h$ in its context. Using the association strengths $strL(\alpha_{ij})$ and attentional cues $strA(o_j)$, the hearer then interprets the utterance using the following equation based on [5]:

$$\rho_{ij} = \omega_L \cdot strL(\alpha_{ij}) + \omega_A \cdot strA(o_j) \tag{5}$$

where ω_L and ω_A are weights between 0 and 1. Throughout both experiments $\omega_L = 1$ is kept constant, the value of ω_A is subject of variation in the first experiment. If the heard word is not in its lexicon, then the hearer will add it to the lexicon in association with all meanings of the objects in the context and $u_{ij} = 1$. If the agents evaluate the feedback, the word is additionally associated with the meaning m_t of the now-known topic o_t with an initial association score $\sigma_{it} = 0.01$. Feedback is provided to the agents with a given probability P_{fbk}, which is subject to variation in the second experiment. When feedback is provided, the agents update $memory_2$using Eq. (3). The hearer will not update $memory_2$ in the case where the topic is not in its context, since in this case it cannot perceive the category of the topic.

3 Experiments

In order to assess the validity of the proposed model we performed several experiments. In the following we present results using two measures: *production coherence* and *interpretation accuracy*. Production coherence is defined as the fraction of agents that produced the same utterance to name objects. Interpretation accuracy is the fraction of agents that could successfully interpret the produced utterances, averaged over the number of games played. These measures were calculated during a testing phase, consisting of 200 games in which the language did not evolved. The test phase took place at the end of each generation. The results presented are averages over ten runs using different random seeds. In the experiments we used a population of 10 agents in 15 generations of 10,000 language games each. During all experiments the context size $n = 4$ was kept constant, while $1 \leq k \leq n$ was chosen randomly each language game.

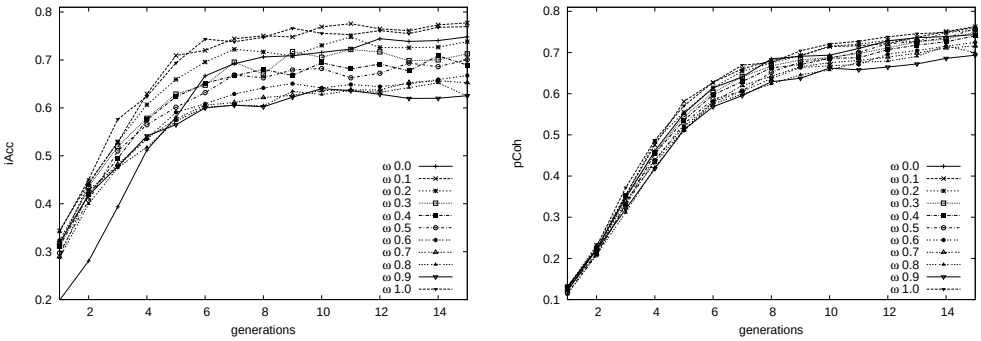

Fig. 2. Results obtained with different values of ω_A. The left figure shows interpretation accuracy, the right one production coherence.

A first set of experiments was aimed at evaluating the effect of considering $strA(o_i)$ in Eq. (5) by varying the weight ω_A between 0 and 1 with intermediate steps of 0.1. In this experiment, corrective feedback was always evaluated, i.e., $P_{fbk} = 1.0$. The results for interpretation accuracy are shown in Fig. 2 (left). In this figure the x-axis represents the number of generations processed. It is interesting to notice that the model also yields good results when the attention cues are not considered, i.e. when $\omega_A = 0$ accuracy increases to one of the highest levels. (Note that this setting reduces the model to the guessing game.) In general, when ω_A increases, the results get poorer – especially after about 5 generations So, it seems that agents get confused when higher values of ω_A are used, i.e., the attentional cues suggest a different interpretation than the lexicon. However when $\omega_A = 1$, accuracy is good again, which suggests that the attentional cues can have a positive impact on the learning process, provided they are sufficiently strong. Coherence (Fig. 2 right) reveals a similar evolution, though the values show less variation. At the end of the simulations, coherence is between 0.70 and 0.75. We have further investigated this aspect with equal contexts, i.e., where in all cases $k = n$. For reasons of space, we cannot report

the results here, however we can say that the results reflect the results presented here. Another set of experiments was performed in order to assess the behaviour of the model when ω_L was varied. The model obtained similar results for all values of this parameter, except when $\omega_L = 0$ the results were considerably worse.

Another aspect we wanted to investigate was the robustness of the proposed model with respect to the amount of feedback received by the agents. We therefore performed experiments with different values of P_{fbk}, while keeping $\omega_A = 1.0$ fixed. Before presenting the results, it is good to recall that when $P_{fbk}= 0$ only $memory_1$ is updated at every language game. In effect this is a CSSL, where the agents do not receive any feedback but have to infer this information by the observation of their contexts. In contrast to earlier CSSL models [10, 16, 20], the learners additionally receive attentional cues from $strA(o_i)$, which makes the model more similar to the one presented by Gong et al. [5]; and the contexts of speaker and hearer are dissimilar, thus the hearer may not have observed the topic.

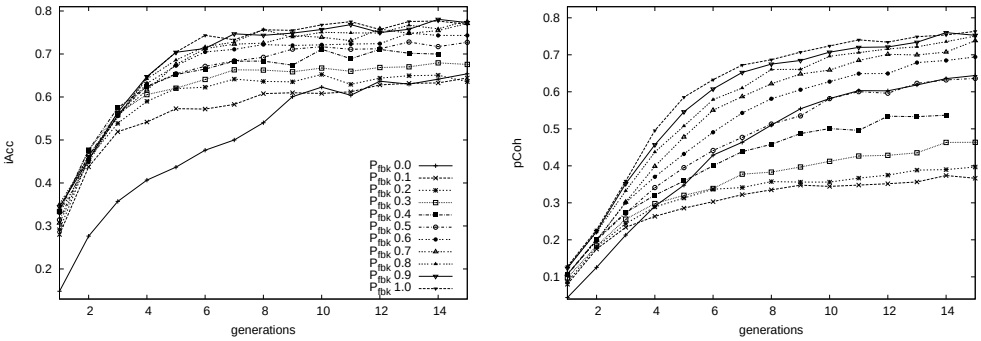

Fig. 3. Results obtained with different values of P_{fbk}. The graphs show interpretation accuracy (left) and production coherence (right).

The results are reported in Figure 3. It can be noticed that even when there was no feedback ($P_{fbk}= 0$), language developed, though it took more generations to reach an acceptable level of interpretation accuracy than for higher values of P_{fbk} (Fig. 3 left). Interpretation accuracy, although acceptable, was still among the lowest. Production coherence, on the other hand, rose towards a value that is among the highest (Fig. 3 right). This is interesting, since the CSSL generally leads to low levels of coherence [20]. Apparently, the attentional cues provide enough information to allow robust word learning, which is in line with the results of experiment 1 when $\omega_A = 1$. With higher values of P_{fbk} agents receive more feedback and can, therefore, develop a lexicon more easily. Recall that when they receive feedback, they update $memory_2$ as well. Clearly this has a positive effect on speed of learning and on interpretation accuracy. It only has a positive effect on production coherence when $P_{fbk}\geq 0.5$ – i.e. when agents received feedback with a high probability. So, although with a low amount of P_{fbk} accuracy was doing reasonably, coherence remained well behind. This suggests that low

amounts of feedback antagonises the attentional cues, since the feedback may provide different information than the attentional cues, but it has a relatively strong effect on the lexicon development through Eq. 3. When P_{fbk} becomes sufficiently high, the feedback is sufficiently strong to drive a coherent lexicon development. Nevertheless, we can conclude that the model is quite robust to various levels of feedback.

4 Conclusions and Future Work

In this paper a new multi-agent model for the evolution of perceptually grounded lexicons is presented. This model combines the guessing game model [13] with the cross-situational statistical learning model [20] and the introduction of environmental attentional cues similar to the models proposed in [5].

Simulations based on the Talking Heads show that the model is quite robust for different levels of attentional cues set on the objects. However, the simulations show that – in general – the more the attentional cues are used in the interpretation by the hearer, the more the hearer tends to get confused. This is primarily due to the unreliability of the attentional cues, which confuses the hearer. Interestingly, the results improve when the weight for the attentional cues becomes one. In this case, the attentional cues are strong enough to form a beneficial account for the language development.

Another important result is that the model is robust to the enforced dissimilarity of the contexts of agents playing a language game. This is interesting, since it shows that the agents do not require explicit meaning transfer (which is the case whenever feedback is present) while the hearers may not have seen the objects speakers are referring to. Clearly, the results improve when more feedback is present. However, when no feedback is present at all, the results exceed some of the results achieved with infrequent use of corrective feedback, thus showing the robustness of cross-situational statistical learning in combination with using stochastic attentional cues.

Future work should investigate more precisely why the model behaves differently for the different parameter settings. For instance, why do the simulations with higher values of $\omega_A < 1$ or lower values of $P_{fbk} > 0$ perform worse than the cases where $\omega_A = 1$ or $P_{fbk} = 0$? It is also interesting to study the effect of varying P_{fbk} with different values of ω_A. In addition, we intend to incorporate the current model in the recently started New Ties project[4], which aims at developing a benchmark platform for studying the evolution and development of cultural societies in very large multi-agent systems. In this project, we will extend the model such that instead of assigning attentional cues randomly, agents will autonomously estimate (or calculate) these cues as part of a Theory of Mind [19].

Acknowledgements

This paper is part of the New Ties project, which is supported by the European Commission Framework 6 Future and Emerging Technologies programme under contract

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

003752. The authors thank Andrew Smith and three anonymous reviewers for providing invaluable comments on earlier versions of this paper.

References

1. N. Akhtar and L. Montague. Early lexical acquisition: The role of cross-situational learning. *First Language*, 19:347–358, 1999.
2. P. Bloom. *How Children Learn the Meanings of Words*. The MIT Press, Cambridge, MA. and London, UK., 2000.
3. M. M. Chouinard and E. V. Clark. Adult reformulations of child errors as negative evidence. *Journal of Child Language*, 30(3):637–669, 2003.
4. T.M. Cover and P.E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, IT-13(1):21–7, January 1967.
5. T. Gong, J. Ke, J. W. Minett, and W. S-Y. Wang. A computational framework to simulate the co-evolution of language and social structure. In *ALife 9*, Boston, MA, U.S.A., 2004.
6. J. R. Hurford. Biological evolution of the saussurean sign as a component of the language acquisition device. *Lingua*, 77,2:187–222, 1989.
7. S. Kirby. Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2):102–110, 2001.
8. M. Oliphant. The learning barrier: Moving from innate to learned systems of communication. *Adaptive Behavior*, 7 (3-4):371–384, 1999.
9. W. V. O. Quine. *Word and object*. Cambridge University Press, 1960.
10. A. D. M. Smith. Intelligent meaning creation in a clumpy world helps communication. *Artificial Life*, 9(2):559–574, 2003.
11. L. Steels. Emergent adaptive lexicons. In P. Maes, editor, *SAB96*, Cambridge, MA, 1996. MIT Press.
12. L. Steels. The synthetic modeling of language origins. *Evolution of Communication*, 1(1):1–34, 1997.
13. L. Steels, F. Kaplan, A. McIntyre, and J. Van Looveren. Crucial factors in the origins of word-meaning. In A. Wray, editor, *The Transition to Language*, pages 214–217. Oxford University Press, Walton Street, Oxford OX2 6DP, UK, 2002.
14. L. Steels and P. Vogt. Grounding adaptive language games in robotic agents. In C. Husbands and I. Harvey, editors, *Proceedings of the Fourth European Conference on Artificial Life*, Cambridge Ma. and London, 1997. MIT Press.
15. M. Tomasello. *The cultural origins of human cognition*. Harvard University Press, 1999.
16. P. Vogt. Bootstrapping grounded symbols by minimal autonomous robots. *Evolution of Communication*, 4(1):89–118, 2000.
17. P. Vogt. Thsim v3.2: The talking heads simulation tool. In *ECAL03*, pages 535 – 544. Springer-Verlag, 2003.
18. P. Vogt and H. Coumans. Investigating social interaction strategies for bootstrapping lexicon development. *Journal of Artificial Societies and Social Simulation*, 6(1), 1 2003.
19. P. Vogt and F. Divina. Language evolution in large populations of autonomous agents: issues in scaling. In *Proceedings of AISB 2005: Socially inspired computing joint symposium*, pages 80–87, 2005.
20. P. Vogt and A. D. M. Smith. Learning colour words is slow: a cross-situational learning account. *Behavioral and Brain Sciences*, page In press, 2005.