Does language shape the way we conceptualize the world?

Joachim De Beule (joachim@arti.vub.ac.be)
Vrije Universiteit Brussel, Artificial Intelligence Lab,
Pleinlaan 2, 1050 Brussels, Belgium

Bart De Vylder (bartdv@arti.vub.ac.be)
Vrije Universiteit Brussel, Artificial Intelligence Lab,
Pleinlaan 2, 1050 Brussels, Belgium

Abstract

In this paper it is argued that the way the world is conceptualized for language is language dependent and the result of negotiation between language users. This is investigated in a computer experiment in which a population of artificial agents construct a shared language to talk about a world that can be conceptualized in multiple and possibly conflicting ways. It is argued that the establishment of a successful communication system requires that feedback about the communicative success is propagated to the ontological level, and thus that language shapes the way we conceptualize the world for communication.

Introduction and Research Question

Language and communication involve many aspects of human cognition including the sensory-motor schema's needed to observe the world, the social abilities for establishing joint attention and communicative intent and the mechanisms responsible for parsing and producing abstract grammatical expressions.

A key issue here is how a population of distinct and only locally interacting agents (language users) can agree upon a global language. It is commonly accepted that at least part of the answer is self-organization: a consensus is reached through repeated peer-to-peer negotiations about how to express some meaning.

A prerequisite for this, which is often neglected, is that the agents already have to agree upon the set of expressible meanings. It is implicitly assumed that all agents conceptualize the world according to some universal ontology.

However, there are strong indications that the way in which observations are conceptualized for language is language dependent and also the result of negotiation between language users. For example, different languages lexicalize color categories differently and it is suggested that color terms might have an influence on color categorization (see for example [Steels and Belpaeme, 2005], [Roberson, 2005]; see also [Levinson, 2001] for evidence on how language appears to shape a language learner's meaning structure.)

We investigate this phenomenon in a population of artificial agents placed in an artificial world that can be conceptualized in multiple and conflicting ways. Agents are equipped with learning mechanisms that allow them to establish a shared language. A prerequisite for a successful communication system is that the agents have mutually compatible conceptualization schemes or ontologies. It is shown that, in turn, feedback on the communicative success has to be propagated to the ontological level in order to obtain compatible ontologies. As such it is shown that a language both depends on and influences an agent's ontology and vice versa.

Related and Previous work

There have been many computational models in which a population of artificial agents evolve a shared language [Cangelosi and Parisi, 2001]. Not so many however have discussed in depth the co-evolution of meaning and form. In the following two exceptions will be discussed briefly.

The Talking Heads Experiment

In the talking heads (TH) and related experiments (see e.g. [Steels, 1998]) a population of robots develop a shared ontology and lexicon to communicate about differently shaped and colored objects by playing language games. Each game two agents are presented with a collection of objects called the context. One of the objects is the topic of the game. Only one of the agents, the speaker, is informed about the topic. He conceptualizes the topic (i.e. construes a meaning describing the topic) and verbalizes the result. The other agent, the hearer, then should locate the topic. If he succeeds the game is a success , otherwise it is a failure.

The current experiment is at a higher level of abstraction and ignores many difficulties that arise when working with real robots. This is done on purpose, as it allows us to precisely control the structure of the world and its influence on language. Also, the focus is here on the co-evolution of ontology and language. Although in the TH setup meaning and form co-evolve as well, there are some important differences.

In the TH an ontological category is defined as a region in some sensory channel. An example of a sensory channel is the horizontal position (HPOS) and an example of a 'left' category is $0 \leq \text{HPOS} < 0.5$. A new category is created by splitting a channel or an existing category in two. This is done when the agent fails to discriminate the topic from the other objects in the context. For example, the left category is not sufficient to discriminate the topic when the context contains two objects with a HPOS less than 0.5. This might be solved by subdividing the left category into two subcategories etc. As such a so called discrimination tree is constructed.

It is clear that, in the absence of noise, an agent's discrimination abilities can be made arbitrarily high by progressively growing the discrimination tree (i.e. introducing more specific categories.) However, this rapidly conflicts with the agent's communicative success: the larger an agent's ontology the larger the lexicon needed to express it and the less learnable the language will be. Therefore, when a certain discriminative success is reached no additional categories are created and infrequently used leaf-categories are deleted.

An important point here is that all agent's ontologies share a common structure. This is because all agents have the same sensor channels and they all use the same top-down mechanism to grow their ontology. This allows them to decide on which categories to prune and ensures, up to a certain level, that all remaining categories are relevant for all agents in the population (see also [Smith, 2003] for the influence of biased meaning creation on communication.) This is an important difference with the currently described experiment in which a category's relevance in the population is not a priori known and it is not clear which categories to add or delete if multiple candidates are at hand.

Color Category Formation

In a more recent study ([Steels and Belpaeme, 2005]) it was already shown that feedback about the success in communicating color categories is most probably needed to explain the way in which humans categorize color for language. In this experiment the observation space is a continues real-valued 3 dimensional color-space and categories consist of prototypes in this space. Again, the more categories in the ontology the bigger its discriminative power will be but the less suitable for language. An important mechanism to control the size of an ontology here makes use of the fact that a distance measure can be defined between any two categories or observations. This allows for example to merge two categories that are close together by replacing them by their mean category. In the current experiment categories are more like predicates that are either true or false for an observed object or event (e.g. (red ?object) or (in-the-past ?event).) Merging such categories is not well defined and other means of controlling the size of the ontology are needed.

In addition, in the color-category experiment categories can be shifted in color space. This is actually the main mechanism by which agents reach a consensus: if some agent defines the meaning of a word to be some point in color space, another agent can shift his own category associated with that word toward that point. Again, a shift operation is not at hand in the current experiment.

Experimental Setup

As in the TH, our experiments consist of presenting a speaker and a hearer with a world scene called context and topic about which they have to communicate. A context consists of a collection of objects. The topic is one of the objects in the context. By letting the agents play language games [Steels, 2001] and have them act according to a fixed interaction protocol we don't have to be concerned with modeling e.g. communicative intent or the establishment of shared attention etc. In this section we explain how contexts are represented and generated, what the architecture is of an agent and how the agents act and interact.

Representation of the World

In figure 1 a scene in a 2 dimensional grid world is shown containing 3 objects. An agent observes a scene through

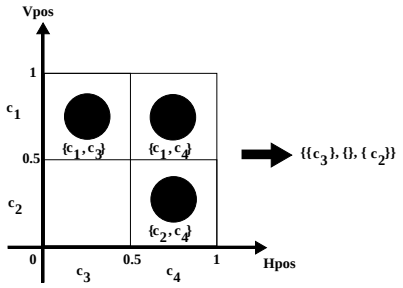

Figure 1: A scene in a 2D-grid world containing three objects. There are two sensory channels Vpos and Hpos, and four categories: $c_1 = 0.5 < \text{Vpos} \leq 1.0$, $c_2 = 0 \leq \text{Vpos} \leq 0.5$ and similar categories for the Hpos channel. The upper left object is categorized by the categories c_1 and c_3 but it is only discriminated by the c_3 category. For the purposes in this paper this scene is completely specified by the description $\{\{c_3\}, \{\}, \{c_2\}\}$.

a collection of sensory channels. Every channel returns a numerical value between 0 and 1 for every object in the scene. On every channel categories are defined which return true or false when applied to such a value. For example, there could be a channel for the horizontal position of an object. An example of a category on this channel is *left=Hpos<0.5*. In this paper we are not concerned with how such categories could come to existence and simply use a collection of random intervals on each channel (see [De Beule, 2004] for an example of how an agent can search for new categories in the domain of time.)

A category is said to categorize an object if it returns true for the object. Thus, in figure 1, the c_1 category categorizes both upper objects. A category is said to be a discriminating category for an object with respect to some context if it categorizes the object but it does not categorize any other object in the context. Thus, in figure 1, both the categories c_1 and c_3 categorize the upper left object but only the c_3 category is discriminating.

More general, let $\mathcal{C}$ be the set of all possible categories (functions of observed objects returning true or false), and let a scene S contain n objects: $S = \{o_1, ..., o_n\}$. Let $D(o_i, S)$ be the subset of $\mathcal{C}$ of all discriminating categories for object o_i with respect to S. All information of a scene S needed for our purposes, like the set of discriminating categories for an object in S, is given by the description obtained by mapping D over

S: $\{D(o_1, S), ..., D(o_n, S)\}$. For example, the scene in figure 1 is completely described by $\{\{c_3\}, \{\}, \{c_2\}\}$. A world is defined as a collection of scenes.

Equation (1) defines a collection of three scenes that will be used to illustrate some things in the following.

$$\begin{aligned} S_1 &= \{\{c_1, c_2\}, \{c_3, c_4\}\} \\ S_2 &= \{\{c_1, c_3\}, \{c_2, c_4\}\} \\ S_3 &= \{\{c_1, c_4\}, \{c_2, c_3\}\}. \end{aligned} \quad (1)$$

For the experiments that follow a large set of world scenes of differing complexity will be required. Scenes containing a variable number of objects are generated as explained, assuming that the agents have 5 sensory channels.

However, two additional constraints are posed on the scenes that are used. First we do not allow that any object in a scene has an empty set of discriminating categories. As such, the scene in figure 1 is not allowed since the upper right object has an empty set of discriminating categories. This is because our agents are only allowed to produce single word utterances expressing single categories. Hence, with these restrictions, a language game that has a topic for which there are no discriminating categories will always fail.

Second, we do not allow equivalent categories in a world: if there are two categories such that if an object is discriminated by one of them then it is also always discriminated by the other, then the two categories are said to be equivalent. A world containing equivalent categories can easily be transformed to one without equivalent categories by keeping only one category for every equivalence class.

In order to compare different worlds we define the complexity of a world as the mean number of (non equivalent) categories by which an object in a scene can be discriminated, averaged over all scenes in the world. For example, Both objects in scene S_1 in (1) can be discriminated by two categories resulting in a complexity value of $(2 + 2)/2 = 2$ for this scene. The other scenes in (1) also have a complexity of 2. Thus, a world consisting of the scenes S_1, S_2 and S_3 has complexity $(2+2+2)/3 = 2$.

Agent Architecture

Agents act either as speaker or as hearer. A speaker is schematically represented in figure 2, it consists of an ontology and a lexicon. An ontology is a mapping $\langle c_i, s_{o,i} \rangle$ between categories c_i and ontological strengths $s_{o,i}$. A category's ontological strength reflects its usefulness for discriminating objects in the world. A lexicon contains a set of lexical entries $\langle c_i, w_i, s_{e,i} \rangle$, each associating a category c_i with a word w_i with a strength $s_{e,i}$. An entry's strength reflects its usefulness for communication.

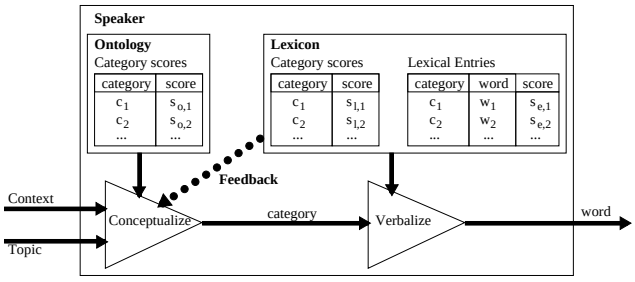

Figure 2: Speaker Agent Architecture.

A speaker is presented with a context (a scene) and a topic (an object in the scene). He first has to select a category from its ontology that uniquely describes the topic and then has to verbalize it according to its lexicon. In this paper, selecting a category for language is called conceptualization. Conceptualization could be done based on ontological strengths. But categories might be useful for other purposes than language and the decision of which category to select when multiple candidates are at hand might depend on the purpose for which it is selected.

This is modeled by also assigning a lexical strength $s_{l,i}$ to every ontological category c_i. The lexical strength of a category is equal to the strength of its strongest lexical entry. If the lexicon contains no entries for a category then its lexical strength is 0. This is represented by the "feedback" arrow in figure 2.

Hence, selecting a category is done according to the modulated strength given in equation (2).

$$s_i = s_{o,i}^{(1-\lambda)} s_{l,i}^{\lambda} \quad (2)$$

The parameter λ models the amount of modulation: if it is 0 a category is selected solely according to the ontological strength, if it is 1 only the lexical strengths contribute.

Thus, to conceptualize the topic, the speaker first determines the set of categories in his ontology that discriminate the topic. The ontological strengths of these are increased by 1. If his ontology does not contain any discriminating categories a random one is added with initial strength 1. The topic is conceptualized by the discriminating category that has the highest modulated strength. The selected category is then verbalized by the word of its strongest lexical entry. If there is no lexical entry for the category a new one is added, associating the category with a new (unique) word and with an initial strength 0.5.

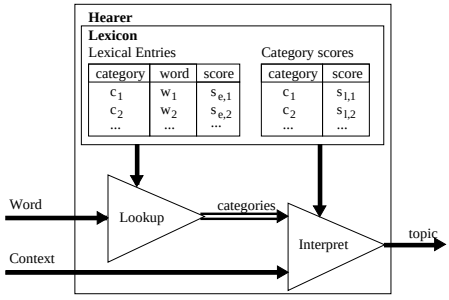

Figure 3: Hearer Agent Architecture.

Figure 3 schematically represents a hearer. A hearer is presented with a word and a context. First he looks up

the set of categories associated with the word. He then filters out those that do not uniquely determine an object in the context. Of the remaining categories he picks the one with the highest lexical strength. The resulting category determines the hearer topic. If the hearer does not know the word the game cannot be completed. If however the game was completed but the hearer determined a different topic from the one presented to the speaker the game fails, otherwise it is a success.

The communicative success of a population of agents playing language games in a certain world is defined as the chance of having a successful game between two random agents presented with a random scene and topic.

Learning

The adoption of new categories and the updating of ontological scores is done as described in the previous section: whenever an agent is unable to discriminate a topic he extends his ontology. And each time a category *could* be used for conceptualization its strength is incremented. As such, the ontology gradually adopts the structure contained in the world.

The outcome of a game has no effect on an agent's ontology, only on its lexicon, and thus possibly also on the lexical category strengths. The goal of an agent is to evolve an ontology and lexicon with which it can successfully communicate with other agents. He should therefore adapt his lexicon to conform himself to the population, i.e. to mimic the other agents, by adding or removing lexical entries and changing their strength. He can do so because each time he interacts with another agent he is actually sampling the population and gains information about it.

When a game is not completed (i.e. the hearer does not know the word), the hearer is presented with the topic. He then also conceptualizes it according to his own ontology and lexicon and associates the speaker's word with it with initial strength 0.5.

Now consider the case of a completed game. Assume that a speaker uttered a word w and that the hearer interprets it to mean category c. If the game was successful, the hearer hypothesizes that the speaker also conceptualized the topic with category c. In addition, he hypothesizes that the speaker *preferably* associates the word w with that category. Note that this might be incorrect, for example when there are multiple categories with which the topic can be discriminated.

For the hearer, using these hypotheses to conform to the population means adapting his lexicon such that if he were himself a speaker he would show the same behavior. In other words he should enforce the score of the lexical entry $\langle w, c\rangle$.

The speaker however does not learn anything about the hearer's preference: it is not because the hearer *understands* a word that he himself would *prefer* to use it. It might even be that he would have preferred another category to conceptualize the topic. Therefore, after a successful interaction the speaker's lexicon is left unchanged (apart from lateral inhibition, see below.)

After an unsuccessfully completed game however, both the speaker and the hearer can conclude that they should have done something else and they both inhibit the strength of the lexical entry used. Finally, whenever a lexical entry reaches a strength of 0 it is forgotten, (i.e. removed from the lexicon.)

Updating the lexicon as described so far is sufficient for the agents to reach successful communication. However, in language games it is common to also inhibit competing entries after a successful interaction. Hence, when a lexical entry $\langle c, w, s\rangle$ is successfully used, the strengths of other lexical entries competing for the same word w or the same category c are inhibited. This is called lateral inhibition and is needed to eliminate synonyms and homonyms and to reduce the size of the lexicon as is illustrated in figure 4.

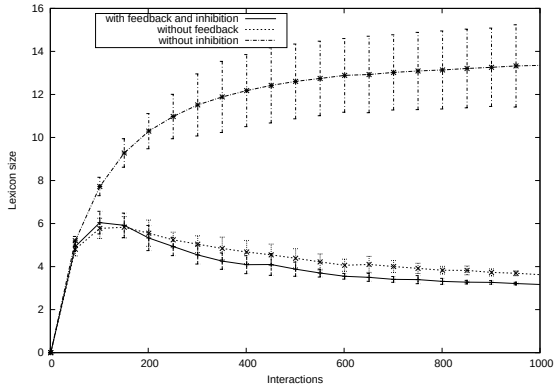

Figure 4: Effect of lateral inhibition and feedback (parameter λ) on the size of the lexicon. All curves show the evolution of the average lexicon size in a population of 10 agents playing language games in the world of equation (1), averaged over 50 independent runs. In the two lower curves the agents used lateral inhibition to reduce the size of their lexicon. In the top and bottom curve the agents used complete feedback to conceptualize a topic (i.e. $\lambda = 1$.) In both these cases the agents reach 100% communicative success.

Table 1 and equations (3) summarize the updating rules that are used in all experiments in this paper unless otherwise stated. These rules can be interpreted as the Rescorla-Wagner/Widrow-Hoff rule as described in [Sutton and Barto, 1981].[1]

$$\begin{aligned} \text{reinforcement:} \quad & s \leftarrow s + p(1-s) \\ \text{inhibition:} \quad & s \leftarrow s - ps \end{aligned} \tag{3}$$

To determine the value of the learning rate p in (3), consider that these equations can be interpreted as implementing a weighted memory: if p is 0 then no learning

[1]The lexicon is then interpreted as a network in two ways: one network for expression with one input for each category and one output for each word and one network for interpretation with switched inputs and outputs. The input of the category/word to be expressed/interpreted is set to 1, the others to 0. The expected output is 1 for the corresponding word/category and 0 for the others.

Table 1: Updates of lexical entry strengths after a successful (left) or unsuccessful (right) but completed interaction between a speaker and hearer.

	Success	Failure
Speaker	lateral inhibition	inhibition
Hearer	reinforcement lateral inhibition	inhibition

is done. If it is 1 then only the information of the last interaction is kept. If it is in between 0 and 1 then the information of all previous interactions contribute to the strength s but with the more recent ones contributing more (every new update the previous contributions are multiplied by $p < 1$.) Thus, the inverse of p is a measure for the length of an agent's memory. An upper bound on $1/p$ can be calculated by stating that in $1/p$ interactions an agent should optimize the chance of interacting with a random agent exactly once (more then once does not provide extra information.) In a population of N_a agents this chance is given by the binomial distribution to be

$$\frac{1}{pN_a}\left(1-\frac{1}{N_a}\right)^{\frac{1}{p}-1}. \quad (4)$$

Optimizing this chance gives $p \simeq 1/N_a$, in words the learning rate should be greater or equal to the inverse of the population size.

This is only a lower bound on the learning rate. Consider for example the case where all agents but one have converged. For that single agent this situation is comparable to a population of only two agents. If the population consists of N_g distinct groups of indistinguishable agents, then a better learning rate is given by the inverse of N_g. For a population size of 10, a learning rate of about $p = 0.3$ was empirically found to be optimal. In all the experiments discussed in this paper this value is used unless otherwise stated.

Notice that, when the modulation parameter λ in equation (2) is positive, feedback on the communicative success is propagated to the conceptualization level of an agent since ontological category scores will be modulated by lexical category scores which in turn are updated according to communicative success. And since the outcome of conceptualization determines the outcome of the verbalization and hence also whether the next interaction will be successful or not, a positive parameter $\lambda > 0$ implements a *closed* feedback loop between the ontological level (conceptualization) and language level (lexicon and communicative success.) This loop implements an alignment mechanism with which the agents can align their conceptualization until they are compatible.

Comparison of Performance

We will now turn to the comparison of populations that do use feedback and populations that don't. In order to compare the performance of two populations we both present them with the same world and let them play language games in parallel. Whenever one of the populations reaches a certain cutoff communicative success, the difference in performance is measured as the difference in communicative success. This difference is taken such that positive values mean that the population with feedback performed better, negative values the opposite. This is illustrated in figure 5 with a cutoff of 0.99.

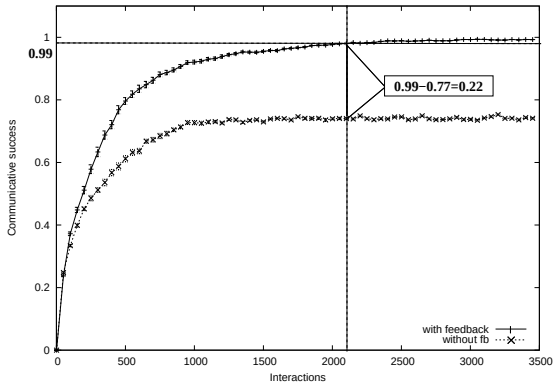

Figure 5: Typical effect of feedback on the communicative success and illustration of how the performance of different populations is compared. Both curves show the evolution of the communicative success of a population of 10 agents playing language games in the world of equation (1), averaged over 50 independent runs. In the top curve the agents used complete feedback to conceptualize a topic (i.e. $\lambda = 1$), whereas in the lower curve no feedback was used ($\lambda = 0$). After approximately 2100 interaction the feedback population reached 0.99% communicative success whereas the other population got stuck at 0.77%. The difference in performance is then calculated to be 0.99-0.77=0.22 in favor of the feedback population.

As a first experiment we measured the influence of the feedback parameter λ on the performance. Populations of varying sizes were 11 times presented with a randomly generated TH-world but each time with their feedback parameter increased by 0.1, starting from 0. Performance differences were measured for different values of the cutoff (the value of the communicative success of the best performing population at the moment that performances are measured.) This experiment was repeated several times using different worlds of varying complexity.

The results are first that positive feedback ($\lambda > 0$) has a significant and positive effect on the performance. No significant difference was found for different feedback values as long as it was positive. This might be an artifact of the way worlds were generated.

Second, the positive effect is more clear for higher cutoffs. Together with the fact that all experiments with a cutoff of 0.99 were completed and thus that at least one population reached almost 100% communicative success, this suggests that populations that do not use feedback

often simply are incapable of reaching 100% communicative success. Figure 5 thus indeed illustrates the typical effect of feedback on the communicative success, although more extensive testing should be done to verify this hypotheses.

Finally, the positive effect is less significant for populations of only two agents, but from three on no clear trend could be detected. This is probably because in certain situations a population of only two agents with incompatible conceptualization schemes can still successfully communicate whereas this is not the case with three or more agents[2].

In the second experiment we measured the influence of the world complexity on the positive effect that feedback has on performance. Based on the results of the previous experiment, we only considered populations of three or more agents and only compare total feedback to no feedback with a cutoff of 0.99. Populations of varying size were presented with worlds of varying complexity. The results are shown in figure 6. Again it is clear that feedback has a positive effect on performance. A value of 0.3 for the performance measure for example means that at the moment that the population with feedback reached 0.99% communicative success, the population without feedback only reached a value of 0.69% and, based on the first experiment, probably will not be able to do better at all.

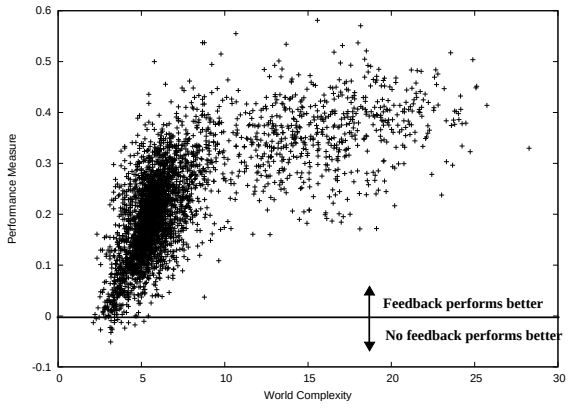

Figure 6: Influence of the world complexity on the positive effect of feedback on performance. Relatively few worlds were generated with a complexity greater than 9 which explains why the plot is less dense in this region.

In addition it can be seen that this positive effect is less significant in less complex worlds. This is not surprising since in the limit of a world with complexity 1 there is only one way to conceptualize an object and no choice has to be made which can be guided by feedback. This also suggests that in order to speed up learning and minimize errors during the language acquisition phase a new language learner should be presented with simple and unambiguous scenes.

[2]An example of a world in which this could happen is given by equation (1).

Conclusion

The main thesis of this paper was to show that successful communication requires that feedback about the communicative success should be propagated to the ontological level.

Even though we have made some major simplifications, both regarding the complexity and structure of the world and the complexity of language, we still found a significant effect: populations that do not use feedback perform significantly worse. This suggests that successful communication requires that language shapes the way we conceptualize the world for communication.

References

[Cangelosi and Parisi, 2001] Cangelosi, A. and Parisi, D., editors (2001). *Simulating the Evolution of Language*. Springer Verlag, London.

[De Beule, 2004] De Beule, J. (2004). Creating temporal categories for an ontology of time. In Rineke Verbrugge, Niels Taatgen, L. S., editor, *Proceedings of the Sixteenth Belgium-Netherlands Conference on Artificial Intelligence*, pages 107–114.

[Levinson, 2001] Levinson, S. C. (2001). Covariation between spatial language and cognition, and its implications for language learning. In Bowerman, M. and Levinson, S. C., editors, *Language acquisition and conceptual development*. Cambridge University Press, Cambridge.

[Roberson, 2005] Roberson, D. (2005). Color categories are culturally diverse in cognition as well as in language. *Cross Cultural Research*, 39(1):56–71.

[Smith, 2003] Smith, A. D. (2003). Intelligent meaning creation in a clumpy world helps communication. *Artificial Life*, 9(2):559–574.

[Steels, 1998] Steels, L. (1998). Synthesising the origins of language and meaning using co-evolution, self-organisation and level formation. In Hurford, J. R., Studdert-Kennedy, M., and Knight, C., editors, *Approaches to the Evolution of Language: social and cognitive bases*. Cambridge University Press, Cambridge, UK.

[Steels, 2001] Steels, L. (2001). Language games for autonomous robots. *IEEE Intelligent Systems*, sept-oct 2001:17–22.

[Steels and Belpaeme, 2005] Steels, L. and Belpaeme, T. (2005). Coordinating perceptually grounded categories through language. A case study for colour. *Behavioral and Brain Sciences*. Accepted as target article.

[Sutton and Barto, 1981] Sutton, R. and Barto, A. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, (88):135–170.