# Joint Attention and Language Evolution

*Johan Kwisthout, Paul Vogt, Pim Haselager, and Ton Dijkstra*

Department of Information and Computing Sciences
Utrecht University
Technical Report UU-CS-2007-039
www.cs.uu.nl
ISSN: 0924-3275

# Joint Attention and Language Evolution

**Johan Kwisthout[*], Paul Vogt[**], Pim Haselager[***], and Ton Dijkstra[***]**

[*]ICS, Utrecht University
P.O. Box 80089, 3508 TB Utrecht, The Netherlands
johank@cs.uu.nl
[**]Communication & Information Sciences, Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands
p.a.vogt@uvt.nl
[***]NICI, Radboud University Nijmegen
P.O. Box 9104, 6500 HE Nijmegen, The Netherlands
w.haselager@nici.ru.nl, dijkstra@nici.ru.nl

## Abstract

This study investigates to what extent more advanced joint attentional mechanisms, rather than only shared attention between two agents and an object, influence the results of language games played by these agents. We present computer simulations which show that adding constructs that mimic follow attention capabilities increases the performance of agents in these language games substantially. Using follow and direct attention mechanisms, but without Theory of Mind-like capabilities, the agents are able to develop a shared lexicon much faster than when using only checking attention or corrective feedback. These results support the hypothesis that language evolution and evolutionary Theory of Mind develop in a co-evolutionary way, and that joint attentional skills are necessary and sufficient prerequisites for both.

## 1. Introduction

An important prerequisite of a successful conversation is the participants' ability to engage in *joint attention* in order to understand each other. This is not a coincidence. For young children, the ability to share attention with an adult concerning a third object or actor is a very important step in their language development. Tests like the Intentionality

Detector or the Eye Direction Detector that examine various aspects of joint attention, have shown that infants acquire joint attention skills at approximately the same age they start to learn their first words (Baron-Cohen, 1995). According to Tomasello (1999), the ability to engage in joint attention may have been the crucial *mechanism for cultural learning*, enabling mankind to rise from stone age to modern culture and technology in relatively short time. To be able to engage in joint attention might be a crucial prerequisite in *language evolution* as well.

Recent computational studies (e.g., Oliphant, 1999; Steels, 2001; Steels & Kaplan, 2002; Vogt & Coumans, 2003) used *language game models* to investigate how agents (either robots or software agents) learned the meaning of words, i.e., developed a common lexicon. In these language games a population of agents, situated in a common but changing environment, repetitively exchange utterances for concepts (like shape or color) present in the current environment, until a common lexicon for these concepts has emerged. Although only a small aspect of language is considered, namely the establishment of a common lexicon in a very simplified simulation setting, the findings can be used as indirect evidence for language evolution. Steels (1999) argued that these simulations provide valuable evidence, because the emerging structures (in this case, the common lexicon) are based on the properties and dynamics of a population of autonomous agents. According to Steels (1999):

> *In such investigations, it becomes quite natural to study language evolution. For example, one can test whether agents with a particular architecture enabling them to construct and acquire a lexicon, indeed arrive at a shared lexicon, whether this lexicon is resistant to changes in the population, whether it scales up to large numbers of meanings and agents, under what conditions shifts in meaning might occur, etc.* (p. 8)

If there are large differences in the effects of basic cognitive social skills (such as feedback and joint attention) on the outcome of these language games, it is plausible to suggest that similar effects play a role in language evolution. For instance, if joint attention is indeed a crucial prerequisite for the establishment of a common lexicon in language games (for instance, when without joint attention, lexicon establishment is found to be far less successful), this suggests that early hominids needed to have these capabilities before more advanced language usage could emerge. In Vogt and Coumans (2003), different types of language games are related to the type of non-verbal interaction strategies the agents were able to use. The *guessing games* used corrective feedback, *observational games* used joint attention as a strategy, and *cross-situational learning games*[1] used no non-verbal interaction strategy at all, but used statistical learning instead. In Vogt and Coumans (2003) these types of games were compared and no significant difference between the effects of corrective feedback and joint attention was found. Significant differences were found, however, in robotic experiments (Vogt, 2000), and in experiments involving grammar induction (Vogt, 2005). The cross-situational learning games, however, led to much worse performance than both guessing games and observational games in all studied cases (Vogt, 2000; Vogt & Coumans, 2003).

The implementation of Vogt and Coumans used rather abstract and simplified notions of joint attention and corrective feedback. In both cases, the meaning intended by the speaker was explicitly transferred (either simultaneously with the verbal interaction in case of joint attention or after evaluation of the game in case of corrective feedback). This explicit transfer, though useful to investigate some properties of language evolution computationally, is unrealistic and even makes communication redundant, especially in the observational game that uses joint attention (Vogt & Coumans, 2003; Smith, 2001). For humans, joint attention concerns a real world object or event, not a private meaning such as a color. In the literature on child language acquisition the term joint attention refers to a set of skills that can be categorized in three distinct stages: checking attention, following attention, and directing attention (Carpenter, Nagell, & Tomasello, 1998). Typically, only the first stage (*checking attention*) is modeled in language games. In this paper, we investigate whether there is a significant difference in performance between these stages, i.e., whether agents with following and/or direct attention skills arrive faster

[1] In Vogt and Coumans (2003), the cross-situational learning games were inappropriately called *selfish games*.

at a common lexicon. We have augmented the cross-situational learning games used in Vogt and Coumans (2003) with skills that represent *checking*, *following* and *directing* attention, and compared the results of the various language games with respect to the cross-situational game without these enhancements.

In the following Section 2, we further discuss the concepts of joint attention, Theory of Mind, and their relation. In Section 3, we discuss previous experiments and the implementation of more advanced stages of joint attention in the observational games. Method, results and discussions are presented in Sections 4, 5, and 6, respectively. Finally, we formulate some conclusions in Section 7.

## 2. Joint attention

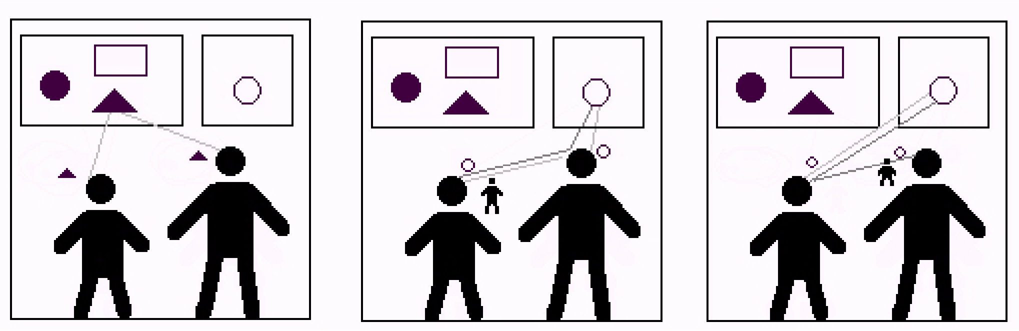
The term *joint attention* has been coined to describe a set of skills and interactions that emerge in infants of about nine months of age. Normally, at this age children begin to follow the gaze of their caregivers and engage with them in more complex social interactions that involve joint attention. The most prominent feature in these skills and interactions is that they are *triadic*: Whereas younger children typically either pay attention to a toy *or* their caregiver, the interactions of older children are usually more sophisticated and involve both the object and the other person (Baron-Cohen, 1995; Tomasello, 2000).

Carpenter et al. (1998) categorized various forms of joint attention (like joint engagement, gaze following, and point following) into three distinct stages, namely *checking* attention, *following* attention, and *directing* attention (see Figure 1). While the *follow* and *direct* stages differ in the passive versus active role of the child, the difference between merely sharing attention and following attention is somewhat more subtle. Carpenter et al. (1998) defined[2] these three stages as follows:

[2] It should be pointed out, that in Carpenter et al. the term *joint(attentional) engagement*, rather then *check attention*, is used to refer to the interactive form of sharing attention as described in this citation. We will use the more common term *check attention* to describe this behavior.

Figure 1

*Checking, following, and directing attention*



Checking attention:

> *By definition, all joint attentional skills involve infants sharing attention with a partner in some manner. We are concerned here, however, with relatively extended episodes of joint attentional engagement in which adult and infant share attention to an object of mutual interest over some measurable period of time (at least a few seconds). The prototypical example of an episode of joint attentional engagement is a situation in which adult and infant are playing with a toy and the infant looks from the toy to the adult's face and back to the toy. (…) Minimally, the infant must be engaged with an object on which the adult is also focused, then demonstrate her awareness of the adult's focus by looking to her face, and then return to engagement with the object. (p.5)*

Following attention:

> *It is difficult to know what infants understand of their social partners as intentional agents when they are looking to them and engaging with them in these extended periods of joint engagement. But when infants begin to follow into the attention or behavior of others in certain specific ways, a much more compelling case can be made that they understand something about the other person as an intentional agent. In particular, infants may follow into the attention of others by following the direction of their visual gaze or manual pointing gesture to an outside object. (p.8)*

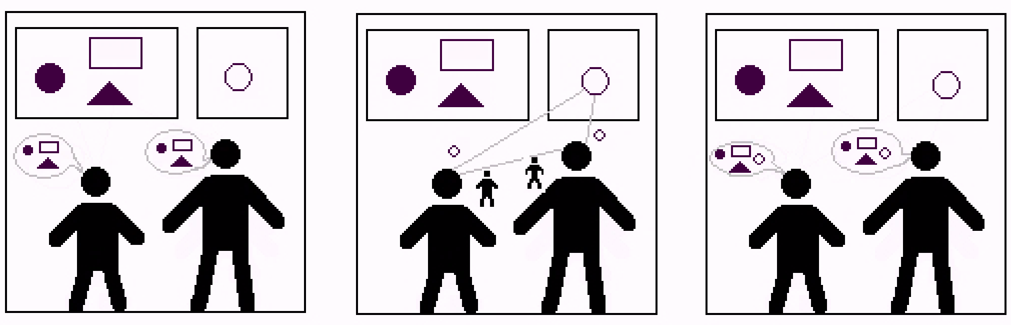
Directing attention:

> *Human infants demonstrate their understanding of adults as intentional agents, not only by following into their attention and behavior, but also by attempting to direct their attention and behavior to outside entities through acts of intentional communication. (p.17)*

These definitions imply that in the *checking* stage the ‘third object’ *is already* within the scope of the two agents (like child and adult), for example because it was physically *given* to the child by the adult to hold it in its hands, whereas in the *follow* and *direct* stage the third object is *brought into* scope, by the adult or the child. In Figure 2, the difference between share and direct attention is sketched. In the direct attention stage, the scope is *extended* when the infant directs the adult’s attention to the rectangle outside the box. The adult, being able to understand the child as an intentional agent, follows the attention of the visual gaze of the infant, bringing the rectangle into the scope of their shared attention.

Figure 2

*Scope of the agents in check versus direct attention*



Note that, in order for this scope-extension to succeed, both agents must be able to employ joint attentional capabilities. One cannot direct if the other cannot follow, and vice versa. In normal development, the child will first acquire follow attention, and later on, direct attention capabilities.

Closely related to joint attention is the concept of *Theory of Mind* (Premack & Woodruf, 1978). Having a Theory of Mind (hereafter ToM) means that one sees other actors as intentional agents like oneself, with comparable beliefs, desires, and intentions. It has been shown that very young children do not have a full-blown ToM. For example, children only pass ToM indicators like the False Belief Test (Wimmer & Perner, 1983) and the Opaque Context Test (Robinson & Apperlyb, 2001) after approximately four and five years of age, respectively. At this age, children know a considerable number of words (Bloom, 2000). Using tests like the Intentionality Detector or the Eye Direction Detector, to evaluate various aspects of joint attention, it has been shown that infants acquire joint attention skills at approximately the same age they start to learn their first words (Baron-Cohen, 1995). They know hundreds of words at 24 months of age, long before the False Belief Test or Opaque Context Test indicate the existence of a workable ToM, as shown in Table 1, which is adapted from Reboul (2003). As Reboul concluded from these data, a child needs some sort of joint attention skills in order to acquire a vocabulary, but from this base ToM and language acquisition develop in parallel rather then serially. It is clearly not the case that a workable ToM is required before the child starts to acquire a vocabulary. Nevertheless, the development of a ToM in the first years—for example, the ability to view other persons as *intentional agents,* demonstrated by complex social skills as social referencing or imitative learning (Tomasello, 1995)—undoubtedly facilitates further vocabulary development.

On the basis of these developmental data, Reboul suggests that language evolution and evolutionary ToM development follow the same pattern. They develop in a co-evolutionary way, rather than serially (ToM preceding language evolution). Basic joint attentional skills are necessary prerequisites for both ToM development and language evolution. Malle (2002) also suggests that ToM and language have evolved “coincidentally concurrent”, as mutual escalations utilizing advances from either side, or driven by a third factor. The hypothesis that ToM and language evolved as mutual escalations is supported by another observation in language acquisition. Although names of simple objects that play a role in the infant’s life are learned during the first years, children only use deictic relations[3] correctly at the age of three or four years, depending on the question whether the speaker's or the listener's perspective was taken (Pan, 2005). Furthermore, various studies suggest that autistic children are particularly impaired in this area (Tager-Flusberg, 1981). This suggests that the usage of these more advanced language constructs could emerge only after some sort of ToM evolved.

In this study, we investigate to what extent adding advanced joint attention mechanisms—without having ToM mechanisms—facilitates language learning in language games. In the next section, we will discuss language games and enhancements to these games that capture the nature of following and directing attention.

Table 1

*Age, Language Development and ToM Development*

| **Age** | **Language development** | **ToM development** |
|---|---|---|
| 0-9 months | | ID and EDD |
| 9-18 months | Going from 6 to 40 words | SAM |
| 24 months | 311 words | Development of ToM |
| 30 months | 575 words | Development of ToM |
| 48 months | Further development of vocabulary | False Belief Test |
| 60 months | Further development of vocabulary | Opaque Context Test |

*Note:* data from Reboul (2003). ID: Intentionality Detector; EDD: Eye Direction Detector; SAM: Shared Attention Mechanism. See Baron-Cohen (1995) for a discussion of these mechanisms.

[3] Deictic relations are relations whose referents depend on the speakers' perspective, like 'X is *behind* Y'. Children typically have difficulties specifying relations as they are experienced by another person, for example 'to my right, and left for those of you watching at home…'.

## 3. Language games

In the language game model, introduced by Steels (1996), a population of agents (either software agents or physical robots) tries to develop a shared lexicon using communicative actions in a particular environment (e.g., a whiteboard with colored geometrical objects), where a lexicon is a set of associations between words (strings of characters) and meanings (features of objects). Such language games are typically played between two agents; one of them (the *speaker*) trying to label a feature of an object in its attentional view, while the other (the *hearer*) tries to identify this feature based on this uttered label. Vogt and Coumans (2003) describe three types of language games, using shared attention (observational game), corrective feedback (guessing game), or no feedback or joint attention at all (cross-situational learning game). Observational games were developed by Oliphant (1999), guessing games became familiar through the Talking Heads experiments (Steels, Kaplan, McIntyre, & van Looveren, 2002), and cross-situational learning games were independently developed by Smith (2001) and Vogt (2000). The current study enhanced the cross-situational learning games with elements that mimic the three stages of joint attention.

In the cross-situational learning games, agents are given a context containing a number of virtual objects, each represented by a number of features, such as shape, color and size, like *color-of-object1*. The value of such a feature is denoted as a meaning, e.g., the value of *color-of-object1* could be denoted as the meaning ‘red’. The speaker selects one of the available meanings available in the context and conveys this using a verbal utterance. The hearer then guesses what meaning in the context could be denoted by this utterance. This guess is based on the co-occurrence frequencies with which the utterance and meanings co-occurred in different contexts or situations; the association with the highest co-occurrence frequency is selected. This cross-situational learning mechanism is similar to that proposed by Siskind (1996). It has been shown mathematically that when learning from an idealized input, cross-situational learning is very robust against context size (i.e., the ratio between context size and lexicon size can be very high), though the time it takes to learn a lexicon increases with increasing context sizes (Smith, Smith, Blythe, & Vogt, 2006). However, the idealized input assumes a strict one-to-one mapping between form and meaning in the predefined lexicon, the input is consistent such that each utterance always co-occurs with at least the intended meaning (or feature) and the input is presented to the learner with a uniform distribution. When one deviates from these idealized assumptions, cross-situational learning appears to be much harder, though not infeasible (Vogt & Coumans, 2003; Vogt & Divina, 2007).

There is increasing evidence that children can and do use cross-situational learning as a mechanism for learning word-meaning-mappings (Akhtar & Montague, 1999; Houston-Price et al., 2005; Klibanoff & Waxmann, 2000; Mather & Schafer, 2004; Smith & Yu, 2007). While other learning mechanisms, like imitation, may also play a role in child language acquisition, we take the stance that cross-situational learning is the basis of all *associative* learning of word-meaning mappings and that joint attention mechanisms, possibly together with other constraints and biases such as mutual exclusivity (Markman, 1989), the principle of contrast (Clark, 1993) and the whole object bias (Macnamara, 1982), merely serve to reduce the learning context.

In the current study, we will assume that all agents are cross-situational learners, but they have—in contrast to previous studies—joint attentional mechanisms to reduce the context size. As it has been shown in various previous studies that the smaller the context size, the faster lexicons converge in a population (Smith et al., 2006; Divina & Vogt, 2006), we predict that in simulations that use joint attention mechanisms, the lexicons will converge faster.  The question is to what extent the three proposed joint attention mechanisms yield better performances and whether there are optimal combinations of mechanisms that agents can use.

## 4. Methods

Simulations were run with a population containing 10 agents, each starting with an empty lexicon. The agents were situated in a virtual environment containing a number of objects, each formed as a 3 dimensional vector that can have 4 different values in each dimension. Each position in one dimension is called a *feature* of the particular object and could be interpreted as, for instance, a color, a shape, or a size. So, in total there are

$4^3$=64 different objects in this environment, composed of 3x4=12 meanings $m_j$ (each feature corresponds directly to one meaning, e.g., the feature “color-of-object-1” could correspond to meaning “red”). Note that this differs from the method used in Vogt and Coumans (2003), where each meaning was represented as an integer, corresponding to a whole object.

Each agent is equipped with a private lexicon that is represented as an association matrix that associates forms $w_i$ with meanings $m_j$. Initially, each agent has an empty lexicon; the lexicons are constructed while playing language games. Each association is given a weight $w_{ij}$ that is calculated as the a posterior co-occurrence probability $P(m_j|w_i)$ as follows:

$$w_{ij} = P(m_j \mid w_i) = \frac{u_{ij}}{\sum_j u_{ij}}$$

In all simulations, each time a language game is played, two agents are selected from the population at random, one is randomly assigned the role of speaker, the other the role of hearer. Four objects are selected arbitrarily from the environment to form the *situation S*. The *context* $C_S$ of this situation is defined as the set of all features $f_j$ of all objects $O_i$ Î $S$. The speaker selects one random object $O_t$ Î $S$ from this context as the *topic* and from this object, it selects one arbitrary feature $f_t$ Î $O_t$ to form the *target*. The speaker then tries to produce an utterance by searching its lexicon for a form that has the highest weight with the target meaning. If no such form is found, the speaker invents a new form as a random string, adds the form-target pair to its lexicon and utters this novel form.

In turn, the hearer tries to interpret the utterance by searching its lexicon for the association of which the meaning is consistent with one of the potential meanings available in the context $C_S$ and that has the highest weight. Depending on the type(s) of joint attention mechanism(s) agents use in a language game, the context $C_S$ is adjusted to form the learning context $C_L$. The hearer adapts its lexicon by increasing the co-occurrence frequencies $u_{ij}$ between the form $w_i$ and all meanings $\mathrm{m_j}$ in the learning context $C_L$ by 1. If an association between form and meaning does not exist, it is added to the lexicon before updating its co-occurrence frequency. The following explains how the joint attention mechanisms can change the learning context:

**Check attention.** When agents use check attention, both the original context $C_S$ and the learning context $C_L$ are set to all meanings representing the topic $O_t$ selected by the speaker, i.e. $C_S = C_{LC} = O_t$. This is done prior to the verbalization, so it also influences the interpretation process.

**Follow attention.** Here the speaker selects a random object $O_r$ from the situation $S$ that contains the same feature as the target, i.e., $f_t$ Î $O_r$ and that is different from the topic $O_t$. This object is then communicated to the hearer, thus modeling the hearer's *following* of the speaker's new attention. After the hearer has interpreted the speaker's utterance, the hearer constructs the learning context by taking the cross-section of the meanings corresponding to this new object and the original context, i.e. $C_{LF}$=$C_S$ Ç $O_r$.

**Direct attention.** This mechanism is slightly more complicated. Here the hearer selects, after interpretation, a random object $O_h$ from the situation $S$ that contains the same feature as the interpreted meaning, i.e., $f_t$ Î $O_h$ and communicates this object to the speaker, thus modeling the *directing* of the speaker's attention to $O_h$. The speaker then signals whether this object contains the intended target or not. If it does, the hearer constructs the learning context $C_{LD}$ by taking the cross-section of the original context with the novel object, i.e., $C_{LD}$ = $C_S$ Ç $O_h$. If the speaker signals that the object does not contain the target, then the context is refined by taking the *complement* of the original context $C_S$ and the new object, i.e. $C_{LD} = C_S - O_h$.

We have carried out eight series of simulations where we varied the different joint attention mechanisms available to the agents. The eight simulation series correspond to the eight different combinations of having none, one or more of the attention mechanisms available, as shown in Table 2. In the different conditions, each language game used all available mechanisms in the order as proposed by Carpenter et al. (1998). So, if all mechanisms were available, the agents would first establish joint attention by checking attention, then the hearer would interpret the speaker's utterance, after which the learning context is first refined by following attention and then by directing attention. Only after the joint attention mechanisms are processed, the hearer adapts the co-occurrence frequencies of the utterance with the meanings that remain in the learning context $C_L$. (Note that the speaker always increments the co-occurrence frequency of the utterance and the target.)

Table 2.

*The eight different simulation series and the attention mechanisms switched off (-) or on (+). The final column shows how the learning context $C_L$ is constructed.*

| | Name | Check attention | Follow attention | Direct attention | $C_L$ |
|---|---|---|---|---|---|
| 1 | **xxx** | - | - | - | $C_S$ |
| 2 | **xfx** | - | + | - | $C_{LF}$ |
| 3 | **xxd** | - | - | + | $C_{LD}$ |
| 4 | **xfd** | - | + | + | $C_{LF}$ Ç $C_{LD}$ |
| 5 | **cxx** | + | - | - | $C_{LC}$ |
| 6 | **cfx** | + | + | - | $C_{LC}$ Ç $C_{LF}$ |
| 7 | **cxd** | + | - | + | $C_{LC}$ Ç $C_{LD}$ |
| 8 | **cfd** | + | + | + | $C_{LC}$ Ç $C_{LF}$ Ç $C_{LD}$ |

It is worthwhile comparing the above types of language games with analogous studies in Vogt and Coumans (2003). As mentioned, all games implement cross-situational learning. However, the game indicated by **xxx** (i.e., the game that does not either checking, following or directing attention) is most similar to the cross-situational game used in Vogt and Coumans, as well as in all other implementations of cross-situational learning (e.g., Smith 2001; De Beule, De Vylder & Belpaeme, 2006; Smith et al,. 2006).

The game indicated by **cxx** is most similar to the observational game, though in Vogt and Coumans (2003) the hearer was informed about the target meaning while here the hearer is informed about the topic, which contains other meanings besides the target. This is more realistic, as we typically cannot, for instance, point to a feature of an object. Although this was also true in the robotic experiments described in Vogt (2000), there the assumption of a whole object bias (Macnamara, 1982) was adopted.

Games **xfx**, **xxd** and **xfd** most closely resemble the guessing game, but with a fundamental difference here. In the guessing game, the hearer first directs the attention to its guess, after which the speaker acknowledges success or failure (similar to **xxd**). Moreover, in case of a failure, the speaker points at the topic so that the hearer can acquire the right association. Although this is similar to following attention, the guessing game does not reduce to **xfd**, but rather to something like **xdf**—the order in which following and directing attention is applied is reversed. However, in the original guessing game the hearer only reinforces (when successful)[4] or inhibits (when unsuccessful) the weight of the used association, rather than adjusting its learning context and adapting the associations of the utterance with all meanings in this context.

We realize that the simulations carried out are still far from reality, as humans do not learn by using only one type of interaction that uses either none, one, or all possible strategies available to them. Instead, humans use different strategies in different interactions, constrained by what is available to them. Moreover, children learn from hearing complex multiword utterances rather than from one word utterances, and they understand a whole range of privately acquired concepts, rather than a limited set of pre-defined meanings. Nevertheless, the current set up of the experiment allows us to investigate—on the basis of the proposed model—the effects of different joint attention mechanisms on the emergence of a lexicon.

[4] In a successful guessing game, the weight of associations that compete with the one used (i.e. that have either the same form or the same meaning) are laterally inhibited in addition to reinforcing the used association.

## 5. Results

Series of simulations were run with each of these different game models, where each language game model was run 100 times with different random seeds for 100,000 language games or until communicative accuracy reached 100% for 10 language games in a row. *Communicative accuracy* is defined as the number of correctly played games averaged over the last 100 games. A game was played correctly if the hearer guessed the target meaning (i.e., feature) intended by the speaker.

We also measured the hearer's *learning context size*, which we define as the number of features (or meanings) in the learning context ($C_L$) after joint attention has been processed and with which the lexicon is learnt. Furthermore, we measured *time of convergence* as the number of games for communicative accuracy to become equal to 1 for ten games in a row. The means and standard deviations of communicative accuracy, context size, and time of convergence are presented in Table 3. Communicative accuracy and time of convergence for the different conditions are also shown in Figures 3 and 4.

Table 3

*Means and standard deviation of communicative accuracy, context size, and time of convergence.*

| | **communicative accuracy** | | **learning context size** | | **time of convergence** | |
|---|---|---|---|---|---|---|
| | mean | std.dev | mean | std.dev | Mean | std.dev |
| **xxx** | 0.2522 | 0.0668 | 8.3252 | 0.0034 | 100,000 | 0 |
| **xfx** | 0.6986 | 0.1178 | 4.5641 | 0.0116 | 97,471 | 13,362 |
| **xxd** | 0.3404 | 0.0737 | 5.4107 | 0.0117 | 100,000 | 0 |
| **xfd** | 0.7298 | 0.1227 | 4.2050 | 0.0977 | 94,641 | 19,887 |
| **cxx** | 0.9184 | 0.0977 | 3 | 0 | 66,147 | 35,461 |
| **cfx** | 1 | 0 | 2.0926 | 0.0159 | 2,403 | 729 |
| **cxd** | 0.9968 | 0.0224 | 2.1240 | 0.0057 | 18,546 | 19,882 |
| **cfd** | 1 | 0 | 2.0650 | 0.0045 | 2,223 | 431 |

Between the various language game models, communicative accuracy differed significantly (F(7,792) = 1473, $p < 0.0001$), as was the case for time of convergence (F(7,792) = 727, $p < 0.0001$). To compare the effects of the check attention mechanism with more advanced mechanisms, we submitted the convergence time scores of the language games to a Two-Factor ANOVA, with check attention (yes/no) and follow/direct attention (none/follow/direct/follow and direct) as the between-subject variables. The most interesting significant result here was the interaction between having or lacking a check attention mechanism, and having or lacking follow and direct attention mechanisms (F(3,792) = 174, p < 0.0001). In the conditions without check attention mechanisms, the communicative accuracy of most game models did not converge to 1 within 100,000 games. Nevertheless, the communicative accuracy was much lower in the **xxx**-game model (0.25) than in the **xfd**-game model (0.73). On the other hand, in the conditions with check attention mechanisms, the communicative accuracy of most games converged to 1 within 100,000 games, but the time of convergence was much slower in the **cxx**-game model (66,147) compared to the **cfx**- (2,403) and **cfd**- (2223) game models, and—to a lesser extent—to the **cxd**-model (18,546).

Figure 3

*Average Time of Convergence*

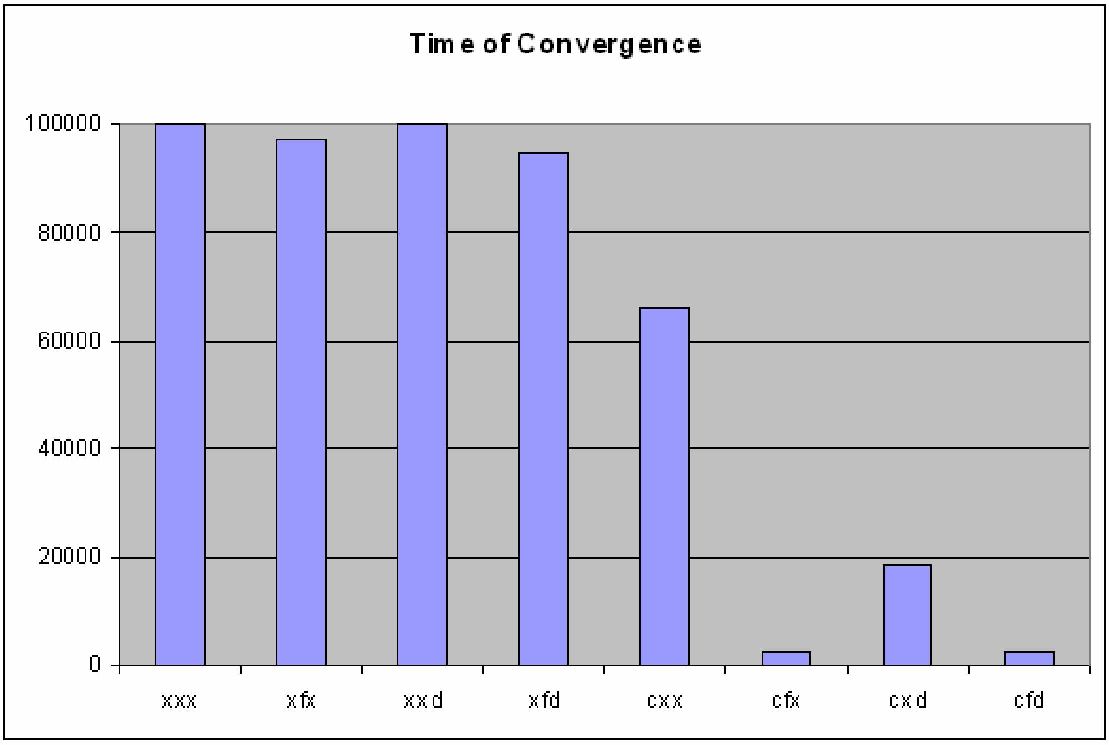
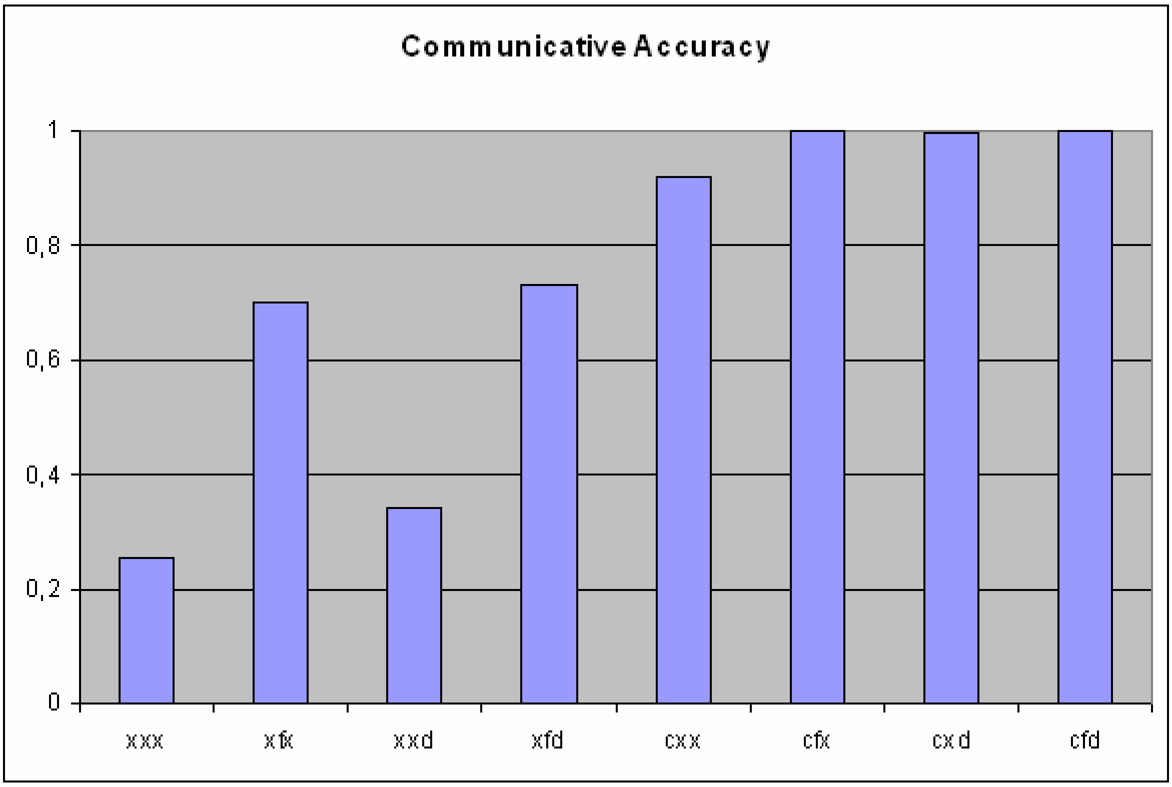


Figure 4

*Communicative Accuracy*



The learning context size differed significantly between the language game models (F(7,792) = 368468, $p < 0.0001$). While the **xxx**-game model had an average context size of 8.3252, all game models which used some kind of joint attention mechanism were able to decrease the context size, to an average ranging from 4.5641 (**xfx**-game model) to 2.0650 (**cfd**-game model).

The value of 3 for the **cxx-**game mode can be understood by realizing that the learning context is set to the 3 features of the topic. Only when attention is further refined through follow attention and/or direct attention, the context size becomes lower.

Interestingly, the strategy that yields lowest context size (i.e., both follow and direct attention) also yields best performance in terms of communicative success and time of convergence, which is consistent with our prediction. But if we compare the results between follow and direct attention, rather than their combination, then the follow attention strategy yields the best performance on all indicators. So, of the different stages of joint attention, follow attention seems to contribute most to the performance of the simulations. Direct attention alone yields better performance than check attention, but performs worse than follow attention.

## 6. Discussion

The results showed dramatic improvements in performance for two of the attention mechanisms: checking attention and following attention. When the checking attention mechanism was absent, none of the conditions yielded a communicative accuracy near 100%. Nevertheless, following and (to a lesser extent) directing attention yielded significant improvements relative to the simulations where either is absent. When the checking attention mechanism was available, they all (nearly) converged to 100% communicative accuracy. Here following and (to a much lesser extent) directing attention affected the time of convergence drastically from around 66,000 language games for checking attention to about 18,000 games for directing attention and 2,500 games for following attention. From these results, it is clear that the use of these joint attentional enhancements, in particular checking and follow attention, have a large impact on these language games.

This conclusion is in line with the developmental data from Reboul (2003). Whereas infants knew only a few to a couple of dozen words at the moment they were developing their joint attentional skills (from nine to eighteen months of age), this number rapidly increased to over three hundred in the following half year, when they were able to use these skills. The catalyst in this rapid increase is thus not the ability to share attention—which is typically acquired after 9-12 months of age—but the ability to follow and direct attention, acquired after 11-15 months of age (Carpenter et al. 1998).

The results suggest that the ability to *follow* attention is more crucial than the ability to *direct* attention. This can be understood on the basis of the context of the language games. In the follow attention enhancements of these games, the listener follows the attention of the speaker, who will try to find a consistent alternative to the object currently in scope of the game. In cases when such an alternative is present, this object will have only one or two feature(s) in common with the current object, thus reducing the learning context of the language game to one or two feature(s). In contrast, in the direct attention enhancements the listener *guesses* an alternative object and seeks feedback from the speaker. This alternative might form a counterexample that does *not* have any features in common with the speaker's topic, and thus the listener must learn from the (typically larger) *complement* set of features, which is less effective (Smith et al., 2006). Thus, while both enhancements facilitate cross-situational learning by decreasing the context size, follow attention in general leaves a smaller subset of the original context to be considered.

The ability to follow attention precedes the ability to direct attention in child development. While our results strongly suggest that the former is more important in initial vocabulary development, it must be pointed out that the concepts we used were distinct from each other, i.e., not hierarchically organized. Clark (1993) reported that up to one-third of the vocabulary of one- and two-year old children consists of overextension (e.g., calling all adult man 'daddy') or underextension (e.g., using 'bird' only for birds that can fly), whereas these unconventional mappings are rare beyond age two-and-a-half. It is reasonable to suggest that direct attention mechanisms are particularly useful in later stages of language development to further specify the right subset of contexts allowed for a particular concept.

Although Vogt and Coumans (2003) use slightly different parameters and measurements[5], their findings for the communicative accuracy of the observational and cross-situational games are comparable with our results for games models **cxx** and **xxx**. In contrast to our simulations, in which **cxx** scored notably better than **xfx**, **xxd** and **xfd**, they found no significant difference between the observational and guessing games. It should be noted, though, that there is a fundamental difference between the guessing

[5] For example, the context size was five, rather than four, and instead of objects with three distinguishable attributes with four values each (thus $4^3 = 64$ meanings), 100 arbitrary meanings were used to describe objects.

game and the **xfx**, **xxd** and **xfd** game models, as discussed in Section 4, which makes comparison difficult.

In Vogt and Coumans (2003), problems were reported regarding the scalability of cross-situational learning in multi-agent systems. The larger the population, the more difficult convergence could be achieved. The current simulations, which were done with a population size of 10, show that adding joint attentional mechanisms to cross-situational learning may solve that problem. Further research is required to investigate the scalability of this model in more detail.

One approach in which the model's scalability is studied, is currently investigated in the context of the NEW TIES project, which aims to study the evolution of an artificial cultural society  (Gilbert et al., 2006). In this project, large populations of virtual robots (i.e., agents that have some embodiment and who are fully autonomous) operate in an environment containing various objects (such as food sources) with various features about which the agents communicate and develop a shared vocabulary. In this environment, the visual context can be rather large, so establishing joint attention is required to achieve communicative accuracy. The model of Vogt and Divina (2007) is currently enhanced to allow more elaborate multimodal dialogues to improve establishing joint attention, including the following and directing attention mechanisms.

## 7. Conclusions

In this study we have investigated whether employing more advanced stages of joint attention (i.e., follow and direct attention, rather than check attention) improve lexical development in simulations with language games. We argue that the crucial distinction between these and the earliest stage of joint attention is the *scope* of the shared attention. While the objects of shared attention in check attention are physically 'put' into scope (e.g., by giving a toy to an infant to hold it in its hands), the scope can be extended in later stages by initiative of the adult (the child following attention) or by the child (directing the attention of the adult). We modeled this *scope extension* by augmenting the agents in the language games with a 'toolbox' of methods that typically require follow attention (the speaker brings another object, also having the desired property, into scope) or direct attention (the hearer inquires whether a specific object also has this property).

This scope extension can (and typically does) reduce the context in which hearers learn word-meaning mappings. As a result, our simulations yield dramatic improvements in performance when the follow attention mechanism is added to language games that use check attention. Obviously, some kind of *mechanism* is needed to reduce the theoretically infinite number of possible word-meaning mappings in language acquisition (Quine, 1960). On the basis of our results we can conclude that follow and—to a lesser extent—direct attention mechanisms are good candidates for such a mechanism.

We have taken the stance that the prime mechanism for associative learning of word-meaning mappings is *cross-situational learning* (Siskind, 1996; Vogt, 2000; Smith, 2001). As previously shown, performance of cross-situational learning can improve substantially when the contexts from which agents learn are smaller (Divina and Vogt, 2006; Smith et al., 2006). The joint attentional mechanisms we have modeled all reduce the learning context, thus improving learning. Hence, our study indicates that cross-situational learning, enhanced with joint attention mechanisms (and possibly other mechanisms too), is a robust and realistic learning mechanism by which children can learn word-meaning mappings.

## References

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

Baron-Cohen, S. (1995). *Mindblindness: An Essay on Autism and Theory of Mind.* Cambridge, MA: MIT Press.

Bloom, P. (2000). *How Children Learn the Meanings of Words.* Cambridge, MA: MIT Press.

Carpenter, M., Nagell, K, & Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development, 63(4).*

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

De Beule Joachim, De Vylder Bart and Belpaeme Tony (2006) A cross-situational learning algorithm for damping homonymy in the guessing game. In L.M. Rocha, L.S. Yaeger, M.A. Bedau, D. Floreano, R.L. Goldstone and E.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 and C. Nehaniv (Eds.), *Symbol Grounding and Beyond: Proceedings of the Third International Workshop on the Emergence and Evolution of Linguistic Communication* (p. 1–15). Berlin: Springer.

Gilbert, N., den Besten, M., Bontovics, A., Craenen, B.G.W., Divina, F., Eiben, et al. (2006). Emerging Artificial Societies Through Learning. *Journal of Artificial Societies and Social Simulation 9(2).*

Houston-Price, C., Plunkett, K., Harris, P. (2005) 'Word-Learning Wizardry' at 1;6. *Journal of Child Language 32(1)* 175–189

Klibanoff, R. S. and Waxman, S. R. (2000) Basic level object categories support the acquisition of novel adjectives: Evidence from preschool-aged children. *Child Development* 7(3): 649-659

Macnamara, J. (1982). *Names for things: a study of human learning*. Cambridge, MA: MIT Press.

Malle, B.F. (2002). The relation between language and theory of mind in development and evolution. In T. Givón and B. F. Malle (Eds.), *The evolution of language out of pre-language (p. 265–284).* Amsterdam: Benjamins.

Markman, E.M. (1989) Categorization and naming in children: problems of induction. Cambridge. MA: MIT Press.

Mather, E. and Schafer, G. (2004) Object-label covariation: A cue for the acquisition of nouns? *Poster presented at the meeting of the International Society of Infant Studies.* Chicago.

Oliphant, M. (1999). The learning barrier: Moving from innate to learned systems of communication. *Adaptive Behavior, 7,(3-4), 371–384.*

Pan, B.A., & Gleason, J.B. (2004). Semantic Development: Learning the Meaning of Words. In Gleason (ed.), *The development of language* (6th ed.). Needham Heights, MA: Allyn & Bacon/Pearson Education.

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

Quine, W.V.O. (1960). *Word and object*. Cambridge, MA: MIT Press.

Reboul, A. (2004). Evolution of Language from Theory of Mind or Coevolution of Language and Theory of Mind? In: *Issues in Coevolution of Language and Theory of Mind*. Retrieved September 20th, 2007, from http://www.interdisciplines.org/coevolution/papers/1.

Robinson, E.J., & Apperlyb, I.A. (2001). Children’s difficulties with partial representations in ambiguous messages and referentially opaque contexts. *Cognitive Development, 16, 595–615.*

Siskind, J.M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition 61(1-2), 39–91.*

Smith, A.D.M. (2001). Establishing Communication Systems without Explicit Meaning Transmission. In J. Kelemen and P. Sosik (Eds.), *Proceedings of the 6th European Conference on Artificial Life, ECAL 2001, LNCS 2159 (p. 381–390).* Berlin: Springer.

Smith, K., Smith, A.D.M., Blythe, R., & Vogt, P. (2006) Cross-situational learning: a mathematical approach. In P. Vogt, Y. Sugita, E. Tuci and C. Nehaniv (Eds.) *Symbol grounding and beyond: Proceedings of Emergence and Evolution of Linguistic Communication III, LNAI 4211.* Berlin: Springer.

Smith, L. B. & Chen, Y. (2007). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition.* In press.

Steels, L. (1996). Emergent adaptive lexicons. In P. Maes (Ed.), From animals to animats 4: *Proceedings of the Fourth International Conference on Simulating Adaptive Behavior*. Cambridge, MA: MIT Press.

Steels, L. (1999). The Puzzle of Evolution. *Kognitionswissenschaft, 8(4), 143–150.*

Steels, L. (2001). Language games for autonomous robots. *IEEE Intelligent Systems, 16(5), 16–22.*

Steels, L., & Kaplan, F. (2002). Bootstrapping grounded word semantics. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition: formal and computational models*. Cambridge, UK: Cambridge University Press.

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

Tager-Flusberg, H. (1981). On the nature of linguistic functioning in early infantile autism. *Journal of Autism and Developmental Disorders, 11(1), 45-56.*

Tomasello, M. (1995). Joint attention as social cognition. In C. Moore and P. Dunham (Eds.), *Joint attention: its origins and role in development.* Lawrence Erlbaum Associates.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition.* Harvard University Press.

Tomasello, M. (2000). The item based nature of children's early syntactic development. *Trends in Cognitive Sciences, 4, 156-163.*

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

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

Vogt, P., & Coumans, H. (2003). Investigating social interaction strategies for bootstrapping lexicon development. *Journal of Artificial Societies and Social Simulation* 6(1).

Vogt, P., & Divina, F. (2007). Social symbol grounding and language evolution. *Interaction Studies 8(1): 31–52.*

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function in wrong beliefs in young children’s understanding of deception. *Cognition, 13, 103–128.*