"Catchline" (i.e. wording at head of first page only) *Connection Science*
Vol. X, No. X, Month 200X, 000–000

RESEARCH NOTE

Joint attention and language evolution

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

[a]*ICS, Utrecht University, Utrecht, The Netherlands;* [b]*ILK, Tilburg University, Tilburg, The Netherlands;* [c]*NICI, Radboud University, Nijmegen, The Netherlands*

This study investigates how more advanced joint attentional mechanisms, rather than only shared attention between two agents and an object, can be implemented and how they influence the results of language games played by these agents. We present computer simulations with language games showing that adding constructs that mimic the three stages of joint attention identified in children's early development (checking attention, following attention and directing attention) substantially increase the performance of agents in these language games. In particular, the rates of improved performance for the individual attentional mechanisms have the same ordering as that of the emergence of these mechanisms in infants' development. These results suggest that language evolution and joint attentional mechanisms have developed in a co-evolutionary way, and that the evolutionary emergence of the individual attentional mechanisms is ordered just like their developmental emergence.

Keywords*:* language acquisition; language games; language evolution; joint attention

*Corresponding author. Email: johank@cs.uu.nl

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, which examine various aspects of joint attention, have shown that infants acquire joint attention skills at approximately the same age as 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. The ability to engage in joint attention may also be a crucial prerequisite for *language evolution*.

In the literature on child development, the term joint attention refers to a set of skills that can be categorised in three distinct stages: checking attention, following attention, and directing attention (Carpenter, Nagell & Tomasello, 1998). These three stages mainly differ in the way joint attention is initiated: checking attention involves a natural sharing of attention without a clear initiator, following attention involves the caregiver directing the attention of the infant to an object, and directing attention involves the infant directing the attention of the caregiver to an object. They also differ in the way objects are brought into the scope of attention: in checking attention the attended object is already in the scope of attention, in following attention and directing attention the attended object is explicitly brought into the scope of attention. Carpenter et al. (1998) found that the order in which the three stages arise is: checking attention (month 9-10), followed by following attention (month 10.5) and only then directing attention (month 12.6). This ordering that has been confirmed in other studies for autistic children (Carpenter, Pennington & Rogers, 2002) and healthy children (Mundy et al., 2007). Moreover, the frequency with which children use these three attentional mechanisms between 9 and 18 months of age predicts how well these children perform at language development tests at 24 months (Mundy et al., 2007).

In this paper we address three questions: 1) How can we model these three joint attentional mechanisms with respect to other models of language evolution? 2) To what extent do the different attentional skills contribute to language emergence in language game models? 3) Why do we find exactly this ordering in the emergence of attentional skills and not another? We assume that joint attentional skills are used to reduce the uncertainty with which the meanings of words can be inferred and that the different skills contribute differently in reducing this uncertainty. Assuming a cross-situational learner (Siskind, 1996), we know that the more uncertainties there are in the learning situations, the longer it takes to learn a set of word-meaning mappings (Smith, Smith, Blythe & Vogt, 2006). Although the ordering of emergence in attentional skills may be accounted for by different factors (e.g., complexity of the required skill), we will assume that those skills that contribute most to language development are most likely to have evolved first. If they have evolved first phylogenetically, it is not unlikely that they also develop first ontogenetically.

The computer model used for the current study is based on the *language game model* that investigate how agents (either robots or software agents) can develop a common lexicon, i.e. how they can develop a shared set of word-meaning mappings (e.g., Oliphant, 1999; Steels, 2001; Smith, 2005; Steels & Kaplan, 2002; Vogt & Coumans, 2003; Vogt & Divina, 2007). In these language games a population of agents, situated in a common but changing environment, repetitively exchange utterances for concepts (like shape or colour) present in the current environment, until a common lexicon for these concepts has emerged. The hard problem that the agents have to solve is the *social symbol grounding problem* (Cangelosi, 2006; Vogt & Divina, 2007):

how can a (large) group of agents arrive at a shared set of symbols? Various studies have shown that a one-to-one bias towards word-meaning mappings is required for learning them, i.e. learning mechanisms acquiring word-meaning mappings seem to require a pressure towards assuming one-to-one mappings (Oliphant, 1999; Smith, 2004). The question remains how such a bias is implemented, and to what extent this bias is a *strict* bias (i.e., using it as a fixed constraint, rather than as a tendency)? Seeing such a bias as a strict one could be wrong, because words can have different meanings and letting agents learn only one-to-one mappings would not allow this. The bias should rather be seen as pressure or tendency towards words having one meaning and vice versa.

In (virtual) robotic experiments agents communicate typically in situations containing many different objects and events, which after categorisation lead to even more meanings (see, e.g., Steels, 2005; Vogt, 2006, for overviews). Various strategies to reduce the complexity of the situations have been investigated. These include either sharing attention by means of pointing to the subject of the communication (Vogt, 2000), using corrective feedback regarding the interpretation of hearers (Steels & Kaplan, 2002, Vogt, 2003), a combination of sharing attention and corrective feedback (Steels, Kaplan, McIntyre & Van Looveren, 2002), or neither pointing nor feedback, but cross-situational learning (Vogt, 2000; Smith, 2005; De Beule, De Vylder & Belpaeme, 2006). Comparisons of various methods indicate that joint attention is very beneficial in terms of speed, that corrective feedback improves the quality of the emerging lexicon and that cross-situational learning only works under certain conditions (Vogt, 2000; Vogt & Coumans, 2003; Vogt, 2005; Vogt & Divina, 2007). One problem is that many of these implementations assume that sharing attention or corrective feedback hands over the meaning (sense) of a word almost explicitly[1], for instance, by assuming a *whole object bias* in which the meaning is assumed to be the meaning relating to the whole object (or referent) as in (Vogt, 2000). Although the whole object bias is realistic (Macnamara, 1982), children do not use it continuously. Moreover, it does not allow one to learn the meaning of, for instance, adjectives referring to features such as colours or shapes.

In the current study we will assume that cross-situational learning is the core learning mechanism that allows agents to infer the meaning of a word from the statistical co-variance of the word with its meaning. The speed with which agents can learn an idealised language using cross-situational learning is proportional to the complexity of the situation in which the agents communicate (Smith et al. 2006). Although under ideal circumstances cross-situational learning works well, a reduction of complexity in learning situations is required in more complex worlds (e.g., Vogt & Coumans, 2003). We investigate how the three joint attentional stages can be implemented to reduce the complexity of the situations. By doing so, we investigate how the different stages influence the emergence of lexicons and why the stages come in the ordering with which they emerge in children.

Although we only consider a small aspect of language development, 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 the kind of simulations such as presented here provides 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

[1] In many simulations the meanings actually are transferred explicitly (Steels, 1996; Oliphant, 1999; Vogt & Coumans, 2003).

> 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 the different joint attentional skills) on the outcome of these language games, it is plausible that similar effects play a role in language evolution. For instance, if checking attention is indeed a crucial prerequisite for the establishment of a common lexicon in language games (for instance, when without checking 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 the following section, we further discuss the concepts of Theory of Mind, joint attention, and their development in children. Next, section 3 discusses how joint attention relates to language evolution and language development. In Section 4, we describe the model and method used in this study. Results are presented in Section 5, followed by a discussion in Section 6. Finally, we formulate some conclusions in Section 7.

2. Theory of Mind and joint attention

What abilities separate mankind from other species? Among other suggestions, like bipedalism and tool fabrication, the ability to use and master a *complex language* and the possession of a *Theory of Mind* (Premack & Woodruf, 1978) are proposed as being unique to mankind. Having a Theory of Mind (i.e. the capacity for 'mind reading' or 'mentalising') means that one sees other actors as intentional agents like oneself, with comparable beliefs, desires and intentions, and that one can understand what other actors are thinking. While having a Theory of Mind (ToM) is necessary to engage in complex communicative behaviours, 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 that evaluate various aspects of joint attention, it has been shown that infants acquire *joint attention skills*, like gaze following and joint engagement, at approximately the same age as 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, although from this perspective ToM and language acquisition develop in parallel rather than serially. It is clearly not the case that a workable ToM is required before the child starts to acquire a vocabulary. Nevertheless, the development towards 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.

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.

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 (specifically: 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[2] correctly at the age of three or four years, depending on whether the speaker's or the listener's perspective was taken (Pan & Gleason, 2004). Furthermore, various studies suggest that autistic children are particularly impaired in this domain (Tager-Flusberg, 1981). This suggests that the usage of these more advanced language constructs could emerge only after some sort of ToM evolved.

The term *joint attention* describes a compound 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).

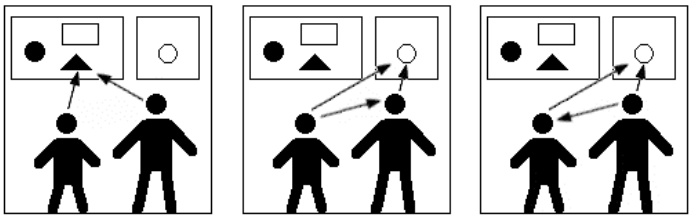

Figure 1. Checking attention (a), following attention (b) and directing attention(c).

[2] 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…'.

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. Figure 1 depicts these three stages. In the checking attention stage, both child and adult share attention to the black triangle and to each other. In the following attention stage, the child *follows* the attention of the adult to the white circle, and in the directing attention stage the child *directs* the attention of the adult to the white circle. While the *following* and *directing* stages differ in the passive versus active role of the child, the differences between checking attention and following or directing attention are subtler. Carpenter et al. (1998) described[3] these three stages as follows:

> 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 behaviour 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 behaviour, but also by attempting to direct their attention and behaviour to outside entities through acts of intentional communication. (p.17)

These descriptions 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 *following* and *directing* stages the third object is *brought into* scope by the adult or the child. In Figure 2, the difference between checking and directing attention is sketched. In the initial stage, the child and the adult share attention to the objects in the box on the left, which is the current scope of their shared attention (the encircled objects in Figure 2a). Through directing attention, the scope is *extended* when the infant directs the adult's attention to the circle in the box on the right (Fig. 2b). The adult, being able to understand the child as an intentional agent, follows the attention of the visual gaze of the infant, bringing the circle into the scope of their shared attention (encircled objects Fig. 2c). Normally, the child will check to see if the adult has followed its direction of attention, so both participants are aware that they share attention. The difference between following and checking attention is similar to the disparity between directing and checking, except that the initiative of shifting attention is taken by the adult instead of the child.

---

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

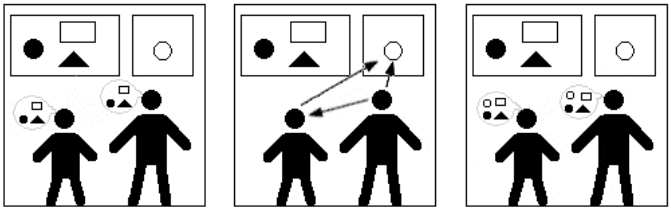

Figure 2. Scope of the agents in *checking* versus *directing* attention.

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 – after having acquired checking attention – first acquire following attention, and later on, directing attention capabilities.

3. Joint attention and language development

Research has indicated that using joint attentional skills correlates well with children's language development. It has been shown experimentally that children who learn new words in a joint attentional setting do better than children who learn them without a joint attentional setting (Tomasello & Todd, 1983). Autistic children, who reveal a different development of joint attentional skills than non-autistic children show correlated differences in language development (Dawson et al., 2004). Even regarding the various stages in joint attention, there appears to be a measurable correlation between the use of these stages of joint attention between 9 and 18 months old infants and the level of language competence at 24 months of age (Mundy et al., 2007). An interesting issue is how joint attention influences language development in children, i.e. what makes joint attention a mechanism that influences the ability to learn language.

The obvious possibility, which we will explore here, is that joint attention allows individuals to reduce the number of hypothetical meanings when learning a word's meaning. According to Quine (1960), each unfamiliar word that we learn can, in principle, mean an infinite number of things (the word "gavagai" expressed when a rabbit scurries by can mean 'rabbit', 'undetached rabbit parts', 'running furry animal', 'dinner', 'it will rain', and so on). So, in order to learn the meaning of a word, one must be able to reduce the number of possible hypotheses substantially. Various mechanisms have been proposed of which joint attention is just one (Bloom, 2000). Others include, for instance, mutual exclusivity (Markman, 1989), the principle of contrast (Clark, 1993), and the whole object bias (Macnamara, 1982). Even though the various mechanisms are not necessarily mutual exclusive, they will tend to fail in minimising the hypothesis set to one.

All these principles and constraints assume an underlying mechanism that stores and manages the associations between words and meanings. Various mechanisms have been proposed, but the most straightforward mechanism is associative –Hebbian– learning. In essence, this associative learning mechanism strengthens the

associations between a word and all meanings that apply in a certain situation (or *context*). When applied over varying situations, a word's meaning tends to co-occur with that word, and the learning mechanism eventually boils out all competing hypotheses. This mechanism, also known as *cross-situational learning* (Siskind, 1996), has long been considered to be an impossible learning mechanism (see, e.g., Bloom, 2000, for a discussion). However, there is increasing evidence that children can and do use cross-situational learning as a mechanism for learning word-meaning mappings (Akhtar & Montague, 1999; Klibanoff & Waxmann, 2000; Mather & Schafer, 2004; Houston-Price, Plunkett & Harris, 2005; Smith & Yu, 2007).

Despite Quine's referential indeterminacy, it has been shown mathematically that 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 super linear with increasing context sizes (Smith et al., 2006). However, these results were achieved with idealised assumptions concerning the input to the learner, in particular it was assumed that

1. there is a strict one-to-one mapping between word and meaning in the input lexicon,
2. the input for the learner is consistent and comes from one source, such that each utterance always co-occurs with the intended meaning (or feature) and
3. the input is presented to the learner with a uniform distribution.

When one deviates from these idealised assumptions, cross-situational learning appears to be much harder, although not infeasible. When cross-situational learning is applied in multi-agent simulations of language evolution and where the number of agents in which the lexicon develops is larger than two, then assumptions 1 and 2 are violated. In such simulations, multiple agents invent different words for a meaning early during the simulation, after which the population needs to converge on a single convention (Baronchelli, Felici, Caglioti, Loreto & Steels, 2006). Several simulations have shown that cross-situational learning alone does not provide a sufficiently powerful one-to-one bias for the lexicon to converge (Vogt & Coumans, 2003; Vogt & Divina, 2007). Adding extra one-to-one biases, such as mutual exclusivity (Smith, 2005) or the mechanism of synonymy damping proposed by De Beule et al. (2006), can overcome this problem. A yet unpublished study by Vogt has shown that when the input to a learner follows a non-uniform Zipfian distribution (Zipf, 1949) through which assumption 3 is violated, the time required to learn –even a small– lexicon runs beyond control when the average context size increases.

In sum, although under ideal circumstances cross-situational learning can work well in situations containing large numbers of hypotheses, under more realistic circumstances the context size from which individuals learn must be well limited. In this study, we will assume that all agents have three joint attentional skills based on those proposed by Carpenter et al. to reduce the context size. Based on previous studies (Smith et al., 2006), we predict that each mechanism has a positive effect on the rate with which lexicons are acquired in the population. The question is which skill yields better performances and whether there are optimal combinations of mechanisms those agents can use.

4. Model and methods

The model is based on the language game model introduced by Steels (1996) in which a population of agents tries to develop a shared lexicon using communicative actions in a particular environment (e.g., a whiteboard with coloured geometrical objects) by engaging in a series of language games. Such language games are typically played between two agents; one of them (the *speaker*) tries to produce a word expressing a feature (or meaning) of an object in its scope of attention, while the other (the *hearer*) tries to identify this feature

based on this uttered word. When the speaker does not know a word, it invents a new random string. When the hearer does not know a word, it acquires the word. When the hearer knows the word, it strengthens the co-occurrence frequencies between the word and the features (or meanings) in the context using cross-situational learning mechanism. The various joint attentional mechanisms are used to construct the context from which the hearers acquire the word-meaning mappings.

Simulations were run with a population containing 10 agents, each starting with an empty lexicon. The agents were situated in a virtual world containing 64 objects, each characterised as a 3 dimensional vector with 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 colour, a shape, or a size. So, in total there are 4^3=64 different objects in this world, constructed from the in total 3x4=12 possible features (or meanings m_j). Note that we assume that each feature corresponds directly to one meaning, e.g., the feature "colour-of-object-1" could correspond to meaning "red".

Each agent is equipped with a private lexicon represented as an association matrix that associates words 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 ω_{ij} that is calculated as the a posterior co-occurrence probability $P(m_j|w_i)$ as follows:

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

Here u_{ij} is the frequency with which word w_i co-occurred with meaning m_j in all previous situations.

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 with a uniform distribution from the world 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 $O_i \in S$. The speaker selects one random object $O_t \in S$ from this context as the *topic* and from this object, it selects one arbitrary feature $f_t \in O_t$ to form the *target*. The speaker then tries to produce an utterance by searching its lexicon for a word that has the highest weight with the target meaning. If no such word is found, the speaker invents a new word as a random string, adds the form-target pair to its lexicon and utters the new word.

In turn, the hearer tries to interpret the uttered word 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) that 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 word w_i and all meanings m_j in the learning context C_L by 1. If an association between word and meaning does not exist, it is added to the lexicon before updating its co-occurrence frequency.

As mentioned, we use the joint attentional mechanisms to reduce the complexity of the learning situation, i.e., these mechanisms are used to construct the learning context C_L. So, how can we translate the joint attentional skills proposed by Carpenter et al. (1998) into the abstraction of the model? Although in the simulations all agents are equally old, we will assume that in the language games, the speaker takes up the role of the adult

and the hearer the role of the child. The following explains how we propose that the joint attention mechanisms can change the learning context:

Checking attention. With checking attention the object is already in the agents' scope of attention, so we assume it *precedes* the exchange of the verbal utterance, which means that both agents share attention to the topic right from the start of the interaction, i.e. $C_S = O_t$. If no further attentional mechanism is used, the learning context is also set to the topic, i.e. $C_L = C_S = O_t$. Note that in the model, the speaker selects the topic. The definition of checking attention assumes that the object is already in the scope of both agents' attention, but, because the topic selection is random, it does not matter when the topic is selected.

Following attention. With following attention, the object is brought into the child's scope of attention by the adult. We assume that this occurs *after* the hearer has interpreted the utterance based on the situation's context C_S. The speaker then selects a random object O_r from the situation S that contains the same feature as the target, i.e., $f_t \in O_r$ and that is different from the topic (i.e. $O_r \neq O_t$). This object is then brought into the hearer's scope of attention and –in case there was no prior joint attention in the game– the features describing this object construct the learning context, i.e. $C_L = O_r$. (If there was prior joint attention –checking attention–then the learning context is constructed as the cross-section of the topic and this additional object, i.e. $C_L = O_t \cap O_r$. If no object is found that contains the intended feature and unequal to the topic, then no attention is modelled and the hearer takes the context of the whole situation as learning context (i.e., $C_L = C_S$). We realise that this is not entirely realistic compared to what humans do (the adult will probably point to the topic), but we decided that a different object is brought into the scope of attention, rather than the topic. We made this choice, because if the speaker can bring the topic into the scope of attention, the mechanism would essentially reduce to checking attention when it comes to learning.

Directing attention. With directing attention, the object is brought into the adult's scope of attention by the child. Again, we assume that this occurs *after* the hearer has interpreted the utterance. The hearer then selects a random object O_h from the situation S that contains the same feature as the interpreted meaning, i.e., $f_I \in O_h$ and brings this object into the speaker's scope of attention. The speaker then provides feedback by signalling whether or not this object contains the intended target, i.e., whether $f_t \in O_h$. We assume that the hearer can use this information to construct the learning context. If the speaker signalled a success, the hearer constructs the learning context C_L as this novel object, i.e., $C_L = O_h$. 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. $C_L = C_S - O_h$. (In the cases where checking attention and/or following attention preceded directing attention in the same language game, yielding a learning context C'_L, then $C_L = C'_L \cap O_h$ or $C_L = C'_L - O_h$.)

It is important to note that we assume that checking attention precedes the verbal interaction, whereas following attention and directing attention occur as a response to an interaction (i.e., occur afterwards). So, checking attention has an impact on interpretation as in the current model: Agents interpret an utterance with a meaning that is in the context (i.e. in the scope of attention). For following attention and directing attention, the scope is initially the entire context C_S, used for interpretation, but learning (i.e. adapting the weights) is carried out on C_L after joint attention has been achieved. This distinction is important, because the simulations are measured based on the ability to interpret an utterance.

An example

To explain the basics of the model, consider the following example. A mother and child are playing with four toy dogs. One is red, striped and furry, the second is green, striped and furry, the third is red, dotted and plastic; and the fourth is yellow, striped and plastic. Suppose that the mother wants to talk about furry things. Further suppose, for the sake of the example, that the child understands phrases like "That is …" and that she knows the word for **dog**, but has never heard any of the words for the colours, textures or materials. So, let us assume that the context of the whole game is limited to **red**, **green**, **yellow**, **striped**, **dotted**, **furry** and **plastic**. When the child would hear the word "furry" in this context, she has no clue, other than to associate this word with all these seven properties.

Now suppose that the child starts playing with the red striped furry dog and the mother says, following her child's attention, "That is furry, isn't it?" Mother and child are *checking attention* and the child, knowing that her model shares her attention to the dog, reduces the learning context to red, striped and furry. The child looks as if she does not understand her mother, which is true because she still does not know whether "furry" means **red**, **striped** or **furry**. In response, the mother draws the child to a green striped furry dog and says "Look, another furry thing!" The child, *following attention*, realises that "furry" does not mean **red**, but either **striped** or **furry**. The child guesses that it means **striped,** and verifies this by *directing* the mother's *attention* to the yellow striped plastic dog while saying "furry?" The mother responds "No, that's not furry." Now the child can infer the meaning of "furry". Because the child already decided that "furry" means either **striped** or **furry**, the object she pointed to was a striped plastic dog and given the negative response, she can infer that "furry" must mean **furry**.

In this example, all three joint attention stages were used shortly after another and in the order of emergence. We model such an interaction as if it is one language game. However, in reality and in the model the joint attentional mechanisms may be used separately. 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 eight different combinations of having none, one or more of the attention mechanisms available, as shown in Table 2. During a simulation, all agents use only one and the same strategy. In the different conditions, each language game used all available mechanisms in the order as proposed by Carpenter et al. (1998), i.e., used in the order checking attention > following attention > directing attention, as in the above example. Only after the joint attention mechanisms are applied, 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	Checking attention	Following attention	Directing attention	C_L
1	**xxx**	-	-	-	C_S
2	**xfx**	-	+	-	O_r or C_S
3	**xxd**	-	-	+	O_h or C_S-O_h
4	**xfd**	-	+	+	$O_r \cap O_h$ or $C_S \cap O_h$ or $O_r \cap C_S$-O_h or C_S-O_h
5	**cxx**	+	-	-	O_t
6	**cfx**	+	+	-	$O_t \cap O_r$ or O_t
7	**cxd**	+	-	+	$O_t \cap O_h$ or O_t-O_h
8	**cfd**	+	+	+	$O_t \cap O_r \cap O_h$ or $O_t \cap O_h$ or $O_t \cap O_r \cap O_t$-O_h or O_t-O_h

We realise that the simulations carried out are still far from reality, as humans do not learn by applying only one type of interaction that uses 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.

It is instructive to note a number of differences between the models of the current paper and those studied before. All current models are based on cross-situational learning. That means that all co-occurrence frequencies between a word and the meanings in the context are increased. In the observational games (e.g., Oliphant, 1999; Vogt, 2000; Vogt & Coumans, 2003), the attended object is the meaning, so the update only strengthens the correct association and weakens incorrect ones. The same holds for the guessing games (e.g., Steels & Kaplan, 2002; Vogt, 2003; Vogt & Coumans, 2003), most closely resemble models containing following and/or directing attention and no checking attention (i.e., **xfx**, **xxd** and **xfd**). The essential difference is that in terms of learning the guessing game reduces to checking attention as the speaker informs the hearer what the topic was in case the hearer guessed wrong (so, the learning context becomes the topic). In following attention, the information is not always given in case no alternative object could be found (if we would allow the speaker to bring the topic into the scope, then following attention **xfx** would effectively become the guessing game). With directing attention, the hearer use the information of its own interpretation, the situation and the speaker's response to construct the learning context, rather than that the speaker provides this as in the guessing game. The learning context may therefore be larger than necessary. In combination with following attention, directing attention can use the additional information to further reduce the learning context size, which does not occur in the guessing game.

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 final 100 games. A game was played correctly if the hearer guessed the target meaning (i.e., feature) intended by the speaker based on the interpretation.

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). 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. When this condition was not reached within 100,000 games, then time of convergence was set to 100,000. 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 checking attention mechanism with more advanced mechanisms, we submitted the convergence time scores of the language games to a Two-Factor ANOVA, with checking attention (yes/no) and following/directing attention (none/following/directing/following and directing) 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 following and directing attention mechanisms ($F(3,792) = 174$, $p < 0.0001$). In the conditions without checking attention mechanisms, the communicative accuracy of most simulations did not converge to 1 within 100,000 games (only the **xfx** and **xfd** models converged occasionally). Nevertheless, the communicative accuracy was much lower in the **xxx** model (0.25) than in the **xfd** model (0.73). On the other hand, in the conditions with checking 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** model (66,147) compared to the **cfx** (2,403) and **cfd** (2,223) models, and—to a lesser extent—to the **cxd** model (18,546). The differences between these models are significant ($F(2,297) = 65$, $p < 0.001$).

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**) to 2.0650 (**cfd**). The differences between the context sizes of the games with combined joint attention mechanisms having checking attention (the **cfx**, **cxd**, and **cfd**) were, however, not significant ($F(2,297) = 2.9$, $p > 0.5$). The value of 3 for the **cxx** 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 following attention and/or directing attention, the context size becomes lower.

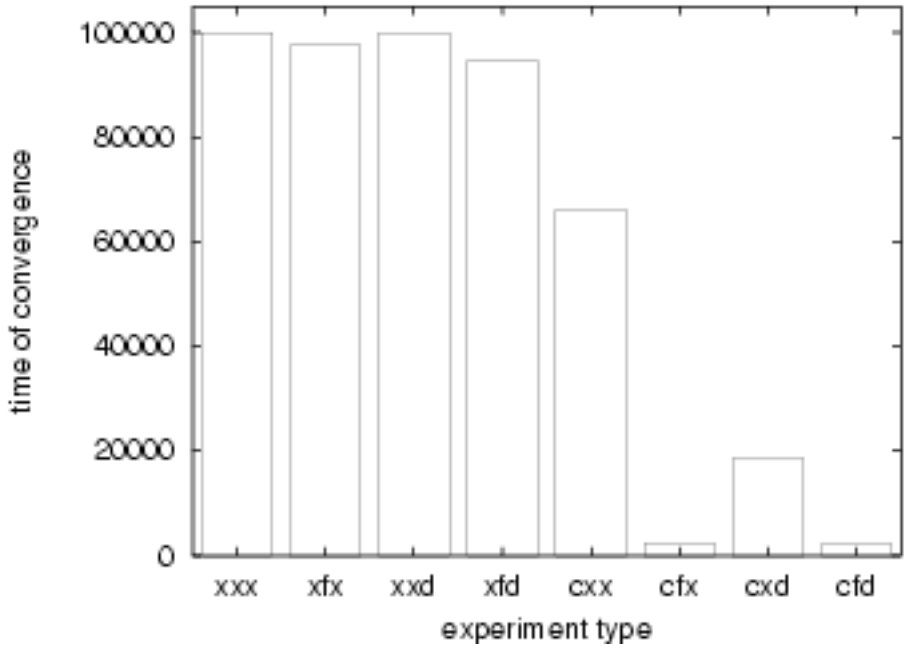

Figure 3. Average time of convergence.

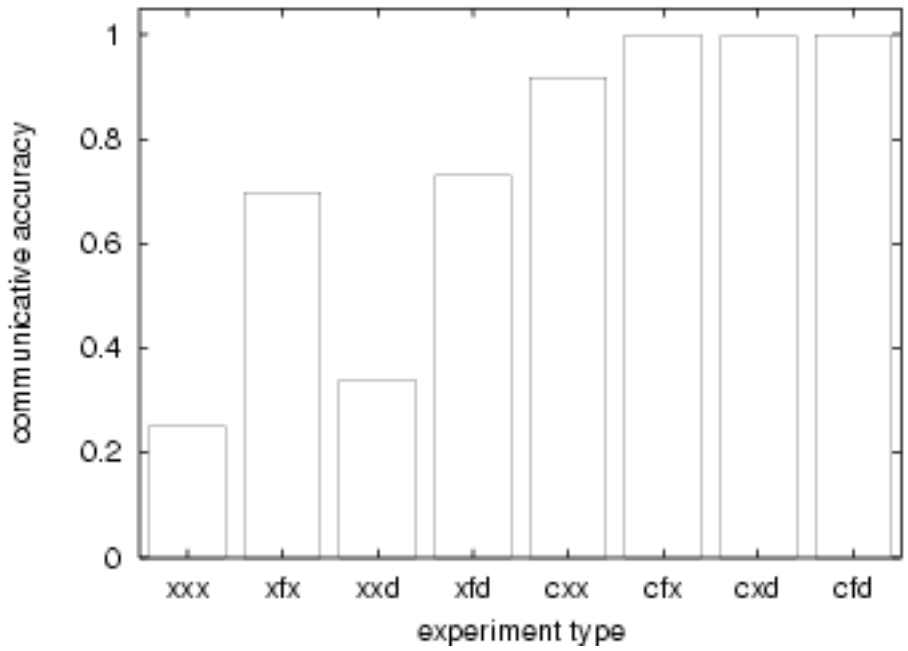

Figure 4. Average communicative accuracy

The strategy that yielded lowest context size (i.e., the one containing all attentional stages) also yielded best performance in terms of communicative accuracy and time of convergence, which is consistent with our prediction. But if we compare the results between following and directing attention (both in combination with checking attention, i.e. **cfx** and **cxd**), then the following attention strategy yielded the best performance on all indicators, despite insignificant differences in context sizes for both models.

When comparing the models that contain only one of the three mechanisms (**cxx**, **xfx** and **xxd**) with each other in terms of improvement on the model without any joint attention (**xxx**), the differences between the three models are most apparent. Clearly, checking attention shows the largest improvement, as it consistently reduces the context size to three. Following attention, which is second best, only does so when there is another object than the topic that has the target feature (recall that this is an artefact of the model, not necessarily realistic). Directing attention comes third and only yields context sizes of three when the guessed object contains the target meaning.

6. Discussion

The simulations 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 attention 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, all four conditions (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 alone to about 18,000 games for directing attention and 2,500 games for following attention.

In general, the improvements in performance correlate well with the decreases in context size. The differences in context size between the simulations that included checking attention and following attention and/or directing attention were not significant, while their differences in performance were. This can be explained as follows: In the directing attention simulation (**cxd**), the child directs the attention. Early in development, the child makes relatively many mistakes, after which the learning context becomes the difference in features between topic and guessed object, which tends to be larger than 1 (the chance that two objects have 1 feature in common is larger than the chance they have 2 features in common). Later on, when the child's language improves, she will more often guess correctly, thus reducing the learning context more frequently to 1. In the following attention model (**cfx**), the learning context is constructed from the intended objects (topic and additional object) throughout the entire simulation, so the chances to learn from a context size 1 is larger. However, in this model the context size can only be reduced in games when an additional object can be found. On average, the context sizes do not differ enough to be significant, but the development in time of their sizes does have an impact on the learning speed.

From our 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. Moreover, the results suggest that the ability to *check* attention is more crucial than the ability to *follow* attention, which in turn is more crucial than the ability to *direct* attention. Sharing attention is –of course– better than not sharing attention, because this narrows down the context size. With checking attention, both participants are aware what the focus of attention is (by definition), so the verbal communication is typically relevant, i.e the utterance refers to the object of attention. To understand the different effects of the three attentional models, let us consider –as a possible interpretation of the model– that following attention is a response of the adult to a child's inability to understand a word's meaning so that the child cannot share attention based on the language alone. The adult will sometimes respond, but not always. When the adult responds, it will draw the child's attention to an object that has the intended meaning, but otherwise the child's 'learning context' remains as uncertain as it

was. However, in case that the child is directing attention, it may be to an object that does not have the intended meaning (in which case the child has to learn from an uncertain situation). So, it is easy to understand that, in terms of context size, checking attention has the largest effect, then following attention and finally directing attention, which – in turn – results in the same ordering regarding the effectiveness of language development.

This conclusion is in line with the developmental data from Reboul (2003) and other data suggesting a clear link between joint attention and language development (e.g., Mundy et al., 2007). Whereas infants knew only a few to a couple of dozen words at the moment they were developing their joint attentional skills (from 9 to 18 months of age), this number rapidly increased to over 300 in the following half year, when they were able to *use* these skills. The catalyst in this rapid increase is thus not just caused by checking attention—which is typically acquired after 9-12 months of age—but also the ability to follow and direct attention, acquired after 11-15 months of age (Carpenter et al. 1998). That the so-called word spurt occurs between 18 and 24 months has likely to do with the further refinement of joint attentional skills, an increased frequency of usage (Mundy et al., 2007) and the applicability of them to language learning, though other factors may play a role too. For instance, the accumulation of linguistic knowledge allows children to use the pragmatics of the language to focus attention.

Because of the parallel ordering of effectiveness and skill emerge in children, it is tempting to suggest a relation here. Let us examine a possible reflection of an ontogenetic ordering of these skills in a phylogenetic ordering of them. Since checking attention has revealed the largest improvement in language development, it is quite possible that such a skill would have evolved first. Would we have run a genetic algorithm selecting for the three skills, then the fitness landscape would have seen the largest increase for a population using checking attention and the second largest increase for following attention. One could therefore expect the three skills to evolve in the order checking attention > following attention > directing attention. However, other effects on fitness, such as the ease with which the different skills can be performed (checking attention appears easier than following attention and directing attention appears most difficult), could significantly influence the evolutionary process. Another issue concerns the possible prediction that –if there are other species than humans that have some form of joint attention– checking attention would be the most likely form to find among them, followed by following attention. However, whether this can be verified is questionable, because the most likely candidates (chimpanzees) do not seem to have joint attentional skills. Even though they jointly engage in certain behaviours and can follow eye gaze and even occasionally point to objects, they miss the aspect of understanding that the other has similar intentions (Tomasello & Carpenter, 2007). Nevertheless, this concerns a much discussed issue in which new discoveries are to be expected that may turn out to be relevant to the prediction formulated here.

Additional research into the nature of these three joint attentional skills, the influence of their subcomponents (e.g. declarative pointing, gaze following, joint engagement, showing etcetera), as well as more advanced components of Theory of Mind on language development and language evolution could shed more light into a better understanding of our ability to learn and use language. Ideally, a model should be built in which all aspects of joint attention are available to agents. This model should then search a whole space of possible frequency distributions with which the different skills are used at different moments in development. The setting that uses frequency distributions and development that most closely resembles those of humans, should reveal a language development similar to that observed in humans. If not, the underlying model of language learning is probably wrong. If it is, the model is a very likely model for human language learning. Such an endeavour would require a combined effort from computer modellers and developmental psychologists. At

the moment, there are insufficient empirical data on frequency distributions and development of joint attentional skills available to build such an integrated model (though the data from Mundy et al., 2007, comes close).

Another type of approach in which it is possible to continue studying the evolution of joint attentional skills in many interesting aspects (possibly even on their ontogenetic and phylogenetic emergence), 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).[4] In this project, large populations of virtual robots (i.e., virtually embodied and situated agents who are –to some extent– 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 currently uses checking attention and following attention skills (directing attention is being constructed) that allow agents to acquire language that allows them to learn rules required to survive in their environment (Vogt & Haasdijk, 2007). It would, for instance, be interesting to let the model evolve the three attentional mechanisms in order to investigate which one tends to evolve first.

7. Conclusions

In this study we have investigated how we can implement the three stages of joint attentional skills found with children (checking attention, following attention and directing attention) in the language game model and how this affects language development. We argue that the crucial distinction between these three stages of joint attention concerns the *scope* of the shared attention. While the objects of shared attention in checking 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 modelled 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. Learning word-meaning mappings in the model is achieved by cross-situational learning, whose performance is known to be correlated with context size. As a result, our simulations yield substantial improvements in performance when one of the joint attentional mechanisms is added to the language game model. We found that checking attention yields the largest improvement, following attention the second largest and directing attention comes last. In exactly this ordering, the analogues mechanisms tend to emerge for human children (Carpenter et al., 1998). We argue that the ordering in performance increase is an indicator for fitness of the various joint attentional stages. Assuming that the most effective mechanism evolved first, we suggest that checking attention may have evolved first, following attention second and directing attention last in human evolution.

As a finale note, the child's growing participation in more complex social interactions provides an important illustration of the embeddedness of cognition. According to a recent approach in cognitive science, addressed

[4] The software running the NEW TIES project, including the code implementing checking attention, following attention and directing attention, can be downloaded and used from http://www.new-ties.org.

under a variety of labels (e.g. situated cognition, enactive cognition, embodied embedded cognition), an organism's bodily interaction with the environment can significantly determine the nature of the cognitive tasks it has to fulfil (van Dijk et al., 2008). Specifically, by creating and/or using structure in the environment many cognitive tasks can be simplified or changed radically in character (e.g. when one does not want to forget posting a letter, one could put it next to one's shoes, thereby changing a memory task into a perceptual one). Such cognition-aiding structures are sometimes referred to as scaffolds (Clark, 1997). Although research into scaffolds often focuses on physical structure, other actors in a social environment provide extremely useful scaffolds for the developing child (Vygotsky, 1978). This investigation can be taken as an illustration of how a developing capacity to use available social scaffolds may help to enhance a child's growing shared lexicon.

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.

Baronchelli, A., Felici, M., Caglioti, E., Loreto, V., Steels, L. (2006) Sharp transition towards shared lexicon in multi-agent systems. *Journal of Statistical Mechanics P06014.*

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)*.

Carpenter, M., Pennington, B. F., and Rogers, S. J. (2002). Interrelations among social-cognitive skills in young children with autism and developmental delays. *Journal of Autism and Developmental Disorders, 32,* 91-106.

Cangelosi, A. (2006) The Grounding and Sharing of Symbols. *Pragmatics and Cognition, 14(2):275-285*

Clark, A. (1997). *Being there*. Cambridge, MA.: MIT Press

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

Dawson, G., Toth, K., Abbott, R., Osterling, J., Munson, J., Estes, A., and Liaw, J. (2004). Early social attention impairments in autism: Social orienting, joint attention, and attention to distress. *Developmental Psychology, 40(2)*: 271-283.

De Beule, J., De Vylder, B., & Belpaeme, T. (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.

Dijk, J. van, Kerkhofs, R., Rooij, I. van & Haselager, W.F.G. (2008). Can there be such a thing as embodied embedded cognitive neuroscience? *Theory & Psychology, 18*(3).

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.

Mundy, P., Block, J., Delgado, C, Pomares, Y., Vaughan Van Hecke, A., Venezia Parlade M. (2007) Individual Differences and the Development of Joint Attention in Infancy. *Child Development 78 (3), 938–954*

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. (2005) Mutual Exclusivity: Communicative Success Despite Conceptual Divergence. In Maggie Tallerman, editor, *Language Origins: Perspectives on Evolution*. Oxford University Press

Smith, K. (2004) The evolution of vocabulary. *Journal of Theoretical Biology*, 228(1):127--142.

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. & Yu, C. (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. (2005) The emergence and evolution of linguistic structure: from lexical to grammatical communication systems. *Connection Science*, 17(3-4):213—230

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.*

Tomasello, M. & Carpenter, M. (2007) Tomasello, M. & Carpenter, M. (2007). Shared Intentionality. *Developmental Science, 10 (1), 121-125.*

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

Vogt, P. (2003) Anchoring of semiotic symbols. *Robotics and Autonomous Systems 43(2): 109-120.*

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

Vogt, P. (2006) Language evolution and robotics: Issues in symbol grounding and language acquisition. In: Angelo Loula, Ricardo Gudwin, Joao Queiroz (Eds.) *Artificial Cognition Systems* Idea Group

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*.

Vogt, P. and Haasdijk, E. (2007) Social learning of skills and language In Acerbi, A., Marocco, D. and Vogt, P. (Eds.) *Proceedings of Int. Workshop on Social learning in embodied agents.*

Vygotsky, L.S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

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.*

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley, Cambridge, MA.