Noname manuscript No.
(will be inserted by the editor)

Anchoring social symbol grounding in children's interactions

Paul Vogt · J. Douglas Mastin

Received: date / Accepted: date

Abstract In this article, we will discuss how computational social symbol grounding (i.e. how shared sets of symbols are grounded in multi-agent models) can be used to study children's acquisition of word-meaning mappings. In order to use multi-agent modelling as a reliable tool to study human language acquisition, we argue that the simulations need to be *anchored* in observations of social interactions that children encounter "in the wild" and in different cultures. We discuss what aspects of such social interactions and cognitive mechanisms can and should be modelled, as well as how we intend to anchor this model to corpora containing features of children's social behaviour as observed "in the wild" to mimic children's (social) environment as reliably as possible. In addition, we discuss some challenges that need to be solved in order to construct the computational model. The resulting SCAFFOLD model will provide a benchmark for investigating socio-cognitive mechanisms of human social symbol grounding using computer simulations.

1 Social symbol grounding

The research project we discuss in this paper aims to investigate how children acquire form-meaning mappings in language, a question that relates to what Vogt and Divina [36] have called *social symbol grounding*. This, in turn relates to Harnad's [13] symbol grounding problem: How do seemingly meaningless symbols (such as words or labels) become meaningful to a cognitive agent (artificial or human) that uses these symbols? Vogt [33] has argued that, when we adopt Peirce's [21] semiotic definition of symbols, the solution essentially boils down to the construction of a semiotic triangle associating a form with a referent mediated by an internalised meaning. Although this is a potentially hard problem for an individual, the challenge becomes even harder when such symbols need to be shared in a communication system such as language. This challenge is referred to as social symbol grounding. The general aim of investigating social symbol grounding is to answer the following question: What social and cognitive mechanisms allow humans (or artificial agents) to associate publicly expressed forms (e.g., words) with the referents that speakers intend? Acquiring such associations is extremely hard, because of three reasons: 1) It may be unknown what the form refers to. 2) The internal 'representations' of referents (i.e. the meanings) may not (nor need) be shared by the communicating individuals [33]. And 3) the number of potential referents is theoretically infinite [22]. When learning word-meaning mappings, this *referential indeterminacy* (i.e., the uncertainty as to what a word refers to) needs to be overcome.

To illustrate this with an example adapted from Quine's [22] famous *gavagai* example, suppose you are doing fieldwork in rural Mozambique where your informants only speak Changana – the local language that you do not yet understand. When you arrive at a family's home and someone comes to you holding a chair and says "utshama", you do not know what this expression means. It can mean many things: 'hello', 'welcome', 'how are you?' or 'nice weather today'. You may look at this person expectantly, and perhaps acknowledge a greeting. Only when this person puts down the chair, gestures towards it and repeats "utshama", you begin to understand that *utshama*'s meaning probably relates

Tilburg centre for Cognition and Communication,
Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
E-mail: p.a.vogt@uvt.nl · E-mail: j.d.mastin@uvt.nl

to the chair; something like 'here's a chair' or 'please, sit down'. In the social interaction with this person you jointly coordinated attention on the chair, and this reduced the referential uncertainty of *utshama*.

Upon the first occurrence of *utshama*, the referential indeterminacy of this word was immense, though common ground [8] concerning social practises upon meeting people and the pragmatic situation of the encounter will have eliminated many unlikely meanings. When the chair was placed in front of you and gestured at, the set of potential meanings for *utshama* was further reduced. However, a few more occurrences in slightly different and/or more refined situations are required for you to learn that *utshama* is the third person singular verb of sitting in the present tense. The cognitive mechanism that could underlie the acquisition of word-meaning mappings in different situations is *cross-situational learning* [26].

Cross-situational learning is a straightforward mechanism that theoretically does not require any heuristics for learning. It works based on multiple exposures of a word in varying situations, where a word's reference is taken as the one that occurs in all (or most) of these situations. There is growing evidence that children and adults can and do use cross-situational learning [27, 28]. Moreover, as shown mathematically, the speed of learning word-meaning mappings is non-linearly proportional to the amount of referential uncertainty (i.e. the number of referents that apply in a situation) [4]. However, computer simulations suggest that cross-situational learning can only explain human word learning when referential uncertainty is substantially reduced [34].

It is well established that humans use a variety of word learning mechanisms (or heuristics) simultaneously to reduce referential uncertainty [29]. These word learning mechanisms include constraints and biases, syntactic cues, and social cues. Constraints and biases, such as the whole object bias [30], mutual exclusivity [18] or the taxonomic bias [17], are cognitive mechanisms that describe ways to exclude potential referents or add preferences to others. Sentential cues [12] make use of information contained in the syntactic structure to guide attention towards a certain referent (e.g., knowledge of a certain verb can direct the listener's attention to an object on which the verb's action is applied). However, as the *utshama* example illustrates, social cues obtained through social interactions, such as joint attention [32] or providing feedback [7] are also essential to establish common ground. The interaction between all these mechanisms provide a scaffold for learning language [40].

Various computational studies have investigated the roles of proposed heuristics on vocabulary development based on cross-situational learning [11,16,36,37,41]. These studies indicate that different heuristics (e.g., mutual exclusivity, principle of contrast, social/gestural cues, joint attention or corrective feedback) have varying effects on the speed and success of word learning. For instance, Kwisthout and colleagues [16] have shown that different forms of joint attention as proposed by Carpenter et al. [6] can reduce referential uncertainty to different extents, thus influencing the learning speed [4, 34]. These findings suggest that when different heuristics occur with different frequencies, then this would have different effects on word learning.

Most of the studies that apply such heuristics do so without being concerned with the statistical distribution of how children experience them "in the wild", and those that do (e.g. [41]) compare the model's vocabulary development with that of different models. However, to assess the model's plausibility more effectively, the results also need to be compared with the vocabulary development of the children whose input data are modelled. Obviously, the studies mentioned have contributed substantially to our understanding of the (socio-)cognitive mechanisms that underlie social symbol grounding, but they have not provided convincing insights into how these mechanisms interact with each other when humans use them to learn word-meaning mappings. The two main reasons for this limitation are: 1) no computational model exists in which the interaction between such mechanisms is studied, and 2) there is the lack of a firm *anchoring*[1] of models in concrete empirical data of human symbol grounding so that a comparison between the model and human performance is impossible [35,38].

The CASA MILA[2] project has been initiated to bridge this gap. The aim of this project is to set up cross-cultural corpora that describe scenarios for simulating infants' language acquisition based on observations taken in their home environments. The corpora will contain statistical properties with which certain (social) behaviours occur, as well as frequencies of certain verbal and non-verbal cues. Agent-based models will then be designed similar to those published in, e.g., [16,37], which implement variants of Steels' language games [31]. The purpose of the present article is to present the outline of the model that we envi-

[1] In contrast to earlier notions of *anchoring* (e.g., [3,9]), we will use this notion to speak about connecting or grounding computational studies in empirical data. We use this term to distinguish from (symbol) grounding.

[2] Cultural And Social Aspects of Multimodal Interactions in Language Acquisition.

sion, called SCAFFOLD, thereby focusing on the different social interactions that infants engage in and the non-verbal cues by which they reduce referential uncertainty. These interactions and non-verbal cues are based on the literature of child language acquisition to develop a model that is consistent with children's social interactions.

2 Anchoring models in empirical data

2.1 Observations of infant behaviour

To anchor computer simulations in empirical data, we require longitudinal corpora of infants' behaviour in their natural environment during daily activities. In addition, we need estimates of these infants' vocabulary development. When testing our SCAFFOLD model based on a given corpus from one infant, we expect that this simulation would reveal a similar vocabulary development as said infant. However, only provided the model sufficiently implements the socio-cognitive mechanisms underlying word learning. Moreover, assuming that there is a unifying cognitive model for word learning, the SCAFFOLD model should be able to explain how individual and cultural differences between the ways children interact with their social environments relate to how these children's vocabularies develop. So, when simulating corpora from multiple infants, we would expect to obtain a similar correlation between input properties and vocabulary development as observed with these different infants. Since no comparable resources were found that fit the criteria we require, we decided to collect the data ourselves from different cultures in urban and rural Mozambique, and also in the Netherlands.

We chose to observe natural behaviour, because experimental studies only capture what a child learner can do, but not necessarily what they actually do. For instance, eliciting information from simulated play between caregivers and children may skew the observations considerably, because non-play activities are ignored [19]. In addition, we are collecting data from different cultures, because focusing only on data from industrialised countries (e.g., the Netherlands) would only provide us interactions as they occur in such a culture. However, language socialisation surrounding infants differ widely across cultures [25]. For instance, whereas industrialised cultures tend to have small nuclear families nowadays, many children in non-industrialised cultures tend to grow up in extended families – often consisting of multiple generations, where it is the norm for a child to have various caregivers (including siblings) [5]. In such cultures, where language socialisation is often less child-oriented [25], one also observes relatively more multiparty interactions [5], as well as more bodily stimulation than cognitive stimulation [15].

We have videotaped infants' natural behaviour longitudinally over one year when the infants where on average 13, 17 and 25 months old, which coincides more or less with the period when they learn their first words. We have collected data from 14 infants from each of the Mozambican communities and from 12 infants in the Netherlands. In addition, we have administered the MacArthur-Bates Communicative Development Inventory [10], adapted to the relevant local languages, to assess the infants' vocabulary development at each of these ages. More details on the data collection methods are described in [19,20,39].

From each video, we have annotated approximately 30 minutes in various dimensions. In particular, we coded the infants' attentional states [1,19], with whom the infants interact, and verbal and non-verbal social cues produced by and addressed to the infants [39]. Before we present what information we have annotated and how they will be used, we describe the envisioned SCAFFOLD model.

2.2 The SCAFFOLD model

The model will be extended from [37] and will consist of a 2-dimensional world that simulates the typical household in which children grow up, containing various objects (e.g., food, toys, furniture), places (living room, kitchen, garden), animals (pets or small livestock) and agents. The agents perceive these items in the world through feature vectors that describe the objects in terms of, for instance, shape, colour, location, etc. The number of objects, places, animals and agents will be derived from the situations encountered in the different observations. In particular, the number of agents in the household, as well as the maturity of these agents, defines the social environment of the infant. Where our observations from the Netherlands typically involve only parent-infant dyads, the Mozambican infants were filmed while various members of the extended family were present.

We distinguish five different types of agents: infant (the target of our study), peers (similar to infants, but not a target), siblings (older immature agents), adults (mature agents) and (a) primary caregiver(s) (mother and/or father). The linguistic and behavioural competence of agents will depend on the maturity of the agents: primary caregivers and adults will have full linguistic and behavioural competence (i.e. they will behave appropriately following rules to be designed); sib-

lings will have complete linguistic and behavioural skills, but may use them inappropriately; peers and infant will have no linguistic competence and only rudimentary behavioural skills at the start of the simulations. Their linguistic competence will be acquired during the simulation by engaging in language games. The number of agents in the simulations and the frequency of social interactions with the different types of agents, as well as groups of agents, will be determined based on the observations.

Social interactions in which infants can learn word-meaning mappings are based on the language game model (cf. [31,36,37]). A language game is situated in a certain context, which contains visible objects, actions and possible internal states. From this context, a speaker agent determines a topic for communication, searches his linguistic memory for a way to produce an utterance, and upon hearing the utterance, the addressee will try to infer the utterance in light of the context. When the addressee is an infant, this agent will attempt to acquire the correct word-meaning mappings based on the utterance and the context. The acquisition of word-meaning mappings proceeds through cross-situational learning by increasing usage frequencies between the uttered word(s) and the meanings of items in the *learning context*. However, when this context contains many items, there is a lot of referential uncertainty. Various word learning strategies will be implemented that reduce the size of the context, thus forming a scaffold from which the cross-situational learning mechanism will proceed (cf. Figure 1).

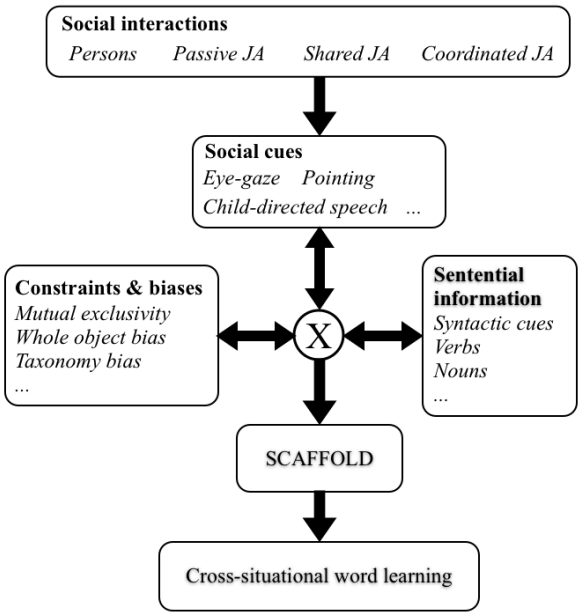

Fig. 1 The SCAFFOLD model. (JA stands for joint attention.)

2.3 Social interactions

2.3.1 Engagement levels

The agents will be implemented such that they are in a certain *attentional state* that controls the way they behave. In these states, agents can act solitarily or socially. The actions they can perform include: move forward, turn left/right, pick up, put down, give, take, play or eat objects. In addition, agents can perform social actions, such as kiss, hug, hit, gesture and talk.

To implement behaviour typically observed among children, we adopted Bakeman and Adamson's [1] attentional states, called *engagement levels*. To account for observed behaviour that did not fit within Bakeman and Adamson's original categories, we extended categorisation with two additional levels: *Observation* and *Shared joint attention*. The reason for adding these two levels was that the original categorisation was based on observations of simulated play, while our analysis of observed natural behaviour revealed there are additional attentional states that occur beyond playing (see [19] for a detailed explanation). The first four engagement levels are states in which infants behave solitary, either being *Unengaged*, *Onlooking* to other agents' non-object oriented behaviour, *Observing* another agent manipulating an object, or manipulating *Objects*.

When encountering another agent (or group of agents), the infant can enter one of five *joint engagement levels*, which simulate different forms of social interactions that control the language games. The first type of joint engagement is called *Persons*, which are interactions where the infant actively communicates with one or more agents. These interactions are about something that either relates to a shared activity (e.g. playing patty cake) or perhaps something more abstract or distal, but not about any concrete external object or event. When the interactions are about an external object or event, the infant engages in one of the three types of joint attention that we analysed: *Passive joint attention*, *Shared joint attention* and *Coordinated joint attention*. All involve the infant interacting with one or more persons and all interactants share their attention to a third object or event. *Passive joint attention* differs from the rest in that only one of the communication partners is aware that the attention is shared. In *Shared joint attention* there is no overt *mutual* interaction goal, which is present in *Coordinated joint attention*. A mutual interaction goal is present when intentions towards an object are shared (cf., [32]), for instance when an object is exchanged. However, when

someone is passing the line of sight of the interactants who are both aware they share the attention and may even comment on that event, there is no clear shared intention and this, thus, classifies as *Shared joint attention*. Also interactions where the intention of one interactant is not understood by the other are considered instances of *Shared joint attention*.

As in [37], all social interactions involve a context, such as visible objects, events (e.g., hugging, kissing, hitting, gesturing, etc.), and internal states (e.g. being hungry). The type of engagement level determines the way the interactants *focus their attention* within the context, thus constructing what we call a *learning context*. If it is a *Persons* interaction, the topic is some action or distal aspect, but not a concrete external object or event. If it is a form of joint attention, the topic is an object or concrete external event. Instances of *Passive* and *Coordinated joint attention* interactions have previously been implemented in simulations [16,37]. The way these were implemented yielded differences in the extent in which referential uncertainties were reduced, thus constructing learning contexts of different sizes. Depending on the type of engagement level, the speaker or addressee (either could be an infant or another agent) can direct or follow the focus of attention either through a verbal or non-verbal cue. For instance, in *Coordinated joint attention*, the speaker (e.g., a sibling) may ask for a particular toy verbally. When the addressee (e.g. the infant) does not understand the speaker, she may signal incomprehension non-verbally, after which the sibling may repeat the request using a non-verbal gesture (e.g. pointing) to direct the infants' attention. Based on such exchanges, the infant can reduce referential uncertainty, thus facilitating cross-situational learning.

Challenges for implementing the various types of joint engagement are: 1) defining natural chains of events that constitute the different engagement levels, 2) ways to control attention towards external objects and events, and 3) to construct learning contexts for *Persons* interactions in which no external object or event is present.

2.3.2 Social cues

It is widely believed that gestures help humans to control joint attention (e.g., [6]), and it has been shown that the usage frequencies of gestures correlate to vocabulary development [24]. So, to implement mechanisms that control attention towards objects or events, thus reducing referential uncertainty, we intend to simulate the use of gestures in communication. In our project, gestures are defined as non-verbal social cues that draw the attention of the communication partner to an object or event. Based on the gesture literature (e.g., [24,42]) we annotate *deictic* gestures, such as eye-gaze, pointing, showing, offering, taking, and reaching for objects, and *non-deictic* gestures, such as conventionals (e.g. waving), iconics (i.e., mimicking some property of an object or event), ritual interactions (e.g., turn-taking games, such as patty cake, dancing, or other playful communicative behaviour), and 'embodiment' (see [39] for more details). The embodiment (or *embody*) gesture [42] is an activity where, typically, the caregiver takes control of the infants' body to demonstrate motor behaviour, such as pushing the infant in a certain direction or guiding their actions to facilitate an activity.

When implementing gestures, we need to define, as realistically as possible, the extent to which the different gestures allow the addressee to identify the intended referent. For example, we need to take into account that distal pointing will less accurately indicate a referent than showing would. Even more challenging would be to implement gestures that *represent* objects or events, such as an iconic drinking gesture that may be used to summon the addressee to drink. We are currently carrying out experiments to estimate how accurately humans can interpret eye gaze and pointing gestures from different distances (cf. [2]). These experiments may be further adapted to assess the accuracy of other types of gestures, such as iconics.

Verbal interactions between infant agents and their social environment are based on linguistic utterances observed in the observations. To this aim, we have annotated all speech produced by and addressed to the infants in the local languages (Changana, Ronga, Portuguese and Dutch) with translations into English. However, instead of treating the utterances as a fixed set of input to infants (as is common in computational models of language acquisition [11,41]), we will use them as possible utterances produced by the competent speakers (siblings, adults and mothers). The rationale behind this is that the dynamics of the simulations are unpredictable, partly because infants' utterances depend on their language learning.

2.4 Cognitive mechanisms

The social cues are aimed to reduce referential uncertainty, but ambiguity may still persist. To further reduce the learning context (or scaffold), cognitive word learning mechanisms that process constraints, biases and sentential information will be implemented, as well as mechanisms that control the way they are used.

The primary constraint is mutual exclusivity [18], which children seem to use, and through which a meaning is discarded from the context when it already is used

to denote a word that did not form part of the utterance. The model will also contain the whole object bias [30] by which the learner will tend to associate novel words to whole objects rather than features of objects. Another bias that we intend to incorporate is the taxonomic bias [17]. According to this bias, the learner will prefer to map novel words to basic-level categories, such as dog or cat instead of animal or German shepherd (cf. [23]).

Sentential information will be applied to information from words that occur in the utterance and whose meanings are already established by the learner [12]. For instance, if the utterance is "eat apple" and the learner may know the meaning of the word apple and additionally knows what apples are typically used for, he may discard actions that do not apply to apples.

To control the interaction between these cognitive word learning strategies and the use of social cues, a selection mechanism will be developed to decide which strategies to use in a certain situation. The plan is to develop a set of heuristics that takes the computational processing complexity of the various strategies into account (the lower this complexity, the more favourable the strategy), but which is constrained by the information available to the agent. For instance, if there is no referent in the context that has a known word, mutual exclusivity cannot be applied. Also, strategies should be applied with care. For example, mutual exclusivity assumes there is no synonymy in the language, but languages contain many synonyms. Hence, it may be better to implement mutual exclusivity as a bias rather than a constraint.

In a way, the combined effort of social cues, constraints, biases and sentential information work like a sieve to filter out irrelevant distracting items from the context. The result of applying these strategies is a scaffold that allows the learner to acquire the correct word-meaning mapping.

When all word learning strategies are modelled, we can simulate different infants interacting with their own social environment as we have observed in our data and monitor their language development. After running these simulations for a prolonged period, we can compare the simulated vocabulary development with that from the real infants. When the simulations predict the infants' vocabulary development reliably under all observed circumstances from Mozambique and the Netherlands, we can suggest that the underlying socio-cognitive model is a plausible model for human social symbol grounding.

3 A long way to go

It will be a major challenge to design the SCAFFOLD model such that simulations reveal similar results as those observed in the three communities we studied. For instance, after analysing the first results from Mozambique, we found that infants' vocabulary development in the urban community are in line with what we would predict based on previous modelling studies (e.g., [16]), but the results obtained from the rural community are not. In particular, we found that in the rural area the amount of *Coordinated joint attention* at 13-months correlates negatively with later vocabulary development [20], which is in stark contrast to what was expected based on the literature [6,32]. In addition, the amounts of speech and co-speech gestures addressed to rural infants at 13-months do not predict later vocabulary development [39], as would have been expected from Western studies [14,24].

To understand these differences, an even deeper analysis of the observed data is required, which can be achieved using the SCAFFOLD model. For instance, we can investigate the role of sibling caregivers in the rural community, who interact increasingly more frequently with the infants, and over time surpass mothers as the most frequent interactant. When a realistic model has been developed that can explain the findings from one community, we need to assess the same model for the other communities. If the model can replicate observations found in those communities, we will have convincing, but not conclusive, evidence that this is a good model to explain human language acquisition.

It is beyond our expectations that such results will be achieved with the first version of the model. It is more likely that it will take years of testing many more models, and/or refinements thereof, before convincing evidence will be obtained. While revising the computer model, we may be able to verify certain hypotheses of why particular individual and cross-cultural differences are observed. On top of that, the model will have to be tested on more data from more cultures and social classes than the three data sets we are currently developing. Moreover, to move beyond the one word stage, additional data is required from older children to investigate the socio-cognitive mechanisms underlying the emergence of grammatical constructions in language.

Once the SCAFFOLD model has been developed and tested, the model and corpora will be made available as open source. This way, the data will provide a challenging benchmark for anchoring and testing various theories of children's social symbol grounding.

Acknowledgements

The research presented in this article is funded by a Vidi grant awarded to the first author by the Netherlands Organisation for Scientific Research (NWO, grant no. 276-70-018). The authors are indebted to Wona Sanana and the Associação Communitario Ambiente de Mafalala for their assistance in recruiting the Mozambican participants, all Mozambican and Dutch research assistants for helping to collect and annotate the data, and all Mozambican and Dutch participants.

References

1. Bakeman, R., Adamson, L.: Coordinating attention to people and objects in mother-infant and peer-infant interaction. Child development pp. 1278–1289 (1984)
2. Bangerter, A., Oppenheimer, D.: Accuracy in detecting referents of pointing gestures unaccompanied by language. Gesture **6**(1), 85–102 (2006)
3. Barwise, J., Perry, J.: Situations and attitudes. The MIT Press, Cambridge, MA (1983)
4. Blythe, R., Smith, K., Smith, A.: Learning times for large lexicons through cross-situational learning. Cognitive Science **34**(4), 620–642 (2010)
5. Brown, P.: The cultural organization of attention. The handbook of language socialization **71**, 29 (2011)
6. Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., Moore, C.: 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) (1998)
7. Chouinard, M.M., Clark, E.V.: Adult reformulations of child errors as negative evidence. Journal of Child Language **30(3)**, 637–669 (2003)
8. Clark, H.H.: Using Language. Cambridge University Press (1996)
9. Coradeschi, S., Saffiotti, A.: An introduction to the anchoring problem. Robotics and Autonomous Systems **43**(2), 85–96 (2003)
10. Fenson, L., Dale, P.S., Reznick, J.S., Thal, D., Bates, E., Hartung, J., Pethick, S., Reilly, J.: The MacArthur Communicative Development Inventories: user's guide and technical manual. Singular Publishing Group, San Diego (1993)
11. Frank, M., Goodman, N., Tenenbaum, J.: Using speakers' referential intentions to model early cross-situational word learning. Psychological Science **20**(5), 578–585 (2009)
12. Gleitman, L.: The structural sources of verb meanings. Language Acquisition **1**, 3–55 (1990)
13. Harnad, S.: The symbol grounding problem. Physica D **42**, 335–346 (1990)
14. Hart, B., Risley, T.: Meaningful differences in the everyday experience of young American children. Paul H Brookes Publishing (1995)
15. Keller, H., Völker, S., Yovsi, R.: Conceptions of parenting in different cultural communities: The case of west african nso and northern german women. Social Development **14**(1), 158–180 (2005)
16. Kwisthout, J., Vogt, P., Haselager, P., Dijkstra, T.: Joint attention and language evolution. Connection Science **20**, 155–171 (2008)
17. Markman, E., Hutchinson, J.: Children's sensitivity to constraints on word meaning: taxonomic versus thematic relations. Cognitive psychology **16**(1), 1–27 (1984)
18. Markman, E.M., Wachtel, G.F.: Children's use of mutual exclusivity to constrain the meaning of words. Cognitive Psychology **20**, 121–157 (1988)
19. Mastin, J.D., Vogt, P.: Analyzing infant engagement: Filling the gaps in research approaches (Submitted)
20. Mastin, J.D., Vogt, P.: Correlations between joint engagement and vocabulary development: A longitudinal, observational study of Mozambican infants from 1;1 to 2;1 (Submitted)
21. Peirce, C.S.: Collected Papers, vol. I-VIII. Harvard University Press, Cambridge Ma. (1931–1958)
22. Quine, W.V.O.: Word and object. Cambridge University Press (1960)
23. Rosch, E.: Principles of categorization. In: E. Rosch, B.B. Lloyd (eds.) Cognition and Categorization. Lawrence Erlbaum Ass. (1978)
24. Rowe, M., Goldin-Meadow, S.: Differences in early gesture explain ses disparities in child vocabulary size at school entry. Science **323**(5916), 951–953 (2009)
25. Schieffelin, B., Ochs, E.: Language socialization across cultures. Cambridge University Press, Cambridge, UK. (1989)
26. Siskind, J.M.: A computational study of cross-situational techniques for learning word-to-meaning mappings. Cognition **61**, 39–91 (1996)
27. Smith, K., Smith, A., Blythe, R.: Cross-Situational Learning: An Experimental Study of Word-Learning Mechanisms. Cognitive Science (2010)
28. Smith, L.B., Yu, C.: Infants rapidly learn word-referent mappings via cross-situational statistics. Cognition **106:3**, 1558–1568 (2008)
29. Snedeker, J.: Word learning. Encyclopedia of neuroscience pp. 503–508 (2009)
30. Soja, N.N., Carey, S., Spelke, E.S.: Ontological categories guide young children's inductions of word meanings: object terms and substance terms. Cognition **38**, 179–211 (1991)
31. Steels, L.: Evolving grounded communication for robots. Trends in Cognitive Sciences **7**(7), 308–312 (2003). DOI 10.1016/S1364-6613(03)00129-3
32. Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H.: Understanding and sharing intentions: The origins of cultural cognition. Behavioral and Brain Sciences **28**(05), 675–691 (2005)
33. Vogt, P.: The physical symbol grounding problem. Cognitive Systems Research **3(3)**, 429–457 (2002)
34. Vogt, P.: Exploring the robustness of cross-situational learning under Zipfian distributions. Cognitive Science **36**, 726–739 (2012)
35. Vogt, P., de Boer, B.: Language evolution: Computer models for empirical data. Adaptive Behavior **18(1)**, 5–11 (2010)
36. Vogt, P., Divina, F.: Social symbol grounding and language evolution. Interaction Studies **8(1)**, 31–52 (2007)
37. Vogt, P., Haasdijk, E.: Modelling social learning of language and skills. Artificial Life **16** (2010)
38. Vogt, P., Lieven, E.: Verifying theories of language acquisition using computer models of language evolution. Adaptive Behavior **18**, 21–35 (2010)
39. Vogt, P., Mastin, J.D.: Child-directed speech and co-speech gestures correlations with vocabulary development in rural and urban Mozambique (In prep)
40. Vygotsky, L.: Mind in society. Harvard University Press (1978)

41. Yu, C., Ballard, D.: A unified model of early word learning: Integrating statistical and social cues. Neurocomputing **70**(13-15), 2149–2165 (2007)
42. Zukow-Goldring, P.: Sensitive caregiving fosters the comprehension of speech: When gestures speak louder than words. Early development and parenting **5**(4), 195–211 (1996)