Active Language Learning Inspired from Early Childhood Information Seeking Strategies

Workshop on Cognitive Architectures for Human-Robot Interaction

Brigitte Krenn
Austrian Research Institute for Artificial Intelligence (OFAI)
Vienna, Austria
brigitte.krenn@ofai.at

Christiana Tsiourti
TU Wien - Automation and Control Institute (ACIN)
Vienna, Austria
christiana.tsiourti@tuwien.ac.at

Friedrich Neubarth
Austrian Research Institute for Artificial Intelligence (OFAI)
Vienna, Austria
friedrich.neubarth@ofai.at

Stephanie Gross
Austrian Research Institute for Artificial Intelligence (OFAI)
Vienna, Austria
stephanie.gross@ofai.at

Matthias Hirschmanner
TU Wien - Automation and Control Institute (ACIN)
Vienna, Austria
hirschmanner@acin.tuwien.ac.at

ABSTRACT

In this paper we introduce an active learning extension to our incremental grounded language learning system implemented on the Pepper robot. This approach is inspired by recent results from child language acquisition research, which shows that children deliberately use gestures like pointing to acquire new information about the world around them. In our system, the Pepper robot learns word-object and word-action mappings by observing a human tutor manipulating objects on a table and verbally describing the actions. Under certain conditions, the robot will interrupt the tutor and actively request information by pointing at an object. We describe our first approach of facilitating active information seeking strategies to enhance our system. We motivate when and how to apply them by reviewing research on question-asking during infancy and toddlerhood.

KEYWORDS

Language acquisition; Active Learning; Robotics;

1 INTRODUCTION

As robots become ubiquitous across diverse human environments such as homes, hospitals and public spaces, they need to understand scenes, and interpret actions and verbal descriptions of humans. Therefore, learning from situated task descriptions is important. This way, robots can adapt to specific situations and learn new tasks from demonstrations. By situated task descriptions we mean situations where a human tutor shows a task to a robot and describes what s/he is doing. This is comparable to what adults do, when interacting with very young children. In developmental language learning, modality rich input is considered to be of particular importance in early language learning, see for instance [28, 37].

One way of improving the efficiency of learning is to make use of active learning strategies. Literature from language learning provides increasing evidence that infants engage in self-guided learning strategies involving metacognition (i.e., the ability to reflect upon their own knowledge states) [9]. Infants communicate their ignorance [17], seek information [3] and otherwise actively direct their learning instead of learning passively. Moreover, infants have been found to learn better when they are given the opportunity to choose what to learn [18].

Infants initiate communication and seek information through different modalities and behaviors, such as non-word vocalizations, gestures, and, eventually words. One of the most salient ways of pre-verbal communication is pointing which appears to be a dyadic or reciprocal mode of engagement [3]. While, previously, the standard functional interpretation of infant pointing [2] has been that it serves either to request an out-of-reach object (i.e., "*I want this*") or to establish joint attention to an object of interest (i.e., "*Look at this*"), more recently, it has been proposed (e.g., [9], [3]) that pointing can also serve an interrogative function. Thus, pointing can also imply that infants communicate their ignorance and request information (i.e.,"What is this?"), with the expectation that an interlocutor will respond with the desired information.

In the current paper, we bring together findings from research on infant crossmodal language learning based on modality rich, tutoring-like input situations and findings from young infants' active information seeking strategies. As a result, we present an extension to a grounded language learning system for simple object manipulation actions which we have developed and implemented on the humanoid robot Pepper. The initial system (henceforth base system) for word-object and word-action learning is realized as an information theoretic model which incrementally learns from real-world multimodal input situations. The extension to the learning model incorporates mechanisms for active information seeking by the robot, inspired by findings from research on meta-cognition, question-asking and learning in infancy and early toddlerhood (9 to 35 months old).

In Section 2, we provide background information on research on metacognition, question-asking and learning in infancy and early toddlerhood. In Section 3 we outline related work on grounded language learning and active learning from the field of robotics. The extended learning model and its implementation on a Pepper robot is presented in Section 4 where also a brief description of the base model is given. Section 5 concludes the contribution and gives an outlook on the planned work.

2 METACOGNITION, QUESTION-ASKING AND LEARNING IN EARLY CHILDHOOD

2.1 Metacognition

Metacognition (i.e., the ability to reflect upon their own knowledge states) has been shown to be an important predictor of learning, both in adults and school-aged children [32]. Previous research with infants and toddlers consistently found strong capacities for learning [32] but poor metacognitive abilities [8]. Specifically, children under the age of four have consistently been shown to experience difficulties in verbally expressing their own state of knowledge [8].

Nevertheless, more recently there has been increasing evidence that infants engage in self-guided learning strategies that may involve metacognition [9]. For example, infants use pointing gestures in an interrogative fashion [3], and learn better when they are given the opportunity to choose what to learn [18]. These findings suggest that young children are aware of knowing some items of information and affirm possessing that knowledge. At the same time, they are aware of lacking other items of information, communicate their ignorance and request information [13].

Harris et al. [13] review evidence that in the course of their second year, children begin to communicate doubt or ignorance in various ways. For instance, a study by Liszkowski et al. [17] provides evidence to support the existence of a culturally shared, gestural, language-independent form of communication. Specifically, results show that children begin to use whole handed pointing around eight months of age, and use index finger pointing at around eleven months. Pointing appears to be a dyadic or reciprocal mode of engagement. The standard functional interpretation of infant pointing has been that it serves either to request an out-of-reach object or to establish joint attention to an object of interest [2]. More recently, it has been proposed that pointing can serve an interrogative function [3]. Thus, an infant's pointing may express not just "*I want this*" or "*Look at this*" but also "*What is this?*" with the expectation that the interlocutor will respond with appropriate information. In addition to pointing, young children also use gaze (e.g., looking toward an available adult), nonverbal gestures (e.g., hand flips, shrugs), vocalization (e.g., "*umm*") and convey explicit statements of their ignorance (e.g., "*I don't know*") either separately or in combination. See for instance [1] who analyzed data from 64 children from the age of 14 months onwards.

The signals used by young children, appear to serve two functions. First, they signal ignorance. Second, when formulated as questions, they also convey information-seeking requests to an interlocutor for supplying missing information [13]. Interestingly, nonverbal forms of metacognition have also been demonstrated in animal species [27]. Among others, honey bees [23] and monkeys [12] have been shown to seek additional information when the available evidence is incomplete, or to indicate their uncertainty by deferring to make a decision when a mistaken response would impose costs and they do not know the best course of action. These findings demonstrate not only that young children and animals can monitor their own uncertainty but also that metacognitive abilities can be expressed without relying on language.

2.2 Initiating Communication

Infants and young children initiate communication and request information from interlocutors through different modalities and behaviors, such as non-word vocalizations, gestures, and, eventually words [35].

Communicative non-word vocalizations are vocal utterances, often accompanied by gesture and/or eye contact with the interlocutor. These vocalizations contain words or speech sounds that do not refer to a specific object or event. For example, a preverbal infant who wants a toy but does not yet have the word "toy" in her verbal repertoire might indicate her desire by pointing to the toy while vocalizing "aaah, aaah, aaah". Likewise, an infant who sees a cat enter the room but does not yet know the word "cat" might indicate awareness of the cat's presence by pointing to it and vocalizing "aaah" [35].

Different kinds of gestures, including movements of the hands, arms, and facial expressions, allow preverbal infants to convey messages or thoughts to their interlocutors [11]. Gestures are classified into different categories occurring in different stages of development. The first type of gestures that appear in infants are deictic gestures (i.e., reaching, pointing). These gestures express a child's intent to request or declare something by referring to a person, object, location, or event through touching it, indicating it, or calling attention to it. Around twelve months of age, infants begin to use representational gestures which communicate a specific meaning [15]. For instance, a flip gesture (i.e., gesture involving the lifting and outward rotation of both hands and the shrugging of the shoulders) has been shown to communicate "I don't know" [1]. In relation to language acquisition, representational gestures appear around the same time as first words and become more complex as children get older.

When used in combination, non-word vocalizations and gestures can lead to a state of joint attention - a shared mental state in which partners in an interaction focus attention respectively on the same objects or events [21]. For instance, an infant who not only points at something of desire or interest but vocalizes while doing so produces a powerful stimulus for a social partner, a stimulus likely to bring about a state of joint attention. For example, while a parent who is engaged in her own activities might fail to notice her infant silently pointing to a toy, when the infant accompanies her point with a vocalization, the parent is much less likely to miss the gesture and more likely to shift attention to the object of infant desire or interest. This triggers the parent's relevant verbal comments (e.g., "This is a toy!") which provide the child with language input adapted to the focus of her attention.

2.3 Question-asking

During the first and second year of life, children are quickly developing their ability to request information [25]. Between 12 and 24 months, children begin to ask questions when they lack knowledge and when they are uncertain about the knowledge they possess [9]. Furthermore, they also begin to consider the availability of reliable informants when posing their queries [3]. By the time they are 5 years old, children also request information when they have identified inconsistencies and contradictions in their understanding of concepts and when they know they are missing information to

complete a given task [34]. Ronfard et al. propose a question-asking model for children composed of four components: (1) initiation, (2) formulation, (3) expression, and (4) response evaluation and follow-up [25]. According to the authors, the four components are present throughout the development of the individual's increasing ability to more explicitly and fully reflect on the processes involved in each of these components and to coordinate them, and as a result to more fully and more efficiently deploy questions as an information seeking strategy. Whereby, the process of asking a question is iterative and dynamic, with the different components interacting and influencing each other.

In the following, we use the components proposed in the model as a means of organising the review of research on question-asking during infancy and toddlerhood. As we are interested in mechanisms of early learning, we focus on research addressing 9 to 35 months old toddlers.

Initiation. Requests for information begin with a realization that information is needed. The earliest evidence that infants (12-16 month old) seek information from an adult in response to a lack of knowledge comes from work by [16] and [3]. Kovács et al. [16] found that 12-month-old infants increased their frequency of pointing across trials when interacting with an experimenter who provided them with novel labels for atypical members of familiar object categories relative to a condition where the experimenter offered them a familiar label in response to their pointing. [3] also found that infant pointing serves an interrogative function. In their study, 16-months-old infants were more likely to point to a novel object when interacting with a knowledgeable informant (i.e., an adult who had correctly labeled objects known to the child) than when interacting with an incompetent informant (i.e., an adult who had mislabeled familiar objects). At around 20-months, infants also begin to seek information not just when they are ignorant but when they lack confidence about what they know [9]. At this age, infants can monitor and communicate their own uncertainty, elicit information, and use that information to improve their performance. Using a nonverbal memory-monitoring paradigm, [9] showed that after training infants were able to strategically ask for help by turning towards and looking at their parents to avoid making mistakes. These findings reveal that infants are capable of monitoring their own uncertainty and non-verbally communicating it, in order to acquire knowledge from others.

Formulation. Once a request for information is triggered (initiation phase), the next step is to formulate a question to request information. The process of formulating a question can be divided into two broad steps: (1) identifying what information to ask for, and (2) phrasing the question so that it can be understood and answered [25]. To date, very little work has examined how infants' and toddlers' prior knowledge and explanatory biases influence the questions they formulate. However, there is some evidence that when infants and toddlers ask a question, they have some (implicit) ideas about what would constitute an appropriate answer. For instance, [7] reports that already 18-months-old children persist in asking a question if they do not receive an informative response, but rarely do so when they receive a response containing the target information.

Expression. While there is a growing body of evidence supporting the fact that infants and toddlers are capable of distinguishing between more and less reliable informants (see [24] for a review), there is much less evidence on how young children use these abilities to selectively decide where to direct their questions. The little evidence that exists, supports the claim that even in infancy the decision to request information is shaped by the infant's evaluations regarding the likelihood of obtaining an informative reply.

Response evaluation and follow-up. There is evidence suggesting that infants and toddlers learn from the responses they receive to their (non-verbal) information seeking requests. Rowe found a robust correlation between the onset and frequency of children's pointing and their subsequent vocabulary size [26]. These findings suggest that children learn labels from receiving informative replies to their pointing gestures. Lucca and Wilbourn found that 18-months-old but not 12-months-old infants showed greater memory for labels of novel objects provided in response to their pointing gestures [18]. One possibility is that children's expectations of their pointing eliciting information develops between 12 and 18 months, and that knowing that they will receive an answer allows them to prepare to encode the answers they receive into memory. An alternative possibility is that this expectation is already present in 12-months-old children but that they lack the memory capacities to encode the answers they receive into memory.

3 RELATED WORK

As motivated in the introduction, in this work we are interested in a use case where people teach a robot about objects and actions in the environment via unconstrained natural language interaction. To improve the efficiency, we propose making use of an active learning strategy, where the robot directs its own learning and acquires the information it needs, by asking questions to the human partner. This research spans natural language learning and active learning. Grounded language learning is concerned with linking the meaning of natural language to machine representations of the physical world [20]. Previous work on learning to ground object names and attributes through dialog with human partners includes [33], [30], [22] [31]. The aforementioned systems rely on learning passively from a human teacher.

One way of improving the efficiency is to make use of active learning strategies, where a robot directs its own learning and asks questions about specific topics. Evidence from experimental investigations (e.g., [6]) suggests that enabling a robot to ask questions that elicit diverse types of input leads to faster and more efficient learning of relevant concepts. [36], related work by [38] and [29] propose the generation of questions to learn objects and visual properties. [19] generate questions to solve ambiguity in the object references and for grasping commands. While the work listed above investigates various ways of employing natural language questions of different linguistic complexity and semantic specificity (e.g. *What is this? What is the object on the left side of the red cube?*) to facilitate the robot's learning process, [5] assess the effect of different non-verbal human-driven feature eliciting strategies, e.g.: in a grocery sorting task, human selection of typical objects per grocery class helps the robot to learn class-relevant features. In contrast, we are interested in a more fundamental design of information seeking

where language capacity and information seeking strategies go hand in hand. At a stage of development where the agent is about to learn word-referent mappings grounded in task scenarios, we therefore leverage insights from the rich corpus of research on language learning, metacognition and question-asking during early infancy (c.f. Section 2). In previous work we have developed a base system enabling a robot to learn word-referent mappings through observation of situated task descriptions provided by a human tutor. The goal of the presented extension is to develop a biologically-inspired question-asking policy that is intertwined and develops with the agent's state of knowledge acquisition and communication capacity.

4 LEARNING MODEL

In the following, we present a first version of an implemented learning model that takes into account findings from the research on metacognition and non-verbal question-asking presented in Section 2. We begin with a brief introduction of the base learning model and then discuss potential extensions for modelling active information seeking and their implementation on a Pepper robot.

4.1 Base Model

The base model relies on sequential input of utterance-situation pairs describing simple actions such as TAKE, PUT and PUSH. The utterance-situation pairs are of the form <*I take the box* – ACTION1 OBJECT1>, <*and put it next to the can* – ACTION2 OBJECT1 OBJECT2>, <*then I push the can to the left of the bottle* – ACTION3 OBJECT2 OBJECT3> representing a multimodal episode comprising the utterance *I take the box and put it next to the can. Then I push the can to the left of the bottle.* and related visual action. (For the algorithm aligning visual actions and utterances see [10], Section 3.2.) Normalized point-wise mutual information (*npmi*) is used as the key measure to compute mappings between words and objects, and words and actions. This way, a lexicon of word-object and word-action pairs is learned incrementally. We use pointwise mutual information (*pmi*) as a measure to compare how much the actual probability of a co-occurrence $p(w, o)$ of a particular word w occurring in the utterance and a certain object 0 occurring in the visual representation in the utterance-situation pairs presented to the learning system differs from what one would expect it to be based on the assumption of independence of the occurrence of w and o, i.e. $p(w)p(o)$. Using the normalized variant *npmi* confines the values between -1 and 1 and thus makes individual values comparable. This allows us to define thresholds above which a word-object mapping will be added to the lexicon. (See [4] for a discussion on *pmi* and *npmi* in the context of collocation extraction. While in collocation extraction, one is interested in the co-occurrence of word pairs, we are interested in the co-occurrence of word-referent (object) pairs.)

The architecture of the model including object tracking, automatic speech recognition (ASR) and automatic speech synthesis (ASS) is shown in Figure 1 and is described in more detail in [14]. At the current stage of development, ASS is used as a control function for verbalising what the robot has already learned, i.e., if the human tutor grabs, moves or puts an object and the words for the respective actions and objects are already learned, the robot verbalises what it sees, such as *grab bottle, move box* etc.

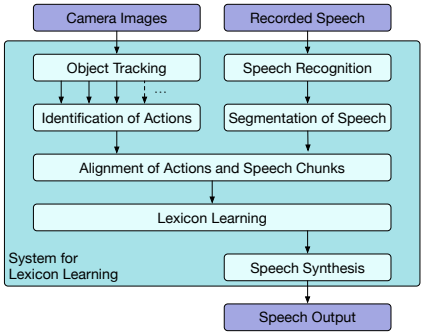

Figure 1: System structure: The tracked object poses and speech input are used for lexicon learning. Descriptions of learned scenes can be synthesized by the robot.

In the course of the interaction with the human tutor, the position of each tracked object is sent to the learning component, as well as the text output of Google ASR. When an object moves in the current setup, we can safely infer that an agent (the tutor) has moved the object volitionally. If a detected object movement and a recognized speech segment co-occur with sufficient temporal overlap, the two can be processed as an established utterance-situation pair. If the object has not been lifted off the surface, one can infer that the action associated with this movement is push, otherwise, there are two actions: take and put, whereby the grabbed object is typically moved or put next to another object. This allows to generate some sort of simple semantics for the perceived event, comprising the action, the object involved and facultatively the goal of the action (which other object the moved object is moved next to).

The speech input may vary greatly (and variation actually facilitates learning), however, at the beginning of the learning process, each word from the utterance will be linked with each object and action referent in the visual situation paired with the utterance. In a first step, each of these links is evaluated against concurring links that contain the same word using *npmi*. If the difference between the link with the highest *npmi* value and the link with the next lower *npmi* value is greater than a given threshold (default: 0.05), we boost the highest link by incrementing a so called 'boost' counter while the 'decline' counter is incremented for all other links. This process is applied for all words independently. In a second step, if a link is outranked by another link with the same referent and the difference between the *npmi* values is greater than the threshold, an extra count is given on the so called 'exclude' counter. This process is applied for all referents (objects) independently.

While 'decline' compares links on the basis of concurring references to a given word, the 'exclude' counter stores information of co-occurrences between words linked to the same referent. Finally, links are assigned to the lexicon if the following conditions are met: (i) the *npmi* value is greater than a given absolute threshold (default: 0.25) and (ii) the ratio between the sum of 'decline' and 'exclude' counts, and the 'boost' count is smaller than a given threshold (default: 0.6). This filter prevents "second best" links reaching the *npmi*

threshold to automatically enter the lexicon. Links that are already established in the lexicon are constantly re-evaluated. If there are concurrent links and one of these links does no longer meet the conditions listed above, it is removed from the lexicon.

In this approach, the robot learns from observing multimodal scenes presented by the human tutor. The TAKE, PUT and PUSH actions have to be performed several times by the tutor with different objects. The extended model described below allows the robot to also actively request for information.

4.2 Extensions Motivated by Infant Metacognition and Question-asking

In the following, we discuss those aspects from Section 2 which we transfer to robot learning, and present an extension to the base learning model which can be realized with the setup of Pepper's perception and action capabilities.

The following procedure describes the currently implemented processing steps where the base learning model is interleaved with mechanisms of active information seeking of the learning agent. In terms of the components of the question-asking model proposed by Ronfard et al. [25], the procedure described below realises (i) the initiation part of active information seeking, i.e., the learning system needs to be aware of its need for information, see step (2) below; and (ii) the formulation part, i.e., the learning system has to identify what information to ask for and how to phrase the question, see step (3) below. So far, we do not account for the 'Expression' component proposed in [25], where the learning system has to identify whether a potential informant is available and can be asked. At the current stage of technical realization, we assume a constructive and supportive tutor who produces adequate names for the objects and does not challenge the system by deliberately providing wrong (not related to the visual situation) or nonsensical utterances. Giving more weight to the utterance-situation pairs stemming from active information seeking as suggested in (3), third bullet point below is a first and preliminary attempt to account for evidence suggesting that the learning of infants and toddlers is boosted by responses they receive to their (non-verbal) information seeking requests, cf. the 'Response evaluation' component in [25].

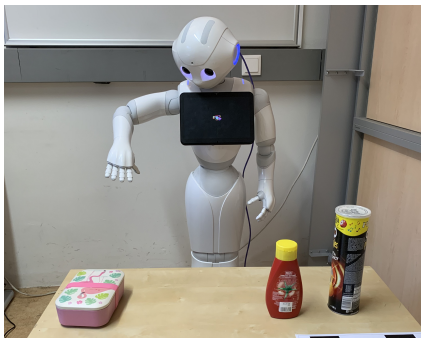

Figure 2: The Pepper robot is pointing at the box to acquire the name of the object from the tutor.

We describe the procedure starting with an empty lexicon, i.e., a state where the agent has not yet learned any word mappings.

(1) For the first n seconds, apply the base learning model. Utterance-situations pairs are derived from the aligned visual information and the verbal description of the human tutor and used as input to the lexicon learning component (cf. Figure 1). The parameter n can be set at the beginning of the learning phase. Currently, we experiment with $n = 20$ seconds.
(2) After n seconds of base learning, the agent checks its lexicon and decides whether active information seeking should be triggered. This is done in the following way:
   - First, the agent checks whether at least one object reference has been successfully learned according to the criteria described in Section 4.1.
   - Second, the agent checks whether at least one of the tracked objects has none or more than one referents assigned to it, and puts these objects in a list of not yet successfully learned objects O.
(3) If $O \neq \{\}$, enter the active information seeking mode:
   - First, randomly select one object $o_j \in O$. (Due to constraints in Pepper's anatomy, we also check if the object is pointable, i.e. the chosen object must be sufficiently apart from the other objects so that Pepper's pointing gesture is distinctive.)
   - Second, trigger an information seeking event involving o_j, such as initiate a pointing gesture at o_j performed by Pepper (Figure 2).
   - Third, for the next m seconds following the information seeking event constrain the agent's expectations regarding the tutor's input; in the current implementation this means that whatever utterance comes from the tutor will be paired with a visual situation comprising only the object referent o_j the agent has pointed at. Add this specific utterance-situation pair as input to the base learning system. We currently experiment with $m = 7$ sec and with giving more weight to the utterance-situation pairs stemming from active information seeking. As a first shot, this is done by repeatedly (up to 3 times) inputting the respective utterance-situation pair to the base learning model.
(4) Go back to / go on in the base learning mode.

5 CONCLUSIONS

We have presented an extension to an incremental cross-modal learning model which is inspired by active information seeking in early infant learning. The model learns word-referent mappings based on visual and verbal information (utterance-scene pairs) provided by a human in task-oriented tutoring scenarios. The human performs basic object handling activities such as grabbing an object, putting / pushing an object next to another object. Currently, the learning model is implemented on the humanoid robot Pepper. Online tests with human tutors interacting with Pepper are underway. We experiment with model parameters, such as for how long the model goes into active learning mode, and whether and to which extent the utterance-scene pairs obtained during active learning mode should receive more weight than utterance-scene pairs obtained during non-active learning. Moreover, we are interested in learning effectivity (i.e., the time / learning iterations required to

learn all objects present in the scenario) with and without active information seeking. Moreover, we will compare different approaches of communicating intent of information seeking strategies such as pointing, combined with gaze or vocalization "umm". Results will be reported at the workshop.

ACKNOWLEDGMENTS

The authors would like to thank Clara Haider for providing the motion algorithms for Pepper. This research is supported by the Vienna Science and Technology Fund (WWTF), project RALLI (ICT15-045) and the CHIST-ERA project ATLANTIS (2287-N35).

REFERENCES

[1] Deborah Teo Bartz. 2017. Young Children's Meta-Ignorance. (jan 2017). https://dash.harvard.edu/handle/1/33051609

[2] Elizabeth Bates, Luigia Camaioni, and Virginia Volterra. 1975. The Acquisition of Performatives Prior to Speech. *Merrill-Palmer Quarterly* (1975).

[3] Katarina Begus and Victoria Southgate. 2012. Infant pointing serves an interrogative function. *Developmental Science* 15, 5 (sep 2012), 611–617. https://doi.org/10.1111/j.1467-7687.2012.01160.x

[4] Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL* (2009), 31–40.

[5] Kalesha Bullard, Sonia Chernova, and Andrea L. Thomaz. 2018. Human-Driven Feature Selection for a Robotic Agent Learning Classification Tasks from Demonstration. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 6923–6930. https://doi.org/10.1109/ICRA.2018.8461012

[6] Kalesha Bullard, Andrea L. Thomaz, and Sonia Chernova. 2018. Towards Intelligent Arbitration of Diverse Active Learning Queries. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 6049–6056. https://doi.org/10.1109/IROS.2018.8594279

[7] Michael M Chouinard. 2007. Children's questions: a mechanism for cognitive development. *Monographs of the Society for Research in Child Development* 72, 1 (mar 2007), vii–ix. https://doi.org/10.1111/j.1540-5834.2007.00412.x

[8] John H. Flavell. 1999. Cognitive Development: Children's Knowledge About the Mind. *Annual Review of Psychology* 50, 1 (feb 1999), 21–45. https://doi.org/10.1146/annurev.psych.50.1.21

[9] Louise Goupil, Margaux Romand-Monnier, and Sid Kouider. 2016. Infants ask for help when they know they don't know. *Proceedings of the National Academy of Sciences of the United States of America* 113, 13 (mar 2016), 3492–6. https://doi.org/10.1073/pnas.1515129113

[10] Stephanie Gross, Matthias Hirschmanner, Brigitte Krenn, Friedrich Neubarth, and Michael Zillich. 2018. Action verb corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

[11] Marianne Gullberg and Kees de Bot (Eds.). 2010. *Gestures in Language Development*. Benjamins Current Topics, Vol. 28. John Benjamins Publishing Company, Amsterdam. https://doi.org/10.1075/bct.28

[12] R R Hampton. 2001. Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences of the United States of America* 98, 9 (apr 2001), 5359–62. https://doi.org/10.1073/pnas.071600998

[13] Paul L Harris, Deborah T Bartz, and Meredith L Rowe. 2017. Young children communicate their ignorance and ask questions. *Proceedings of the National Academy of Sciences of the United States of America* 114, 30 (jul 2017), 7884–7891. https://doi.org/10.1073/pnas.1620745114

[14] Matthias Hirschmanner, Stephanie Gross, Brigitte Krenn, Friedrich Neubarth, Martin Trapp, and Markus Vincze. 2018. Grounded Word Learning on a Pepper Robot. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. ACM, 351–352.

[15] Jana M. Iverson, Olga Capirci, and M.Cristina Caselli. 1994. From communication to language in two modalities. *Cognitive Development* 9, 1 (jan 1994), 23–43. https://doi.org/10.1016/0885-2014(94)90018-3

[16] Ágnes Melinda Kovács, Tibor Tauzin, Ernő Téglás, György Gergely, and Gergely Csibra. 2014. Pointing as Epistemic Request: 12-month-olds Point to Receive New Information. *Infancy* 19, 6 (nov 2014), 543–557. https://doi.org/10.1111/infa.12060

[17] Ulf Liszkowski, Penny Brown, Tara Callaghan, Akira Takada, and Conny De Vos. 2012. A prelinguistic gestural universal of human communication. *Cognitive Science* 36, 4 (2012), 698–713.

[18] Kelsey Lucca and Makeba Parramore Wilbourn. 2018. Communicating to Learn: Infants' Pointing Gestures Result in Optimal Learning. *Child Development* 89, 3 (may 2018), 941–960. https://doi.org/10.1111/cdev.12707

[19] I. Lutkebohle, J. Peltason, L. Schillingmann, B. Wrede, S. Wachsmuth, C. Elbrechter, and R. Haschke. 2009. The curious robot - Structuring interactive robot learning. In *2009 IEEE International Conference on Robotics and Automation*. IEEE, 4156–4162. https://doi.org/10.1109/ROBOT.2009.5152521

[20] Cynthia Matuszek. 2018. Grounded Language Learning: Where Robotics and NLP Meet. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, California, 5687–5691. https://doi.org/10.24963/ijcai.2018/810

[21] C. Moore and P. J. Dunham. 1995. Joint Attention: Its Origins and Role in Development. (1995). https://philpapers.org/rec/MOOJAI-2

[22] Natalie Parde, Adam Hair, Michalis Papakostas, Konstantinos Tsiakas, Maria Dagioglou, Vangelis Karkaletsis, and Rodney D. Nielsen. 2015. Grounding the Meaning of Words through Vision and Interactive Gameplay. *undefined* (2015). https://www.semanticscholar.org/paper/Grounding-the-Meaning-of-Words-through-Vision-and-Parde-Hair/b3edfadc8896ae18ede23c214b143390a288ccba

[23] Clint J Perry and Andrew B Barron. 2013. Honey bees selectively avoid difficult choices. *Proceedings of the National Academy of Sciences of the United States of America* 110, 47 (nov 2013), 19155–9. https://doi.org/10.1073/pnas.1314571110

[24] Diane Poulin-Dubois and Patricia Brosseau-Liard. 2016. The Developmental Origins of Selective Social Learning. *Current Directions in Psychological Science* 25, 1 (feb 2016), 60–64. https://doi.org/10.1177/0963721415613962

[25] Samuel Ronfard, Imac M. Zambrana, Tone K. Hermansen, and Deborah Kelemen. 2018. Question-asking in childhood: A review of the literature and a framework for understanding its development. *Developmental Review* 49 (sep 2018), 101–120. https://doi.org/10.1016/J.DR.2018.05.002

[26] Meredith L Rowe. 2008. Child-directed speech: relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of child language* 35, 1 (feb 2008), 185–205.

[27] J. David Smith. 2009. The study of animal metacognition. *Trends in Cognitive Sciences* 13, 9 (2009), 389–396. https://doi.org/10.1016/j.tics.2009.06.009 arXiv:NIHMS150003

[28] Sumarga H Suanda, Linda B Smith, and Chen Yu. 2016. The Multisensory Nature of Verbal Discourse in Parent–Toddler Interactions. *Developmental Neuropsychology* 41, 5-8 (2016), 324–341.

[29] Yuyin Sun, Liefeng Bo, and Dieter Fox. 2014. Learning to identify new objects. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 3165–3172. https://doi.org/10.1109/ICRA.2014.6907314

[30] Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J. Mooney. 2016. Learning Multi-Modal Grounded Linguistic Semantics by Playing. *undefined* (2016). https://www.semanticscholar.org/paper/Learning-Multi-Modal-Grounded-Linguistic-Semantics-Thomason-Sinapov/0a0ee20a75d82172bfb930dc5f7d8e07a22ac3a6

[31] Andrea Vanzo, Jose L. Part, Yanchao Yu, Daniele Nardi, and Oliver Lemon. 2018. Incrementally Learning Semantic Attributes through Dialogue Interaction. *undefined* (2018). https://www.semanticscholar.org/paper/Incrementally-Learning-Semantic-Attributes-through-Vanzo-Part/caab6304ed64c461d3376db2827058b0f9b423cc

[32] Marcel V. J. Veenman, Bernadette H. A. M. Van Hout-Wolters, and Peter Afflerbach. 2006. Metacognition and learning: conceptual and methodological considerations. *Metacognition and Learning* 1, 1 (apr 2006), 3–14. https://doi.org/10.1007/s11409-006-6893-0

[33] Adam Vogel, Karthik Raghunathan, and Daniel Jurafsky. 2010. Eye Spy: Improving Vision through Dialog. *AAAI Fall Symposium: Dialog with Robots* (2010). https://www.semanticscholar.org/paper/Eye-Spy

[34] Alexandra M. Was and Felix Warneken. 2017. Proactive help-seeking: Preschoolers know when they need help, but do not always ask for it. *Cognitive Development* 43 (jul 2017), 91–105. https://doi.org/10.1016/J.COGDEV.2017.02.010

[35] Breanna M Winder, Robert H Wozniak, Meaghan V Parladé, and Jana M Iverson. 2013. Spontaneous initiation of communication in infants at low and heightened risk for autism spectrum disorders. *Developmental psychology* 49, 10 (oct 2013), 1931–42. https://doi.org/10.1037/a0031061

[36] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. 2018. Visual Curiosity: Learning to Ask Questions to Learn Visual Recognition. (oct 2018). arXiv:1810.00912 http://arxiv.org/abs/1810.00912

[37] Chen Yu and Dana H Ballard. 2007. A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing* 70, 13 (2007), 2149–2165.

[38] Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2017. Learning how to learn: an adaptive dialogue agent for incrementally learning visually grounded word meanings. (sep 2017). https://doi.org/10.18653/v1/W17-2802 arXiv:1709.10423