Models of Cross-Situational and Crossmodal Word Learning in Task-Oriented Scenarios

Brigitte Krenn, Sepideh Sadeghi, Friedrich Neubarth, Stephanie Gross, Martin Trapp, Matthias Scheutz,

***Abstract*—We present two related but different cross-situational and cross-modal models of incremental word learning.**

Model 1 is a Bayesian approach for co-learning object-word mappings and referential intention which allows for incremental learning from only a few situations where the display of referents to the learning system is systematically varied. We demonstrate the robustness of the model with respect to sensory noise, including errors in the visual (object recognition) and auditory (recognition of words) systems. The model is then integrated with a cognitive robotic architecture in order to realize cross-situational word learning on a robot.

A different approach to word learning is demonstrated with Model 2, an information-theoretic model for object- and action-word learning from modality rich input data based on point-wise mutual information. The approach is inspired by insights from language development and learning where the caregiver/teacher typically shows objects and performs actions to the infant while naming what the teacher is doing. We demonstrate the word learning capabilities of the model, feeding it with crossmodal input data from two German multimodal corpora which comprise visual scenes of performed actions and related utterances.

I. Introduction

IF robots are to interact naturally and learn from humans in the future, mechanisms are needed to enable robots to learn new activities based on observations and linguistic instructions. Two questions regarding early word learning in infants are of particular interest for our work on grounded word learning for artificial agents: (1) the multimodal nature of early infant language acquisition where visual activity and linguistic cues are processed in parallel, (2) the types of words that are learned first and why. Regarding the former, see Gogate [1] who discusses the multisensory nature of communication where speech, visual and motor stimuli concur. Suanda et al. [2] show that parent-toddler communication is rich in multisensory input where parent discourse is closely tied to visual stimuli including what the parent has in her/his hands, the child currently grabs and has in her/his focus of visual attention (see also [3], [4], [5], [6]). Recent work by Nomikou et al. [7] shows evidence for the relation between caretakers' action-language synchrony in the input to six months old infants and the infants' later production of verbs.

Regarding the latter question which words are learned first, there is a broad discussion in the literature about the acquisition of nouns and verbs in young infants. Recent evidence suggests that very young infants across languages are already able to learn word-action mappings. Gogate and Maganti [8] show that preverbal infants (8-9 months) of a noun-friendly language background such as English map novel words onto actions long before they talk. This effect temporarily diminishes in postverbal infants (12-14 months) learning a noun-friendly language. These findings contradict earlier work such as Gentner [9], [10] who theorizes that for English learning (a noun-friendly language) toddlers learn nouns before verbs.

Gogate and Hollich [11] provide evidence for the following effects in word learning during infants' first three years: They suggest that in the first year infants' word mapping ability (onto actions and objects) emerges from learning words for referents that are most concrete or imageable. For instance, Nomikou et al. [7] found that the verbs used by caregivers in early interactions are tightly co-ordinated with ongoing actions and frequently in response to infant actions. In addition, there is evidence from studies on intention awareness and compliance that infants from six months on are sensitive to the intentionality of others' actions, see for instance [12], [13], [14].

Furthermore Gogate and Hollich explicate that in the second year the dominance hierarchy of lexical categories in the ambient language differently constrains the developing infants' attention to nouns or verbs. This might explain the noun bias of English learning toddlers identified by Gentner, or findings by Childers and Tomasello [15] which showed that it was easier for two-year-old English speaking children to recall new nouns than new verbs and they also produced 3 times more nouns than verbs. Evidence for the effects of the ambient language on word learning also comes from earlier work. See for instance Brown [16] for early verb learning in the Mayan language Tzeltal, or Choi [17] for Korean.

In the third year, according to Gogate and Hollich, children use what they have learned about their native language to make guesses about new words. Children from noun-friendly languages overcome their noun-bias. Understanding of linguistic cues leads them to flexibly learn the correct referents for verbs.

In addition, Chen et al. [18] found 6-8 month olds from English and Mandarin language environments could discriminate action changes but not object changes, whereas 17-19 months olds were able to discriminate both. Based on cross-linguistic comparisons of Chinese-, English-, and Japanese-speaking children, Imai et al. [19] provide evidence that both universally shared cognitive factors and language-specific linguistic factors matter for early word learning.

Various computational models of word learning have recently been proposed to demonstrate the acquisition of word-

B. Krenn, F. Neubarth, S. Gross and M. Trapp are with the Austrian Research Institute for Artificial Intelligence, Vienna, Austria (e-mail: firstname.lastname@ofai.at); S. Sadeghi and M. Scheutz are with the School of Engineering, Tufts University, Medford, MA (e-mail: firstname.lastname@tufts.edu).

to-meaning mappings. They typically either rely on co-occurrence statistics of words and meaning elements, see for instance Yu and Ballard [20], Frank et al. [21], or use information theoretic measures to model the association between words and referents. See for instance Kachergis et al. [22] who use entropy as the core measure to model familiarity and uncertainty for learning word-referent pairings, or Roy [23] who uses mutual information for grounding objects. While these works concentrate on noun referent learning, our Model 2 demonstrates both noun and verb referent learning. Another approach is presented in Alishahi and Fazly [24] who use the knowledge about lexical categories (a combination of semantic information derived from WordNet and morphosyntactic category such as verb, noun) in cross-situational word learning.

The approaches to word-referent mapping presented in the present paper differ from the above mentioned research in several ways. While [20], [21] are based on batch learning, both approaches presented in Section III of this paper focus on incremental learning where the lexicon is constantly updated, when the system is provided with new input. Thus they are comparable to [22] who also present an incremental model.

Model 1, the word-object mapping model described in Section III-A builds upon [21], in particular, the joint acquisition of the speaker's referential intention and word meanings, and transforms their approach into a mechanism for incremental learning where only a few scenes need to be seen in order to learn word-object mappings. Model 2 (described in Section III-B), in contrast, focuses on modality rich input, comparable to what Yu and Ballard [20] and Frank et al. [21] call social cues or Suanda et al. [2] address as crossmodal input. While Yu and Ballard employ a statistical machine translation model [25] for mapping between words and meaning, and Frank et al. focus on a Bayesian approach to co-learning words and referential intentions, an information-theoretic approach to word-referent learning is realized in Model 2 utilizing normalized pointwise mutual information as a key measure and some additional weighting mechanisms in order to decide which word-object or word-action link enters and remains in the lexicon. In this respect, Model 2 is closer to Roy [2003] or Kachergis et al. [2012] Roy demonstrated audio-visual mappings between object classes and segments from spontaneous speech in child-caregiver interactions. In contrast, Model 2 goes beyond noun-object learning, however, using data from adult teachering situations. Kachergis et al. tuned their model towards replicating the learning effects resulting from word learning experiments with adult humans presented with pictures of unusual objects while hearing spoken pseudo words (in each trial two pictures and 2 pseudowords were presented). In contrast, both our models are geared to learning word referent mappings from full-blown natural language utterances related to visual situations. While Alishahi and Fazly assume that children have already formed some lexical categories each of which contain a set of word forms before word learning starts, we do not assume any prior categorical knowledge in either of our word learning approaches. In the object-word learning experiments using Model 1 (described in Section IV-A), a word learning situation comprises an utterance and a scene represented by the list of visual objects. In the action-word mapping experiments employing Model 2 (described in Section IV-B), each learning situation consists of an utterance, an action label and labels for those objects which are under visual attention. These are objects the speaker holds in her/his hands, objects (A) which are moved, and objects (B) next to which object A is moved.

While the input data for the learning experiments presented in Section IV-A2 are obtained from real-world perceptual inputs to a robot's vision and speech recognition systems, the input data used in Section IV-B stem from two multimodal task corpora, the Action Verb Corpus (Section II-A) and the MMTD Corpus (Section II-B). We consider situated task-oriented communication in an teacher-learner setting as well suited for modelling natural language learning in robots. This is motivated by evidence showing that this kind of communication is rich in multimodal cues and thus comparable to parent-young infant communication [2], [1].

II. Multimodal Task Corpora

We start by introducing two corpora, the Action Verb Corpus (AVC, [26]) and Dataset 1 of the OFAI Multimodal Task Description Corpus (henceforth MMTD, [27])[1] , which are used by Model 2 (described in Section III-B) for learning word-action and object mappings. The data sets are geared towards modelling natural language learning in robots and inspired by early human language learning research, in particular by evidence for modality richness of the input to the infant's learning system [2], [1], [3], [4], [5], [6]. Our data comprise situated task-oriented communication where adult human teachers show and describe in natual language simple tasks such as moving a bottle next to a box and uttering something like *I take the bottle and put it next to the box.* This way, modality rich and highly redundant input to developing and testing our artificial learning systems was produced.

AVC consists of multimodal data from 12 humans (8 male, 4 female) performing in total 500 simple actions (Take, Put, and Push). MMTD is a collection of tasks where a human teacher arranges and rearranges pieces of fruit on a table, and explains what (s)he is doing to a camera for an anonymous learner to replicate the task. The corpus comprises scenes from 22 teachers resulting in 196 actions combined with related utterances serving as input to the learning model.

Both corpora comprise audio and video data. In addition, AVC also contains motion data including the 3D-coordinates of hand, wrist and elbow joints, and object positions. While the video recordings for the MMTD Corpus combine the perspectives of the teacher and the learner, and a view of the whole scene, the visual data in the AVC corpus are restricted to the 1st person perspective, which is comparable to a robot setting where the scene is perceived through the robot's eyes. Moreover, the corpora are annotated for information such as transliterations of the utterances, part-of-speech tags, related lemmas (base forms of words without inflectional information), location of the teacher looks (eye-head gaze),

[1]The corpora can be downloaded from http://www.ofai.at/research/interact/MMTD.html and http://www.ofai.at/research/interact/avc.html.

specific object being moved, whether a hand touches an object, whether an object touches the ground/table. The object-related information is used for determining which objects in the visual scene are in focus while an action is performed. All annotation tiers are time-aligned using Praat[2] for the transcription of the audio and using Elan[3] for aligning audio, video, and annotation tiers. In the following sections, the set-ups for data collection and the annotations for the two corpora are described.

A. Action Verb Corpus (AVC)

1) Setup for Collecting Data: Three objects were positioned on a table: a bottle, a can and a box. The user sat (or stood) in front of the table wearing an Oculus Rift DK2 Virtual-Reality headset[4] with a Leap Motion sensor[5] for hand tracking mounted, see Fig. 1 left side. A camera (Microsoft Kinect) was positioned opposite the user and directed at the table for object tracking. The user performed different actions defined by visual instructions and verbally described what he/she was doing. The user's speech and performed actions were recorded.

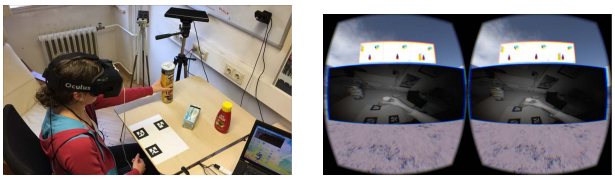

Fig. 1. Action Verb Corpus: Experimental Setup (left side). View through the Oculus including the instructions (right side).

The Leap Motion is a stereo infrared camera which is specialized on hand tracking. The Software Development Kit (SDK) provided detailed information of the position of the various joints of the user's arm down to the separate finger bones. We used the Leap Motion mounted on a VR headset to obtain the best available tracking performance. The Oculus Rift DK2 was worn by the user and provided the head pose of the user. The head pose was used to transfer the tracking data of the Leap Motion to a fixed coordinate system. In addition, the instructions for the current task were also displayed in the Oculus Rift above the camera images, see Fig. 1 right side. In this manner, the user was able to look at the instructions without moving his/her head, e.g. to look at printout versions of the instructions. Additionally, the setup forced the user to direct the Leap Motion to his/her hands because otherwise he/she would not have been able to see what he/she was doing. This behaviour was necessary for satisfying hand tracking performance.

For object tracking, the red, green, blue (RGB) as well as depth (D) data of the Kinect camera was recorded as a ROS bag[6] on a separate machine running Ubuntu. The object tracker from the V4R Library was used on the recorded data.[7] Models of the objects were created beforehand with the RTM-Toolbox.[8] The offline tracking enables the best possible tracking results because the object tracker can be tuned for a specific recording. Besides the position and orientation of the object, two Boolean variables were saved: object is in contact with the table and object is in contact with a hand. The former is set automatically depending on the object's position, the latter is currently annotated manually.

2) Annotation: Apart from the (low-level) representations resulting from the hand-arm and object trackers including per frame the 3D positions of the joints in the elbow, wrist and knuckles of the teacher's left and right hand as well as the object positions, the data were further annotated for: (i) two kinds of transliteration: the one as close as possible to speech preserving speech related signals in the utterance such as hesitations, interruptions, and corrections; the other one close to written text in order to apply computational linguistic tools such as part-of-speech taggers, stemmers, phrase chunkers and parsers, which are typically trained on written text; (ii) part-of-speech tags; (iii) canonical forms of inflected words (lemmas); (iv) hand touches object; (v) object touches surface/ground; (vi) object A moves next to object B. Except for the transliterations and whether a hand touches an object, all tiers were automatically annotated and manually corrected.

B. The OFAI Multimodal Task Description Corpus (MMTD)

1) Setup for Collecting Data: The teacher stood in front of a table with the following objects placed on it: an empty sheet of paper and a plate with three pieces of fruit – a banana, a pear and a strawberry, see Fig. 2. The task for the teachers was to take the pieces of fruit one after the other from the plate and arrange them on the piece of paper, and describe what they were doing to the camera (cam 1) for a prospective listener/learner. As regards examples for action related utterances see: *ich nehme dann die Erdbeere* ('I take then the strawberry', sample utterance related to a TAKE-action) *und lege sie vor mir auf die rechte Seite neben die Banane* ('and put it in front of me on the right side next to the banana', sample utterance related to a PUT-action).

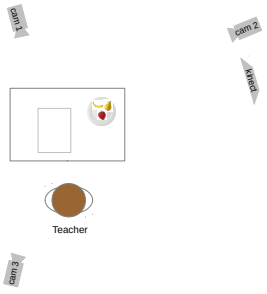

Fig. 2. Schematic setup of Task 1 of the MMTD Corpus.

2) Annotation: The data had been originally annotated for transcriptions and transliterations of the teacher's utterances, parts-of-speech, eye gaze and gestures of the teacher, and for the specific objects present on the scene referenced in

[2]http://www.fon.hum.uva.nl/praat/
[3]https://tla.mpi.nl/tools/tla-tools/elan/
[4]thttps://www.oculus.com/dk2/
[5]https://www.leapmotion.com
[6]A bag is a file format in ROS for storing ROS message data. Cf. http://wiki.ros.org/Bags. The Robot Operating System (ROS) is a flexible framework for writing robot software. Cf. https://www.ros.org/about-ros/.
[7]http://www.acin.tuwien.ac.at/forschung/v4r/software-tools/v4r-library/
[8]http://www.acin.tuwien.ac.at/forschung/v4r/software-tools/rtm/

the teacher's utterance. For the present word learning experiments, additional information was manually annotated, i.e., whether the left or right hand of the teacher touched an object (touchRightHand, touchLeftHand); whether an object touched the ground/table (touchGround); whether object A was moved next to object B (moveObject). The data were manually annotated based on the synchronised input streams of the video cameras (cam 1 to cam 3) and the recordings of the utterances. As illustrated in Fig. 3 mentioning an object in speech almost always correlated with an action that involves the object. Furthermore, if we identify an action such as TAKE as involving touch by hand and loss of connection to the ground, and PUT as movement and gain of connection to ground with a potential placement near another object, we observed that these feature combinations representing basic actions co-occur with the corresponding verbs in the linguistic description.

The present data relate to stimuli used in developmental psychology for studying word-referent learning in early childhood. See for instance [18], [28], [8] who pair verbal stimuli with situated action to investigate noun and verb learning. Whereas these studies work with nonsense words in highly controlled laboratory settings, Nomikou et al. [7] use a more naturalistic setting of mothers interacting with their child while changing diapers in order to investigate the relationship between action-language synchrony and verb learning. Analysing the multimodal data, they found amongst others that the verbs the mothers used in the interaction with their child were tightly coordinated with the ongoing action. These findings in turn strenghten our decision to use in our learning experiments the kind of task description data available from MMTD and AVC where action and action-related utterance are tightly coupled.

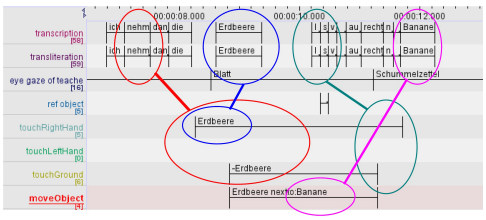

Fig. 3. MMTD: Sample annotation and crossmodal relations; circles and lines indicate relations between modalities.

III. Models of Object and Action Learning

In this section, we present two models of object-word learning and action-word learning, focusing on different aspects of the learning models and the input being processed.

First, we present an **incremental model for cross-situational word-referent learning for words with concrete object references (Model 1).** The specialty of the model is that it distinguishes between referential and non-referential use of words. A word is referential, when it refers to an object present in the current situation and non-referential otherwise. For illustration, see the sequence of situations in Table I which allow the model to incrementally learn mappings between the word *knife* and the object KNIFE and the word *cup* and the object CUP from only a few varying situations (henceforth scenes). In section III-A, the model details are presented. The model is then examined with respect to its robustness to noisy input from speech recognition and computer vision (section IV-A1), embedded in a robot architecture and run on a PR2 robot (section IV-A2).

input situation	utterance	visual scene
situation 1:	*look at the knife*	KNIFE, CUP, BOWL
situation 2:	*knife*	KNIFE, BOWL
situation 3:	*look at the knife*	KNIFE, CUP, BOWL
situation 4:	*cup*	KNIFE, CUP
situation 5:	*look at the cup*	CUP, BOWL
...	...	...

TABLE I
WORD-OBJECT LEARNING FROM VARYING OBSERVATIONS OF INTENTIONAL LANGUAGE USE.

Second, we present a **model for action learning from modality rich input data** (**Model 2**, Section III-B). While Model 1 uses referential intention as key concept, and is geared towards learning from only a few varying situations, Model 2 focuses on learning from highly redundant input, whereby redundancy comes from modality richness and repetitiveness of the data. Fig. 3 illustrates crossmodal relations. The AVC corpus comprises 78 basic TAKE, PUT and PUSH situations, and MMTD comprises 202 basic TAKE and PUT situations, each situation combining visual action and related utterance. See Table II for sample input situations. As each utterance relates to an action in the visual scene, utterance, action label and labels for the objects in the visual focus are input to the learning algorithm.

AVC	
Visual Scene	Utterance
TAKE TEAHORIZONTAL	ich nehme die Schachtel (I take the box)
PUT TEAHORIZONTAL PRINGLES	und stelle sie links neben die Dose (and put it to the left of the can)
TAKE KETCHUP	ich nehme die Flasche (I take the bottle)
PUT KETCHUP PRINGLES	und stelle sie rechts neben die Dose (and put it to the right of the can)
PUSH PRINGLES KETCHUP	ich schiebe die Dose vor die Flasche (I push the can infront of the bottle)
PUSH KETCHUP TEAHORIZONTAL	ich schiebe die Flasche hinter die Schachtel (I push the bottel behind the box)
PUSH TEAHORIZONTAL KETCHUP	ich schiebe die Schachtel neben die Flasche (I push the box next to the bottle)

MMTD	
Visual Scene	Utterance
TAKE BANANA PLATE	*und ich nehme jetzt die Banane* (and I take now the banana)
PUT BANANA PAPER	*und lege sie in die Mitte vom Blatt* (and put it in the middle of the sheet)
TAKE STRAWBERRY PLATE	*dann nehme ich die Erdbeere* (then I take the strawberry)
PUT STRAWBERRY BANANA	*und lege sie neben die Banane* (and put it next to the banana)
TAKE STRAWBERRY PLATE	*und dann nehme ich die Erdbeere vom Teller* (and then I take the strawberry from the plate)
PUT STRAWBERRY BANANA	*und packe sie auch auf das leere Blatt neben die Banane* (and put it too on the empty sheet next to the banana)

TABLE II
SAMPLE INPUT SITUATIONS AS DERIVED FROM AVC AND MMTD: VISUAL SCENE AND RELATED UTTERANCE

A. Model 1: Cross-Situational Object-Word Learning

Here, we present the details of our current word learning model (Model 1) which is limited to learning words which refer to concrete objects in the scene. The input to the model are word learning situations. Each of which consists of an utterance-scene pairing where the utterance is an un-ordered set of words and the scene is a list of objects present in the scene. Building on Frank et al. [21], the learner assumes that in each situation, the teacher uses the generative process illustrated in Fig. 4 to produce an utterance (W_s) corresponding to the current scene (O_s) and using a context appropriate portion (L) of the full lexicon to generate the referential words. *Lexicon* refers to a many-to-many mapping between words and objects, and a *referential word* refers to a noun with a concrete object referent. In each situation, the model has to infer which words are referential and what object they refer to. *Referential intentions* of the teacher (I_s) which refer to the objects that are present in the scene and which the teacher is talking about, determine the space of possible referents for each referential word in the utterance (W_s).

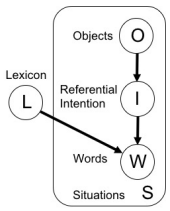

Fig. 4. The graphical model describing the generation of words (W_s) from the intention (I_s) and lexicon (L), and the generation of the intention (I_s) from the objects present in the scene (O_s), where s indexes the situation. The plate indicates multiple copies of the model for different situations (utterance-scene pairs). Image from [21].

In each situation, the model uniformly samples a subset from the power-set of all the objects present in the situation (O_s) representing the referential intention(s) of the teacher (I_s). Each word in the utterance is assumed to be referential with probability γ and non-referential with probability $1-\gamma$. The probability of non-referential use (P_{NR}) of a referential word (words in the model lexicon) is set to $\kappa < 1$ (to penalize the non-referential use of referential words), and is set to 1 for non-referential words. The probability of referential use of a referential word in reference to a particular object (P_R) is the probability of the word being chosen uniformly from the set of all words linked to that object in the lexicon.

In each situation the model tries to reverse the generative process described in Fig. 4 to discover a context-appropriate portion of the full lexicon used by the speaker, where context refers to the entities (words and objects) in the current situation. In doing so, in each situation, the model infers a *mini-lexicon* as a context-appropriate portion of the full lexicon. The model uses its current knowledge of the full lexicon and co-occurrence statistics accumulated across situations for hypothesis generation (generation of hypothetical mini-lexica) and hypothesis testing (inferring the best mini-lexicon). The best mini-lexicon found in each situation then will be integrated into the full lexicon inferred by the learner. These steps will be described in more detail in Section III-A1.

Inferring the best mini-lexicon in each situation, requires finding the MAP (maximum a posteriori) mini-lexicon by marginalizing over all possible referential intentions, since I_s is unobserved. The model finds the MAP mini-lexicon according to the Bayes equation and the probability distribution that it defines over unobserved mini-lexica (L) and the relevant corpus of situations (C) including the current situation as well as the extracted ones from the lexicon (which share some entity with the current situation). The extracted situations are made of the existing (in the lexicon) mappings for each word and object in the current situation.

$$P(L|C) \propto P(C|L)P(L) \tag{1}$$

We use $P(L) \propto e^{-\alpha \cdot |L|}$ serving as a soft mutual exclusivity constraint to produce a preference for one-to-one mappings in the mini-lexicon inferred in each situation. Marginalizing over all possible intentions in each situation we can rewrite the likelihood term $P(C|L)$ as:

$$P(C|L) = \prod_{s \in C} \sum_{I_s \subseteq O_s} P(W_s|I_s, L)P(I_s|O_s) \tag{2}$$

Assuming that $P(I_s|O_s) \propto 1$ and that the words of the utterance are generated independently, we can rewrite the term $P(W_s|I_s, L)$ as:

$$P(W_s|I_s, L) = \prod_{w \in W_s} [\gamma \cdot \sum_{o \in I_s} \frac{1}{|I_s|} P_R(w|o, L) + (1-\gamma) P_{NR}(w|L)] \tag{3}$$

We employ the equations above in each situation, to find the MAP mini-lexicon which describes the generation of situations in C, including the current situation as well as the relevant ones extracted from the full lexicon. We use "lexicon" and "full lexicon" interchangeably in the rest of this paper.

Our model departs from the previous model which is fully Bayesian and a batch learning algorithm assuming full access to all observations [21]. Our learning algorithm is incremental, memory-limited (memory of observations) and only locally (in the context of single situations) Bayesian. A lack of access to all datapoints is not a barrier for convergence of our learning algorithm [29], [30], [31], [32]. The number of computations upon receiving a new situation to update the lexicon depends on the input situation and the number of learned mappings for the existing items (words and objects) in the current version of model lexicon. Since the model is memory-limited, the number of such word learning situations is limited and since the number of items in an utterance and scene are limited too, the number of computations remain fixed as the size of data grows, allowing for scalability as well as online processing of data.

1) Incremental Learning Algorithm: We use the incremental and memory-limited learning algorithm proposed in [29] which remains tractable as the size of data grows. The learning algorithm is truly incremental as it sees each situation only once and performs no iteration over data. Furthermore, its memory of past observations is limited to the word-object mappings stored in the lexicon. The algorithm uses context-appropriate word-object mappings available in memory for hypothesis generation (generation of hypothetical mini-lexica) and hypothesis evaluation (inferring the MAP mini-lexicon). This allows for quick hypothesis generation and hypothesis testing while keeping the Bayesian inference tractable as the amount of data grows. Bayesian inference to infer the MAP mini-lexicon, is only applied locally with limited but

relevant evidence available in memory (relevant to the current observation). Our learning algorithm has two components: (1) inferring the MAP mini-lexicon in each situation, (2) integrating the new mini-lexicon in the current full lexicon, while applying mutual exclusivity constraints. The process of inferring the MAP mini-lexicon, subsequently has two distinct components: (1) generating mini-lexicon proposals, and (2) scoring the generated mini-lexica. Scoring is performed by computing the relative posterior probability of the mini-lexicon proposals based on Equation (1). Generating mini-lexicon proposals is guided by stochastic search techniques.

To summarize, in each situation, the learning algorithm infers a mini-lexicon as an approximation to the context-appropriate portion of the full lexicon used by the speaker. This mini-lexicon is then integrated into the model's current full lexicon by adding the mappings in the best mini-lexicon to the lexicon, and removing the existing alternative mappings from the lexicon, applying a strict mutual exclusivity constraint between situations. We also apply a soft mutual exclusivity constraint, in each situation, through the use of the mini-lexicon prior probability function which is exponential in the size of mini-lexicon and produces a preference for smaller mini-lexica. For more details about the learning algorithm, refer to Sadeghi et al. [29]. To learn more about the application of the same learning algorithm in extended versions of the graphical model used in the present paper refer to Sadeghi and Scheutz [33], [30].

B. Model 2: Crossmodal Action-Word Learning – Actions and Related Objects

In the following, we present a model of crossmodal word learning (Model 2) where action verbs and words referring to the objects involved in the actions are incrementally learned from modality rich input.

1) Data Preparation: The input data for the action verb and action plus object learning experiments presented in Section IV comprise a series of situations. Each situation may consist of one or two events with a different action (e.g., *schieben* "push" alone, or *nehmen* "take" with subsequent *stellen/legen* "put"). Taking the situation as a multimodal perceptual frame, it makes sense to use the term event for a single occurrence of an action that can be perceptually individuated and aligned with a unique utterance. An event comprises an action together with all objects involved in that action. To identify the multimodal sequences, i.e., time series on the annotation tiers related to one action/event, the description episodes of MMTD and AVC were automatically segmented and aligned.

Such an alignment is not straightforward since in MMTD teachers often start to describe the action/event before actually performing it. Segmentation in MMTD and AVC can be facilitated through speech pauses, and as an idiosyncrasy of MMTD – which originally was not designed for word learning – by identifying connectors such as *und* ("and") or *dann* ("then"). The algorithm is attentive to actions and speech chunks which at least temporally overlap with the action sequence. Based on pauses in the speech signal and temporal co-occurrence between speech chunks and performed action, the algorithm finds the sequence of speech chunks that should be aligned with the current action. For a discussion of the alignment of speech and action (acoustic packaging) in infant-directed speech and beyond see [34]. For more details on the segmentation and alignment process applied to MMTD and AVC see [26].

A scene in itself is a list of expressions referring either to objects or actions or both. Thus, by defining different types of scenes, we are able to model different learning strategies. The result of the alignment process is a list of utterance-scene pairs, comprising the whole corpus.

2) A Model for Incremental Action Learning: Assuming that the correlations between words and objects or actions occurring in the same utterance-scene pair are sufficiently high, we designed a word-learning algorithm that sequentially processes each event and checks if word-referent pairs can be assigned to the lexicon. In this manner, the lexicon is incrementally filled with entries, but if a word is mapped to multiple referents or a referent is mapped to multiple words, certain links will also be "unlearned", and hence removed from the lexicon.

The key measure for assessing the significance of a given word-referent pair (a 'link') is pointwise mutual information (pmi). In order to be able to use this value for comparisons between concurring links, one has to employ the normalized pmi (the quotient of pmi and the self-information h with $h(w,r) = log_2\left(\frac{1}{p(w,r)}\right)$). We also tested the potential of the conditional probabilities $p(w|r)$ and $p(r|w)$ to support the decision whether a link should be added to or expelled from the lexicon. These measures, however, were not useful.

For each event, the full set of potential links is created by combining each word (unified list, lowercase) with each referent from the scene (object, action or both), and the statistical values ($npmi$) are updated for these links.

The algorithm proceeds as follows: For all current referents that are linked to a given word, if the difference between the link with the highest $npmi$ and links with lower $npmi$ values is greater than a given threshold ($par.bdif$, default: 0.05), the first link will get an extra count on its 'boost' value, the others on their 'decline' value.

In a second step, for all referents that have occurred so far, the links to words of the current utterance are re-evaluated. Similarly to the 'decline' value, if a link is outranked by another link and the difference between $npmi$ values is greater than the given threshold ($par.bdif$), an extra count on its 'exclude' value is given. While 'decline' compares links on the basis of concurring referents, the 'exclude' value stores information of co-occurrences between words. There is an option to inhibit the 'exclude' value if the two words predominantly co-occur as bigrams (skipping over functional words, such as articles or the preposition *von* "of") in a significant number of co-occurrences (e.g., *Mitte [von dem] Blatt* "center [of the] sheet"; parameter: $par.minbigr$, default: 0.25). If the conditions listed below are met, this particular link is included in the lexicon.

- The $npmi$ value is greater than a given absolute threshold ($par.npmilex$, default: 0.25), and
- the ratio between the sum of the 'decline', 'exclude' and 'boost' values is smaller than a given threshold ($par.boostlex$, default: 0.6).

On the other hand, links already in the lexicon need to be constantly re-evaluated. In each processing step, the links pertaining to the event (that have updated values), are examined, and if several conditions are met, the link is removed from the lexicon. This is important since especially at the beginning of the learning procedure, erroneous links have a higher chance to enter the lexicon. The conditions are the following:

- There are concurrent links, and
- the $npmi$ value is smaller than a given threshold ($par.npmilex$, 0.25), and
- the ratio between the sum of the 'decline', 'exclude' and 'boost' values is equal or greater than a given threshold ($par.boostlex$).

The model not only learns what the best candidate link is, but it is also able to model multiple connections between words and referents. Two words can refer to the same object (or action). This is typically the case with synonyms (e.g., 'bottle', 'flask'), but also hypernyms can be used instead of a given word (e.g., 'that thing'). Additionally, more than one word can be used to refer to an object, e.g. in MMTD, *Blatt Papier* – 'sheet of paper' consisting of a measure noun (*Blatt*, 'sheet') and a common noun (*Papier*, 'paper'). Hypernyms, by definition, refer to several objects (or actions).

IV. Object- and Action-Word Learning Experiments

In this section, we present a number of experiments on object- and action-word learning with varying input data, and discuss their results. We start with object-word learning experiments to demonstrate the robustness of Model 1 (Section III-A) to sensory noise such as noise in vision (e.g., errors in object recognition) and noise in speech (e.g., misrecognition of words), cf. Section IV-A1. Next, we show how the model can be embedded in a subset of the cognitive robotic DIARC architecture [35], and demonstrate how a robot, using the model, can learn new words through real-time interactions with a human teacher. (A high-level view of the DIARC configuration used can be seen in Fig. 5). This is followed by experiments with Model 2 on action word learning from crossmodal input data based on inputs from AVC and MMTD, where the input data to the learning model presented in Section III-B are varied as follows: Full form utterance and related actions and objects (represented as concept labels) are presented to the learning system. This is contrasted with input to the learning system where utterances are paired with either action labels or object labels alone but not both.

A. Model 1: Experiments in Object-Word Learning

1) Sensory noise evaluation: We examined the robustness of the model to noise in vision and speech recognition by systematically adding noise to the inputs from these two components and evaluating the mean F-score of the best lexicon found by the model, averaged over 10 runs. For the purpose of comparison, we implemented several incremental models of cross-situational word learning (association frequency (Equation 4), conditional probability $P(object|word)$, conditional probability $P(word|object)$) mainly to provide a baseline expectation for the results produced by an incremental model.

$$P(word, object) = \frac{Count(word, object)}{\sum_i \sum_j Count(word_i, object_j)} \quad (4)$$

The best lexica found by the non-Bayesian models consisted of a number of word-object pairs with the highest heuristic (e.g., $P(object|word)$) score. We varied the number of links included in the best lexicon found by these models and reported the lexicon with the best F-score. Fig. 6(a) demonstrates the behaviour of the model under noise. Fig. 6(b), Fig. 6(c), and Fig. 6(d) demonstrate the behaviour of non-Bayesian incremental models under noise. As can be seen, our model exhibits more robustness to noise compared to other models, as the least mean F-score value reported for our model (0.76) is much higher than that of other models (0.28, 0.55, 0.2).

2) Proof of concept embodied model: For this experiment we embedded our model in a subset of DIARC architecture [35] and replaced the simulated speech recognition and visual object detection components used in the sensory noise model evaluations with components capable of processing raw speech [36] and vision data. Additionally, we integrated a speech production component (allowing the robot to provide verbal feedback) and a robot manipulation component (allowing the robot to point to target objects in the environment).

The robot demo, available at https://vimeo.com/210659339, illustrates two types of interactions between the embodied model and the human teacher: (1) training and (2) testing. Testing interactions are marked with the word "point" at the start of the utterances made by the human interactor (e.g., "point to the X") and are used to examine the robot's knowledge of words (e.g., the word X). If the robot has at least one word-object mapping for the word X in its lexicon, it uniformly draws one of those mappings and points to the object in the drawn mapping while uttering "here it is". Otherwise, the robot responds "I don't know what that is". The robot uses other interactions (training interactions) to update its lexicon. The robot starts with an empty lexicon (no known word-object mappings). The human interactor then starts to teach new words through a series of word learning situations (utterance-scene pairs), using single word utterances (e.g., "knife") as well as complete sentences (e.g., "look at the knife"). The human interactor changes the scene by taking away and putting back objects on the table. See also Table I for sample inputs.

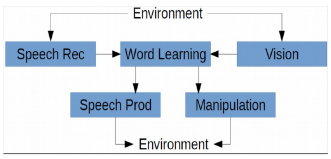

Fig. 5. High-level DIARC architecture for proof-of-concept demonstration. ('Rec' stands for Recognition, 'Prod' for Production.)

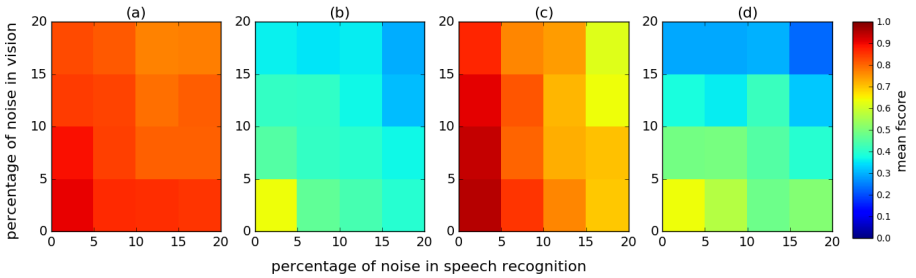

Fig. 6. The heatmap of mean F-score values (averaged over 10 runs) for the lexica found by (a) the incremental model [29], (b) the association frequency model, (c) the conditional probability $P(object|word)$ model, and (d) the conditional probability $P(word|object)$ model, under different noise conditions.

B. Model 2: Experiments in Object- and Action-Word Learning

In this section, results from crossmodal word learning experiments mapping verbs onto action concepts and nouns onto action related objects are presented. The respective gold standard lexica comprising the mappings between lexical form and concept (action or object) label are listed in Table III.[9] The mappings were extracted from the manual annotations of action-related speech and visual action in MMTD and AVC. Examples for lexical forms are *nimm, nehmen, nehm, nehme* (take). The related concept label for each of these forms is TAKE. The gold standard lexica comprise only word-concept mappings for those lexical forms that recur at least three times in the input to the learning system, as the model is not designed to learning from singletons. Therefore the gold standard lexicon for AVC comprises 3 verb-action mappings and 3 noun-object mappings, whereas in MMTD there are 7 verb-action and 8 noun-object mappings. It assumes that each inflected form is learned as a word in its own right, in order to avoid employing an external lemmatizer (a computer program that reduces inflected word forms to their base form).

In the objects only condition, the multimodal input to the learning system comprises pairs of utterance and a list of objects which are in the visual focus of the speaker, i.e., the object(s) which are manipulated by the speaker, while explaining the current activity, or which are landing sites or close to landing sites of the moved object. In other words, the input comprises an utterance and only objects which are in the visual field of attention. See for instance the input pair <*und ich nehme jetzt die Banane*, BANANA PLATE> where BANANA represents the object taken and PLATE the object/location from which the banana is taken. In the actions only condition, the input comprises pairs of utterance and action label, <*und ich nehme jetzt die Banane*, TAKE>. In the action+object condition, the input consists of the utterance, and the labels representing the action and the objects in visual focus, <*und ich nehme jetzt die Banane*, TAKE BANANA PLATE>. See Table II, for more examples of word learning situations comprising action, object(s) and utterance. For all conditions, ten learning runs have been performed randomly changing the succession of input sequences, and the results are then averaged per condition. The respective mean values

[9]For a description and discussion of "gold standard" in corpus annotation see [37].

AVC

actions	objects
nehme – TAKE	dose – PRINGLES
schiebe – PUSH	flasche – KETCHUP
stelle – PUT	schachtel – TEAHORIZONTAL

MMTD

actions	objects
lege – PUT	banane – BANANA
legen – PUT	birne – PEAR
gelegt – PUT	erdbeere – STRAWBERRY
nimm – TAKE	blatt – PAPER
nehmen – TAKE	blattes – PAPER
nehm – TAKE	papier – PAPER
nehme – TAKE	papiers – PAPER
	teller – PLATE

TABLE III
GOLD STANDARD LEXICA FROM AVC AND MMTD.

for precision, recall and F1-score are presented in Table IV. In order to assess the required number of input sequences to stabilize the lexicon, Fig. 7 and 8 show plots of mean F1-scores (individual and averaged runs) against the number of events seen by the learner.

	AVC (78 input scenes)			MMTD (202 input scenes)		
condition	precision	recall	F1-score	precision	recall	F1-score
action	90.0% (±12.9%)	100.0% (±0.0%)	0.943 (±0.074)	39.5% (±6.0%)	27.2% (±4.5%)	0.321 (±0.050)
object	100.0% (±0.0%)	100.0% (±0.0%)	1.000 (±0.000)	63.7% (±10.6%)	46.2% (±6.0%)	0.533 (±0.066)
action + object	88.6% (±6.0%)	100.0% (±0.0%)	0.938 (±0.032)	59.4% (±2.1%)	38.0% (±3.2%)	0.463 (±0.028)

TABLE IV
RESULTS FROM LEARNING WORD-ACTION AND WORD-OBJECT MAPPINGS FROM CROSSMODAL INPUT DATA DERIVED FROM AVC AND MMTD. LISTED ARE THE MEAN VALUES AND STANDARD DEVIATIONS (IN PARENTHESES) FROM 10 LEARNING-TESTING RUNS PER MODEL AND CONDITION. THE F1-SCORE IS THE HARMONIC MEAN OF PRECISION AND RECALL, $F1 = 2 * \frac{precision*recall}{precision+recall}$.

Because of the random selection of input sequences per run the results differ slightly even after averaging over 10 runs. We ran Model 2 over several packs of 10 runs resulting in persistently better lexical learning from AVC than from MMTD. This result can be partially explained by the differences in lexical variation in the input to learning (in other words, in the training corpora). The utterances in the AVC corpus comprise only one lexical form for each action and object referent, whereas in MMTD there are three lexical forms referring to PUT-actions, four for TAKE-actions, two different words, *Blatt, Papier* (sheet, paper) with two different lexical forms each referring to the sheet of paper being part of the task setup (*blatt, blattes; papier, papiers*). See Table III for a summary of all lexical forms referring to the actions and objects in the AVC corpus and in MMTD. If there are more potential links between words and a particular concept, then the scores for each mapping decrease and leave room for erroneous mappings. See for instance the relation between the personal pronoun *sie* "her/it", and the concept PUT, which is erroneously learned from MMTD, but not from AVC, even though in both data sets PUT actions frequently co-occur with *sie* (64 times in a total of 106 PUT actions in MMTD and 24 times in a total of 24 PUT actions in AVC). This situation arises due to the combination of TAKE and PUT actions in both corpora, resulting in utterances such as *ich nehme die Erdbeere und lege sie neben die Banane* ('I take the strawberry and put it next to the banana', MMTD) or *ich nehme die Schachtel und stelle sie links neben die Dose* ('I take the box and put it to the left of the can', AVC).

While in AVC a PUT action is always accompanied by one and the same word/verb *lege* (put), PUT in MMTD is accompanied by a variety of verbs including *lege* (62), empty verb (8), *platziere* (7), *packe* (6), *kommt* (4), *platziert* (3), *zu platzieren* (3), *tu* (2), *absetzen* (1), *tun* (1), *zu liegen kommt* (1), *sich ergibt* (1), *sein* (1), *liegt* (1), *verschiebe* (1), *vertauscht* (1), *ordnen* (1). In addition, out of 64 occurrences of *sie* in total in MMTD, 62 co-occur with a PUT action, whereas out of 33 occurrences of *sie* in total in AVC, 23 co-occur with a PUT action, 9 with a PUSH action, 1 is an alignment error.

This results in significantly different $npmi$ values for the primary candidate *lege* and the personal pronoun *sie*. In MMTD, these are roughly on par, which prohibits the algorithm to decide which of the two is the correct one. In AVC the $npmi$ value for *lege* is almost twice as high as the one for *sie*. In that case the latter is assigned an 'exclude' penalty for its link to the PUT action, which prohibits (correctly) this link from being added to the lexicon.

Another problem for learning object-word mappings in MMTD was the more complex linguistic realizations for locations. For instance: The teachers very often described that they would put a piece of fruit in the center of the sheet of paper (*in die Mitte des Blattes*, or *in die Mitte vom Papier*, etc.). Therefore, *Mitte* "center" acquires the highest $npmi$ values with PAPER. *Blatt* "sheet" receives a lower $npmi$ value and also enters the lexicon, but *Papier* "paper" does not. This problem was circumvented in AVC where the users were asked to move objects by taking one and putting it next to another one. Another possibility to avoid the effect is to ask users to use utterances such as "I take X and put it there". We used this kind of utterances in learning experiments that ran under live conditions on a Pepper robot [38].

V. SUMMARY AND CONCLUSION

In this paper, we presented and assessed two related but different models of incremental word learning. Model 1 realizes an incremental version of the Bayesian approach for cross-situational word-referent learning for words with concrete object references introduced in [21]. The model focuses on the joint acquisition of referential intention and word meaning from only a few varying situations. Model 2 implements an information-theoretic approach to learning from modality rich input data based on normalized pointwise mutual information. (For a discussion of variants of pointwise mutual information see [39].) The proposed model focuses on learning mappings of actions onto verbs and of action-related objects onto nouns. To do so, highly redundant input data were used, whereby redundancy came from modality richness and repetitiveness of the data. The data input to the learning model was derived from two multimodal corpora, the Action Verb Corpus (AVC) and MMTD. All learning scenarios were based on situated task-oriented communication in teacher-learner settings. Using this kind of data for learning was inspired by research on early infant learning, where parent-child interactions produce modality rich and redundant input, for instance, when parents name objects their infants currently interact with; see for instance [34], [2], [1], [3], [4], [5], [6]. While our data stem from adult interactions, we are aware that child-directed speech has many more facets to it than producing massively redundant input to the learning system, [40], [41].

We have illustrated the difference in capacities of the two proposed models in a number of experiments. We have demonstrated the robustness of Model 1 with respect to sensory noise such as noise in vision (e.g., errors in object recognition) and noise in speech (e.g., misrecognition of words), based on simulated data, and demonstrated how a robot (PR2), using the model (embedded in a subset of the cognitive robotic DIARC architecture), can learn new words through real-time interactions with a human teacher. We have tested Model 2 with crossmodal input data from AVC and MMTD, whereby the input data to the learning model varied as follows: (i) full form utterance and a list of related actions and objects in visual attention; (ii) utterance paired with action labels only; (iii) utterance paired with object labels only. The results showed that for Model 2 lexical learning from AVC was easier than learning from MMTD, i.e., the values for precision, recall and F1-scores were persistently higher for AVC than for MMTD. We attribute this to differences in complexity of the utterances accompanying the visual scenes, whereby MMTD shows more lexical variation, – several word (forms) refer to a single object or action, it has on average longer sentences than AVC (8.6 versus 5.9) and also a higher structural complexity – for instance the linguistic realisation of location expressions such as put something in the middle of the sheet of paper. Moreover, it has elliptical constructions where action verbs are missing. All this impedes co-occurrence-based word-referent mapping.

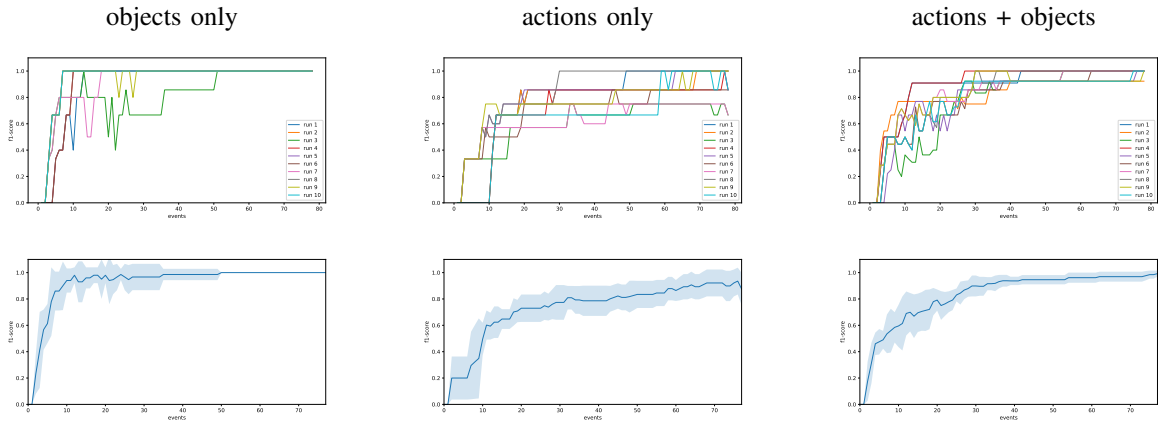

Fig. 7. Model 2 learning from AVC: Incremental plots F1 (upper row) and mean values (lower row) for objects and actions only, and for actions plus objects.

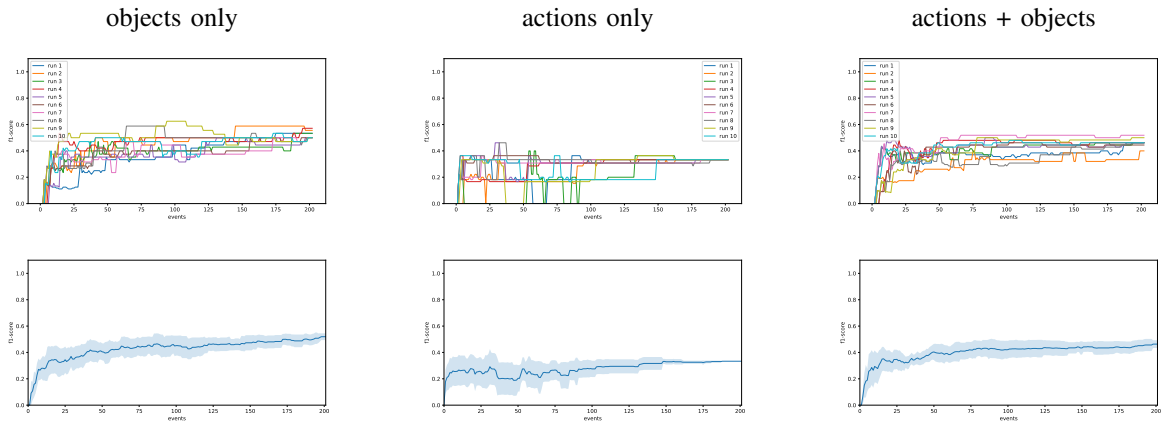

Fig. 8. Model 2 learning from MMTD: Incremental plots F1 (upper row) and mean values (lower row) for objects and actions only, and for actions plus objects.

In future research, to overcome the sensitivity to full forms versus lemmas, mechanisms for exploiting the similarity between morphologically related word forms need to be developed and integrated in the learning algorithms. Currently, we are working on expanding the word learning frameworks to capture syntactic information. First, the models need to differentiate nouns from verbs in parallel with the distinction between action and object referents obtained from the multimodal input. On the basis of a verb-noun distinction, word order can be taken into account. For example, Sadeghi and Scheutz [33], [30] have already demonstrated that Model 1 can be expanded to allow for learning word referents as well as language word order. The knowledge of word order allows the learner to parse the input sentences and to infer a mapping from concepts to grammatical functions, in order to understand who has done what to whom. Furthermore functional words such as articles, pronouns, auxiliary verbs etc. need to be incorporated into the models. Under the assumption that languages have closed class lexica of functional words (or morphemes in the case of agglutinative languages), these items will no longer behave as distractors in the learning procedure, but enhance the model by indicating the category of a syntactic phrase (e.g., articles indicate noun phrases), serving as place-holders for referential expressions (pronouns) or help in identifying slots within the syntactic structure (auxiliaries). From this we not only expect a substantial improvement for the learning algorithms, but also an essential advance in modelling learning procedures on the basis of natural language input.

ACKNOWLEDGMENT

This research is supported by the Vienna Science and Technology Fund (WWTF), project RALLI – Robotic Action-Language Learning through Interaction (ICT15-045) and the CHIST-ERA project ATLANTIS (2287-N35). We wish to thank the anonymous reviewers for their time and conscientious efforts to improve the readability and cross-disciplinary understandability of the article.

REFERENCES

[1] L. Gogate, "Development of early multisensory perception and communication: From environmental and behavioral to neural signatures," *Developmental Neuropsychology*, vol. 40 (5-8), pp. 269–272, 2016.
[2] S. H. Suanda, L. B. Smith, and C. Yu, "The multisensory nature of verbal discourse in parent–toddler interactions," *Developmental Neuropsychology*, vol. 41, no. 5-8, pp. 324–341, 2016.
[3] M. C. Frank, J. B. Tenenbaum, and A. Fernald, "Social and discourse contributions to the determination of reference in cross-situational word learning," *Language Learning and Development*, vol. 9, no. 1, pp. 1–24, 2013.

[4] L. J. Gogate, L. E. Bahrick, and J. D. Watson, "A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures," *Child development*, vol. 71, no. 4, pp. 878–894, 2000.
[5] M. Harris, D. Jones, and J. Grant, "The nonverbal context of mothers' speech to infants," *First Language*, vol. 4, no. 10, pp. 21–30, 1983.
[6] D. Messer, "The redundancy between adult speech and nonverbal interaction: A contribution to acquisition," *The transition from prelinguistic to linguistic communication*, pp. 147–165, 1983.
[7] I. Nomikou, M. Koke, and K. J. Rohlfing, "Verbs in mothers' input to six-month-olds: Synchrony between presentation, meaning, and actions is related to later verb acquisition," *Brain sciences*, vol. 7, no. 5, p. 52, 2017.
[8] L. Gogate and M. Maganti, "The origins of verb learning: Preverbal and postverbal infants' learning of word–action relations," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 12, pp. 3538–3550, 2017.
[9] D. Gentner, "Why nouns are learned before verbs: Linguistic relativity versus natural partitioning," *Center for the Study of Reading Technical Report; no. 257*, 1982.
[10] ——, "Why verbs are hard to learn," *Action meets word: How children learn verbs*, pp. 544–564, 2006.
[11] L. Gogate and G. Hollich, "Early verb-action and noun-object mapping across sensory modalities: A neuro-developmental view," *Developmental neuropsychology*, vol. 41, no. 5-8, pp. 293–307, 2016.
[12] A. L. Woodward, "Infants selectively encode the goal object of an actor's reach," *Cognition*, vol. 69, no. 1, pp. 1–34, 1998.
[13] J. K. Hamlin, K. Wynn, and P. Bloom, "Social evaluation by preverbal infants," *Nature*, vol. 450, no. 7169, p. 557, 2007.
[14] V. Reddy, K. Liebal, K. Hicks, S. Jonnalagadda, and B. Chintalapuri, "The emergent practice of infant compliance: An exploration in two cultures." *Developmental Psychology*, vol. 49, no. 9, p. 1754, 2013.
[15] J. B. Childers and M. Tomasello, "Two-year-olds learn novel nouns, verbs, and conventional actions from massed or distributed exposures." *Developmental psychology*, vol. 38, no. 6, p. 967, 2002.
[16] P. Brown, "Children's first verbs in tzeltal: Evidence for an early verb category," *Linguistics*, vol. 36, no. 4, pp. 713–754, 1998.
[17] S. Choi, "Verbs in early lexical and syntactic development in korean," *Linguistics*, vol. 36, no. 4, pp. 755–780, 1998.
[18] J. Chen, T. Tardif, R. Pulverman, M. Casasola, L. Zhu, X. Zheng, and X. Meng, "English-and mandarin-learning infants' discrimination of actions and objects in dynamic events." *Developmental psychology*, vol. 51, no. 10, p. 1501, 2015.
[19] M. Imai, E. Haryu, H. Okada, L. Lianjing, and J. Shigematsu, "17 revisiting the noun-verb debate: A cross-linguistic comparison of novel noun and verb learning in english-, japanese-, and chinese-speaking children," *Action meets word: How children learn verbs*, p. 450, 2006.
[20] C. Yu and D. H. Ballard, "A unified model of early word learning: Integrating statistical and social cues," *Neurocomputing*, vol. 70, no. 13, pp. 2149–2165, 2007.
[21] M. C. Frank, N. D. Goodman, and J. B. Tenenbaum, "Using speakers' referential intentions to model early cross-situational word learning," *Psychological science*, vol. 20, no. 5, pp. 578–585, 2009.
[22] G. Kachergis, C. Yu, and R. M. Shiffrin, "An associative model of adaptive inference for learning word–referent mappings," *Psychonomic bulletin & review*, vol. 19, no. 2, pp. 317–324, 2012.
[23] D. Roy, "Grounded spoken language acquisition: Experiments in word learning," *IEEE Transactions on Multimedia*, vol. 5, no. 2, pp. 197–209, 2003.
[24] A. Alishahi and A. Fazly, "Integrating syntactic knowledge into a model of cross-situational word learning," in *Proceedings of the Cognitive Science Society*, vol. 32, no. 32, 2010.
[25] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.
[26] S. Gross, M. Hirschmanner, B. Krenn, F. Neubarth, and M. Zillich, "The Action Verb Corpus," in *Proceedings of the 11th International Conference on Language Resources and Evaluation*. Paris, France: ELRA, 2018, pp. 2147–2151.
[27] S. Gross and B. Krenn, "The OFAI Multimodal Task Description Corpus," in *Proceedings of the 10th International Conference on Language Resources and Evaluation*. Paris, France: ELRA, 2016, pp. 1408–1414.
[28] L. Gogate, M. Maganti, and L. E. Bahrick, "Cross-cultural evidence for multimodal motherese: Asian indian mothers' adaptive use of synchronous words and gestures," *Journal of experimental child psychology*, vol. 129, pp. 110–126, 2015.
[29] S. Sadeghi, M. Scheutz, and E. Krause, "An embodied incremental bayesian model of cross-situational word learning." in *Proceedings of the 2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (icdl-epirob)*, 2017.
[30] S. Sadeghi and M. Scheutz, "Early syntactic bootstrapping in an incremental memory-limited word learner," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.
[31] ——, "Sensitivity to input order: Evaluation of an incremental and memory-limited bayesian cross-situational word learning model," in *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, 2018.
[32] S. Sadeghi, B. Oosterveld, E. Krause, and M. Scheutz, "Acquisition of word-object associations from human-robot and human-human dialogues," in *Proceedings of the 2019 IEEE International Conference on Robotics and Automation (ICRA 2019)*, 2019.
[33] S. Sadeghi and M. Scheutz, "Joint acquisition of word order and word referent in a memory-limited and incremental learner," in *Proceedings of the 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 2017.
[34] N. R. George, F. Bulgarelli, M. Roe, and D. J. Weiss, "Stacking the evidence: Parents use of acoustic packaging with preschoolers," *Cognition*, vol. 191, p. 103956, 2019.
[35] M. Scheutz, G. Briggs, R. Cantrell, E. Krause, T. Williams, and R. Veale, "Novel mechanisms for natural human-robot interactions in the diarc architecture," in *Proceedings of AAAI Workshop on Intelligent Robotic Systems*, 2013, p. 66.
[36] P. Lamere, P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, M. Warmuth, and P. Wolf, "The cmu sphinx-4 speech recognition system," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*, vol. 1. Citeseer, 2003, pp. 2–5.
[37] L. Wissler, M. Almashraee, D. M. Díaz, and A. Paschke, "The gold standard in corpus annotation." in *IEEE GSC*, 2014.
[38] M. Hirschmanner, S. Gross, B. Krenn, F. Neubarth, M. Trapp, and M. Vincze, "Grounded word learning on a pepper robot," in *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 2018, pp. 351–352.
[39] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," *Proceedings of GSCL*, pp. 31–40, 2009.
[40] P. F. Dominey and C. Dodane, "Indeterminacy in language acquisition: the role of child directed speech and joint attention," *Journal of Neurolinguistics*, vol. 17, no. 2-3, pp. 121–145, 2004.
[41] R. M. Golinkoff, D. D. Can, M. Soderstrom, and K. Hirsh-Pasek, "(baby) talk to me: The social context of infant-directed speech and its effects on early language acquisition," *Current Directions in Psychological Science*, vol. 24, no. 5, pp. 339–344, 2015.