Robots That Say "No" Affective Symbol Grounding and the Case of Intent Interpretations

Frank Förster, Joe Saunders, and Chrystopher L. Nehaniv

***Abstract*—Modern theories on early child language acquisition tend to focus on referential words, mostly nouns, labeling concrete objects, or physical properties. In this experimental proof-of-concept study, we show how nonreferential negation words, typically belonging to a child's first ten words, may be acquired. A child-like humanoid robot is deployed in speech-wise unconstrained interaction with naïve human participants. In agreement with psycholinguistic observations, we corroborate the hypothesis that affect plays a pivotal role in the socially distributed acquisition process where the adept conversation partner provides linguistic interpretations of the affective displays of the less adept speaker. Negation words are prosodically salient within intent interpretations that are triggered by the learner's display of affect. From there they can be picked up and used by the budding language learner which may involve the grounding of these words in the very affective states that triggered them in the first place. The pragmatic analysis of the robot's linguistic performance indicates that the correct timing of negative utterances is essential for the listener to infer the meaning of otherwise ambiguous negative utterances. In order to assess the robot's performance thoroughly comparative data from psycholinguistic studies of parent–child dyads is needed highlighting the need for further interdisciplinary work.**

***Index Terms*—Developmental robotics, human–robot interaction, language acquisition, negation, pragmatics, psycholinguistics, social robotics.**

I. Introduction

Early productive vocabularies of infants are often said to be dominated by nouns that refer to perceptible objects such as toys, foods, or animals. Thus many of modern psycholinguistic studies on early word acquisition focus on these types of words (see [1]–[5]).

Usage-based theories of language development [6] emphasize the importance of joint reference between caretaker and child to external objects and processes, and detail the mechanisms behind these triadic joint-attentional frames (caretaker-child-object), which are cognitively more complex than direct dyadic interaction (see also [7]).

A similarly strong focus on words referring to concrete physical objects, processes, actions, and object properties such as color or size is prevalent in language-oriented research in developmental robotics [8]–[25] and agent-based approaches in evolutionary linguistics [26]–[29]. There, symbol grounding [30], the linking of symbols with data or concepts derived from the robot's own embodiment, is employed for robots to "make sense" of these linguistic entities. Words and simple grammatical constructions are linked with the robot's sensorimotor stream using various techniques that range from neural networks [23] to logic-based approaches [22].

However, the very earliest productive vocabularies of infants at the onset of speech, the first ten words, are typically dominated by nonreferential, social words such as "hi," "bye," "yes," and "no" [31]–[34]. Negation words are already used at this stage for various communicative functions such as rejection or nonexistence [35]–[37]. These words cannot be linked to sensorimotor data in the same way as referential words as they do not refer to perceptible entities outside the speaker. In this paper, we describe and evaluate a dyadic word acquisition mechanism involving so called *intent interpretations* (see [38], [39], and below) which may explain how children come to learn the meaning of the earliest negation words, in English chiefly the word *no*. This mechanism may be best seen as complementing rather than as a replacement for the dominant mechanism in the developmental literature centered around triadic joint-attentional frames [6]. We hypothesize that the proposed dyadic behavioral mechanism may also explain children's initial acquisition of the meaning and use of other nonreferential terms such as emotion and other mental state words such as "sad," "happy," or "like."

A. Negation, Affect, and Intent Interpretations

Research into the acquisition of early linguistic negation emphasizes the importance of affect [38]. Yet neither the cognitive requirements nor the learning mechanisms driving the acquisition of negation are understood in sufficient detail for roboticists to enable machines to engage in this aspect of human speech.

In this paper, we describe one of the two related experiments both of which operationalize the overarching hypothesis that affect plays a pivotal role in the acquisition of early negation words. This serves as the dual-purpose of both improving our understanding of potential acquisition mechanisms in the

Manuscript received November 25, 2016; revised February 17, 2017 and May 22, 2017; accepted July 1, 2017. Date of publication November 27, 2017; date of current version September 7, 2018. This work was supported by the EU Integrated Project "Integration and Transfer of Action and Language in Robots" through the European Commission under Contract FP-7-214668. *(Corresponding author: Frank Foerster.)*

The authors are with the Adaptive Systems Research Group, University of Hertfordshire, Hatfield AL10 9AB, U.K. (e-mail: f.foerster@herts.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCDS.2017.2752366

human language development as well as laying the groundwork for a new generation of the symbol grounding systems that go beyond the grounding of words and phrases with external physical referents. We reimplemented a language acquisition system very similar to another such system which had been successfully employed for the acquisition of nouns, adjectives, and two-word utterances [19]–[21]. This type of acquisition system relies on a tight temporal coupling between social mechanisms such as the establishment of joint attention on one hand, and some form of embodied machine learning on the other. In other words, these systems depend on statistical regularities in the human conversational behavior. One such regularity which was exploited in the aforementioned studies on noun acquisition is the circumstance that prosodically salient object names and properties are generally produced within a small time window around the moment when joint attention on the respective referent between the two speakers is established. The study presented within the present publication relies on a similar temporal regularity between affective displays on part of the robot and the production of *intent interpretations* on part of the participant qua teacher. Intent interpretations seem to be particularly prevalent in the context of conversationally asymmetric dyads: a conversationally fluent or "strong" interaction partner and a conversationally "inept" or "weak" speaker or learner. In the case of human–human dyads this would typically be parent and infant, in our case the dyad consists of a teacher-participant and a child-like humanoid. In the more recent developmental literature, the words and terms which we found to be prosodically emphasized within these interpretations such as "want," like, or "feel like," are often referred to as mental state words. They are often discussed in conjunction with developmental accounts of theory of mind [40]–[45]. However, negation words are typically not counted amongst mental state words.

Intent interpretations are rarely mentioned in the contemporary accounts of early language acquisition but have been hypothesized some decades ago to play a vital role for infants to learn how to express their intent [38], [39]. Intent interpretations are characterized by the conversationally stronger partner providing the infant with words for their emotions and intentions by the way of interpreting their bodily displays linguistically. Building upon this more general idea we hypothesized that in the case of negative emotional or volitional displays, indicating states such as sadness or unwillingness, these interpretations would frequently contain negation words as in "no, you do not like that" or in a simple *no*. In order for our language acquisition system described below to pick up on these negation words we have to make one additional assumption: negation words are either marked prosodically as salient within the respective utterance or the utterance is a one-word utterance, i.e., it solely consists of a negation word. The so called *rejection* experiment described below was thus set up as an operationalization of the hypothesis that negative intent interpretations (NIIs), performed by the conversationally strong partner, constitute a major source of negation words for the language learner, and these can be grounded using affective in addition to sensorimotor associations.

B. Overview of the Experiment

The *rejection experiment* was a single-blind study, the main purpose of which was to test the impact of the robot's motivated behavior upon participants' linguistic productions in general and the elicitation of NIIs in particular. Moreover, we were testing the suitability of these intent interpretations for the purpose of learning symbol grounding via our language acquisition system. The name of the experiment derives from the observation that NIIs are oftentimes produced as a response to rejective behavior on part of the conversationally weak interactor.

Naïve participants were asked to teach the humanoid iCub [46] words for objects that were located on a table between them and the humanoid (Fig. 1). They were further asked to talk to the robot as if it were a prelinguistic child and told that it would express preferences for certain objects. The participants' instructions were largely identical to an earlier experiment on noun acquisition executed in the same premises and with the same robot which we will henceforth refer to as Saunders *et al.*'s [21] *experiment*. This previous work did not involve affective displays on part of the robot and focused on symbol grounding via sensorimotor association. Data from this experiment will be used for comparative purposes in the later part of this paper.

Differing from Saunders *et al.*'s [21] experiment, the robot in this paper is equipped with a motivation system which triggers emotional facial expressions and matching body behaviors. Simultaneously, it feeds its affective state into a symbol grounding system. During each experimental session all objects are assigned valences that trigger the robot's motivational state toward the object: positive, neutral, or negative.

All of the robot's behaviors are object-directed and underlying it all is a gazing behavior where the robot's gaze switches between one or more objects and the participant's face. The particular gaze durations vary between behaviors (see Section II). When presented with preferred objects the robot smiles and holds out its hand toward them, palm facing upward, such that participants can put objects into its hand (*reaching* behavior).

Inspired by the infants' reactions to being fed with disliked food, the robot, when presented with unpreferred objects, looks at these for a set short-time period, starts to frown, and executes a combined head-eye avoidance movement if the object is located within a center area of its visual field (*rejection* behavior). If the object, after execution of the avoidance movement, still resides within the center area of the robot's visual field another avoidance move is executed, which may result in overall head movements that resemble a head shake. If a participant presents the robot with an object with neutral valence, it displays a neutral facial expression and alternates its gaze between the participant's face and object without approaching nor avoiding the latter (*watching* behavior). At the very start of the experiment, before participants select the first object, and between presentations of particular objects, the robot has a neutral facial expression and executes a *looking around* behavior: it switches its focus between all objects located on the table and the participants' face. Changes between behaviors are triggered by changes in

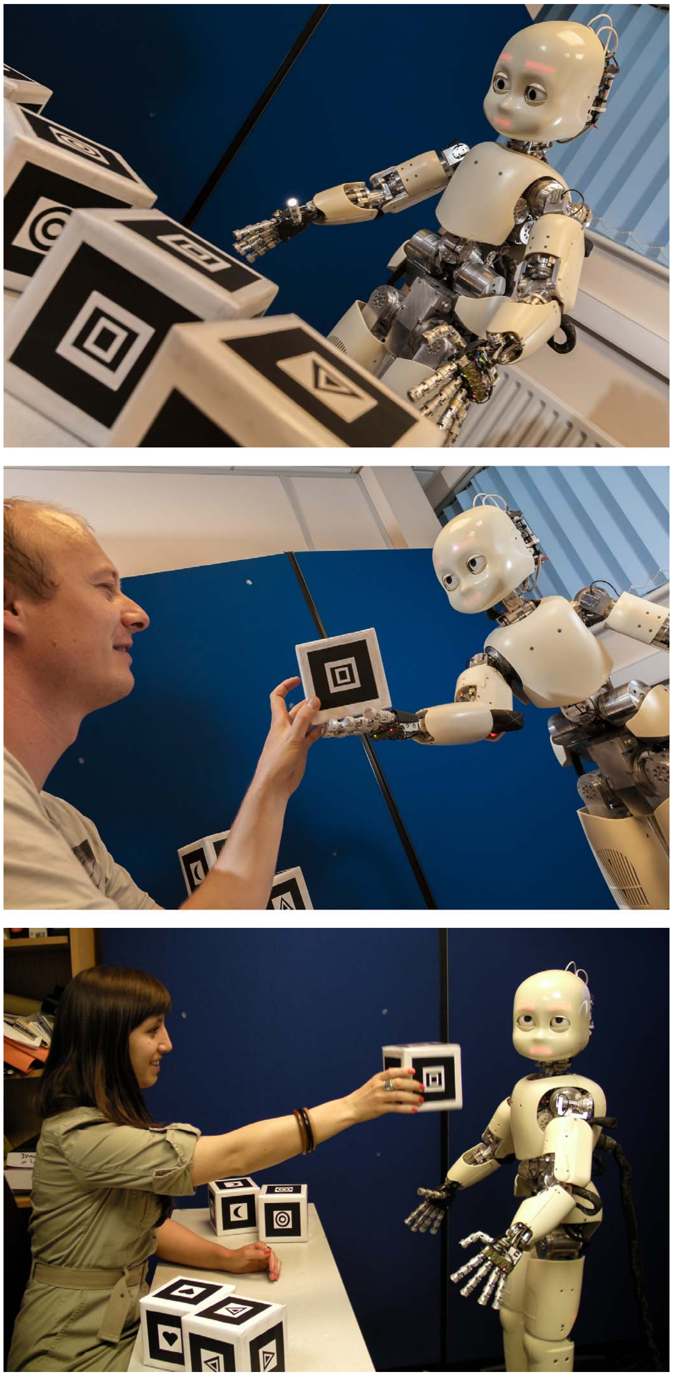


Fig. 1. Experimental setup and depiction of the behaviors executed by the iCub robot during the interaction sessions. Participants and robot face each other with the objects that are to be taught being located between them on a table. Top: *looking around* behavior, executed if no object is selected by the participant. Middle: *reaching* behavior, executed if the robot perceives the selection of an object with positive valence by the participant. Bottom: *avoidance* behavior, executed if an object with negative valence is being perceived as selected.

the robot's motivation system which, in turn, are triggered by the valences of objects picked up or put down by participants in their attempts to teach the robot words for these objects.

The experiment was split into five sessions of five minutes in length, as the word learning required some offline processing which was performed in between sessions. Both participants' speech as well as robot's sensorimotor-motivational (smm) data were recorded and timestamped during each session. Upon termination of each session this data was processed in a semiautomatic manner: first, the prosodically most salient word is extracted from each utterance through acoustic analysis. Subsequently the *smm* data that was recorded during the production of the respective utterance is associated with each extracted word, and the salient words are henceforth grounded in the robot's own percepts and motivation. The entirety of grounded words is finally added to the robot's *embodied lexicon*, where a separate lexicon is constructed for each participant. Each embodied lexicon is initially empty and gets populated exclusively with grounded words extracted from the respective participant's speech over the course of the interactive sessions. No designer knowledge in terms of a set of preselected words is incorporated, so the robot can only acquire and express words which the participant has used during previous sessions. Thus, the embodied lexicon takes a different developmental trajectory with each teacher.

At the start of a follow-up session, this lexicon is loaded into a memory-based learning system [47] and certain trigger behaviors such as *grasping* for, *rejecting*, or simply *watching* a presented object lead to the robot querying its embodied lexicon: it matches its current *smm* state against the lexicon and retrieves the word which best matches this state. This process is performed as often as complete *smm* vectors are received from the perceptual and motivational systems, which operate at about 30 Hz. A thresholding mechanism maintains a score for each retrieved word: the score of the currently retrieved word is increased whereas all other word scores are decreased. As soon as the score of a word reaches a predefined threshold, this word is sent to the speech synthesizer: the robot speaks. Upon speaking, all scores are reset to 0 and the retrieval resumes on a reduced lexicon where the just-synthesized word is temporarily removed from the lexicon until either another word reaches the speaking threshold or changes in the *smm* state occur, thus potentially allowing multiword utterances (see also Section II).

II. Materials and Methods

A. Study Design

The experiment was split into five sessions per participant. Each participant completed five sessions of approximately 5 min each with at least one day in between sessions. This time gap was required in order to complete the post-processing of the speech recordings. All participants were wearing headsets during the interaction and their speech was recorded. We further videotaped each session. Apart from the participant, one to two more people were in the room: an operator, who started up the robot and monitored it during the session, and a helper who placed the boxes back on the table as Deechee, the robot, was prone to drop them. In a few sessions, the helper was absent and the operator took on both roles. Participants were seated opposite the robot with a table separating the two. The five objects, whose names were to be taught, were cardboard boxes of approximately 10-cm side length, with black-and-white shapes printed on each side of each box (see Fig. 1). For

TABLE I
OBJECT-BOUND MOTIVATION VALUES PER SESSION FOR REJECTION EXPERIMENT

	session 1	session 2	session 3	session 4	session 5
triangle	1	-1	1	-1	0
moon	0	1	-1	1	-1
square	-1	0	1	-1	1
heart	1	-1	0	1	-1
circle	-1	1	-1	0	1

TABLE II
CONSTANTS FOR HUMAN–ROBOT INTERACTION, ALL VALUES IN SECONDS

variable	*value*	*description*
face_time	0.8	duration of iCub looking at face when *pickup* detected and motivation ≥ 0
object_time	3	duration of iCub looking at object when *pickup* detected and motivation ≥ 0
dwell_time_face	1.2	duration of iCub looking at face when no *pickup* detected
dwell_time_object	2	duration of iCub looking at *obj* when no *pickup* detected
maxIdleTime	3	perceptual timeout for high level percepts: if no objects or faces are perceived for *maxIdleTime*, iCub looks back at the table

any particular box, the shapes were identical on every side. The shapes were a star, a heart, a square, a crescent moon, and a triangle. After having read and signed the instructions and the consent form, participants were seated in the room and asked to count down in the following way: "three, two, one, start." They were told in advance, that "start" would constitute the start time of the session and that the operator would signal when the five minutes were over. Upon "start," the operator pressed a button, which subsequently produced a time stamp within the *body memory* of the robot.

As mentioned above the setup was devised to test the hypothesis whether intent interpretations might be sufficient for the acquisition of early types of negation such as rejective utterances or *motivation-dependent denial*. In order to establish the situational context in which NIIs would most likely occur we ensured that participants could not avoid presenting disliked objects to the robot. We made it impossible for participants to initially know which objects the robot would like by permuting the object-valence mapping for each session (see Table I). The various time constants controlling the robot's gaze behavior were chosen as listed in Table II.

B. Recruiting and Distribution of Participants

We recruited ten participants, the majority British, with one South-Asian, one U.S.-American, and one South-African speaker, the latter having lived for several decades in the U.K. The participants were balanced by gender. Most of the participants were recruited from the campus and were either students or employees, only three of them had no affiliation with the university. Of the ten participants, four had children, and two other participants came from big families and were involved in raising children. Yet another participant had worked as a teacher for small children, and another one stated having had reasonable exposure to children via friends. The two remaining participants answered that they had no experience with children.[1]

Participants were remunerated with £20 each after completion of all five sessions. This paper was approved by the University of Hertfordshire Ethics Committee for Studies Involving Human Participants for the ITALK project under protocol number 0809/99 and with an extension granted under protocol number 1112/42.

C. Instructions to Participants

The instructions given were very similar to the ones used in Saunders *et al.*'s [21] experiment, namely that their task was to teach Deechee about the available objects. Moreover, participants were told to imagine Deechee to be a small child of approximately two years and, further, that Deechee had preferences for particular objects, that it may like, dislike, or would feel neutral about. The first of these instructions was given in order to increase the likelihood of participants assuming a simplified speech register akin to child-directed speech (CDS [48]–[50]). The second instruction about Deechee having preferences, was given in order to prepare participants for Deechee's emotional displays. It is unclear whether this instruction was necessary, and whether this prime had an impact upon participants' way of speaking. The fact that the experiment was about the acquisition of linguistic negation was not mentioned to the participants.

D. Behavioral Architecture

The behavioral architecture depicted in Fig. 2 generating both humanoid's bodily and linguistic behavior consists of the following modules or components. Each module is executed as a separate process and the modules are only loosely coupled in that they communicate with each other asynchronously via the exchange of so called bottles—a simple messaging service provided by the YARP robot middleware [51]. Due to space limitations each module is only sketched, for more details see [52].

The *perception system* gathers and processes percepts of all modalities. The visual processing is done via a modified version of the system developed by Rüsch *et al.* [53]. For the given experiments, this system was limited to face and object detection. Furthermore, a custom detector was developed which signals when an object has been picked up from or put down on the table.

The *motivation system* generates the robot's motivational state. This is done based on either predefined or randomly chosen object valences and dependent upon the presence of external pressure upon the robot's arm (diminished agency). In the rejection experiment discussed in the present publication, no external pressure was employed and the object valences

[1] We did not ask participants explicitly whether they have had prior experience with robots or programming, but we expect programming knowledge with four of them as they had been either Ph.D. students or researchers themselves at the time of the experiment. One other participant worked as an actuary and had a strong mathematical background. Another two participants had interacted with robots prior to the experiment within other experiments of the research group.

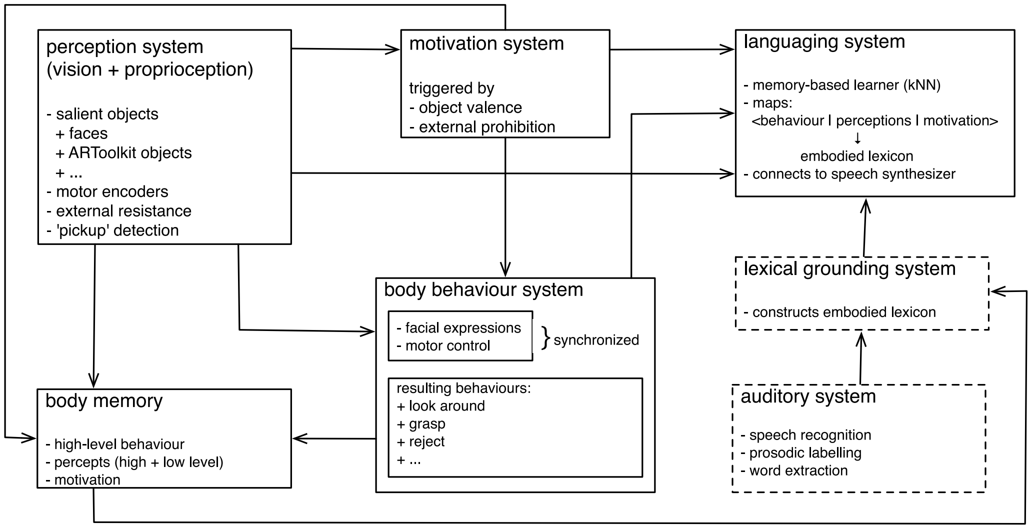


Fig. 2. Functional overview of robotic architecture for language acquisition. Solid lines indicate components that are active during experimental sessions ("online") and dotted lines indicate components that work offline.

were fixed in each session to the values shown in Table I. The motivational state is modeled as scalar value between -1 and 1, with -1 being negative, $+1$ being positive, and a small band around 0 signifying neutral (see valence toward perceived objects [54, Ch. 6]). We chose a model as simple as possible for two reasons. First, we were reluctant to introduce more complexity into the overall architecture if not absolutely necessary. Second, it was unclear how a dimensional emotion model [55] could be mapped in a principled manner to the robot's limited facial expressions and bodily behavior.

The *body behavior system* generates the humanoid's bodily behavior including its facial expressions (see Algorithm 1). It receives input from both perception and motivation system. The central design tenet was to make the robot act as believably as possible, as opposed to making its movements as accurate as possible. A consequence of this tenet is that the system always produces some kind of behavior. Toddlers do not freeze and therefore neither should the robot. Five different behaviors have been implemented: 1) *idle*; 2) *looking around*; 3) *reaching for object*; 4) *rejecting*; and 5) *watching*. Each behavior has a unique behavior id which is broadcasted to the other subsystems whenever a change of behavior occurs.

A change away from the robot's "base behavior" (*looking around*) is triggered first by the participant picking up an object from the table, and, second, by the robot's object-bound valence toward the respective object. The valence modulates the robot's motivational state which in turn modulates its behavior: the robot reaches for objects it likes, it displays rejective behavior for those it "does not like," and watches objects it "feels neutral" about.

During the execution of any of its behaviors, the robot switches its focus between objects and the participant's face. The duration of each such atomic gaze action is listed in Table II.

The *body memory* saves high-level and low-level perceptual data as well as behavior ids and the robot's motivational state to a file which is subsequently used for symbol grounding (see *lexical grounding system* below).

Algorithm 1 Outline of the "Behavioral Loop" for the Robot's Body Behavior for Rejection Experiment. Notice That "offer_detected()" Is Based on Information Broadcast by the Perception System, and "valence()" is Based on Information Pertaining to the Motivation System

```
while negation behavior module is running do
    if ! headController→connected() then
        behavior = IDLE
    else
        if ! offer_detected() then
            behavior = LOOK_AROUND
        else
            getObjectID(oid)
            if valence(oid) > neutralThreshold then
                behavior = REACH_FOR_OBJECT(oid)
            else if valence(oid) < -neutralThreshold then
                REJECT(oid)
            else
                WATCH(oid)
            end if
        end if
    end if
end while
```

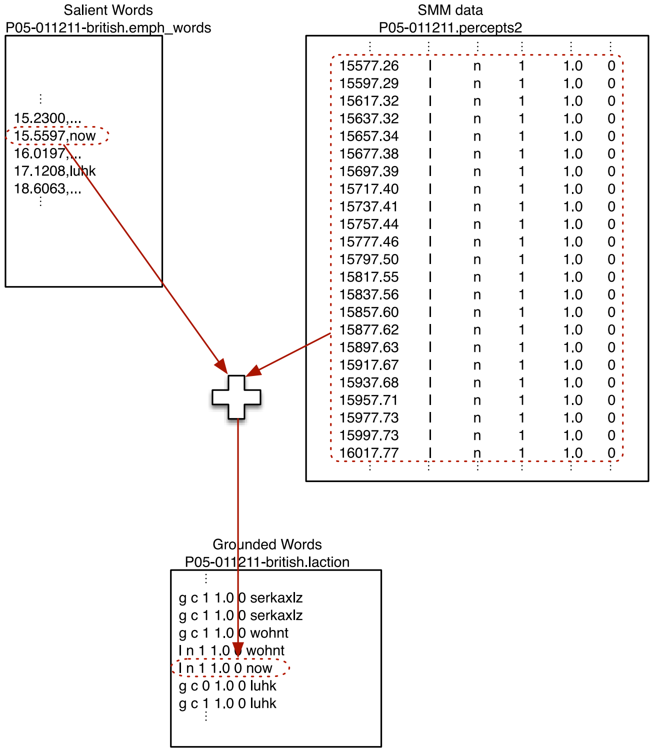
The *auditory system* comprises all processes involved in the extraction of words from the participants' speech. The three functional components may be distinguished: 1) speech recognition and word alignment; 2) prosodic labeling; and 3) word extraction. All of these were developed by Saunders *et al.* [19] and used for this paper in the same manner as they were within the studies described in [21].

Due to the low accuracy of standard speech recognition software for the types of speech as they occur in our experiments, a semiautomated system was employed which relies on the manual transcription of the recorded speech. The output of this system is a timed phonetic transcription.

The latter is manually realigned relative to the original speech recording before the prosodic labeling can be applied. Utterance boundaries are determined as those pauses between words which exceed the average pause duration between all words. Additional utterance boundaries are introduced by a subsequent additional segmentation based on word duration: boundaries are set after words whose duration is larger than the first standard deviation of all words in that utterance (see [19] for details). The outcome of this process is a sequence of utterances consisting of prosodically annotated words which serves as basis for the word extraction. Within the presented experiments exactly one word is extracted per utterance: the prosodically most salient one. Prosodic salience is calculated as $f_0 * energy * d_w$, with f_0 being the maximum fundamental frequency, *energy* the maximum energy, and d_w the duration of the word. All factors were normalized prior to the application of this formula.

The *lexical grounding system* performs the grounding of the lexical items produced by the auditory system. It is executed offline, that is after each experimental session. This module takes as input both smm file that had been recorded by the body memory during that session, and salient word file generated by the auditory system after the same session. Both files are subsequently merged into one file in which each salient word is associated with the *smm* data that was recorded at the time during which the corresponding utterance was produced. In the current implementation we made the decision to eliminate duplicate entries, that originate from the same utterance, in order to keep the lexicon at a manageable size. The grounding process is depicted in Fig. 3. The resulting file, the bottom file in Fig. 3, is merged with the embodied lexicon from previous sessions, if existent, to form an updated embodied lexicon which is subsequently used in the followup session of the respective participant by the languaging system.

The *languaging system* is responsible for the robot's linguistic productions. It reads on startup the *embodied lexicon* which is generated offline by the *lexical grounding system* and matches the current *smm* state of the robot against the lexicon. The matching is done via a k-nearest-neighbor implementation, the Tilburg Memory-Based Learner [56], and with $k = 3$. This choice was made due to its successful use in the experiments of [19], but there is no reason why this particular system could not be replaced with another associative learning system. In order not to fall victim to the curse of dimensionality [57], only high-level percepts and the robot's motivational state was used as basis for matching (Fig. 4). The matching is performed whenever a complete new *smm*-vector is available, which is approximately every 30 ms with the given architecture and hardware. A thresholding mechanism is applied akin to the one employed by Saunders *et al.* [19] which determines which words are uttered at what time.



```
⋮
15.2300,...
15.5597,now
16.0197,...
17.1208,luhk
18.6063,...
⋮
```

```
   ⋮         ⋮    ⋮    ⋮    ⋮    ⋮
15577.26     l    n    1    1.0  0
15597.29     l    n    1    1.0  0
15617.32     l    n    1    1.0  0
15637.32     l    n    1    1.0  0
15657.34     l    n    1    1.0  0
15677.38     l    n    1    1.0  0
15697.39     l    n    1    1.0  0
15717.40     l    n    1    1.0  0
15737.41     l    n    1    1.0  0
15757.44     l    n    1    1.0  0
15777.46     l    n    1    1.0  0
15797.50     l    n    1    1.0  0
15817.55     l    n    1    1.0  0
15837.56     l    n    1    1.0  0
15857.60     l    n    1    1.0  0
15877.62     l    n    1    1.0  0
15897.63     l    n    1    1.0  0
15917.67     l    n    1    1.0  0
15937.68     l    n    1    1.0  0
15957.71     l    n    1    1.0  0
15977.73     l    n    1    1.0  0
15997.73     l    n    1    1.0  0
16017.77     l    n    1    1.0  0
   ⋮         ⋮    ⋮    ⋮    ⋮    ⋮
```

```
⋮
g c 1 1.0 0 serkaxlz
g c 1 1.0 0 serkaxlz
g c 1 1.0 0 wohnt
l n 1 1.0 0 wohnt
l n 1 1.0 0 now
g c 0 1.0 0 luhk
g c 1 1.0 0 luhk
⋮
```

Fig. 3. Grounding of (salient) words. The grounding process associates lexical entries, in our case prosodically salient words, with the concurrently occurring smm data. In our system, the salient word is propagated across the entire duration of the utterance such that the time stamps, visible in the salient-words-file (top-left) mark the start and end of the respective utterance within which the word was produced. Time stamps for utterance boundaries are symbolized by "..." Also notice that we remove duplicates of grounded words that would ensue from the same utterance. In the given example, this means that due to the lack of change within the *smm* data during the production of the utterance the potentially up to 23 ensuing identical grounded words are collapsed into one (bottom).

For each word that is returned by the matching process a counter is maintained which is incremented for each match and decreased for each nonmatch with a lower bound of 0 for all counters. If the same word is returned as best match two or more times in direct succession an additional increment is added to its counter. Once a counter reaches a set threshold its respective word is synthesized and uttered by the robot and the counter is reset to 0. The numerical value of the threshold was chosen empirically such that the resulting speech frequency appeared neither too fast nor too slow. However, in contrast to the single global threshold employed by Saunders *et al.* [19] we adapted the threshold to the robot's motivational state. The robot is likely to speak most (lowest threshold) when it likes an object, it speaks slightly less (medium threshold) when it dislikes an object, and it speaks the least frequently (highest threshold) when it does not care about an object (motivation = 0). In order to avoid excessive repetition of the same word during the duration of an identical *smm* context we employed a so called *differential lexicon* which suppresses the last uttered word in the lexicon during constant *smm* context. As soon as some dimension of the current *smm* vector changes, the full, unconstrained lexicon becomes active again. For further details of each of the functional subcomponents of the languaging system (see [52]).

E. Taxonomy of Negation Types

On the *pragmatic level* we constructed two taxonomies of negation types, one based on the robot's speech, and another one based on participants' speech and subsequently classified

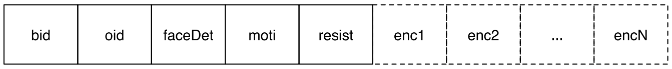


Fig. 4. *smm* vector. Solid line indicates dimensions that were used within experiments; *bid*: behavior id, *oid*: object id, *faceDet*: face detected, *moti*: motivation value, *resist*: resistance detected, and *encX*: encoder reading #X. Within the rejection experiment described in this paper the resistance value is constantly 0 due to participants not physically manipulating the robot's arm.

the negative utterances of both parties according to these taxonomies. We developed two instead of just a single taxonomy in order to keep the robot's taxonomy as close to Pea's taxonomy of early negation as possible such that the analytical results would be comparable (see [58]). Pea's taxonomy only covers children's but not their parents' productions and therefore is often times insufficient for classifying parental negative expressions.

The first taxonomy to be constructed was the taxonomy for robot utterances from an observer perspective and is based on Pea's taxonomy of early negation types [38] which originates in studies on the earliest forms of linguistic negation in children's speech. Subsequently the taxonomy for participants' negation types was constructed. In both taxonomies conversational adjacency constitutes the top-level criterion which naturally affords the construction of conversationally paired types (see [59]). *Negative agreements* on the part of the robot, for example, are preceded by *negative questions* on the participants' part (conversational adjacency), and NIIs uttered by participants are preceded by what is perceived as negative or rejective behavioral displays on the robot's side (behavioral adjacency). Additional, that is, stand-alone types were introduced where this was deemed necessary. All types may be considered speech act theoretical [60], [61] and conversation analytical [59], [62] hybrids in the sense that not only the type of illocutionary force is taken as a distinguishing criterion but also conversational adjacency. The taxonomies were constructed based on a data set consisting of data from the herein presented *rejection* experiment as well as the related *prohibition* experiment not discussed in this publication.[2] The taxonomies therefore contain more negation types than those relevant for the present publication. The complete taxonomies are described in [52], but short descriptions and examples of those negation types most frequently produced within the rejection experiment are given below.

Upon construction of the taxonomies, the latter were applied by the two coders to the negative utterances in the experiments in order to determine their negation type. Prior to the coding of the negative utterances for their type the coders were asked to code the robot's negative utterance set for felicity from the observers' perspective, i.e., they had to judge whether they, as fluent English speakers, would rule a given robot utterance as adequate or apparently meaningful in the given conversational context (see notions of felicity in [60] and [63]). We chose to code for felicity prior to coding for type in order to prevent the criteria used when judging for type from influencing the intuitive judgment regarding felicity. The second coder annotated about 20% of each the participants' and robot's negative utterances which had been randomly selected from the full utterance set from both rejection and prohibition experiment (see also the coding scheme printed in [52, Appendix] for details). Participants' utterances were coded only for type, the robot's utterances were additionally coded for felicity as described. The second coder was employed in order to evaluate the goodness of the constructed taxonomies in terms of intercoder agreement as well as the reliability of the felicity judgments [64]. The intercoder agreements for robot negation type, robot negation felicity, and human negation type in terms of κ values were 0.46, 0.41, and 0.74, respectively. The former two are on the lower end of Rietveld and van Hout's scale [64], [65] for *moderate agreement*, but would be considered too low on Krippendorff's [66] scale. The κ-values resulting from the coding of human negation types on the other side indicate *substantial agreement* between the two coders. We subsequently investigated ways to optimize the robot's taxonomy such that one could reasonably expect a better κ-value when using the optimized taxonomy. Prior to describing the outcome of these attempts we need to introduce those negation types mentioned within this paper. These include the ones most frequently produced during the experiments. In the following those types found in human participants' speech are qualified with "[H]." The types found in the robot's speech are marked "[R]." In the examples question marks indicate the intonational contour of a question, full stops the contour of an assertion.

NII [H] are linguistic interpretations produced by conversationally stronger partners, in the case of children typically mothers or fathers, referring to the emotional and/or volitional states of the child or, in our case, the robot [38, p. 179]. In other words, the conversationally strong partners produce negative expressions that fit the motivational or volitional state of the conversationally weaker partner. Typically the semantics of these expressions is negative as well, i.e., the participant expresses that she thinks the robot does *not want* or *not like* either a particular object or does *not want* or *not like* to perform a particular action such as holding the box.

Examples would be utterances such as "no, you do not like the circle" but also a simple "no?" when not a proper question, i.e., if it appears that an answer is not expected.

Negative motivational questions (NMQs [H]) are questions containing lexical or grammatical negatives which refer to the motivational or emotional state of the addressee. They are therefore similar to NIIs, the only difference being that they are deemed to be proper questions, i.e., the speaker does expect a response by the addressee. They may refer to the motivational state directly such as "are you not feeling well today?" but may also refer to preferences or volitional stances such as "you do not like the heart?" or "you do not want to play?"

Truth-functional denials (TFD [H]) are denials of a truth-functional assertion, i.e., an assertion whose truth is

[2] While it will be reported elsewhere, the prohibition experiment was designed to test a further hypothesis on the developmental origin of linguistic negation, namely that the first negation words to be uttered by a child can be traced back to prohibitive utterances on part of the caretaker. This hypothesis is complementary to the hypotheses of this paper. But in order to clarify the relationship between the prohibition and rejection experiments, we would like to point out that the experimental setup of the former was designed as an extension of the present experiment in order to render the results comparable.

independent of both speaker's as well as addressee's likes, dislikes, capabilities, or perspectives. Examples would be "no, it is not a heart" in reply to the assertion that something was a heart, or a simple *no* in reply to a "it is raining outside." We may expect this type of negation to be used in abundance in a teaching scenario where factual knowledge is what is being taught such as the scenario given in our experiment.

Truth-functional negations (TFN [H]) in our taxonomy capture all kinds of TFN which are not TFDs. TFN is in this sense a residual class that captures all nonadjacent truth-functional utterances, be they negative assertions, suggestions, speculations, or guesses about state of affairs, which are in essence truth-functional. We also allotted negative normative assertions about prevailing rules and norms to this category (second example).

Examples: "It will not rain today" and "in England, you must not drive on the right-hand side."

Negative agreements (A [H+R]) are given if one of the conversation partners produces a negative utterance of some kind and the other agrees with it by uttering a negative as well. Often but not always the latter negative is a repetition of the negative expression used by the first speaker. The first speaker's utterance can have an "assertional" intonation contour or a question contour.

Examples: A: "do not you like carrots?" B: "no" and A: "so you do not like carrots." B: "no."

Motivation-dependent Denials [R] are negative responses to *motivation-dependent questions* or *assertions*. This means that they depend on the current motivational state of the addressee, her likes, dislikes or preferences.

Example: A asks B "do you want a beer?" B answers with no, or with "no, I do not drink alcohol."

Rejections (*R*) [R] are very similar to *motivation-dependent denials*, the main difference being that an utterance of the latter type is adjacent to another utterance of the conversation partner. *Rejections* on the other hand are reactions to nonlinguistic offers or proposals of some kind. Our "definition" of linguistic *rejection* appears to be more narrow than the one employed by Pea. Our definition was thought to sharpen the distinction between the two types.

Example: A physically offers something to B (for example by extending his hand toward B, palm facing upward containing the object in question). B rejects it with a simple "no thanks" or an even simpler *no*.

Negative tag question (NTQ [H]) are negative grammatical constructions that are attached to the end of the utterance. They consist of the negated auxiliary verb of the main clause, if there is one, plus a personal pronoun. They appear to be at the very fringe of "negative meanings" and may not be considered a proper negation type. Despite it being doubtful whether they are semantic negatives, we included them nonetheless as they occurred frequently and clearly contain a grammatical negative.

Examples: "You do like it, do not you?" "You have been to Paris, have not you?"

Two attempts to optimize the robot's taxonomy of negation words were made, where the first involved a set of automatic optimization attempts by way of merging negation types.[3] This optimization attempt indicated that fusions of TFD with *rejection* as well as *negative agreement* with *self-prohibition*, a type only occurring in the prohibition experiment, were necessary for us to reasonably expect a sufficient intercoder agreement. The fusion of the former is highly problematic from a developmental point of view as the two types are firmly distinct in accounts of negation in early child language [37], [38], [67]. Such a fusion would therefore render our taxonomy incomparable to taxonomies constructed to delineate human language development. We therefore did not fuse the indicated types and focused instead on the reasons for the coders' lack of more than moderate agreement in the context of the robot's linguistic negation. The second attempt consisted of coder interviews designed to determine the reasons for disagreement between the particularly disagreed upon types (see [52, Secs. 5.3.6–5.3.8] for details). This means that in the context of the robot's negative utterances we have to be careful when considering numerical results based on judgments from only one coder. The good news in this context is that there was no indication that any one of the two coders would have judged the robot's negative speech systematically more felicitous as compared to the other.

F. Data Analysis

In total approximately 4 h of participants' speech were analyzed originating from 50 experimental sessions, five per participant.[4] The participants' speech was analyzed on three different levels while paying particular attention to negative utterances, words or types: 1) *utterance level*; 2) *word* or *corpus level*; and 3) *pragmatic level*. Additionally we applied the utterance and corpus level analyses to the speech data gathered in Saunders *et al.*'s [21] experiment for the purpose of comparison. The latter consists of approximately 1 h and 35 min of speech originating from nine participants and five sessions per participant. The shorter duration is due to each session only being 2 to 3 min long compared to the approximately 5-min duration per session in our rejection experiment.

On the *utterance-level* we measured the mean length of utterance, the speech frequency in utterances per minute (*u/min*), and the number of distinct words. These measurements form the basis of Fig. 5. The same utterance boundaries were adopted as determined by the auditory system for the purposes of word extraction. All of these measures were calculated based on the full set of utterances of both experiments as well as on the subset of negative utterances only.

Negative utterances within the speech recordings were detected in the following manner. Based on the transcripts we first printed a complete list of all distinct words in the complete corpus of participants' speech and manually selected all negative words such as no, "do not," "not," etc. (see Table III

[3] The optimization algorithm performed all possible mergers and calculated the to-be-expected κ value under the assumption that a coder who under the initial taxonomy had decided that an utterance u would be of type t_1, and assuming t_1 were to be merged with some other type t_2 to a joint type t_{12}, the coder would then categorize u as being of the new type t_{12}.

[4] The transcripts of participants' speech are available at http://uhra.herts.ac.uk/handle/2299/18196.

TABLE III
List of Negation Words Used for Analyses. All Negation Words Listed Here Were Selected From a Complete List of Words Obtained by Accumulating the Words From the Transcripts of the Experiments. A Trailing "(2)" Signifies a Second Phonetic Variant of the Same Lexical Word

no	don't	don't (2)	not	didn't	didn't (2)
isn't	won't	can't	can't (2)	wouldn't	doesn't
doesn't (2)	couldn't	wasn't	weren't	haven't	hasn't
mustn't	cannot	shouldn't	nono	neither	

for the complete list). Based on this list of negative words all negative utterances were extracted from the transcript by automatic means, where negative utterances are those utterances which contain at least one negative word. Subsequently additional steps were undertaken to temporally align the textual artifacts with the video recordings for the pragmatic analysis.

On the *corpus-level* we constructed separate corpora from participants' speech transcripts for both our and Saunders *et al.*'s [21] experiment yielding the *rejection*, and Saunders *et al.*'s [21] *corpora*. Moreover separate corpora for only those words marked as prosodically salient by the auditory system were assembled resulting in the salient word only subcorpora [see Fig. 5(c)]. Prosodically marked salient words are particularly relevant to our robot's language acquisition process as only they become part of its embodied lexicon and thus come to constitute its active vocabulary (see also [20], [21]).

III. Results

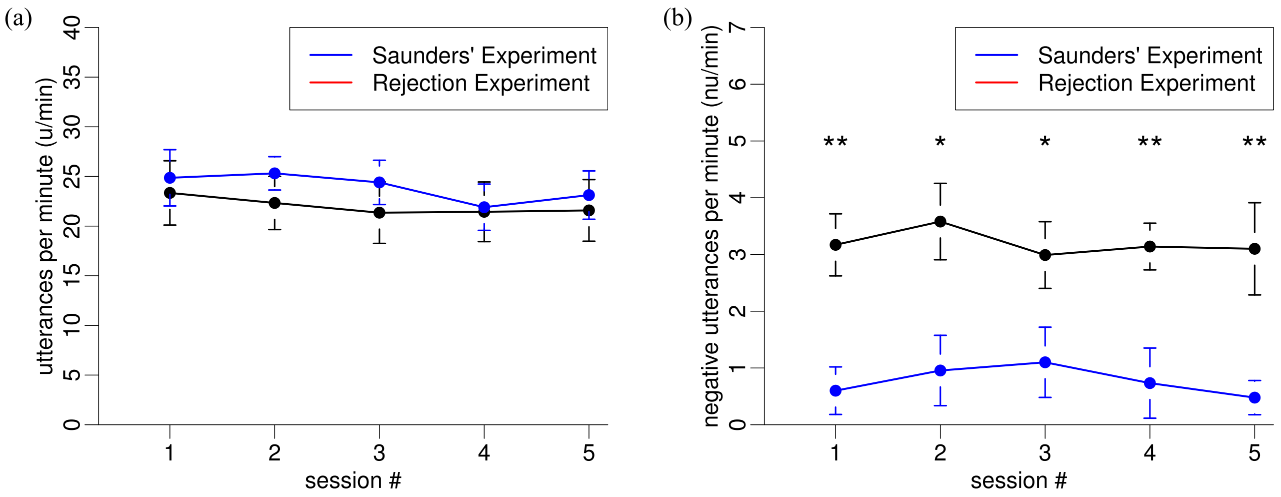
On the *utterance level*, we performed two-sample *t*-test analyses comparing participants' utterances per minute and negative utterances per minute on a per-session basis. Whereas there is no significant difference in the overall production rate of the participants in the rejection as compared to those in Saunders *et al.*'s [21] experiment [Fig. 5(a)], every 6–7 utterances of their speech contain a negation word, a rise of 332% compared to participants' speech in Saunders *et al.*'s [21] experiment [see Fig. 5(b)]. The preponderance of negative utterances leads to a stark rise of negation words in the set of prosodically salient words, the words that enter the robot's embodied lexicon. The quantitative rise of negation words is amplified by the fact that *no* has a comparatively high prosodic saliency rate in the negation experiments, both relative to Saunders *et al.*'s [21] experiment as well as relative to the average word saliency within the rejection experiment [Fig. 6(b)]. As a result *no* becomes the second most-frequent salient word in the corpus of salient words of the rejection experiment [Fig. 5(c)]. A pragmatic analysis of these negative utterances shows that NII are at 31%[5] the most frequent negation type, closely followed by the NMQs (30%), both of which are directly linked to the robot's affective displays. The absolute production rate of TFDs ranks third (22%). Yet the overall lower saliency marking rate of the latter type (29%) as compared to the ones of the two motivational types (NII: 54% and NMQ: 49%) means that the vast majority of negation words in the robot's active vocabulary, in terms of exemplars propagated into the data set of its memory-based learner, originate from the two motivational or affective types. Some intent interpretations and motivational questions were performed in non-negative ways. For example some participants at times lexicalized the robot's negative affective display with emotion words such as *sad* as in "why are you sad?" instead of the more common negative "you do not like it?" We cannot provide the reader with a numerical value in terms of the percentage of such positive intent interpretations as we focused our pragmatic analysis on negative utterances only but we find it important to point out that not all intent interpretations which "describe" a negative motivational state are necessarily negative.

In comparison *no* ranks at place 50 in Saunders *et al.*'s [21] experiment and accounts for less than 0.5% of the corpus [Fig. 5(c)]. The effect of the motivational displays of the robot is thus a vastly higher production rate of negative utterances as compared to the nearly identical experimental setup of Saunders *et al.* [21] where the robot does not display an affective or motivational stance.

Fig. 7(a) displays the mean of the robot's vocabulary size across the five sessions. Shown are both its total as well as its active vocabulary size. With *total vocabulary* we refer here to all grounded words contained in its *languaging* module which are loaded during the start of each session and form the basis of its speech. We counted only lexically unique words but point out that its lexicon file will contain many duplicates with identical or differing *smm* associations. The lower plot shows its active vocabulary as it was displayed by speech during the respective session. As can be seen from the graphs, the robot only ever uttered a fraction of the words in its vocabulary. This could be caused by words that differ lexically but are identical in terms of their smm grounding, and where one of these smm-identical words is numerically dominant. If this is the case and robot finds itself in the respective *smm* context, the numerically more strongly represented word will overshadow the other words such that it is likely that only this word will be uttered in the respective session. The robot's average talkativeness within each session is more or less constant as can be seen in Fig. 7(b) and will strongly depend on the chosen speaking thresholds of the *languaging* module (see Section II.D).

The pragmatic analysis of the robot's negative utterances with regard to their felicity or adequacy, i.e., its "pragmatic" learning success, yielded a success rate of 66.1% (see [52, Tab. 5.51]). An additional temporal analysis correlating instances of NIIs with the robot's motivational state yielded that in roughly two thirds of times (66.2%) the robot was in a negative state when participants produced the respective instance of a NII. Only in 8.9% of all cases was the NII uttered when the robot was in a positive state and in the remaining 24.9% of cases the NII co-occurred with the robot being in a neutral state. The circumstance that in the majority of times participants (co-)produced NII while the robot displayed and was in a negative motivational state provides the potential ground for an association between negative word and negative motivation. For NMQs the temporal correlation between productions of utterances of this type and negative state is weaker and might explain the "pragmatic misfires" of the robot.

[5] This and all the following percentage values are rounded to the nearest 0.1%.

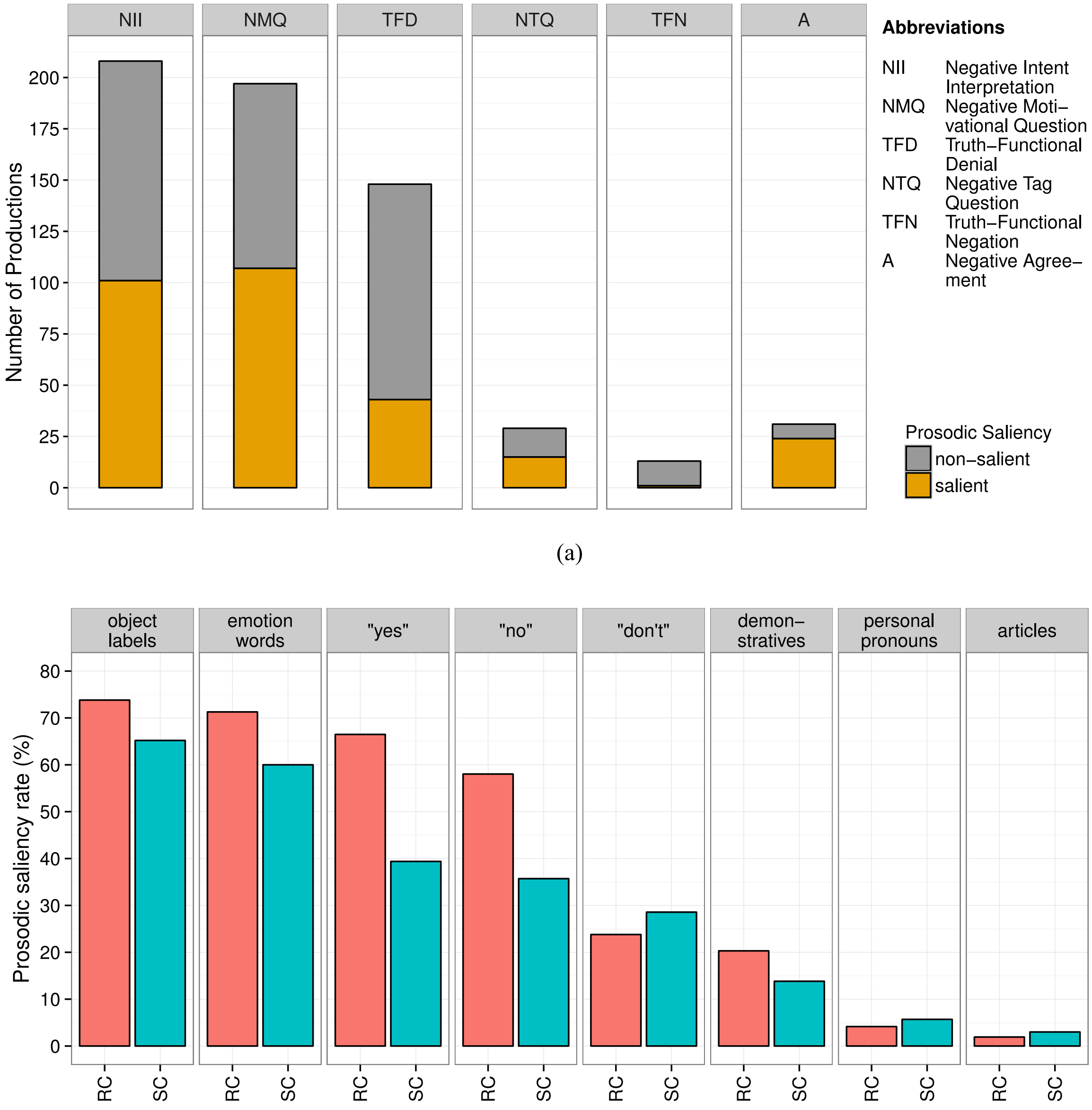


(c)

	All words				Salient words only			
	Rejection Corpus		Saunders et al. Corpus		Rejection Corpus		Saunders et al. Corpus	
rank	word	%	word	%	word	%	word	%
1	you	7.13	a	8.71	square	4.97	blue	6.91
2	the	5.63	this	4.55	no	4.64	red	5.54
3	like	3.31	blue	4.31	triangle	3.95	circle	5.15
4	a	2.72	is	4	heart	3.8	heart	4.75
5	this	2.7	and	3.9	moon	3.53	green	4.36
6	no	2.39	red	3.75	circle ↑	3.53	arrow	3.56
7	one	2.27	green	3.55	like	3.2	cross	3.48
8	square	1.93	the	3.29	circles	2.69	side ↑	3.48
9	do ↑	1.93	that's	2.54	squares	2.42	box	2.82
10	to	1.78	you	2.41	it	2.36	shape	2.42
11	it ↑	1.78	it's	2	yes	2.28	and	2.11
12	that	1.73	heart	1.99	one	2.13	moon ↑	2.11
13	moon	1.62	circle	1.85	right	1.88	square	2.07
14	heart	1.6	arrow	1.84	this	1.82	this	2.02
15	is	1.47	side	1.81	ok	1.73	star	1.85
16	triangle	1.45	cross	1.49	again	1.69	ok	1.76
17	circle	1.32	here	1.39	oh	1.57	is ↑	1.76
18	it's	1.22	we	1.38	good	1.55	small	1.54
19	don't	1.15	on	1.37	Deechee	1.46	right ↑	1.54
+1	see	1.12	no (*50*)	0.35	Deechee^2	1.3	no (*32*)	0.44

Fig. 5. Impact of motivated behavior on linguistic production of participants—ten and nine participants in the rejection and Saunders *et al.*'s [21] experiment, respectively. (a) General production rates in terms of *utterances per minute* between Saunders *et al.*'s [21] experiment (upper blue) and rejection experiment (lower black) differ only marginally (mean ± SEM). (b) Yet the production rate of negative utterances per minute is significantly higher in the rejection (middle black) as compared to Saunders *et al.*'s [21] experiment (lower blue). $^*P < 0.05$, $^{**}P < 0.01$ (two sample t-test). (c) Twenty most frequent words plus negation words ordered by rank and based on combined word corpora of all participants' speech. The relative abundance of negative utterances is mirrored on this level: *No* belongs to the ten most frequent words in the rejection corpus, compared to its rank 50 in Saunders *et al.*' [21] corpus. *No* ranks even higher in the corpus of prosodically salient words due to its high saliency and the word therefore enters the robot's vocabulary frequently. Arrows (↑) indicate rank-equality between the stated and the next higher entry. Negation words are emphasized with gray background. The "+1" row contains the 20th most-frequent words unless a different rank is specified in brackets. The superscript^2 indicates a second phonetic variant of a lexically identical word.

In 40.1% of cases, the robot was in and displayed a negative state, in 40.9% in a neutral state and in 19% in a positive state. Thus, in terms of the potential for establishing a (statistical) association between a negative motivational state and negation word, NIIs appear to be clearly better suited as compared to NMQs. Given that instances of both types are not mutually exclusive but were produced in the same sessions, the potential for establishing such an association is still given. Yet due to the prominent co-occurrence of neutral motivational states and negative words in the case of NMQs, pragmatic misfires or misuses of such words are to be expected as witnessed in our experiments. This is particularly the case when the core word learning mechanism is modeled as a mainly associative one (with *associative* in the Hebbian sense). Unfortunately, due to the absence of quantitative psycholinguistic data on the felicity rates of infants' use of negation words we cannot conclude how good the robot's felicity rates are in comparison. In other words, we do not know how good our robot is in its use of *no* and similar words as compared to children because we do not know how good toddlers are at using them felicitously. We do know that children in certain stages overgeneralize nouns [68] or grammatical constructions [69], [70]. Yet no comparable,



(a)

(b)

Fig. 6. (a) Frequency of human utterances classified as being of the stated negation types (pragmatic level) and percentage of utterances falling under the respective type with salient negation word (only types with > 5% of total number of negative utterances). (b) Prosodic saliency rates of selected words and word groups. *No* has a considerably higher rate for being prosodically marked as salient in the rejection experiment as compared to Saunders *et al.*'s [21] experiment, RC: rejection corpus, and SC: Saunders *et al.*'s [21] corpus.

let alone quantitative data on the successful use of *no* and other negation words exists.

IV. Discussion

The observation that the majority of NII productions are well-aligned to the interlocutor's display of affect may not be surprising to conversation analysts who observed humans to be sensitive to differences of several hundred milliseconds [71] in the conversational moves of other speakers. Yet such tight interactive couplings have only rarely been considered in robotic language acquisition. Symbol grounding up to this point, if any human teacher was employed at all, is in most cases a very carefully conducted process where typically the constructor of the system or another trained person acts as language teacher. In these cases, either due to their privileged knowledge of the inner workings of the system, or by mere training, these teachers utter the right words, or at least the right kind of words, at the right time (see [10], [23]). Roy and Pentland [16] provided a rare counterexample where recorded unconstrained CDS is used to train a symbol-grounding system. Nevertheless the vast majority of symbol-grounding systems so far depended on these somewhat artificial ways of establishing linguistic or interactional regularities for the learning to be successful. The presented rejection experiment shows that certain aspects of human conversational behavior are regular enough to serve as learning

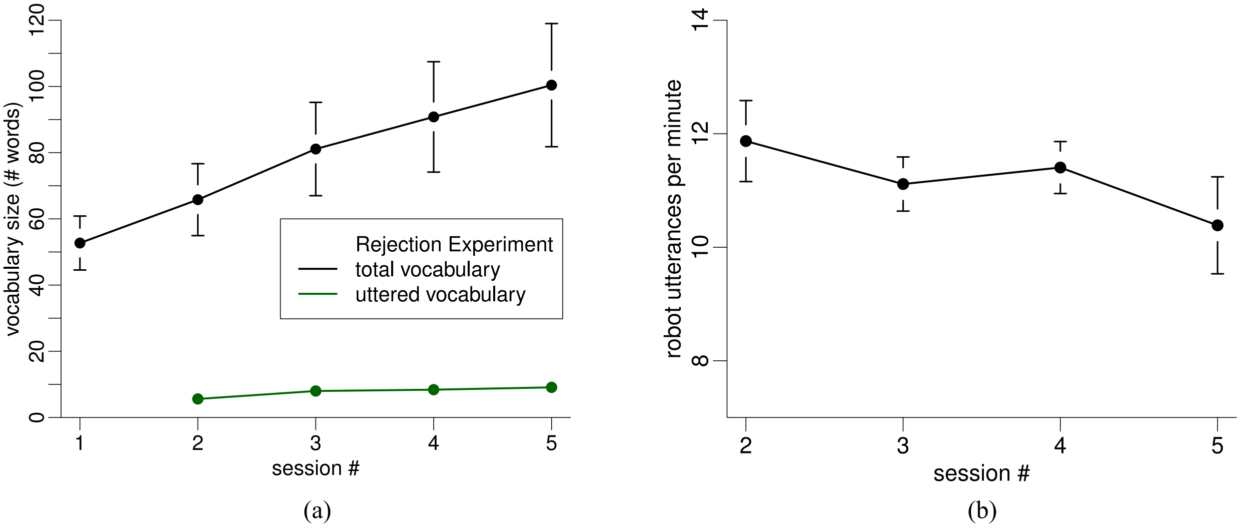


Fig. 7. Mean robot vocabulary size and number of robot utterances per minute in the rejection experiment. (a) Top black line displays the total vocabulary size of the robot in terms of the number of lexically unique grounded words extracted from participants' speech (mean ± SEM). This is the vocabulary which is loaded into the robot's languaging module at the beginning of the subsequent session and the sole basis of its speech. Note that the actual vocabulary file may also contain lexical duplicates with differing smm data as well as potential real duplicates where both lexical words and the associated smm data are identical. The bottom green line displays the number of unique words that the robot actually uttered in the respective session (mean ± SEM). (b) Mean number of robot utterances/words per minute by session across all participants (mean ± SEM).

resource for some of children's earliest words. In particular, we demonstrated that the use of early negation words may be learned by exploiting these conversational and pragmatic regularities. The latter may be a symptom or side product of what Levinson [72] has termed the "interaction engine," a set of cognitive abilities and behavioral dispositions unique to our species which form a prerequisite to all language acquisition. Intent interpretations, originally proposed by Pea [38] Ryan [39] as source for the acquisition of volition words, have indeed the potential to act as a central interactional resource for learning how to negate.

Given that our knowledge of intent interpretations "in the wild," that is within parent–child dyads, are only anecdotal rather than quantitative, only detailed psycholinguistic observations including recordings of their precise timing would provide us with the means to properly evaluate and contrast our observations.

Another source of insight into the adequacy of the robot's "negative" linguistic behavior is the analysis of the sources of coder disagreement [52]. As mentioned in Section II, our automatic optimization attempt yielded the developmentally nonsensical suggestion to fuse *rejection* with TFD in order to increase the intercoder agreement. The qualitative analysis subsequently indicated that the main reason for disagreement or uncertainty amongst coders as to which of these two types an utterance belongs to was the following: both coders were frequently uncertain when deciding whether a given negative word, typically *no*, would constitute an "in game" move, in which case it would most probably be classified as TFD, or whether it constituted a "meta" move, which means that the coder judged the robot unwilling to play the "naming game" at all, a case of *rejection*. A contributing factor to this uncertainty is the lack of natural timing of the robot's utterances. Rather than speaking within the hypothetically important time frame of 1 s after an adjacent utterance or producing a communicatively equivalent behavioral move [71], the robot speaks when it is "confident enough" that a given *smm*-word association holds. This apparent violation of conversational time constraints thus impacts directly upon semantic or pragmatic judgments with regards to the meaning of a given negative utterance. We refer to [52] for a more elaborate discussion of reasons for coders' confusion, but would like to emphasize the importance of timing upon the meaning of a word: there are words, in this paper mainly *no*, whose meaning for the addressee as well as for observers and coders of the conversation changes or becomes ambiguous if certain time thresholds are exceeded. Beyond it being a merely methodological issue this observation provides support for the view that the meaning of *no* is pragmatic rather than strictly semantic (see Montgomery's [45] discussion of pragmatic ends as source of word meaning for mental terms).

V. Future Work

Given the exploratory nature of this paper the architectural choices were driven by a design variant of Occam's razor: we chose the simplest algorithms whenever given a seemingly arbitrary choice. The rationale behind this strategy was to keep the number of variables in the learning architecture as small as possible in order not to confuse the eventual analysis with a plethora of variables. The two areas where this choice may have been too simplistic are the chosen 1-D model of affect or motivation, and the choice of a mere association-style core learning algorithm, *k*-nearest-neighbor, for grounding words in smm data. This means that apart from participants' prosodic emphasis of important words, no social signals are evaluated, nor could they, in principle, be incorporated into the learning algorithm in any straightforward manner. It is conceivable that the transition toward the more powerful class of reinforcement learning algorithms would improve the robot's learning success. The introduction of such signals could render the robot sensitive to socially derived reinforcement signals as well as internal reinforcement signals potentially derived from the robot's successful use of words. The ability of assessing one's own success or felicity may also provide an indicator as to when words can be "crystallized" or condensed into "concepts"—the current choice of a lazy learner as core learning algorithm does not lend itself easily for the purpose of concept formation. Moreover, a choice of reinforcement learning in this context appears to be more in line with established psychological models of learning, where the term "associative learning," in stark contrast to its use in computer science, already implies the presence of reinforcement signals rather than referring to the reinforcement-free style of algorithms akin to Hebbian learning [73].

Another factor in the cases of unsuccessful language use was that the robot, unbeknownst to participants, was real-time deaf, i.e., it did not know when it was being spoken to. This prevented it from replying in a timely manner other than by chance. Our analysis of the reasons for coder's uncertainty underlines the importance of timing for speech production. The particular meaning of *no* is dependent upon its type, conversational adjacency is an essential criterion in judgments for type, and judgments about adjacency are time-sensitive. Therefore observers' as interlocutors' inference of the meaning of *no* is sensitive to the production timing of the utterance. In the future, we intend to evaluate timing issues more generally by conducting a formal conversation analysis of the conversations recorded within the experiments.

In terms of future improvement of the learning architecture the indicated manifold degrees of freedom above suggest that the most promising and effective approach may be temporal high-density studies of interactions of parent–child dyads. Apart from informing future design choices, such studies, if conducted with sufficient temporal resolution, i.e., on the level of a few hundreds of milliseconds, could show whether children's emotional and volitional displays reliably trigger parental intent interpretations, thereby either confirming or rebuffing our findings based on participant-robot dyads. In terms of assessing the robot's linguistic performance, a comparative study is needed in order to assess children's performance when using negation words. Such a study should start at the time of their onset of speech and continue for several months charting their felicity in using *no* and other similarly pragmatic words. To date no quantitative insight about children's (un-)felicitous use of such words exists akin to the analysis which we performed on our data set. Most helpful would be precise descriptions of the type and timing of eventual pragmatic errors qua misfires, and their impact upon the further course of the respective interaction. In particular parental reactions to such misfires, as well as eventual real-world pragmatic results of the misfire would help in constraining speculations about the potential presence of social as well as internal reinforcement signals.

Insights from such observations would greatly facilitate future modeling efforts for word learning algorithms in that they could narrow the search space of potential algorithmic modifications to ecologically valid choices.

Another practical issue for improvement concerns the speech recognition technology which we employed to extract words from participants' speech. Our experiments were rendered rather time-consuming by the need to employ manual transcriptions due the word recognition rate not being sufficient otherwise. In this context recent improvements in automatic speech recognition (ASR) and statistical language modeling such as deep learning [74], [75], heuristic ASR systems [76], [77], and ASR systems designed to cope with certain noise levels [78] promise improvements for future experiments in human–robot interaction learning systems.

VI. Conclusion

This paper provides support for the hypothesis that the earliest forms of linguistic negation, in English predominantly *no*, may originate in parental NIIs. These intent interpretations are typically triggered by bodily displays of negative affect or volition, displays of "not liking" and "not wanting" something, respectively, with these interpretations being temporally tightly coupled with the triggering displays. Via the described implementation we gave an example that social mechanisms akin to the ones encountered in CDS may be elicited and exploited by an artificial embodied system for machine learning purposes. These intent interpretations are a rarely documented interactional phenomenon in the contemporary developmental literature, yet in this paper we found them to be a potential resource of meaning for grounding negative words such as *no* in affect.

Our architecture is the first to extend symbol grounding beyond the realm of sensorimotor-data to encompass affect. Recent psychological studies indicate that affect may also play an important role in the grounding of so called abstract concepts other than negation [79]. Affective grounding thus may help artificial agents to tackle concepts other than negation. Although associative learning in the Hebbian sense combined with certain social mechanisms may suffice to acquire a certain linguistic skill level, more powerful learning algorithms such as reinforcement learning might be needed in the long run, for example to stabilize the robot's active vocabulary. Yet in order to answer the question which core learning mechanism underlies children's word learning in the one-word stage and in the context of negation words, more detailed psycholinguistic studies are needed with a particular focus on the precise timing of both behavioral and linguistic conversational moves between parent and child.

References

[1] G. J. Hollich *et al.*, "Breaking the language barrier: An emergentist coalition model for the origins of word learning," *Monographs Soc. Res. Child Develop.*, vol. 65, no. 3, pp. 1–135, 2000.

[2] R. Bedford *et al.*, "Failure to learn from feedback underlies word learning difficulties in toddlers at risk for autism," *J. Child Lang.*, vol. 40, no. 1, pp. 29–46, 2013.

[3] H. B. Hani, A. M. Gonzalez-Barrero, and A. S. Nadig, "Children's referential understanding of novel words and parent labeling behaviors: Similarities across children with and without autism spectrum disorders," *J. Child Lang.*, vol. 40, no. 5, pp. 971–1002, 2013.

[4] A. Borovsky and J. Elman, "Language input and semantic categories: A relation between cognition and early word learning," *J. Child Lang.*, vol. 33, no. 4, pp. 759–790, 2006.

[5] J. P. Bunce and R. M. Scott, "Finding meaning in a noisy world: Exploring the effects of referential ambiguity and competition on 2·5-year-olds' cross-situational word learning," *J. Child Lang.*, vol. 44, no. 3, pp. 650–676, 2017.

[6] M. Tomasello, *Constructing a Language*. Cambridge, MA, USA: Harvard Univ. Press, 2003.

[7] J. D. Mastin and P. Vogt, "Infant engagement and early vocabulary development: A naturalistic observation study of Mozambican infants from 1;1 to 2;1," *J. Child Lang.*, vol. 43, no. 2, pp. 235–264, 2015.

[8] A. Cangelosi, E. Hourdakis, and V. Tikhanoff, "Language acquisition and symbol grounding transfer with neural networks and cognitive robots," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Vancouver, BC, Canada, 2006, pp. 1576–1582.

[9] A. Cangelosi, "Grounding language in action and perception: From cognitive agents to humanoid robots," *Phys. Life Rev.*, vol. 7, no. 2, pp. 139–151, 2010.

[10] P. Dominey and J. Boucher, "Learning to talk about events from narrated video in a construction grammar framework," *Artif. Intell.*, vol. 167, nos. 1–2, pp. 31–61, 2005.

[11] P. F. Dominey and J.-D. Boucher, "Developmental stages of perception and language acquisition in a perceptually grounded robot," *Cogn. Syst. Res.*, vol. 6, no. 3, pp. 243–259, 2005.

[12] K. Gold, M. Doniec, C. Crick, and B. Scassellati, "Robotic vocabulary building using extension inference and implicit contrast," *Artif. Intell.*, vol. 173, no. 1, pp. 145–166, 2009.

[13] C. Lyon *et al.*, "Embodied language learning and cognitive bootstrapping: Methods and design principles," *Int. J. Adv. Robot. Syst.*, vol. 13, p. 105, Jan. 2016.

[14] N. Mavridis and D. Roy, "Grounded situation models for robots: Bridging language, perception, and action," in *Proc. AAAI Workshop Modular Construct. Human Like Intell.*, 2005, pp. 32–39.

[15] A. F. Morse *et al.*, "Modeling u-shaped performance curves in ongoing development," in *Proc. 33rd Annu. Meeting Cogn. Sci. Soc. Expanding Space Cogn. Sci.*, 2011, pp. 3034–3039.

[16] D. K. Roy and A. P. Pentland, "Learning words from sights and sounds: A computational model," *Cogn. Sci.*, vol. 26, no. 1, pp. 113–146, 2002.

[17] D. Roy, "Grounding words in perception and action: Computational insights," *Trends Cogn. Sci.*, vol. 9, no. 8, pp. 389–396, 2005.

[18] D. Roy, "A mechanistic model of three facets of meaning," in *Symbols and Embodiment: Debates on Meaning and Cognition*, A. C. Graesser, M. de Vega, and A. M. Glenberg, Eds. Oxford, U.K.: Oxford Univ. Press, 2008, ch. 11.

[19] J. Saunders, C. L. Nehaniv, and C. Lyon, "The acquisition of word semantics by a humanoid robot via interaction with a human tutor," in *New Frontiers in Human-Robot Interaction*, K. Dautenhahn and J. Saunders, Eds. Philadelphia, PA, USA: John Benjamins, 2011, pp. 211–234.

[20] J. Saunders, H. Lehmann, Y. Sato, and C. L. Nehaniv, "Towards using prosody to scaffold lexical meaning in robots," in *Proc. IEEE Int. Conf. Develop. Learn. (ICDL)*, Frankfurt, Germany, 2011, pp. 1–7.

[21] J. Saunders, H. Lehmann, F. Förster, and C. L. Nehaniv, "Robot acquisition of lexical meaning—Moving towards the two-word stage," in *Proc. IEEE Int. Conf. Develop. Learn. Epigenet. Robot. (ICDL)*, San Diego, CA, USA, 2012, pp. 1–7.

[22] J. M. Siskind, "Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic," *J. Artif. Intell. Res.*, vol. 15, no. 1, pp. 31–90, 2001.

[23] Y. Sugita and J. Tani, "Learning semantic combinatoriality from the interaction between linguistic and behavioral processes," *Adapt. Behav.*, vol. 13, no. 1, pp. 33–52, 2005.

[24] J. Zhong, M. Peniak, J. Tani, T. Ogata, and A. Cangelosi, "Sensorimotor input as a language generalisation tool: A neurorobotics model for generation and generalisation of noun-verb combinations with sensorimotor inputs," *ArXiv e-prints, arXiv:1605.03261 [cs.RO]*, 2016. [Online]. Available: http://adsabs.harvard.edu/cgi-bin/bib_query?arXiv:1605.03261

[25] J. Tani, *Exploring Robotic Minds: Actions, Symbols, and Consciousness As Self-Organizing Dynamic Phenomena*. New York, NY, USA: Oxford Univ. Press, 2016.

[26] L. Steels and J.-C. Baillie, "Shared grounding of event descriptions by autonomous robots," *Robot. Auton. Syst.*, vol. 43, nos. 2–3, pp. 163–173, 2003.

[27] L. Steels, "Modeling the cultural evolution of language," *Phys. Life Rev.*, vol. 8, no. 4, pp. 339–356, 2011.

[28] L. Steels, "Grounding language through evolutionary language games," in *Language Grounding in Robots*, L. Steels and M. Hild, Eds. Boston, MA, USA: Springer, 2012, pp. 1–22.

[29] L. Steels, "Agent-based models for the emergence and evolution of grammar," *Philosoph. Trans. Roy. Soc. B Biol. Sci.*, vol. 371, no. 1701, 2016, Art. no. 20150447.

[30] S. Harnad, "The symbol grounding problem," *Physica D Nonlin. Phenomena*, vol. 42, nos. 1–3, pp. 335–346, 1990.

[31] L. Fenson *et al.*, "Variability in early communicative development," *Monographs Soc. Res. Child Develop.*, vol. 59, no. 5, pp. 1–185, 1994.

[32] M. Hao *et al.*, "Developmental changes in the early child lexicon in mandarin Chinese," *J. Child Lang.*, vol. 42, no. 3, pp. 505–537, 2015.

[33] A. Schults, T. Tulviste, and K. Konstabel, "Early vocabulary and gestures in estonian children," *J. Child Lang.*, vol. 39, no. 3, pp. 664–686, 2012.

[34] D. Bleses *et al.*, "Early vocabulary development in Danish and other languages: A CDI-based comparison," *J. Child Lang.*, vol. 35, no. 3, pp. 619–650, 2008.

[35] A. Gopnik, "Three types of early word: The emergence of social words, names and cognitive-relational words in the one-word stage and their relation to cognitive development," *First Lang.*, vol. 8, no. 22, pp. 49–70, 1988.

[36] V. Volterra and F. Antinucci, "Negation in child language: A pragmatic study," in *Developmental Pragmatics*, E. Ochs, Ed. New York, NY, USA: Academic Press, 1979.

[37] S. Choi, "The semantic development of negation: A cross-linguistic longitudinal study," *J. Child Lang.*, vol. 15, no. 3, pp. 517–531, 1988.

[38] R. Pea, "The development of negation in early child language," in *The Social Foundations of Language & Thought: Essays in Honor of Jerome S. Bruner*, D. Olson, Ed. New York, NY, USA: W. W. Norton, 1980, pp. 156–186.

[39] J. Ryan, "Early language development: Towards a communicational analysis," in *The Integration of a Child Into a Social World*, M. P. M. Richards, Ed. Cambridge, U.K.: Cambridge Univ. Press, 1974.

[40] V. Slaughter, C. C. Peterson, and M. Carpenter, "Maternal mental state talk and infants' early gestural communication," *J. Child Lang.*, vol. 36, no. 5, pp. 1053–1074, 2009.

[41] J. Olson and E. F. Masur, "Infants' gestures influence mothers' provision of object, action and internal state labels," *J. Child Lang.*, vol. 38, no. 5, pp. 1028–1054, 2011.

[42] N. Budwig, "A developmental-functionalist approach to mental state talk," in *Language, Literacy, and Cognitive Development: The Development and Consequences of Symbolic Communication*, E. Amsel and J. P. Byrnes, Eds. Mahwah, NJ, USA: Lawrence Erlbaum Assoc., 2002, pp. 59–86.

[43] K. Nelson, "Language pathways into the community of minds," in *Why Language Matters for Theory of Mind*, J. W. Astington and J. A. Baird, Eds. Oxford, U.K.: Oxford Univ. Press, 2005, pp. 26–49.

[44] D. K. Symons, "Mental state discourse, theory of mind, and the internalization of self—Other understanding," *Develop. Rev.*, vol. 24, no. 2, pp. 159–188, 2004.

[45] D. E. Montgomery, "The developmental origins of meaning for mental terms," in *Why Language Matters for Theory of Mind*, J. W. Astington and J. A. Baird, Eds. Oxford, U.K.: Oxford Univ. Press, 2005, pp. 106–122.

[46] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, "The iCub humanoid robot: An open platform for research in embodied cognition," in *Proc. 8th Workshop Perform. Metrics Intell. Syst.*, Gaithersburg, MD, USA, 2008, pp. 50–56.

[47] W. Daelemans and A. van den Bosch, *Memory-Based Language Processing*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[48] C. E. Snow, "Mothers' speech research: From input to interaction," *Talking to Children: Language Input and Acquisition*. Cambridge, U.K.: Cambridge Univ. Press, 1977, pp. 31–49.

[49] C. Gallaway and B. J. Richards, *Input and Interaction in Language Acquisition*. London, U.K.: Cambridge Univ. Press, 1994.

[50] E. L. Newport, "Motherese: The speech of mothers to young children," in *Cognitive Theory*, vol. 2, J. N. Castellan, D. B. Pisoni, and G. R. Potts, Eds. Hillsdale, NJ, USA: Lawrence Erlbaum Assoc., 1977, pp. 177–217.

[51] G. Metta, P. Fitzpatrick, and L. Natale, "YARP: Yet another robot platform," *Int. J. Adv. Robot. Syst.*, vol. 3, no. 1, pp. 43–48, 2006.

[52] F. Förster, "Robots that say 'no': Acquisition of linguistic behaviour in interaction games with humans," Ph.D. dissertation, School Comput. Sci., Univ. Hertfordshire, Hatfield, U.K., Sep. 2013.

[53] J. Rüsch *et al.*, "Multimodal saliency-based bottom-up attention—A framework for the humanoid robot iCub," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Pasadena, CA, USA, 2008, pp. 962–967.

[54] F. J. Varela, E. Thompson, and E. Rosch, *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA, USA: MIT Press, 1991.

[55] K. R. Scherer, T. Bänziger, and E. Roesch, *A Blueprint for Affective Computing: A Sourcebook and Manual*. Oxford, U.K.: Oxford Univ. Press, 2010.

[56] (Oct. 5, 2012). *TiMBL*. [Online]. Available: http://ilk.uvt.nl/mblp/

[57] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.

[58] F. Förster, C. L. Nehaniv, and J. Saunders, "Robots that say 'no,'" in *Proc. 10th Eur. Conf. (ECAL)*, Budapest, Hungary, Sep. 2009, pp. 158–166.

[59] I. Hutchby and R. Wooffitt, *Conversation Analysis: Principles, Practices and Applications*. Cambridge, U.K.: Polity Press, 1998.

[60] J. L. Austin, *How to Do Things With Words*. Cambridge, MA, USA: Harvard Univ. Press, 1975.

[61] J. R. Searle, *Speech Acts: An Essay in the Philosophy of Language*. Cambridge, U.K.: Cambridge Univ. Press, 1969.

[62] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[63] S. C. Levinson, *Pragmatics*. Cambridge, U.K.: Cambridge Univ. Press, 1983.

[64] B. Di Eugenio, "On the usage of Kappa to evaluate agreement on coding tasks," in *Proc. LREC*, vol. 1. Athens, Greece, 2000, pp. 441–444.

[65] T. Rietveld and R. van Hout, *Statistical Techniques for the Study of Language and Language Behaviour*. Berlin, Germany: Mouton de Gruyter, 1993.

[66] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*. Newbury Park, CA, USA: Sage, 1980.

[67] L. Bloom, *Language Development: Form and Function in Emerging Grammars*. Cambridge, MA, USA: MIT Press, 1970.

[68] S. A. Gelman, W. Croft, P. Fu, T. Clausner, and G. Gottfried, "Why is a pomegranate an apple? The role of shape, taxonomic relatedness, and prior lexical knowledge in children's overextensions of apple and dog," *J. Child Lang.*, vol. 25, no. 2, pp. 267–291, 1998.

[69] M. Bowerman, "The 'no negative evidence' problem: How do children avoid constructing an overly general grammar?" in *Explaining Language Universals*. Oxford, U.K.: Basil Blackwell, 1988, pp. 73–101.

[70] P. J. Brooks, M. Tomasello, K. Dodson, and L. B. Lewis, "Young children's overgeneralizations with fixed transitivity verbs," *Child Develop.*, vol. 70, no. 6, pp. 1325–1337, 1999.

[71] G. Jefferson, "Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation," in *Conversation: An Interdisciplinary Perspective*, D. Roger and P. Bull, Eds. Clevedon, U.K.: Multilingual Matters, 1989, ch. 8, pp. 166–196.

[72] S. C. Levinson, "On the human 'interaction engine,'" in *Roots of Human Sociality: Culture, Cognition and Interaction*, N. J. Enfield and S. C. Levinson, Eds. Oxford, U.K.: Berg, 2006, ch. 1, pp. 39–69.

[73] R. A. Rescorla and A. R. Wagner, "A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement," in *Classical Conditioning II: Current Research and Theory*, vol. 2, A. H. Black and W. F. Prokasy, Eds. New York, NY, USA: Appleton-Century-Crofts, 1972, pp. 64–99.

[74] A. L. Maas *et al.*, "Building DNN acoustic models for large vocabulary speech recognition," *Comput. Speech Lang.*, vol. 41, pp. 195–213, Jan. 2017.

[75] W. De Mulder, S. Bethard, and M.-F. Moens, "A survey on the application of recurrent neural networks to statistical language modeling," *Comput. Speech Lang.*, vol. 30, no. 1, pp. 61–98, 2015.

[76] X. L. Aubert, "An overview of decoding techniques for large vocabulary continuous speech recognition," *Comput. Speech Lang.*, vol. 16, no. 1, pp. 89–114, 2002.

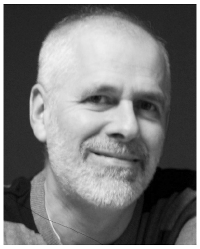
[77] B. Lecouteux and D. Schwab, "Ant colony algorithm applied to automatic speech recognition graph decoding," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 2122–2126.

[78] H. Kallasjoki, J. F. Gemmeke, and K. J. Palomäki, "Estimating uncertainty to improve exemplar-based feature enhancement for noise robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 2, pp. 368–380, Feb. 2014.

[79] S.-T. Kousta, G. Vigliocco, D. P. Vinson, M. Andrews, and E. D. Campo, "The representation of abstract words: Why emotion matters," *J. Exp. Psychol.*, vol. 140, no. 1, pp. 14–34, 2011.

Frank Förster received the Ph.D. degree from the University of Hertfordshire, Hatfield, U.K., in 2013.

He is Research Fellow in the Adaptive Systems Research Group, University of Hertfordshire. From 2014 to 2015, he was a Research Assistant at the Multimedia and Vision Laboratory, Queen Mary University of London, London, U.K. He has carried out research on language acquisition in robotics via linguistically unrestricted human–robot interaction. His current research interests include language acquisition in its widest sense, pragmatics, social and developmental robotics, HRI, and regularities and coordination mechanisms underpinning human and human–robot face-to-face interaction.



Joe Saunders received the Ph.D. degree from the University of Hertfordshire, Hatfield, U.K., in 2007.

He is a Senior Research Fellow in the Adaptive Systems Research Group, University of Hertfordshire. He has carried out research on robot adaptation and learning at both physical and linguistic levels.

Chrystopher L. Nehaniv received the B.Sc. degree (honors) from the University of Michigan, Ann Arbor, MI, USA, in 1987, with a focus on mathematics, biology, linguistics, and cognitive science, and the Ph.D. degree in mathematics from the University of California at Berkeley, Berkeley, CA, USA, in 1992.

He is a Ukrainian–American Mathematician and Computer Scientist. He has held academic and research positions at the University of California at Berkeley, the University of Aizu, Aizuwakamatsu, Japan (full professorship from 1993 to 1998), and in Hungary. Since 2001, he has been a Research Professor of Mathematical and Evolutionary Computer Science at the University of Hertfordshire, Hatfield, U.K., where he plays leading roles in the Algorithms, Biocomputation, and Adaptive Systems Research Groups. He has authored over 200 publications and edited 30 volumes in computer science, mathematics, interactive systems, artificial intelligence, and biosystems and coauthored a monograph on algebraic theory of automata networks. His current research interests include interaction, development, and enaction in biological and artificial systems, the notion of first- and second-person experience (including phenomenological and temporally extended experience in ontogeny) in humanoids and living things, as well as on mathematical and computer algebraic foundations and methods for complex adaptive systems.