A multimodal corpus for the evaluation of computational models for (grounded) language acquisition

Judith Gaspers[a], Maximilian Panzner[a], Andre Lemme[b], Philipp Cimiano[a], Katharina J. Rohlfing[c], Sebastian Wrede[b]

[a]Semantic Computing Group, CITEC, Bielefeld University, Germany
{jgaspers|mpanzner|cimiano}@cit-ec.uni-bielefeld.de
[b]Research Institute for Cognition and Robotics (CoR-Lab), Bielefeld University, Germany
{alemme|swrede}@cor-lab.uni-bielefeld.de
[c]Emergentist Semantics Group, CITEC, Bielefeld University, Germany
kjr@uni-bielefeld.de

Abstract

This paper describes the design and acquisition of a German multimodal corpus for the development and evaluation of computational models for (grounded) language acquisition and algorithms enabling corresponding capabilities in robots. The corpus contains parallel data from multiple speakers/actors, including speech, visual data from different perspectives and body posture data. The corpus is designed to support the development and evaluation of models learning rather complex grounded linguistic structures, e.g. syntactic patterns, from sub-symbolic input. It provides moreover a valuable resource for evaluating algorithms addressing several other learning processes, e.g. concept formation or acquisition of manipulation skills. The corpus will be made available to the public.

1 Introduction

Children acquire linguistic structures through exposure to (spoken) language in a rich context and environment. The semantics of language may be learned by establishing connections between linguistic structures and corresponding structures in the environment, i.e. in different domains such as the visual one (Harnad, 1990). Both with respect to modeling language acquisition in children and with respect to enabling corresponding language acquisition capabilities in robots, which may ideally be also grounded in their environment, it is hence of great interest to explore i) how linguistic structures of different levels of complexity, e.g. words or grammatical phrases, can be derived from speech input, ii) how structured representations for entities observed in the environment can be derived, e.g. how concepts and structured representations of actions can be formed, and iii) how connections can be established between structured representations derived from different domains. In order to gain insights concerning the mechanisms at play during language acquisition (LA), which enable children to solve these learning tasks, models are needed which ideally cover several learning tasks. For instance, they may cover the acquisition of both words and grammatical rules as well as the acquisition of their grounded meanings. Complementarily, data resources are needed which enable the design and evaluation of these models by providing suitable parallel data.

Aiming to provide a basis for the development and evaluation of LA models addressing the acquisition of rather complex and grounded linguistic structures, i.e. syntactic patterns, from sub-symbolic input, we designed a German multimodal input corpus. The corpus consists of data of multiple speakers/actors who performed actions in front of a robot and described these actions while executing them. Subjects were recorded, i.e. parallel data of speech, stereo vision (including the view-perspective of the "infant"/robot) and body postures were gathered. The resulting data hence allow grounding of linguistic structures in both vision and body postures. Among others, learning processes that may be evaluated using the corpus include: acquisition of several linguistic structures, acquisition of visual structures, concept formation, acquisition of generalized patterns which abstract over different speakers and actors, establishment of correspondences between structures

Proc. of 5th Workshop on Cognitive Aspects of Computational Language Learning (CogACLL) @ EACL 2014, pages 30–37,
Gothenburg, Sweden, April 26 2014.

from different domains, acquisition of manipulation skills, and development of appropriate models for the representations of actions.
This paper is organized as follows. Next, we will provide background information concerning computational models of LA. In Section 3, we will then describe the corpus design and acquisition, including the desired properties of the collected data, corresponding experimental settings and technical implementation. We will then present the resulting data set and subsequently conclude..

2 Background

To date, several models addressing LA learning tasks have been proposed and evaluated using different copora. Yet, these models typically focus on a subset or certain aspects of the LA learning tasks mentioned in the previous section, often assuming other learning tasks, e.g. those of lower complexity, as already solved by the learner. For instance, models addressing the acquisition of grammatical constructions and their meaning (Kwiatkowski et al., 2012; Alishahi and Stevenson, 2008; Gaspers and Cimiano, in press; Chang and Maia, 2001) typically learn from symbolic input. In particular, assuming that the child is already able to segment a speech signal into a stream of words and to extract structured representations from the visual context, such models typically explore learning from sequences of words and symbolic descriptions of the non-linguistic context. Models addressing the acquisition of word-like units directly from a speech signal (Räsänen, 2011; Räsänen et al., 2009) have also been explored. These, however, typically do not address learning of more complex linguistic structures/constructions.
Taken together, lexical acquisition from speech and syntactic acquisition have been mainly studied independently of each other, often assuming that syntactic acquisition follows from knowledge of words. However, learning processes might actually be interleaved, and top-down learning processes may play an important role in LA. For instance, with respect to computational learning from symbolic input, it has been shown that knowledge of syntax can facilitate word learning (Yu, 2006). Children may, for instance, also make use of syntactic cues during speech segmentation and/or word learning, but models addressing lexical acquisition from speech have to date mainly ignored syntax (Räsänen, 2012). Models addressing the acquisition of syntactic patterns directly from speech provide a basis for exploring to what extent learning mechanisms might be interleaved in early LA. Moreover, they allow to investigate the possible role of several top-down learning processes which have to date been little explored.
Several corpora comprising interactions of children with their caregivers have been collected. A large such resource is the CHILDES data base (MacWhinney, 2000), which contains transcribed speech. Data from CHILDES have been often used to evaluate models learning from symbolic input, in particular models for syntactic acquisition from sequences of words; additional accompanying symbolic context representations have been often created (semi–)automatically. Moreover, multimodal corpora containing caregiver-child interactions have been recorded and annotated (Björkenstam and Wirn, 2013; Yu et al., 2008), thus also allowing to study the role of social interaction and extra-linguistic cues in language learning. By contrast, in this work we aim to provide a basis for developing and evaluating models which address the acquisition of syntactic patterns from speech. Hence, allowing to derive generalized patterns, linguistic units as well as the objects and actions they refer to have to re-appear in the data several times. Thus, in line with the CAREGIVER corpus (Altosaar et al., 2010) we did not record caregiver-child interactions but attempted to approximate speech used by caregivers with respect to the learning task(s) at hand. However, the focus of the CAREGIVER corpus is on models learning word-like units from speech. Thus, a number of keywords were spoken in different carrier sentences; speech is accompanied by only limited non-linguistic context information in the corpus. In contrast to CAREGIVER, we did not restrict language use directly and recorded parallel context information from different modalities, focusing not only on the acquisition of word-like units from speech and word-to-object mapping but moreover on the acquisition of simple syntactic patterns and mapping language to actions.

3 Corpus design and acquisition

In this section, we will first describe the desired properties of the corpus. Subsequently, we will present the corresponding experimental settings, used stimuli and procedure, the technical imple-

mentation of the robot behavior and the data acquisition as well as the resulting corpus.

3.1 Desired properties

Our goal was to design a corpus comprising multimodal data which supports the evaluation of computational models addressing several LA learning tasks, and in particular the acquisition of grounded syntactic patterns from sub-symbolic input only as well as the development of components supporting the acquisition of language by robots. Thus, the main focus was to design the corpus in such a way that the data acquisition scenario was simplified enough to allow solving the task of learning grounded syntactic patterns from sub-symbolic input with the resulting data set (which of course contains much less data when compared to the innumerable natural language examples children receive when acquiring language over several years). In particular, since the acquisition of rather complex structures should be enabled using sub-symbolic information, several (repeated) examples for contained structures were needed, allowing the formation of generalized representations. Thus, we opted for a rather simple scenario. Specifically, the following properties were taken into account:

- Rather few objects and actions were included that could moreover be differentiated rather easily from a visual point of view. However, in order to reflect differences between actions, these differed i) with respect to the number of their referents as well as ii) with respect to their specificity to certain objects. In particular, we included actions which could be performed on different subsets of the objects, ranging from specificity to one certain object to being executed with all of the objects.

- Objects and actions reappeared several times, yielding several examples for each of them. Repeated appearance is an essential aspect, since the formation of generalized representations starting from continuous input requires several observations in order to allow abstraction over observed examples/different actors and speakers.

- The scenario was designed such that it encouraged human subjects to use rather simple syntactic patterns/short sentences. Yet, language use was in principle unrestricted in order to acquire rather natural data and to capture speaker-dependent differences. This also reflects the input children receive in that parents use rather simple language when talking to children.

- Data were gathered from several human subjects in order to allow for the evaluation of generalization over different speakers (with different acoustic properties and different language use, e.g. different words for objects, different syntactic patterns with different complexity, etc) as well as over different actors in case of actions, since children interact with different people and are able to solve this task. Moreover, generalization to different speakers/actors is also important with respect to learning in artificial agents which should preferably not be operable by a single person only.

- Parallel data were gathered in which objects and actions were explicitly named when they were used. This is an important aspect because the corpus should allow learning connections between vision, i.e. objects and actions, and speech (segments) referring to these objects/actions, i.e. (sequences of words) and syntactic patterns. It reflects the input children receive in that caregivers also explain/show objects directly to their children and may show them how to use objects/perform actions in front of them (Rolf et al., 2009; Schillingmann et al., 2009).

We opted for the collection of parallel data concerning vision and body postures for human tutors. Hence, the corpus allows grounding of linguistic structures in both vision and body postures. Including body postures moreover allows the evaluation of algorithms showing manipulation skills which is of interest with respect to learning in robots.

We used stereo vision to allow computational learners to reliably track object movement and interaction using both visual and depth information. With respect to vision, four cameras with two different perspectives were used: two static external cameras as well as the robot's two internal moving cameras. The latter basically mimics the "infant" view, i.e. while the external cameras were static,

the robot moved its eyes (and thus the cameras) and focused on the tutor's hand performing the actions, thus reflecting how a child may focus her/his attention to the important aspects of a scene/a performance of her/his caregiver.

3.2 Participants

A total of 27 adult human subjects participated in data collection (7 male, 20 female, mean age: 26). Subjects were paid for their participation.

3.3 Experimental setting

Human subjects performed pre-defined actions and simultaneously described their performances in front of the robot iCub (Metta et al., 2008); Fig. 1 depicts a human subject interacting with iCub. While interacting with iCub, human subjects' behavior was recorded. In particular, the following data were recorded simultaneously:

Figure 1: A human subject interacting with iCub.

- Speech/Audio (via a headset microphone)
- Vision/Video, static perspective (via two cameras, allowing for stereo vision)
- iCub-Vision/Video, iCub's (attentive) perspective (via iCub's two internal cameras, again allowing for stereo vision)
- Body postures (via a Kinect).

An experimental sketch showing the experimental setting including the positions of the human subject and iCub, as well as camera and Kinect positions, is illustrated in Fig. 2. As can be seen, the human subject was placed directly opposite to iCub. The two external cameras and the Kincet were placed slightly sloped opposite to the subject. Subjects were instructed about which actions should be performed via a computer screen which was operated by an experimentator.
In order to encourage subjects to perform the tutoring task rather naturally, i.e. just like they were

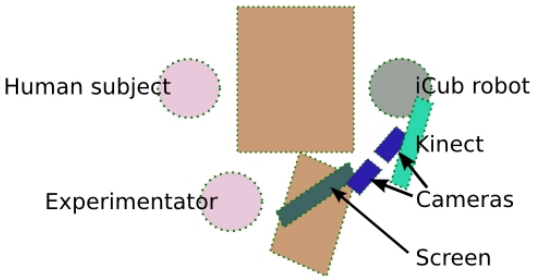

Figure 2: Experimental sketch.

interacting with a human (child), iCub provided feedback (Nagai and Rohlfing, 2009; Fischer et al., 2011). In particular, a gazing behavior was implemented to make the robot appear attentively following the tutoring.

3.4 Stimuli

Data were gathered in the framework of a toy cooking scenario. In particular, subjects prepared several dishes in front of iCub using toy objects. Specifically, 21 toy objects were chosen such that

Figure 3: Utilized objects.

they were rather easy to differentiate with respect to color and/or form. The chosen objects were: pizza, pita bread, plate, bowl, spaghetti, pepper, vinegar, red pepper, lettuce leafs, tomato, onion, cucumber, cheese, toast, salami, chillies, egg, anchovy, cutting board, knife, and mushrooms. The objects are depicted in Fig. 3. Moreover, six different actions were chosen which could be executed using these objects. Again, the goal was to support rather easy identification visually (with respect to their trajectories). The chosen actions were: *showing an object*, *cutting an object (egg or tomato) into two pieces (with knife)*, *placing an object onto another one (plate, pizza, cutting board, toast)*, *putting an object into another one (bowl, pita bread)*, *pour vinegar*, and *strew pepper*. Thus, most actions were object-specific to a

certain degree, i.e. they were to be executed with a certain subset of the objects each. The *show* action was to be executed using each of the objects. Furthermore, 20 different dishes, i.e. preparation processes each consisting of a sequence of actions, were created (four dishes including salad, pizza, pita bread, spaghetti and sandwich/toast, respectively). This was done in order to gather rather fluent/consistent courses of action and rather fluent communication in case of descriptions. For instance, one sequence for preparing a salad started as follows: *showing bowl*, *showing lettuce leafs*, *putting lettuce leafs into bowl*, *showing cutting board*, *showing knife*, *showing tomato*, *putting tomato onto cutting board*, *cutting tomato into two pieces*, *putting tomato pieces into bowl*, etc.

3.5 Procedure

Subjects first prepared one dish while not being recorded in order to get familiar with the task. They were instructed to perform presented actions and to describe their performance simultaneously. Moreover, they were asked to name objects and actions explicitly, since a goal of the corpus is to allow learning connections between speech, vision and body postures. Subjects were not asked to use particular words or phrases, but were free to make own choices. For instance, when being exposed to a picture of the pita bread, they were supposed to explicitly name the pita bread. Yet, they were free to choose a suitable word (or sequence of words), e.g. "Pita", "Pitatasche", "Teigtasche", "Dönertasche", "Brottasche", etc.

Actions to be performed were presented to the subjects via a computer screen; either one action was presented or – in most cases – two actions were presented at once to be executed one after another. In most cases two actions were presented in order to gain more fluent communications and courses of action. In no case more than two actions were presented together because we wanted subjects to focus on performance and not on remembering a certain course of action. Actions were presented only in the form of pictures in order to elicit rather natural language use. In particular, as mentioned previously, subjects could choose freely how to name objects and actions. An example for a screen/picture showing two actions to be performed one after another is presented in Fig. 4. An experimentator operated the screen, i.e. guided the subjects through the sequences of

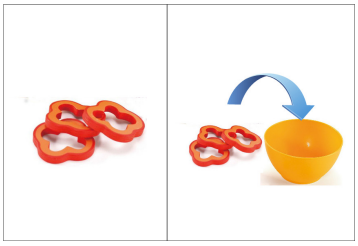

Figure 4: Example screen showing the actions *show red pepper* and *put red pepper into bowl*.

actions. Subjects participated for approximately one hour; only subject's actual performances were recorded, yielding approximately 20–30 minutes of usable material per subject.

3.6 Robot behavior

As mentioned previously, a gazing behavior was implemented to make the robot appear attentively following the tutoring. In particular, the robot's gaze followed a subject's presentation of an action by gazing at her/his right wrist. At times when subjects did not move their hands (to present actions) the robot was looking around, i.e. it gazed at random targets. In the following, the implementation of the robot behavior will be described in more detail.

The experimental setup shown in Fig. 2 allows the system to observe a person in front of the robot iCub. While the presentation task was performed by the person, the robot was supposed to gaze at the right wrist of this person. Via the Kinect data it was possible to acquire the body posture of the robot's interaction partner. We extracted the location of the wrist and represented the Cartesian position in the coordinate system of the robot. This position was then used as the target to generate the head and eye movements. The movement was executed by the *iKinGaze* module available in the iCub software repository (Pattacini, 2010).

Next to this "tracking" behavior of the robot we also used a "background" behavior. The "background" behavior then drew randomly new targets $\mathbf{x}_{targ}$(in meter) from the uniform distribution $\Omega \in [-1.5, -1, 5] \times [-0.2, 0.2] \times [0.2, 0, 4]$ in front of the robot. After convergence to the target the behavior waited for $t = 3$ seconds before a new target was drawn. The switch from "background" behavior to "tracking" behavior was triggered if new targets arrived from the Kinect-based tracking component. This behavior stayed active

as long as targets were received. If no targets were arriving during $t = 2$ sec. after the gazing converged on the last target, the "background" behavior took over. Due to the difference in distance between targets, the motion duration was different as well. Therefore, time delays were added to the target generation, which resulted in a more natural behavior of the robot gazing.

4 Acquired data

In order to record synchronized data from the external sensors, the robot system and the experimental control software, we utilized a dedicated framework for the acquisition of multimodal human-robot interaction data sets (Wienke et al., 2012). The framework and the underlying technology (Wienke and Wrede, 2011) allows to directly capture the network communication of robotics software components. Through this approach, system-internal data from the iCub such as its proprioception and stereo cameras images can be synchronously captured and transformed into an RETF[1]-described log-file format with explicit time and type information. Moreover, additional recording devices such as the Kinect sensors, the external pair of stereo cams or the audio input from a close-talk microphone are captured directly with this system and stored persistently. An example of the acquired parallel data is provided by Fig. 5 while Table 6 summarizes the technical aspects of the acquired data.

The applied framework also supports the automatic export and conversion of synchronized parts of the multimodal data set to common formats used by other 3rd party tools such as the annotation tool ELAN (Sloetjes and Wittenburg, 2008) used for ground truth annotation of the acquired corpus. In this experiment, we additionally captured the logical state of the experiment control software which allowed us to efficiently post-process the raw data and, e.g., automatically provide cropped video files containing only single utterances. A logical state corresponds to the image seen at the screen by a human subject at a certain time, showing the action(s) to be performed.

The acquired corpus contains in total 11.45 hours / approx. 2.3 TB of multimodal input data recorded in 27 trials. Each trial was recorded in about 1 hour of wallclock time and cropped to 20–30 minutes of effective parallel data. While in 5

[1]Robot Engineering Task-Force, cf. http://retf.info/

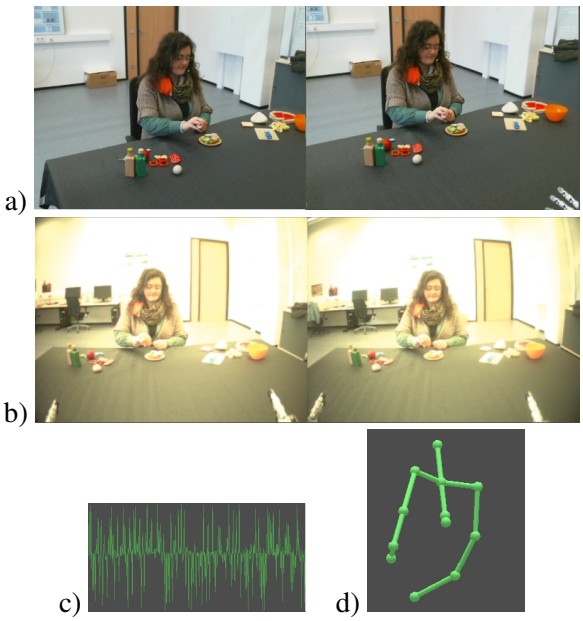

Figure 5: Example of acquired parallel data comprising a) visual data from two static cameras, b) visual data from two cameras contained in the robot's eyes, c) audio and d) body posture data recorded by the Kinect. In this example the subject is preparing a sandwich, and currently strewing pepper onto it.

cases not all of the parallel data streams are available due to difficulties with the robot and the wireless microphones, we decided to leave this data in the corpus to evaluate machine learning processes addressing learning from one or a subset of the modalities only, e.g. blind segmentation of a speech stream.

From the data logs, we exported audio (in AAC format) and the 4 synchronized video (with H.264 encoding) files (MP4 container format) for each trial with an additional ELAN project file for annotation. This annotation is currently carried out; a screenshot of acquired data and corresponding annotations in ELAN is depicted in Fig. 7. It comprises annotation of errors, as well as starting and end points for both presented actions and spoken utterances. In particular, in case of speech word transcriptions are added, while in case of vision actions are annotated in the form of predicate logic formulas. Hence, once the corpus is preprocessed, it is also suitable for the evaluation of models learning from symbolic input with respect to data from one or more domains. For instance, one could explore the acquisition of syntactic patterns from speech by providing parallel visual context

#	Device	Description	Data type	Frequency	Dimension	Throughput
1	Cam 1	Scene video	`rst.vision.Image`	$\approx$ 30 Hz	$640 \times 480 \times 3$	$\approx$ 28 MB/s
2	Cam 2	Scene video	`rst.vision.Image`	$\approx$ 30 Hz	$640 \times 480 \times 3$	$\approx$ 28 MB/s
3	Mic 1	Speech	`rst.audition.SoundChunk`	$\approx$ 50 kHz	1-2	$\approx$ 0.5 MB/s
4	iCub Cam 1	Ego left	`bottle/yarp::sig::Image`	$\approx$ 30 Hz	$320 \times 240 \times 3$	$\approx$ 7 MB/s
5	iCub Cam 2	Ego right	`bottle/yarp::sig::Image`	$\approx$ 30 Hz	$320 \times 240 \times 3$	$\approx$ 7 MB/s
6	Kinect	Body posture	`TrackedPosture3DFloat`[2]	$\approx$ 30 Hz	36	$\approx$ 6 kB/s
7	Control	Logical state	string	$\approx$ 0.05 Hz	-	$\approx$ 5 B/s

Figure 6: Description of acquired data streams, type specifications, average frequency, data dimension and throughput as measured during recording.

information either in sub-symbolic form or in the form of predicate logic formulas.

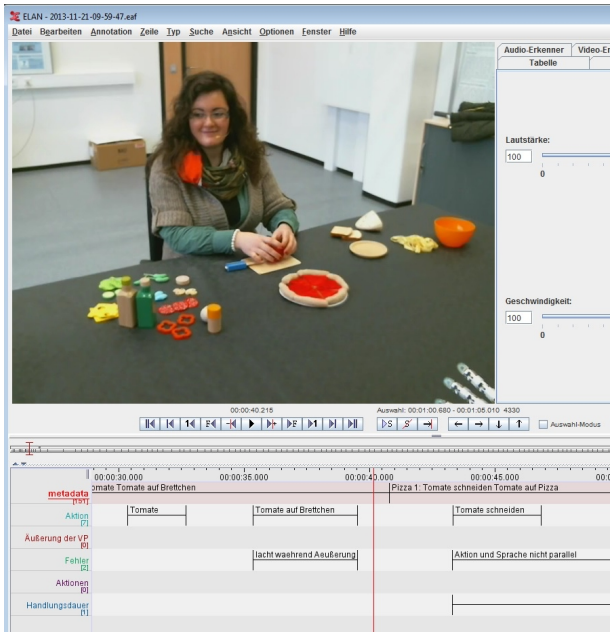

Figure 7: Example of acquired data and corresponding annotations in ELAN.

Word transcriptions for utterances for the whole data set are not yet available. According to the experimentators' impressions, most subjects indeed used, as desired, rather short sentences. Furthermore, a few subjects tried to vary their linguistic descriptions, i.e. to use different sentences for each description. Thus, the corpus appears to cover not only several examples of rather simple linguistic constructions with variations across speakers, but moreover input examples with a rather large degree of linguistic variation for a single speaker, hence providing examples of more challenging data.
We will make the corpus available to the public once post-processing is completely finished.

5 Conclusion

In this paper, we have described the design and acquisition of a German multimodal data set for the development and evaluation of grounded language acquisition models and algorithms enabling corresponding abilities in robots. The corpus contains parallel data including speech, visual data from four different cameras with different perspectives and body posture data from multiple speakers/actors. Among others, learning processes that may be evaluated using the corpus include: acquisition of several linguistic structures, acquisition of visual structures, concept formation, acquisition of generalized patterns which abstract over different speakers and actors, establishment of correspondences between structures from different domains and acquisition of manipulation skills.

Acknowledgments

We are deeply grateful to Jan Moringen, Michael Götting and Stefan Krüger for providing technical support. We wish to thank Luci Filinger, Christina Lehwalder, Anne Nemeth and Frederike Strunz for support in data collection and annotation. This work has been funded by the German Research Foundation DFG within the Collaborative Research Center 673 *Alignment in Communication* and the Center of Excellence *Cognitive Interaction Technology*. Andre Lemme is funded by FP7 under GA. No. 248311-AMARSi.

References

Afra Alishahi and Suzanne Stevenson. 2008. A computational model of early argument structure acquisition. *Cognitive Science*, 32(5):789–834.

Toomas Altosaar, Louis ten Bosch, Guillaume Aimetti, Christos Koniaris, Kris Demuynck, and Henk van den Heuvel. 2010. A speech corpus for modeling language acquisition: Caregiver. In *Proceed-*

ings of the International Conference on Language Resources and Evaluation.

Kristina Nilsson Björkenstam and Mats Wirn. 2013. Multimodal annotation of parent-child interaction in a free-play setting. In *Proceedings of the Thirteenth International Conference on Intelligent Virtual Agents*.

Nancy C. Chang and Tiago V. Maia. 2001. Learning grammatical constructions. In *Proceedings of the 23rd Cognitive Science Society Conference*.

Kerstin Fischer, Kilian Foth, Katharina J. Rohlfing, and Britta Wrede. 2011. Mindful tutors: Linguistic choice and action demonstration in speech to infants and to a simulated robot. *Interaction Studies*, 12(1):134–161.

Judith Gaspers and Philipp Cimiano. in press. A computational model for the item-based induction of construction networks. *Cognitive Science*.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Tom Kwiatkowski, Sharon Goldwater, Luke Zettlemoyer, and Mark Steedman. 2012. A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk*. Mahwah, NJ.

Giorgio Metta, Giulio Sandini, David Vernon, Lorenzo Natale, and Francesco Nori. 2008. The iCub humanoid robot: an open platform for research in embodied cognition. In *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, pages 50–56, New York, NY. ACM.

Yukie Nagai and Katharina J. Rohlfing. 2009. Computational analysis of motionese toward scaffolding robot action learning. *IEEE Transactions on Autonomous Mental Development*, 1:44–54.

Ugo Pattacini. 2010. *Modular Cartesian Controllers for Humanoid Robots: Design and Implementation on the iCub*. Ph.D. thesis, RBCS, Istituto Italiano di Tecnologia, Genova.

Okko Räsänen, Unto K. Laine, and Toomas Altosaar. 2009. Computational language acquisition by statistical bottom-up processing. In *Proceedings Interspeech*.

Okko Räsänen. 2011. A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events. *Cognition*, 120:149176.

Okko Räsänen. 2012. Computational modeling of phonetic and lexical learning in early language acquisition: existing models and future directions. *Speech Communication*, 54:975–997.

Matthias Rolf, Marc Hanheide, and Katharina J. Rohlfing. 2009. Attention via synchrony. making use of multimodal cues in social learning. *IEEE Transactions on Autonomous Mental Development*, 1:55–67.

Lars Schillingmann, Britta Wrede, and Katharina J. Rohlfing. 2009. A computational model of acoustic packaging. *IEEE Transactions on Autonomous Mental Development*, 1:226–237.

Han Sloetjes and Peter Wittenburg. 2008. Annotation by category: Elan and iso dcr. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Johannes Wienke and Sebastian Wrede. 2011. A Middleware for Collaborative Research in Experimental Robotics. In *IEEE/SICE International Symposium on System Integration (SII2011)*, Kyoto, Japan. IEEE.

Johannes Wienke, David Klotz, and Sebastian Wrede. 2012. A Framework for the Acquisition of Multimodal Human-Robot Interaction Data Sets with a Whole-System Perspective. In *LREC Workshop on Multimodal Corpora for Machine Learning: How should multimodal corpora deal with the situation?*, Istanbul, Turkey.

Chen Yu, Linda B. Smith, and Alfredo F. Pereira. 2008. Grounding word learning in multimodal sensorimotor interaction. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.

Chen Yu. 2006. Learning syntax-semantics mappings to bootstrap word learning. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society (2006) Key: citeulike:5276016*.