AUTONOMOUS ACQUISITION OF NATURAL LANGUAGE

Eric Nivel,[1] Kristinn R. Thórisson,[1,6] Bas R. Steunebrink,[5]
Haris Dindo,[2] Giovanni Pezzulo,[4] Manuel Rodriguez,[3] Carlos Hernandez,[3] Dimitri Ognibene,[4]
Jürgen Schmidhuber,[5] Ricardo Sanz,[3] Helgi P. Helgason,[1] Antonio Chella[2] & Gudberg K. Jonsson[7]

[1]*Reykjavik University / CADIA,* [2]*Universita degli studi di Palermo / DINFO,* [3]*Universidad Politecnica de Madrid / ASLAB,* [4]*Consiglio Nazionale delle Ricerche / ISTC,* [5]*Scuola Universitaria Professionale della Svizzera Italiana / IDSIA,* [6]*Icelandic Institute for Intelligent Machines,* [7]*University of Iceland / Human Behavior Laboratory*

ABSTRACT

An important part of human intelligence is the ability to use language. Humans learn how to use language in a society of language users, which is probably the most effective way to learn a language from the ground up. Principles that might allow an artificial agents to learn language this way are not known at present. Here we present a framework which begins to address this challenge. Our auto-catalytic, endogenous, reflective architecture (AERA) supports the creation of agents that can learn natural language by observation. We present results from two experiments where our S1 agent learns human communication by observing two humans interacting in a realtime mock television interview, using gesture and situated language. Results show that S1 can learn multimodal complex language and multimodal communicative acts, using a vocabulary of 100 words with numerous sentence formats, by observing unscripted interaction between the humans, with no grammar being provided to it a priori, and only high-level information about the format of the human interaction in the form of high-level goals of the interviewer and interviewee and a small ontology. The agent learns both the pragmatics, semantics, and syntax of complex sentences spoken by the human subjects on the topic of recycling of objects such as aluminum cans, glass bottles, plastic, and wood, as well as use of manual deictic reference and anaphora.

KEYWORDS

Autonomy, knowledge acquisition, natural language, communication

1. INTRODUCTION

One of the most useful skills to evolve in humans is the ability to use language. This skill builds on several identifiable sub-skills, such as auditory timbre discrimination, sequence learning, fine motor control, and context-anchored learning, whose combination honed over generations has lead to the diverse use of language observed in modern human society. The best way to learn language for a human is in a social context, where multiple examples of its use can be observed, with numerous examples of successful and unsuccessful variations relative to the users' goals, exceptions and contextualized cues and usage help define concepts; and where the effect of language use on oneself and other language users occurs naturally, and where implicit and explicit "experiments" of language use can be made. If our aim is to create an artificial agent that masters the numerous facets and subtleties of language the same is probably true: The agent should be situated in some kind of human social context, where it can learn how to use language. The principles

necessary for such an agent to be constructed are not known at present, and no architecture exists that comes close to supporting the construction of such an agent.

In this paper we present work on an agent that learns language by observation. The mobility and sensing of a situated agent is necessarily constrained by its body; such an agent can neither be assumed to process every input available in its environment nor to follow every thought to its ultimate conclusion: Real-world agents do not have the resources to accomplish *all* the jobs they ideally should, given their goals, due to limited computing and memory capacity. These assumptions have some fundamental implications for the design of our cognitive architecture AERA (auto-catalytic, endogenous, reflective architecture), based on a new constructivist methodology (Thórisson 2012) that puts the autonomy of the agent as a top priority, thus doing away with the allonomic[1] view on which all common software development methodologies are based, where the human programmer provides the system with all its algorithms. Similar to Wang's NARS (Wang 2011, 2006), AERA is designed around assumptions of self-bootstrapping from incomplete knowledge and insufficient resources (Thórisson 2013, Nivel et al. 2012, Wang 2011). An AERA-based agent is provided with only a small object ontology, a handful of top-level goals, and optionally a couple of domain-related goals to help with the bootstrapping. Due to our agents being situated in a social interaction scenario and being engineered to learn continuously, their knowledge is acquired incrementally over time, based on their own experience.

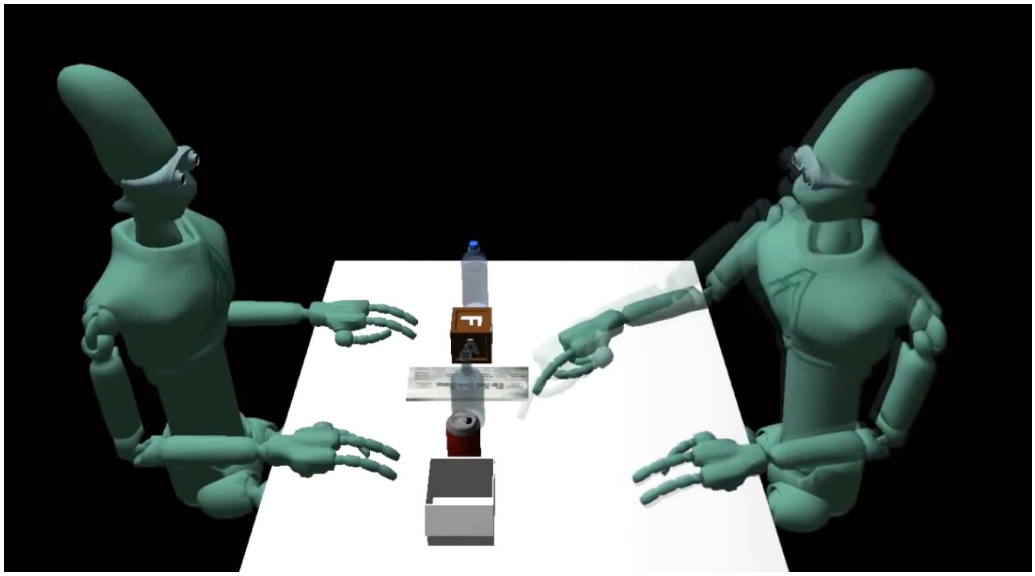

Figure 1. The realtime interaction between a human and the S1 agent, in the form of a simplified TV interview, is conducted in a virtual environment via live tracking of multimodal behavior and speech.

Dethroning allonomic methodologies means that we must let go of the idea that we, the designers, provide our system with task-specific algorithms, which would mean that we would pick the most important tasks the system is to perform, and proceed to implement by hand all the necessary information the system needs to perform them. Instead we must focus on developing principles for the system itself to *invent algorithms*. And we must go even further, for high levels of autonomy means that the system we target must

1 'Allonomy' is the opposite of autonomy; allonomic controllers may impart some level of autonomy to what they control while not being autonomous themselves.

constantly be learning, by training itself on appropriate tasks and subtasks, after it leaves the lab. The term "algorithm" may not be entirely appropriate for what our autonomous system is learning, because even on sequential repeats of the same task the system may be modifying how it does it (cf. Wang 2006), from the smallest to the largest subtask. High levels of autonomy mean high levels of domain independence, so we also cannot allow ourselves to provide the system mainly with domain-specific knowledge. In fact, in a constructivist approach the system development task becomes that of designing a meta-control scheme that, instead of providing hand-coded solutions to specified tasks and subtasks, must give the system enough flexibility and initiative to propose subgoals on its own, based on the drives (highest-level goals) provided by the system's designers.

2. NATURAL LANGUAGE ACQUISITION

Our approach to knowledge representation has its roots in non-axiomatic term logic. Since knowledge in our agent is established on the basis of its experience, truth cannot be absolute but is bound to only be established to a certain degree and within a certain time interval. In our approach the simplest term thus encodes an observation, and is called a fact (or a counter-fact indicating the absence of an observation). A fact carries a payload (the observed event), a likelihood value in [0, 1] indicating the degree to which the fact has been ascertained and a time interval in microseconds, the period within which the fact is believed to hold (or, in the case of a 'counter-fact', the period during which the payload has not been observed). Facts have a limited life span, corresponding to the upper bound of their time interval. Payloads are terms of various types, some of which are built in the AERA Executive, the most important of these being atomic state, composite state, prediction, goal, command, model, success/failure, and performance measurement. Additionally, any type can be defined by the programmer, and new types can be created by I/O devices at runtime.

Except when the agent is in initial stages of bootstrapping (which should only happen once for each new environment or domain), a lot of its knowledge will be composite, that is, relationships and combinations between small "atomic" knowledge "bricks". In the case of natural language, sentences are structured out of sequences of words, with fairly complex relationships and rules (generally called 'grammar'); words are constructed out of phonemes[2] (and letters, which have a rather complex relationship to phonemes).

To extract such knowledge from observing language-using humans in the real world the agent must have the ability to work with partially correct hypotheses about the "rules"[3] that guide the process of constructing a sentence with a particular meaning. To this end a language-learning agent would need to represent its experience as contextualized knowledge structures of some kind, with variable levels of complexity, which would allow it to change the relationships between the knowledge structures previously acquired in various ways at various levels of granularity. For instance, an incorrectly represented abstraction of how to pair nouns and verbs so that others understand what we mean might be erradicated when more examples of the various ways of its pairing are observed. The speed would be dependent on the efficiency of the agent's processes for this purpose, and this is our task here: To implement a system that can produce the necessary hypotheses for how "the world works" – in our case how natural language is used – and representing it in a way that allows modification in a way that moves the system towards increased accuracy. In this respect our work is compatible with e.g. that of Dominey & Boucher (2005), who demonstrated a robot learning language from

2 'Phoneme' is a construct hypothesized by humans; here it is used as a shorthand for the already-categorized sounds that can be used to convey meaning in a human natural language in a modular way. (Our agent is of course not bound to such human-hypothesized concepts, as it generates its own knowledge based on its own experience and capabilities.)

3 The effective ("correct") use of natural language might be formalizable as explicit rules, but natural language is primarily a vehicle for getting things done, and as such may not be so unlike any task with complex contextual dependencies and relationships between its atomic operands.

limited domain and language-specific knowledge; our work goes further by proposing general principles for extracting meaning from observation, as described below (see also Nivel et al. 2013).

With an aim of generality we wanted to find a representation amenable for representing all kinds of experience, that could be used for reasoning operations, and that could scale reasonably well by growing with cumulated experience – a homogenous representational scheme. Knowledge in our approach is composed of *states* (be they past, present, predicted, desired or hypothetical) and of executable code – called *models*. Models are capable of generating knowledge, for example predictions, hypotheses, and goals, and are executed by a virtual machine, the Executive. States and models are low-granularity, as low granularity in this respect better supports knowledge plasticity than high granularity due to the fact that modifications to small parts are less likely to have detrimental, unforeseen side-effects. Low granularity also means that higher levels of combinatorics are leveraged. Models are at a granularity level equivalent to SOAR's production rules (cf. Laird 2012), but while their surface structure has some similarity to these (e.g. having a right-hand side and a left-hand side and directly supporting reasoning), they differ significantly in many respects including supporting parallel execution, simultaneous forward and backward chaining, and having strong representation of time. Representing time is of course necessary for producing timed behavior; for natural language time must be manipulatable at several scales, from the a large-scale composite operation (e.g. achieving a mission such as doing a TV interview) to intermediate-size actions (e.g. what utterance will elicit a desired answer/information from an interlocutor) to the smallest levels of individual operations (e.g. producing a prediction), which is a necessary requirement of any system that must (a) perform in the real world and (b) model its own operation with regards to its expenditure of (temporal) resources. Considering time values as intervals allows us to encode the variable precision and accuracy needed to deal with the real world, for example, sensors do not always perform at fixed frame rates and so modeling their operation may be critical to ensure reliable operation of their controllers and models that depend on their input, and the precision for goals and predictions may vary considerably depending on both their time horizons and semantics. Last, since acquired knowledge can never be certain, one can assume that "truth" – asserting that a particular fact holds – can only be established for specific periods with varying degrees of temporal uncertainty.

3. AUTO-CATALYTIC ENDOGENOUS REFLECTIVE ARCHITECTURE

To be equipped to learn natural language in situ in human society, an artificial agent needs to be endowed with many complex cognitive functions, including the ability to direct its own attention to the right things at the right time (cf. Ognibene et al. 2013) and relate sounds to contextual actions and cues. As it bootstraps its language knowledge (from possibly meager beginnings) it needs to be capable of classifying events based on its own incomplete knowledge of the world at any point in time, in a way that it can easily update its knowledge experience is gained. In our approach all knowledge is represented in a way that relates it to the passage of time. The architecture has no special sub-components that manage learning, attention, planning, dialogue, and so on; instead, communicative learning, planning, and execution are emergent processes that result from the same set of low-level processes: These are essentially the execution of fine-grained programs – models – that are automatically generated, are reusable and shared system-wide, collectively implementing functions that span across the entire scope of the system's operation in its environment. For example, models generate both goals and predictions, some other programs monitor their success or failure and are thus able to reinforce the system's confidence about their effectiveness. Learning a skill consists of learning models and their context and sequence of execution, which in AERA result from the assessment of models' performance and the detection of novelty (which is how the triggering of new models is implemented), both of which are low-level processes in our system. Bad models are discarded and/or replaced by better ones. High-level

processes (planning, attention, learning) influence each other reciprocally: For example, learning better models and sequences thereof improves planning; having good plans means that a system will direct its attention to more (goal-)relevant states, and this means in turn that learning is more likely to be focused on changes that impact the system's mission (e.g. correct identification of novelty), which on average increases its chances of success. These high-level processes are dynamically coupled, via the low-level processes, as they both result from the execution of the models.

In our approach cognitive control results from the continual value-driven scheduling of *reasoning jobs*. In this approach high-level cognitive processes are grounded directly in the core operation of the machine resulting from two complementary control schemes. The first is top-down: Scheduling allocates resources by estimating the global value of the jobs at hand, and this judgment results directly from the products of cognition – goals and predictions. These are relevant and accurate to various extents, depending on the quality of the knowledge accumulated so far. As the latter improves over time, goals and predictions become more relevant and accurate, thus allowing the system to allocate its resources with a better judgment; the most important goals and the most useful/accurate predictions are considered first, the rest being saved for later processing or even discarded, thus saving resources. In that sense, cognition controls resource allocation. The second control scheme is bottom-up: Resource allocation controls cognition. Shall resources become scarce (which is virtually always the case in the system-environment-mission triples we target), scheduling narrows down the system's attention to the most important goals/predictions the system can handle, trading scope for efficiency and therefore survivability – the system will only pay attention to the most promising (value-wise) inputs and inference possibilities. If the resources become more abundant the system will start considering goals and predictions of less immediate value.

AERA is data- and event-driven, meaning that the execution of code is triggered by matching patterns with inputs. The bootstrap code – the initial resource for the system – contains (among other things) drives and top-level models and top-level goals (drives). ('Code' refers to models (which constitute executable knowledge) that have either been given (as part of the bootstrap code) or learned by the system.) A drive is an "innate" goal given by the programmer whose semantics can also be those of a constraint; it is a goal whose payload is a *fact* that cannot be observed directly – think for example of the drive "keep operating successfully": the environment does not produce explicit direct evidence of its achievement, but several indicators can be combined to infer it. A top-level model is hand-crafted for giving the system a way to entail the success (or failure) from an observable (such an observable could be "your owner gives you a reward"). As an AERA-based system is event-driven, drives and top-level models form together the system's motivation, providing a top-down impetus for the system's running, while sensors provide an influx of data, driving its operation bottom-up. A complete description of AERA can be found in Nivel et al. (2013).

4. EXPERIMENTS WITH NATURAL COMMUNICATION & LANGUAGE

The goal of the two experiments, E1 and E2, described here was to assess the ability of our first agent, S1 implemented in AERA, to learn the pragmatics, semantics, and syntax of human natural communication. We wanted an appropriately complex task that put a measure on S1's capability to autonomously disentangle a wide variety of causal relationships, sufficient to convince us about the generality of its knowledge acquisition and generalization capabilities. Human natural multimodal communication contains a wide variety of data types at two orders of magnitude of time. We defined a scenario that included considerable spatio-temporal and language behavior complexity: a dyadic mock-television interview. In the experimental setup two humans interact for some time, allowing S1 to observe their behavior and interaction; S1's task is to learn how to conduct the interaction in exactly the same way as the humans do, in either role of interviewer or interviewee. In E1 the interviewer asks the interviewee to pick up objects and move them to

new locations on the table between them (Table 1), the interviewee moves the objects as requested but does not speak – a kind of put-that-there with learning (Bolt 1980); in E2 the interviewer asks numerous questions about the recyclability of the objects on the table between them, the interviewee giving informed answers to these (see Table 2). In E2 both interviewer and interviewee use deictics of various kinds and some forms of body language (see Table 3).

The knowledge given to S1 is represented as a small set of primitive commands for its drivers (arm joints and speech output) and categories of sensory data (speech, prosody, and joints/geometry), along with a few top-level goals such as "pleasing the interviewer" (operationally defined as the interviewer saying "thank you" or asking a new question) and "getting the interviewee to speak" (operationally defined as production of speech). The full specification of the seed for the two experiments can be found in Nivel & Thórisson (2013).

S1 observes the real-time interaction between the two humans in a virtual equivalent of a video-conference: The humans are represented as avatars in a virtual environment – to allow the interaction to proceed naturally, without any artificial protocols, each human sees the other as a realtime avatar on their screen. Their head and arm movements are tracked with motion-sensing technology (with sub-centimeter, sub-second accuracy and lag-time), their speech recorded with microphones. Signals from the motion-tracking are used to update the state of their avatars in real-time, so that everything one human does is translated virtually instantly into movements of her graphical avatar on the other's screen. Between the avatars is a desk with objects on it, visible to both participants. This is the case in both the human-human condition and the human-agent conditions (agent taking either role). In both experiments we had S1 observe the humans until it accurately predicted all major event types observed in the dialogue (~2.5 minutes for E1 and ~20 hours for E2). We then had S1 interact with the humans for a sufficiently long period to produce videos (~10 minutes for E1, ~15 minutes for E2) that could be analyzed for t-patterns (Magnusson 2000); recordings of S1 interacting in either role with one of the humans (same as who participated in the human-human scenario) thus formed the basis for data analysis.

4.1 Experiment 1 (E1)

The objects that the interaction revolves around are: two *blue cubes*, one *red cube*, one *red sphere*, one *blue sphere*. The seed containing all initial (hand-coded) knowledge consisted of a set of primitive commands (move hand, grab, release, point at) and a set of dimensions for the input space (object type, color, actor's role, speech). The seed also includes initial knowledge that models the consequences of invoking the primitive commands: these models are for example explaining how the position of the system's hand is affected by invoking the command move hand and how a hand and an object are linked together after invoking the command grab. The natural language used in E1 consisted of a fixed set of sentence fragments (see Table 1). The seed for S1 in E1 is described in detail in Nivel & Thórisson (2013).

Table 1. The words and word order used in E1. The human participants were asked to "interact normally" to achieve their tasks (meaningless sentences – e.g. a sentence starting with "Take it ..." as a first sentence in an interaction, which had no prior referent for the ellipsis – were not observed). We did not provide our S1 agent with any grammar or words.

Words	**Word Order**
verbs: put, take *nouns:* sphere, cube *adjectives:* blue, red *adverb:* there *determiners:* a, the *pronoun:* it *conjunctions:* and, ... *interjection (ack):* thank you	*Utterance:* (Part1), Part2 *Part1:* take, [a | the] noun], (conj) *Part1:* take, [it | [a | the] noun], (conj) *Part2:* put, [it | [a | the] [blue | red] noun], there, ..., thank you *(Silence of some measurable length is indicated as "..."; parenthesis means that an element is optional.)*

Results show that the performance of S1 in E1 matches the human-human scenario very closely, and S1 only needed to observe the humans for around 2.5 minutes before its performance was error-free in either role. A subsequent inspection of S1's realtime performance for 10 minutes, in real-time interaction with humans under the same operating conditions as in the human-human scenario, revealed no mistakes, restarts, or self-corrections in the interaction on behalf of S1 – it performed flawlessly and completely error-free. The system acquired and generalized interaction skills to a sufficient level to allow it to perform 100% error-free communication of the same nature and complexity as that observed in the human-human interaction. S1 learned the sequences of orders ("take a blue cube..." then wait for the interviewee to comply before adding "...and put it there." and pointing with a finger to a location on the table), and it learned to do this with a series of different targets (e.g. a blue cube first, then a red sphere), as demonstrated by the human actors – the latter of which results from the hierarchization of control via model affordances. S1 identified the causal relationship between deictics and utterances (e.g. "there" correlated with pointing gestures) – this is an example of learned structural hierarchy in the form of composite states – as well as ellipsis ("put it there"). The pronoun "it" was learned to identify the object that draws the most attention (in terms of learned job priority), i.e. the target of the most valuable goals (picking an object is a learned pre-condition on the next step, moving it to some location, to earn the reward) – this is an example of value-driven resource allocation steering cognition (and vice-versa); it matches exactly how humans used ellipsis in the observed interactions.

4.3 Experiment 2 (E2)

Given the success of E1, in E2 we increased the complexity of its task as follows: The scenario included all communicative behavior of E1, with a considerably increase in both spatial and language complexity. In particular, the language component in E2 included much longer and more complex sentences, and both interviewee and interviewer generated speech. The vocabulary was 100 words; S1 was given no kind of grammar, nor a list of permissible words.[4] On the desk between the interviewer and interviewee were a set of (virtual) objects: *aluminum can*, *glass bottle*, *plastic bottle*, *cardboard box*, *newspaper* and *painted wooden cube*. As before, the task of the participants is to talk about these objects, in particular, the interviewer's task is to ask the interviewee about the materials which the various objects are made of, and the pros, cons, cost, and methods for recycling them. As in E1, the interviewee must understand the utterances of the interviewer to a sufficient degree to produce the desired actions,[5] in this case long explanations about the pros and cons o recycling various kinds of materials, using deictic references, ellipsis, and standard human dialogue and turntaking skills.

S1 observes two humans interacting; participants in all sessions were not trained actors. They interacted according to the targeted set of behaviors (see Table 2 and Table 3). The sequence of their actions and the use of multimodal deictics was free-form and real-time, the interaction semi-improvised. The human participants tried to not make mistakes, but occasional errors were unavoidable as all sessions were live and non-scripted, of several minutes each. As before, for the natural language no formal grammar definitions were produced or given to S1, nor a list of permissible sentences. The sessions proceeded as in E1; we had S1 observe until it could perform in either role without making any mistakes.

4 Due to the number of commission errors in the speech recognizer, however, its output was filtered by the set of 100 words.

5 While the humans in the experiments are not trained actors and their behavior is not stylized, their interaction was nevertheless correct in all major aspects – all question-answer pairs are correct and consistent. S1 thus does not have to deal with incorrect language, which would undoubtedly bring the observation time well above 20 hours.

Table 2. Some examples of the unscripted sentences produced by the E2 human participants in realtime dialogue.

Which releases more greenhouse gasses when produced, [an aluminum can or a glass bottle | an aluminum can or a plastic bottle | a plastic bottle or a glass bottle]?
What [else | more] can you [tell, tell me, tell us, say] about [this | that | it]?
There are many types of plastic.
Tell [me | us] about this [object | thing | one].
More energy is needed to recycle a plastic bottle than a can of aluminum.
Compared to recycled plastic, new plastic releases fifty percent more greenhouse gasses.
More energy is needed to recycle a glass bottle than a can of aluminum.
A glass bottle takes one million years to disintegrate completely in the sun.
Glass is made by melting together several minerals.
A recycled aluminum-can pollutes (only) five percent of what a new [can | one] pollutes.
Recycling an aluminum-can costs only five percent of a new one.
Compared to recycling, making new paper produces thirty-five percent more water pollution.
This is a cube made from unpainted wood.

The results of E2 are summarized in Table 3. In E2 S1 learned everything that it observed in the human-human interactions, and can perform an equivalent interview in either role of interviewer and interviewee.[6] The full socio-communicative repertoire exemplified in E1, with additional complexity in deictic gestures and grammar, acquired autonomously by S1 after an observation period of approximately 20 hours, has been correctly learned, with no mistakes in its subsequent application, including timing of all actions.

5. CONCLUSIONS & FUTURE WORK

We have demonstrated an implemented architecture that can learn autonomously many things in parallel, at multiple time scales. The results show that the AERA-based S1 agent can learn complex multi-dimensional tasks from observation from only a small ontology, a few drives (high-level goals), and a few initial domain models to support autonomous bootstrapping on a complex task. Human dialogue is an excellent example of the kinds of complex tasks current systems are incapable of handling autonomously, and to our knowledge no prior architecture has demonstrated comparable results (cf. Franklin et al. 2013, Laird 2012, Wang 2011). The fact that no difference of any importance can be seen in the performance between S1 and the humans in simulated face-to-face interview is an indication that the resulting architecture holds significant potential for further advances, and that our methodology (Nivel et al. 2013, Thórisson 2012) is a way for escaping the constraints of current computer science and engineering software methodologies when aiming for artificial general intelligence and increased systems autonomy. However, in its current incarnation AERA is entirely dependent on observation, as learning is exclusively triggered by unexpected goal achievement, or a prediction that turns out to be wrong – i.e. by surprise. This limits the acquisition of knowledge to phenomena that are directly observable – hidden causation is difficult for the current system to figure out, as are other kinds of inexplicit relations (similarity, equivalence, etc.). One of the main directions of our planned near-future work is set toward building more prototypes to assess the generality and scalability of our system. Elsewhere we have argued that curiosity results from the need to overcome the limitations imposed by the scarcity of inputs (Steunebrink et al. 2013); we plan to expand the types of programs to implement a richer

6 Videos of the interaction can be found on www.humanobs.org and on youtube.com on channel CADIAvideos.

set of inferences from which curious behaviors can be devised and planned, whenever the system has resources to spare.

Table 3. Summary of results obtained in Experiment 2 (E2). S1 has learned how to conduct an interview with a human, and can perform flawlessly in either role of interviewer and interviewee after around 20 hours of observation, producing grammatically, semantically, and pragmatically correct utterances in interactions spanning tens of seconds. Our S1 agent was not provided with any grammar or words.

Category	What Has Been Learned	Result
Interview gross structure	S1 has learned how to structure dialogue in an interview, as observed in the human-human interaction. S1 has learned roles of both interviewer and interviewee from observation, having been only provided with the top-level goals for either, and can perform them both. S1 also learned to use interruption to keep the interview within the allowed time limits.	S1 can conduct dialogue with a human efficiently and effectively, as interviewer and interviewee, in a way that is virtually identical to human-human interaction. Appropriate and correct actions taken, given the behavior of either role.
Turn-taking	S1 has learned the basic skills of turn-taking form observation, as plainly obvious in the videos, and clearly demarcated in turn-taking patterns shown by t-pattern analysis. In E2 the interview includes gestures and speech for both roles. Turn-taking is slightly slower-paced than typical human-human interaction.	S1 efficiently and effectively takes turns, asking questions at the right times (as interviewer) and answers timed correctly (as interviewee). The style and action repertoire is precisely that observed in the human-human condition.
Explicit manual deictics	S1 has learned to use three kinds of deictics: pointing by finger, by palm, and picking up and putting down an object in synchrony with speech. Successful resolution of a manual deictic gesture by the interviewer allows interviewee to produce correct answer to questions, and to use it reciprocally when replying.	Both the timing and form of the gestures is appropriate for the context. Resolution of a manual deictic gesture by the interviewer allows interviewee to place objects in the right location, and to pick out a referenced object out of the five.
Ellipsis	Use of pronoun "it" and "the [*X*]″ (e.g. "Take the cube" in the beginning of a new instruction) is correctly used to reference (as interviewer) / interpreted (as interviewee) an object mentioned earlier.	S1 has learned to use ellipsis in both sentence interpretation and generation. Successful resolution of ellipsis by S1 as interviewee allows it to place objects in the right location, and to pick out a referenced object out of the five.
Sentence construction	S1 constructs all sentences correctly Correct combination of dialogue events to allow correct uses of pronoun and adverb, supporting disambiguation/ indication of what should be done.	S1 can construct sentences in either role of interviewer and interviewee, based on those observed in the human-human interaction. The sequence of words is produced using generalized models acquired autonomously from observing the human interaction.
Constructing proper answer to questions	When the interviewer asked a question, not only were the gestures and speech interpreted for the correct response, the reply constructed was appropriate to the question.	Given the numerous valid questions that can be asked in E2, S1 replies with an appropriate and correct utterance.

ACKNOWLEDGEMENT

This work was supported by the European Project HUMANOBS – Humanoids that Learn Socio-Communicative Skills By Observation (FP7 STREP – Cognitive Robotics, Grant number 231453), by Nascence (FP7-ICT-317662), SNF grant #200020-138219, and by grants managed by Rannís, Iceland. We are grateful to H. H. Thorisson and G. S. Valgardsson for the design of the S1 avatar and Th. Bryndis Thorisdottir for the content of the interview in E2.

REFERENCES

Bolt, R. A. (1980). "Put-That-There": Voice and Gesture at the Graphics Interface. *Computer Graphics*, 14(3), 262-70.

Dominey, P. F. & J-D. Boucher (2005). Developmental Stages of Perception and Language Acquisition in a Perceptually Grounded Robot. *Cogn. Syst. Res.,* **6**(3):243-259.

Franklin, S., T. Madl, D'Mello, K. Sidney & J. Snaider (2013). LIDA: A Systems-level Architecture for Cognition, Emotion, and Learning. *Transactions on Autonomous Mental Development.*

Goertzel, B., C. Pennachin, S. Araujo, F. Silva (2013). A General Intelligence Oriented Architecture for Embodied Natural Language Processing.

Laird, J. E. (2012). *The Soar Cognitive Architecture.* MIT Press, Cambridge, MA.

Magnusson, M. S. (2000) Discovering hidden time Patterns in Behavior: T-patterns and their detection. Behavior *Research Methods, Instruments & Computers*, 32, 93-110.

Nivel, E. & K. R. Thórisson (2013). Seed Specification for AERA S1 in Experiments 1 & 2. Reykjavik University School of Computer Science Technical Report, RUTR-SCS13005.

Nivel, E., K. R. Thórisson, B. R. Steunebrink, H. Dindo, G. Pezzulo, M. Rodriguez, C. Hernandez, D. Ognibene, J. Schmidhuber, R. Sanz, H. P. Helgason, A. Chella & G. K. Jonsson. Bounded Recursive Self-Improvement. RU-SCS130006 Technical Report (ArXiV: 1312.6764).

Ognibene, D., E. Chinellato, Miguel Sarabia & Yiannis Demiris (2013). Contextual action recognition and target localisation with active allocation of attention on a humanoid robot. Bioinspiration & Biomimetics, 8(3), 035002. doi:10.1088/1748-3182/8/3/035002.

Steunebrink, B. R., J. Koutnik, K. R. Thórisson, E. Nivel & J. Schmidhuber (2013). Resource-Bounded Machines are Motivated to be Efficient, Effective, and Curious. In K-U Kühnberger, S. Rudolph and P. Wang (eds.), Proceedings of the Sixth Conference on Artificial General Intelligence (AGI-13), 119-129, Beijing, China.

Thórisson, K. R. (2012). A New Constructivist AI: From Manual Construction to Self-Constructive Systems. In P. Wang & B. Goertzel (eds.), *Theoretical Foundations of Artificial General Intelligence*, 4:145-171. Atlantis Thinking Machines.

Thórisson, K. R. (2013). Reductio ad Absurdum: On Oversimplification in Computer Science and its Pernicious Effect on Artificial Intelligence Research. In A. H. M. Abdel-Fattah &K.-U. Kühnberger (eds.), *Proceedings of the Workshop Formalizing Mechanisms for Artificial General Intelligence and Cognition* (Formal MAGiC), Beijing, China, July 31st, 31-35. Institute of Cognitive Science, Osnabrück.

Wang, P. (2006). *Rigid Flexibility: The Logic of Intelligence*. Springer, Dordrecht. 2006.

Wang, P. (2011). The assumptions on knowledge and resources in models of rationality. *International Journal of Machine Consciousness*, 3(1):193-218.