Published in IEEE TAMD Vol. 3-2, pp109-112, 2011 (Editorial for the special issue "Grounding Language in Action")

Grounding language in action

Katharina J. Rohlfing, *CITEC, Bielefeld University* and Jun Tani, *RIKEN Brain Science Institute*

THE topic of this Special Issue is that action and language are interwoven. Driven by traditional approaches that prevail in our education, we might be surprised about the connection between language and action, since we are inclined to view language as a symbolic system as it connects entities in the world with the corresponding conceptions that a perceiver has in mind. Action, on the other hand, was considered to be an event in the world that has to be perceived first. Only then, could it also be labeled by a perceiver, so a particular conception of it could be represented in the mind. The research from perspective on how cognition develops, however, contributed to findings suggesting that what we know about action, language and interaction emerge in parallel and have an impact on each other ([1], [2], [3]). This parallel development seems to provide a ground for further mental growth. For a system to develop, it requires different processes to interact not only with each other but also with the physical world. This coupling is also a valuable source of intelligence [2] as it provides knowledge that shapes the system's performance without actually being part of the system.

It has been expected that synthetic modeling studies combined with robotics experiments can contribute to the understanding of the dynamic development on the system level [4] [5] [3]. Robots may learn to understand meanings of words by associating their references to related experiences of sensory-motor patterns rather than to other words, as in a classical dictionary. For example, a *cup* might be understood by referencing to the visuo-proprioceptive flow associated with the action of grasping it. This type of robotics experiments have been conducted by various research groups utilizing different computational schemes.

More specifically, Roy and his colleagues developed a robotic manipulator that is able to translate spoken commands into situated actions [6]. By maintaining a dynamic mental model of its immediate physical environment, adjectives describing object properties are grounded in sensory expectations relative to specific actions. Verbs, in turn, are grounded in sensory-motor control programs. This model allows us to infer that words that appeared in the command sentences are grounded in the corresponding sensory-motor reality. In contrast to the situation in Roy's experiment, in which each single command sentence is corresponding to a specific action event, Dominey and his colleagues [7] as well as Iwahashi and his colleagues [8] conducted experiments on introducing continuous interactions between robots and users in a cooperative tutoring tasks while having dialog between them. Such studies are well motivated by research concerned with social learning [9]. Weng and his colleagues [10], [11] conducted experiments on task transfer to robots in which the trainer shaped the behaviors of the agent interactively, continuously, and incrementally through verbal commands. The experiment results showed that the robot could learn new and more complex sensory-motor tasks by transferring sensory-motor skills learned in earlier periods of open-ended development. This result may correspond to the scaffolding strategies which have been known to be essential for the cognitive development process in humans.

The aforementioned studies have illustrated well how to ground symbolically represented linguistic structures into sensory-motor related analog patterns by utilizing various computational schemes. They, however, do not explain how language and sensory-motor competencies can co-develop through their mutual interactions by self-adapting sub-symbolic level activities in the systems. Regarding this issue, the connectionist approaches [12] of utilizing various neural network models with emphasis on self-organization in sub-symbolic level activities have already shown promising results in the cognitive neuroscience contexts: Two decades ago, Elman [13] provided evidence suggesting that simple recurrent neural network (RNN) models can learn to extract syntactic knowledge from exemplar sentences. In this study, it was shown that syntactic structures can be acquired with generalization by self-organizing their distributed representation on the sub-symbolic level. Moreover, Miikkulainen [14] showed that RNNs can also learn to extract semantic structures from exemplars.

By following these results, there have been some attempts ([15], [16]) to integrate linguistic competency with sensory-motor competency by utilizing connectionist models including RNN. Cangelossi and Rita [15] showed that a simple feed forward network can learn mapping of command word inputs to corresponding motor outputs. Furthermore, it was shown that new action concepts can be created and transferred by combining prior-learned words in novel ways. Sugita and

K. J. Rohlfing, is with Emergentist Semantics Group of Bielefeld University, Center of Excellence Cognitive Interaction Technology (CITEC), Universitätsstr. 21—23, 33615 Bielefeld Germany (e-mail: rohlfing@techfak.uni-bielefeld.de).

J. Tani is with the Laboratory for Behavior and Dynamic Cognition at the RIKEN Brain Science Institute, 2-1 Hirosawa, Wako-shi, Saitama, 351-0198 Japan phone: 048-462-111 ext. 7411; fax: 048-467-7248; (e-mail: tani@brain.riken.jp).

Tani [16] proposed a connectionist model that consists of a linguistic RNN and a sensory-motor RNN which are connected through some neural populations. They interact with each other in the course of learning command sentences consisting of verbs and object nouns, as like "push object-X" or "touch object-Y", and associated action programs in terms of sensory-motor patterns. Employed in a robot learning experiment, the model showed that action concepts can be acquired using a compositional representation for possible combinations of verbs and object nouns. It was also shown that a certain level of generalization is achieved by which the network can recognize even unlearned combinations of verbs and object nouns. It is considered that such generalization is due to the compositional structure which has been self-organized in the distributed activities of neural units.

These research results by the connectionist approach suggest that linguistic concepts may not be merely grounded in sensory-motor competency. Instead, linguistic concepts may co-develop with sensory-motor programs by having dense interactions between them. The action concepts might be self-organized structurally on sub-symbolic levels by utilizing learning signals originated from both linguistic and behavioral modalities, rather than being allocated in discrete symbols. This argument is related to motor theories like the mirror neuron theory of action understanding ([17], [18], [19]). It has been recently proposed that conceptual knowledge is grounded in sensory-motor systems [20], i.e. "when a person hears or reads text involving action, there is activation of the motor system in his or her brain, which corresponds to the referential semantic content of the description" [21, p. 46].

How the mutual bootstrapping between language and action competency can be used for mental development is shown in developmental studies, in which language has been recognized as playing an influential role in establishing concepts about objects or events. More specifically, labels or words for objects were found to highlight the commonalities between objects [22] and situations [23], facilitate object categorization ([24], [25]), have the power to override the perceptual categories of objects [26] and provide additional information, on which basis infants perceive regularities and orders in actions and events ([27], [28]) and reason about physical events [29]. According to Gelman [30], via labels, we transfer expert knowledge to novices. Labels can be, therefore, considered as just another feature of objects ([25], [30, p. 128]).

In summary, in developmental cognitive psychology, it is acknowledged that language is a powerful social signal. However, further discussions are needed to develop a clearer picture about how this signal is integrated into a cognitive architecture: Is language an augmenter, spotlight or an inducer (to mention only few possible functions) [31]? Here, we need more insights from the cognitive modeling to prove and improve our understanding.

In the current trends in the synthetic modeling approach on the problem of grounding language in action, there seems to be a substantial gap between the connectionist approach and the computational approach. The connectionist approach – which focuses on developmental aspects in sub-symbolic levels – is still far from scaling to the human level cognitive competency. On the other hand, in various computational approaches, although we often witness quite sophisticated demonstrations of cognitive capability of hearing, speaking and acting by robots, which seems to be akin to human, it is still debatable that how much the employed computational programs inside represent the principles of real human cognitive mechanisms. The synthetic modeling for grounding language in action could be ultimately difficult research topics, because it requires the most sophisticated integrations or co-developments among various cognitive mechanisms, not limited to language and action, but also including attention, theory of minds and emotional control. Therefore, the future research would be extremely challenging as well as attractive because it is a trial of uncovering the most mysterious part in human cognition. Such challenge has been initiated recently in various projects including an European Community project, ITALK [3]. We look forward to hearing the first step results from such enterprises.

In this Special Issue, we aimed at extending our insights into how language is grounded in action by a series of papers from different disciplines concerned with development and learning.

In the first article, **Iris Nomikou and Katharina Rohlfing** provide a naturalistic approach to explore the ecology of very early mother-child interactions and exemplify in this way that cross-modal information is present from very early on and might provide the basis for later signal semantics. More specifically, the article about "Language *does* something", describes interactions between 3-month-old infants and their mothers during an everyday activity. It is shown that making use of multimodal sources seems to be a common practice in such naturalistic interactions as German mothers vocalize in a tight relationship with their actions which effect in language being perceivable and tangible to the infants. In addition, these findings provide a complex picture on intermodal learning, in which the phenomenon of synchrony appears as a dynamic one: Once semantics of the action is taken into account, it manifests itself in a variety of different types.

Making use of multimodal sources seems to be not only a common practice in interactions with infants but is shown to be essential for learning, as argued in "Temporal, environmental, and social constraints of word-referent learning in young infants: A NeuroRobotic model of multimodal habituation" by **Richard Veale, Paul Schermerhorn and Matthias Scheutz**. Based on empirical and neuroscientific evidences about the abilities of young infants, the authors model learning of a word-referent association. Implemented on a robot, the effect of temporal synchrony is investigated by exposing the system to different synchronous and asynchronous conditions. Furthermore, without making assumptions about the (pre-)existence of auditory or visual categories, it is tested

whether the system is capable of habituating to multimodal stimuli. The results demonstrate that the robot's responses to manipulations of the relative timing of the presentation of auditory and visual stimuli are consistent with those of young human infants. Furthermore, the results demonstrate that synchrony and temporal contiguity are necessary for the looking-time biasing effects of habituation to occur.

The relationship between action and language can shape attentional processes. In "Emergence of declaratives in artificial communicating systems", **Uno Ryoko, Davide Marocco, Stefano Nolfi and Takashi Ikegami** examined how behavior coordination of multiple robot agents controlled by simple neural network models can be evolved by using abstract communication signals. Their experiment results suggest that two types of joint attention (JA) emerge through evolution. In the so-called instrumental JA, attention is used as a tool to achieve action goals. On the other hand in the so-called participatory JA, the attention is used to establish JA itself. The study provides unique observations on how proto-linguistic communication of JA – which is prerequisite of language – can be developed in sub-symbolic level in the course of multiple agents interactions.

Grounding can be viewed not only in terms of a temporal relationship between language and action. In this phenomenon of grounding, the semantics of one of the signals can also be considered. How exactly the semantics of the verbal behavior is synchronized with the performance of a particular action and how it can be used as an element of social learning, is shown in the article by **Meredith Meyer, Bridgette Hard, Rebecca Brand, Molly McGarvey, and Dare Baldwin** about "Acoustic Packaging: Maternal Speech and Action Synchrony". In this semantically-based approach, descriptions of ongoing performed actions were found to be more synchronous with the actions themselves than other types of utterances. This work extends the current findings on Acoustic Packaging and proposes that rather than being simply aligned with performed action, utterances directly related to actions are paired selectively with the movement units in an interaction with a child, and their prosody conveys action-relatedness to adult speakers.

As these papers show, an important aspect of action and language is that they expand temporarily. But what then is time? This question is addressed in "Are we there yet? Grounding temporal concepts in shared journeys" by **Ruth Schulz, Gordon, Wyeth and Janet Wiles**. Even though one might think that time – similar to space – are both foundations of any intelligent system, the authors show that time and space are not directly perceived and need to be constructed from sequences of perceptions. In this paper, it is tested whether cognitive maps constructed by individual agents from their own journey experiences can be applied to learning temporal concepts and an associated lexicon, on which basis an agent can answer the question of "How long" did it take to complete a journey. Using evolutionary language games for specific and generic journeys, the authors established a successful communication based on representations of time, distance and amount of change (motion). Their results show that even though spatial and temporal terms are not identical, they can be learned using similar language evolution methods.

In thinking about grounding, we should consider not only the fundamental temporal relationship between language and action but also characteristics that allow a system to form complex structures. In the paper "An experiment on behaviour generalization and the emergence of linguistic compositionality in evolving robots", **Elio Tuci, Tomassino Ferrauto, Arne Zeschel, Gianluca Massera and Stafano Nolfi** investigated how behavioral and linguistic compositionality can be developed, sharing similar motivations with the aforementioned study by Sugita and Tani [14]. The authors applied genetic algorithm to evolve their proposed neural network models controlling simulated robots instead of applying supervised training as Sugita and Tani did. Their simulation experiments showed that compositionality of combining 3 verbs and 3 object nouns can be developed. It was also shown that generalization to recognize unlearned word combinations and to generate their corresponding actions is achieved in some evolutionary run cases. It is interesting to see that compositionality with generalization can be achieved by means of exploratory-based adaptation applied to simple neural network models.

REFERENCES

[1] L. B. Smith, and E. Thelen, „Development as a dynamic system," *Trends Cogn. Sci.*, no. 7, 343–348, 2003

[2] L. B. Smith "Cognition as a dynamic system: Principles from embodiment," *Developmental Review*, no. 25, pp. 278-298.

[3] A. Cangelosi, G. Metta, G. Sagerer, S. Nolfi, C.L. Nehaniv, K. Fischer, J. Tani, B. Belpaeme, G. Sandini, L. Fadiga, B. Wrede, K. Rohlfing, E. Tuci, K. Dautenhahn, J. Saunders, and A. Zeschel, "Integration of action and language knowledge: A roadmap for developmental robotics," *IEEE Trans. Autonom. Mental Develop.*, vol. 2, no. 3, pp. 167-195, 2010.

[4] L. Steels and P. Vogt, "Grounding adaptive language games in robotic agents," in *Proc. 4th Eur. Conf. Artif. Life*, I. Harvey and P. Husbands, Eds. Cambridge, MA: MIT Press, 1997, pp. 474–482.

[5] D. Roy, "Grounding words in perception and action: Insights from computational models," *Trends Cogn. Sci.*, vol. 9, no. 8, pp. 389–396, 2005a.

[6] D. Roy, K.Y. Hsiao, and N. Mavridis, "Mental imagery for a conversational robot," *IEEE Trans. Syst. Man Cybern., B Cybern.*, vol. 34, pp. 1374–1383, 2004.

[7] P.F. Dominey, A. Mallet, and E. Yoshida, "Real-time spoken-language programming for cooperative interaction with a humanoid apprentice," *Int. J. Humanoid Robot.*, vol. 6, no. 2, pp. 147–171, 2009.

[8] N. Iwahashi, "Interactive learning of spoken words and their meanings through an audio-visual interface," *Trans. IEICE*, vol. E91-D, no. 2, pp. 312-321, 2008.

[9] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H.Y. Moll, "Understanding and sharing intentions: The origins of cultural cognition," *Behav. Brain Sci.*, vol. 28, pp. 675–735, 2005.

[10] Y. Zhang and J. Weng, "Task transfer by a developmental robot," *IEEE Trans. Evol. Comput.*, vol. 11, no. 2, pp. 226–248, 2007.

[11] J. Weng, "Developmental robotics: Theory and experiments," *Int. J. Humanoid Robot.*, vol. 1, no. 2, pp. 199-236, 2004.

[12] D. Rumelhart, G. Hinton, and R. Williams, "Learning internal representations by error propagation," In *Parallel distributed processing*, D. Rumelhart and J. Mclelland, Eds. Cambridge, MA: MIT Press, 1986, pp. 318–362.

[13] J. L. Elman, "Finding structure in time," *Cogn. Sci.*, 14: 179–211,1990.
[14] R. Miikkulainen, "*Subsymbolic Natural Language Processing: An Integrated Model of Scripts Lexicon, and Memory,"* Cambridge, MA: MIT Press, 1993.
[15] A. Cangelosi and T. Riga, "An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots," *Cogn. Sci.*, vol. 30, no. 4, pp. 673-689, 2006.
[16] 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.
[17] G. Rizzolatti, L. Fogassi, and V. Gallese, „Neurophysiological mechanisms underlying the understanding and imitation of action," *Nature Reviews Neurocience*, no. 2, pp. 661-670, 2001.
[18] V. Gallese, C. Kevsers, and G. Rizzolatti, "A unifying view of the basis of social cognition," *Trends Cogn. Sci.*, no. 9, pp. 296-403, 2004.
[19] M. Tettamanti, G. Buccino, M. C. Saccuman, V. Gallese, M. Danna, P. Scifo, F. Fazio, G. Rizzolatti, S. F. Cappa, and D. Perani „Listening to action-related sentences activates fronto-parietal motor circuits," *Journal of Cognitive Neuroscience*, no. 17, vol. 2, pp. 273–281, 2005.
[20] D. Kemmerer, J. Gonzalez Castillo, T. Talavage, S. Patterson, and C. Wiley, "Neuroanatomical distribution of five semantic components of verbs: Evidence from fMRI," *Brain & Language*, no. 107, pp. 16-43, 2008.
[21] Taylor & Zwaan, "Action in cognition: The case of language," *Language and Cognition*, no. 1, vol. 1, pp. 45-58, 2009
[22] S. R. Waxman, and S. A. Gelman, "Early word-learning entails reference, not merely associations," *Trends Cogn. Sci.*, vol. 13, no. 6, pp. 258-263, 2009.
[23] S. Choi, "Early development of verb structures and caregiver input in Korean: Two case studies," *International Journal of Bilingualism*, no. 3, pp. 241-265, 1999.
[24] M. T. Balaban & S. R. Waxman, "Do words facilitate object categorization in 9-month-old infants?," *Journal of Experimental Child Psychology*, no. 64, pp. 3-64, 1997.
[25] F. Xu, "The role of language in acquiring object kind concepts in infancy," *Cognition*, no. 85, pp. 223-250, 2002.
[26] K. Plunkett, J.-F. Hu, and L. B. Cohen, "Labels can override perceptual categories in early infancy," *Cognition*, no. 106, pp. 665-681, 2008.
[27] R. J. Brand and D. A. Baldwin, "Motionese and Motherese: Two avenues for supporting infant action processing, " presented at the X. International Congress for Studies in Child Language, Berlin, Germany, 2005, Paper
[28] R. J. Brand and S. Tapscott, "Acoustic packaging of action sequences by infants," *Infancy*, vol. 11, no. 3, pp. 321–332, 2007.
[29] Y. Gertner, R. Baillargeon, C. Fisher and J. D. Simons, "Language Facilitates Infants' Physical Reasoning", presented at the Biennial Meeting of the Society of Research in Child Development, Denver, USA, 2009, Paper.
[30] S. A. Gelman, "Learning from others: Children's construction of concepts," *Annual Review of Psychology*, no. 60, pp. 115-140, 2009.
[31] P. Wolff and K. J. Holmes, "Linguistic relativity," Advanced Review in Cognitive Science, 2010.

Katharina J. Rohlfing received her Master's degree in German language, Media Studies, and Philosophy from the University of Paderborn in 1997. As a member of the Graduate Program "Task-oriented communication", she received her PhD from Bielefeld University in Linguistics in 2002. In 2006, she became a Dilthey Fellow (Funding initiative "Focus on the Humanities") with the project on the "Symbiosis of Language and Action" supported by the Volkswagen Foundation. She is head of the Emergentist Semantics Group within the Center of Excellence Cognitive Interaction Technology (CITEC) of Bielefeld University. She is interested in the development of meaning and the interface between cognitive development and language acquisition. Her research contributions include studies on multimodal parent-child interactions and situational factors contributing to language understanding in children.

Jun Tani received the B.S. degree in mechanical engineering from Waseda University, dual M.S. degree in electrical engineering and mechanical engineering from University of Michigan and Dr. Eng. from Sophia University. He started his research career in Sony Computer Science Laboratory in 1990. He has been appointed as a team leader in Laboratory for Behavior and Dynamic Cognition, Brain Science Institute, RIKEN in Tokyo since 2000. He was also appointed as a visiting associate professor in University of Tokyo from 1997 to 2002. He is interested in neuroscience, psychology, phenomenology, complex adaptive systems and robotics.