frontiers
RESEARCH TOPICS

ACTION AND LANGUAGE INTEGRATION IN COGNITIVE SYSTEMS

Hosted by
Angelo Cangelosi

frontiers in
NEUROROBOTICS

FRONTIERS COPYRIGHT STATEMENT

Cover image provided by Ibbl sarl, Lausanne CH

ISSN 1664-8714
ISBN 978-2-88919-038-6
DOI 10.3389/978-2-88919-038-6

ABOUT FRONTIERS

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

FRONTIERS JOURNAL SERIES

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing.

All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

DEDICATION TO QUALITY

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.

By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

WHAT ARE FRONTIERS RESEARCH TOPICS?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area!

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

ACTION AND LANGUAGE INTEGRATION IN COGNITIVE SYSTEMS

Hosted By
Angelo Cangelosi, University of Plymouth, United Kingdom

Recent theoretical and experimental research on action and language processing in humans and animals clearly demonstrates the strict interaction and co-dependence between language and action. This has been demonstrated in neuroscientific investigations (e.g. Cappa&Perani, 2003; Pulvermuller 2003; Rizzolatti&Arbib, 1998), psychology experiments (e.g. Glenberg&Kaschak, 2002; Pecher&Zwaan 2005), evolutionary psychology (e.g. Corballis 2002) and computational modelling (e.g. Cangelosi&Parisi 2004; Massera et al. 2008). All these studies have important implication both for the understanding of the action basis of cognition in natural and artificial cognitive systems, as well as for the design of cognitive and communicative capabilities in robots (Cangelosi et al. 2005).

The journal "Frontiers in Neurorobotics" is seeking submissions of new articles in the topic of action and language integration both in natural cognitive systems (e.g. humans and animals) and in artificial cognitive agents (robots and simulated agents). Manuscripts can regard new theoretical and computational investigations, as well as new neuroscientific and psychological investigations. Review articles in this topic are also welcome.

Table of Contents

frontiers in
NEURОROBOTICS

EDITORIAL
published: 01 May 2012
doi: 10.3389/fnbot.2012.00002

Editorial of e-book on action and language integration

Angelo Cangelosi*

Centre for Robotics and Neural Systems, Plymouth University, Plymouth, UK
**Correspondence: a.cangelosi@plymouth.ac.uk*

Increasing theoretical and experimental research on action and language processing in humans and animals clearly demonstrates the strict interaction and co-dependence between language and action. This has been extensively demonstrated in neuroscientific investigations (e.g., Rizzolatti and Arbib, 1998; Cappa and Perani, 2003; Pulvermuller, 2003), psychology experiments (e.g., Glenberg and Kaschak, 2002; Pecher and Zwaan, 2005; Barsalou, 2008), evolutionary psychology (e.g., Corballis, 2002), and computational modeling (e.g., Cangelosi and Parisi, 2004; Massera et al., 2007; Cangelosi, 2010). All these studies have important implication both for the understanding of the action basis of cognition in natural and artificial cognitive systems, as well as for the design of cognitive and communicative capabilities in robots (Cangelosi et al., 2010).

The journal "Frontiers in Neurorobotics" published a collection of articles on the topic of action and language integration both in natural cognitive systems (e.g., humans and animals) and in artificial cognitive agents (robots and simulated agents). These articles are now collected in an e-book, for wider dissemination. This set of chapters provides an up to date overview of current advances in the grounding of language into sensorimotor knowledge. The first chapters primarily focus on experimental evidence from cognitive psychology (Symes et al., 2010), cognitive neuroscience studies (Borghi et al., 2010), and comparative experimental/simulation studies (Greco and Caneva, 2010). Two chapters then use neural network simulation for motor chains for sentence processing (Chersi et al., 2010) and a computational model of gaze planning in word recognition and reading (Ferro et al., 2010). Finally, four chapters use cognitive systems and robotics methodologies to investigate general principles of action–language grounding (Parisi, 2010), teleological representations of action and language for human–robot interaction experiments (Lallee et al., 2010), verbal and non-verbal communication in neurorobotics models (Bicho et al., 2010), and action bases of action words (Marocco et al., 2010).

EXPERIMENTAL STUDIES

Borghi et al. (2010) focus on language comprehension as an embodied simulation of actions. This hypothesis is supported by embodied and grounded cognition theories (Barsalou, 2008; Pezzulo et al., 2011) and the neural underpinnings in neural substrates involve canonical and mirror neurons (Rizzolatti et al., 1996). Borghi et al. review their recent behavioral and kinematic studies to characterize, and evidence, the relationship between language and the motor system. This review leads to three consistent findings: (i) the simulation evoked during sentence comprehension is fine-grained, and shows sensitivity to the different effectors used to perform actions; (ii) linguistic comprehension also relies on the representation of actions in terms of goals and of the chains of motor acts necessary to accomplish them; and (iii) the goals are modulated by both the object features the sentence refers to, as well as by social aspects such as the characteristics of the agents implied by sentences. The authors also explicitly discuss the implications of these studies for embodied robotics.

Symes and colleagues present a cognitive psychology study on the integrating action and language through biased competition. This is based on previous psychological investigations that have demonstrated that planning an action biases visual processing, as in Symes et al.'s (2008) findings reporting faster target detection for a changing object amongst several non-changing objects. This new experimental study investigates how this effect might compare to, and indeed integrate with, effects of language cues. Using the same change-detection scenes as in Symes et al. (2008), two effective sources of bias are identified: (i) action primes, and (ii) language cues. For example, a sentence as "Start looking for a change in the larger objects" cues object size, and these successfully enhanced detection of size-congruent targets. Additional experiments explore the biases' co-occurrence within the same task, such as action prime (participants plan a power or precision grasp) *and* a language (a sentence) cue preceding stimulus presentation. Experimental results support the authors' predictions from the biased competition model by Desimone and Duncan (1995), in particular reliably stronger effects of language, and concurrent biasing effects that were mutually suppressive and additive.

Greco and Caneva (2010) focus on compositional symbol grounding for motor patterns. They propose a new comparative experimental/simulative paradigm to study the learning of compositional grounded representations for motor patterns. In a psychology experiment, participants learn to associate non-sense arm motor patterns, performed in three different hand postures, with non-sense words. Two experimental conditions are carried out: (i) in the *compositional* condition, each pattern was associated with a two-word (verb–adverb) sentence; (ii) in the *holistic* condition, each pattern was associated with a unique word. Experimental results show that the compositional group achieved better results in naming motor patterns, especially for patterns where hand postures discrimination was relevant. In order to ascertain the differential effects of memory load and of systematic grounding, neural network simulations were also carried out. After a basic simulation reproducing the default participants' performance, in some simulations the number of stimuli (motor patterns and words) was increased and the systematic association between words and patterns was disrupted, while keeping the same number of words and compositionality. Simulation results show that in both conditions the advantage for the compositional condition significantly increased. This indicates that the advantage for the compositional condition may be related to systematicity rather than to mere informational gain. Overall, both experimental and simulation data support the hypothesis of a shared action/language compositional motor representation.

NEURAL NETWORK STUDIES

Chersi et al. (2010) investigate the relationship of language to motor chains for sentence processing. As in Borghi et al. (2010), they also start from embodied theories of language grounding in the sensorimotor system, and language understanding as a process based on a mental simulation process (Jeannerod, 2007; Gallese, 2008; Barsalou, 2009). This hypothesizes that during action words and sentence comprehension the same perception, action, and emotion mechanisms implied during interaction with objects are recruited. Their aim is to identify the precise dynamics underlying the relation between language and action, e.g., to disentangle experimental evidence reporting both either facilitation or interference effects between language processing and action execution. This chapter presents a new neural network reproducing experimental data on the influence of action-related sentence processing on the execution of motor sequences. Chersi et al.'s modeling framework is based on three main principles: (i) the processing of action-related sentences causes the resonance of motor and mirror neurons encoding the corresponding actions; (ii) a varying degree of crosstalk exists between neuronal populations depending on whether they encode the same motor act, the same effector, or the same action-goal; (iii) neuronal populations' internal dynamics, which results from the combination of multiple processes taking place at different time scales, can facilitate or interfere with successive activations of the same or of partially overlapping pools. Interactions between sensory and motor modalities are modeled as a crosstalk between neuronal pools in motor and mirror chains. Results show also that the neural dynamics governing the activation of the pools can qualitatively reproduce the timings observed in behavioral experiments.

Ferro et al. (2010) propose a computational model of gaze planning in word recognition And the theory that reading is an active sensing process. Their computational model of gaze planning during reading consists of two main components: (i) a *lexical representation network*, acquiring lexical representations from input texts from the Italian CHILDES database; (ii) a *gaze planner* capable to recognize written words by mapping strings of characters onto lexical representations. Thus the model implements an active sensing strategy that selects which characters of the input string are to be fixated, depending on the predictions dynamically made by the lexical representation network. The analyses investigate the developmental trajectory of the system in performing the word recognition task as a function of both increasing lexical competence, and correspondingly increasing lexical prediction ability.

NEUROROBOTICS STUDIES

Parisi (2010) discusses a general neural modeling approach to language grounding in robots, consistent with the same literature on embodiment and grounding theories. The paper proposes a neural model of language according to which the robot's behavior is controlled by a neural network composed of two sub-networks: (i) the network controlling non-linguistic interaction between the robot and its environment; and (ii) a network for the processing of linguistic comprehension and production. Parisi reviews results of a number of computational simulations and suggests that the model can be extended to account for variety of language-related phenomena such as disambiguation, the metaphorical use of words, the pervasive idiomaticity of multi-word expressions, and mental life as talking to oneself. This modeling approach implies a view of the meaning of words and multi-word expressions as a temporal process that takes place in the entire brain and has no clearly defined boundaries. This can be further extended to emotional words, considering that an embodied view of language should consider not only the interactions of the robot's brain with the external environment, but also the interactions of the brain with what is inside the body such as motivational and emotional processes.

Lallee et al. (2010) link embodied and teleological representations of action and language for humanoid robotic experiments with the iCub platform. In this chapter the authors extend their framework for embodied language and action comprehension to include a teleological representation of goal-based reasoning for novel actions. Both from a theoretical perspective, and via human–robot interaction experiments with the iCub robot, they demonstrate the advantages of this hybrid, embodied–teleological approach to action–language interaction. Lallee et al. first demonstrate how embodied language comprehension allows the system to develop a set of representations for processing goal-directed actions such as "take," "cover," and "give." A crucial component of the new approach is the representation of the subcomponents of these actions, which includes state–action–state (SAS) relations between initial enabling states, and final resulting states for these actions. Robotic experiments demonstrate how grammatical categories including causal connectives (e.g., because, if–then) can allow spoken language to enrich the learned set of SAS representations. The study also examines how this enriched SAS repertoire enhances the iCub's ability to represent perceived actions in which the environment inhibits goal achievement.

Bicho et al. (2010) employ a dynamic neural field architecture for human–robot interaction and the integration of verbal and non-verbal communication. Specifically they investigate how a group of people coordinate their intentions, goals, and motor behaviors whilts performing joint action tasks. Their model is inspired by experimental evidence about the resonance processes in the observer's motor system, and their involvement in our ability to understand actions of others and to infer their. Bicho et al. develop a control architecture for human–robot collaboration that exploits perception–action linkage as a means to achieve more natural and efficient communication grounded in sensorimotor experiences. The architecture consists of a coupled system of dynamic neural fields. These represent a distributed network of neural populations that encode in their activation patterns goals, actions, and shared task knowledge. Human–robot experiments consist of verbal and non-verbal communication for a joint assembly task in which the human–robot pair has to construct toy objects from their components. This dynamic neural field architecture sustain the robot's capacity to anticipate the user's needs and goals and to detect and communicate unexpected events that may occur during joint task execution.

Marocco et al. (2010) presents new experiments with a simulated model of the humanoid robot iCub (Tikhanoff et al., 2011) to investigate the embodied representation of action words. The simulated iCub robot is trained to learn the meaning of action words (i.e., words that represent dynamical events that happen in time) such as "push," "hit." The words are learned by physically interacting with the environment and linking the robot's effects of its own actions (proprioception) with the behavior observed on the objects, before and after the action. The control system of the robot is an artificial neural network trained to manipulate an object through a Back-Propagation-Through-Time algorithm. Results show that the robot is able to extract the sensorimotor contingency of a particular interaction with an object and to reproduce its dynamics by acting on the environment. Moreover, in the absence of linguistic input, the robot is capable of associating a certain temporal sensorimotor dynamics to the learnt action words.

CONCLUSION

The collection of chapters in this volume provides a variety of methodological approaches to the experimental investigation and the neural network and cognitive robotic modeling of action and language integration. The studies address different phenomena linked to language grounding, such as sentence processing and comprehension, reading and word recognition, action word learning, compositionality of action and language representations, and language acquisition through interaction with the environment. All studies offer further support the existing evidence and theoretical stances of the grounding of language in action and perception, and the contribution of embodied cognition and mental simulation in language processing. Moreover, the multi-methodological contributions proposed in the volume and the close link between experimental data and computational and robotic modeling allows the fine investigation of behavioral, cognitive, and embodiment factors in the grounding of language in sensorimotor knowledge.

REFERENCES

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.* 59, 617–645.

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 1281–1289.

Bicho, E., Louro, L., and Erlhagen, W. (2010). Integrating verbal and nonverbal communication in a dynamic neural field architecture for human–robot interaction. *Front. Neurorobot.* 4:5. doi: 10.3389/fnbot.2010.00005

Borghi, A. M., Gianelli, C., and Scorolli, C. (2010). Sentence comprehension: effectors and goals, self and others. An overview of experiments and implications for robotics. *Front. Neurorobot.* 4:3. doi: 10.3389/fnbot.2010.00003

Cangelosi, A. (2010). Grounding language in action and perception: from cognitive agents to humanoid robots. *Phys. Life Rev.* 7, 139–151.

Cangelosi, A., Metta, G., Sagerer, G., Nolfi, S., Nehaniv, C. L., Fischer, K., Tani, J., Belpaeme, B., Sandini, G., Fadiga, L., Wrede, B., Rohlfing, K., Tuci, E., Dautenhahn, K., Saunders, J., and Zeschel, A. (2010). Integration of action and language knowledge: a roadmap for developmental robotics. *IEEE Trans. Auton. Ment. Dev.* 2, 167–195.

Cangelosi, A., and Parisi, D. (2004). The processing of verbs and nouns in neural networks: insights from synthetic brain imaging. *Brain Lang.* 89, 401–408.

Cappa, S. F., and Perani, D. (2003). The neural correlates of noun and verb processing. *J. Neurolinguistics* 16, 183–189.

Chersi, F., Thill, S., Ziemke, T., and Borghi, A. M. (2010). Sentence processing: linking language to motor chains. *Front. Neurorobot.* 4:4. doi: 10.3389/fnbot.2010.00004

Corballis, M. C. (2002). *From Hand to Mouth: The Origins of Language*. Princeton, NW: Princeton University Press.

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222.

Ferro, M., Ognibene, D., Pezzulo, G., and Pirrelli, V. (2010). Reading as active sensing: a computational model of gaze planning in word recognition. *Front. Neurorobot.* 4:6. doi: 10.3389/fnbot.2010.00006

Gallese, V. (2008). Mirror neurons and the social nature of language: the neural exploitation hypothesis. *Soc. Neurosci.* 3, 317–333.

Glenberg, A., and Kaschak, K. (2002). Grounding language in action. *Psychon. Bull. Rev.* 9, 558–565.

Greco, A., and Caneva, C. (2010). Compositional symbol grounding for motor patterns. *Front. Neurorobot.* 4:111. doi:10.3389/fnbot.2010.00111

Jeannerod, M. (2007). *Motor Cognition: What Actions Tell the Self.* Oxford: Oxford University Press.

Lallee, S., Madden, C., Hoen, M., and Dominey, P. F. (2010). Linking language with embodied and teleological representations of action for humanoid cognition. *Front. Neurorobot.* 4:8. doi: 10.3389/fnbot.2010.00008

Marocco, D., Cangelosi, A., Fischer, K., and Belpaeme, T. (2010). Grounding action words in the sensorimotor interaction with the world: experiments with a simulated iCub humanoid robot. *Front. Neurorobot.* 4:7. doi: 10.3389/fnbot.2010.00007

Massera, G., Cangelosi, A., and Nolfi, S. (2007). Evolution of prehension ability in an anthropomorphic neurorobotic arm. *Front. Neurorobot.* 1:4. doi: 10.3389/neuro.12.004.2007

Parisi, D. (2010). Robots with language. *Front. Neurorobot.* 4:10. doi: 10.3389/fnbot.2010.00010

Pecher, D., and Zwaan, R. A. (eds). (2005). *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking.* Cambridge: Cambridge University Press.

Pezzulo, G., Barsalou, L. W., Cangelosi, A., Fischer, M. H., McRae, K., and Spivey, M. J. (2011). The mechanics of embodiment: a dialog on embodiment and computational modelling. *Front. Psychol.* 2:5. doi: 10.3389/fpsyg.2011.00005

Pulvermuller, F. (2003). *The Neuroscience of Language: On Brain Circuits of Words and Serial Order.* Cambridge: Cambridge University Press.

Rizzolatti, G., and Arbib, M. A. (1998). Language within our grasp. *Trends Neurosci.* 21, 188–194.

Rizzolatti, G., Fadiga, L., Gallese, V., and Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Brain Res. Cogn. Brain Res.* 3, 131–141.

Symes, E., Tucker, M., Ellis, R., Vainio, L., and Ottoboni, G. (2008). Grasp preparation improves change-detection for congruent objects. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 854–871.

Symes, E., Tucker, M., and Ottoboni, G. (2010). Integrating action and language through biased competition. *Front. Neurorobot.* 4:9. doi: 10.3389/fnbot.2010.00009

Tikhanoff, V., Cangelosi, A., and Metta, G. (2011). Language understanding in humanoid robots: iCub simulation experiments. *IEEE Trans. Auton. Ment. Dev.* 3, 17–29.

Received: 29 March 2012; accepted: 10 April 2012; published online: 01 May 2012.

Citation: Cangelosi A (2012) Editorial of e-book on action and language integration. Front. Neurorobot. ***6****:2. doi: 10.3389/fnbot.2012.00002*

ORIGINAL RESEARCH ARTICLE
published: 07 June 2010
doi: 10.3389/fnbot.2010.00009

Integrating action and language through biased competition

Ed Symes*, Mike Tucker and Giovanni Ottoboni

School of Psychology, University of Plymouth, Plymouth, UK

Edited by:
Angelo Cangelosi, University of Plymouth, UK

Reviewed by:
Claudia Gianelli, Università di Bologna, Italy
Daniele Caligiore, Institute of Cognitive Sciences and Technologies, Italy

***Correspondence:**
Ed Symes, School of Psychology, University of Plymouth, Drake Circus, Plymouth, Devon PL4 8AA, UK.
e-mail: esymes@plymouth.ac.uk

Several recent psychological investigations have demonstrated that planning an action biases visual processing. Symes et al. (2008) for example, reported faster target detection for a changing object amongst several non-changing objects following the planning of a target-congruent grasp. The current experimental work investigated how this effect might compare to, and indeed integrate with, effects of language cues. Firstly a cuing effect was established in its own right using the same change-detection scenes. Sentences cued object size (e.g., "Start looking for a change in the larger objects"), and these successfully enhanced detection of size-congruent targets. Having thereby established two effective sources of bias (i.e., action primes and language cues), the remaining three experiments explored their co-occurrence within the same task. Thus an action prime (participants planned a power or precision grasp) *and* a language cue (a sentence) preceded stimulus presentation. Based on the tenets of the biased competition model (Desimone and Duncan, 1995), various predictions were made concerning the integration of these different biases. All predictions were supported by the data, and these included reliably stronger effects of language, and concurrent biasing effects that were mutually suppressive and additive.

Keywords: biased competition, top-down and bottom-up interaction, action intentions, language cues, change detection

INTRODUCTION

"Indeed, the general point is that attention greatly reduces the processing load for animal and robot. The catch, of course, is that reducing computing load is a Pyrrhic victory unless the moving focus of attention captures those aspects of behavior relevant for the current task…" (Arbib et al., 2008, p. 1461).

The present paper examines how current behavioral targets that are defined explicitly through language, or implicitly through action intentions, might serve to bias object representations and ultimately selective attention. Moreover, the experimental work investigates how the biasing effects of these two different sources might integrate. The biased competition model (Desimone and Duncan, 1995) serves as the theoretical backdrop and is used to generate experimental predictions regarding the integration of language and action as sources of representational bias. The data reported later on support all of the predictions made by the model.

In the real world, action intentions or action plans typically refer to object-related goal states that can be broken down into various stages. For example, the specific intention to turn on a lamp may require planning to walk across the room and planning to reach toward and grasp its switch. At all of these stages, the relevant action plan depends to an extent on the goal object itself, whether it be its location (thus implicating walking direction), or its intrinsic properties (thus implicating grasp aperture, for instance). In the experimental world of the present studies however, we interchangeably use the terms action intentions or action plans to simply refer to a pre-activated motor system. Schütz-Bosbach et al. (2007) refer to similar states of preparedness between selecting and executing an action as "motor attention." This pre-activation is not necessarily related to a target object (indeed in the change detection task used, an action is prepared but the target object itself is unknown!) All that is relevant therefore, is the prepared state of the motor system.

AN OVERVIEW OF THE BIASED COMPETITION MODEL

According to the biased competition model, all visual inputs compete for neuronal representation in multiple visual brain regions (Desimone and Duncan, 1995; Desimone, 1998; Duncan, 1998). Such competition occurs at all stages of visual processing, but most strongly at the level of the neuron's receptive field (e.g., higher visual areas such as V4). Competition is automatic and ongoing, occurs with and without directed attention, and is characterized by suppressive interactions between stimuli. Enhanced amplitude and duration of responses to one object are associated with decreased responses to others. Evidence of these suppressive interactions comes in part from single-cell monkey studies that demonstrate smaller responses for pairs of stimuli falling within a neuron's receptive field, than for those stimuli presented alone (Luck et al., 1997; and see Beck and Kastner, 2009 for a recent review). Importantly, this competition can be resolved by spatially directed attention. Attending to one stimulus in a pair biases competition between objects. Reynolds et al. (1999) reported that monkey neuronal responses were weighted in favor of the attended stimulus of the pair, such that response levels resembled those evoked by that stimulus when it was presented alone (i.e., the suppressive influence of the non-attended stimulus was counteracted). Related findings have been reported in humans using fMRI; here, language instructions explicitly told participants where to direct their attention (Kastner et al., 1998).

Contrast gain control has been proposed as one possible mechanism that can account for this attentional biasing (Reynolds et al., 2000). Contrast gain control increases the effective salience of the attended stimulus. Monkey V4 neurons for example, responded to an attended stimulus with increased sensitivity, as if its physical contrast had increased (Reynolds et al., 2000). Directly adjusting the physical luminance contrast of a stimulus produced equivalent responses (in the absence of attention, V4 neurons were preferentially driven by the higher contrast stimulus in a pair, Reynolds and Desimone, 2003). According to the contrast gain account, any effect that attention has on competition depends on where the stimulus falls on the contrast-response function; attention should not increase neuronal sensitivity, for instance, when a high contrast stimulus is at the saturation point on the contrast-response function (Reynolds et al., 2000).

ACTION INTENTIONS MIGHT ALSO ACT AS A BIASING SIGNAL

As well as the biasing effects of *spatially* directed attention, the biased competition model also predicts that priming neurons responsive to current behavioral targets can bias competition. Actively searching for a red object, for instance, should bias competition in favor of red objects by preactivating "red" feature coding neurons. These magnified signals suppress the signals from neurons that are selective for other colors (Duncan, 1998). Some authors have recently proposed that biased competition could be the mechanism that underlies cases of enhanced visual processing following action planning (see Bekkering and Neggers, 2002; Hannus et al., 2005; Symes et al., 2008, 2009). Indeed, there is a steadily growing body of behavioral research that suggests that planning an action of some sort (an action intention) affects a range of visual processes. These include selection (e.g., Bekkering and Neggers, 2002; Fischer and Hoellen, 2004; Hannus et al., 2005; Linnell et al., 2005); attentional capture (Welsh and Pratt, 2008), motion perception (Lindemann and Bekkering, 2009); detection through feature weighting (Craighero et al., 1999; Symes et al., 2008, 2009); and detection through dimensional weighting (Fagioli et al., 2007; Wykowska et al., 2009). In such cases, action intentions may serve as the behaviorally relevant prime that preactivates neurons responsive to current behavioral targets.

One example of this that is of particular relevance to the current study comes from Symes et al. (2008). The authors used a change detection paradigm as a tool for investigating visual processing following action-based priming. Change detection provides an effective means of measuring the locus of focused attention (Simons and Rensink, 2005), and the flicker paradigm used can be conceived of as a spatiotemporal version of the static extended displays in visual search experiments (Rensink, 2005). In the flicker paradigm, two pictures that are identical in all but one respect (i.e., the change) cycle back and forth, separated by a blank "flicker" that eradicates visual transients associated with the change. Changes can be surprisingly hard to detect (so-called "change blindness"), often taking several cycles. In one experiment (Symes et al., 2008, Experiment 1b), participants searched for an unknown target amongst 12 graspable objects in a photographed array (the target was one object, such as an apple, being alternated with another size-matched object such as an orange). Prior to the onset of the scene, participants prepared and maintained a grasp plan (either a whole hand "power grasp," or a thumb and forefinger "precision grasp"). Target detection time was faster when the intended grasp (e.g., power or precision) was compatible with the target (e.g., large or small). Thus planning an action biased object representations and ultimately selective attention of action-appropriate object features. The experiments reported later adopted this same methodology to investigate the interacting influences of action intentions ("action primes") and language-directed attention ("language cues").

INTERACTING INFLUENCES IN BIASED COMPETITION

Although they have almost exclusively been investigated in isolation, in everyday life top-down and bottom-up influences on competition are likely to interact (Beck and Kastner, 2009). Reynolds and Desimone (2003) report physiological data that captures an instance of top-down and bottom-up interaction, with the bottom-up bias coming from luminance contrast and the top-down bias coming from directed attention. Neuronal responses were recorded to a stimulus pair consisting of a "good" grating (the neurons responded well to its horizontal orientation) and a "poor" grating (the neurons responded poorly to its vertical orientation). When *attending* to either stimulus in the pair, attention and contrast were *additive* influences. Attending to the lower contrast poor stimulus afforded a slight reduction in the response to the pair (i.e., the poor stimulus gained some control over the neural response). Attending to the higher contrast poor stimulus afforded it almost complete control over the response to the pair (effectively eliminating the influence of the good, featurally preferred stimulus).

RATIONALE AND PREDICTIONS FOR THE CURRENT STUDY

Based on the above considerations of biased competition, we make the following observations and broad predictions regarding the current behavioral study, which explores the interaction of two top-down influences on visual object representation – one explicit (language cues) and the other implicit (action intentions).

1) As was the case when using action primes (e.g., Symes et al., 2008, 2009), cuing "large" or "small" objects through a single source – this time language instructions – should bias competition in favor of congruently sized objects (presumably by preactivating "large" or "small" feature coding neurons). We expected to find proxy evidence of this in faster detections on trials where the cue and target were size-congruent (i.e., trials with valid cues).
2) Since the language cues specified the targets of directed attention, they were expected to produce a strong biasing influence that is comparable to the influence of directed attention (e.g., Reynolds and Desimone, 2003). Relatedly, the biasing influence of action primes is expected to be weaker. This influence arises without directed attention (see Experiment 2a of Symes et al., 2008), and in this sense is more comparable to the bottom-up influence of contrast, which also arises without directed attention (e.g., Reynolds and Desimone, 2003).
3) In line with the suppressive interactions predicted by the biased competition model, we expected that when there were multiple concurrent weighting sources their effects on object representations would compete.
   a. Firstly we expected to find biasing effects for each source that reflected their different relative strengths as described in prediction 2 above.

b. These concurrent effects should be *smaller* than when they are found independently (i.e., they are mutually suppressive).
c. However, when one weighting source is sufficiently stronger than another, it may even suppress the effect of the weaker source completely. Indeed, the findings of Symes et al. (2009) support this prediction – when bottom-up target saliency was high, action primes were ineffective (see also Wykowska et al., 2009).

4) Reynolds and Desimone's (2003) data revealed that top-down and bottom-up biases produced similar cellular responses that were *additive*. It therefore follows that the biasing effects of two top-down signals (language and action) should also be additive. Thus we expected the best performance on trials when cue, prime, and target were all congruent.

GENERAL METHOD

All experiments were approved by the University of Plymouth's Human Ethics Committee, and informed consent was obtained from all participants.

LANGUAGE CUES AND ACTION PRIMES

Using a flicker paradigm, the experiments reported below attempted to enhance change detection by weighting size-related features of the target in a top-down manner. This weighting was achieved using language cues and action primes. The language cues actually specified overt searching behaviors. *Partially valid* cues (Experiments 1, 3, and 4), which were valid for half of the time, instructed participants how to *start* their search (e.g., "Start looking for a change in the larger objects"). Participants were told that if they could not find the target easily, they should include noncued objects in their search (i.e., they only had to *start* their search based on the cue). *Completely valid* cues however (Experiment 2), were always valid, and instructed participants how to conduct their search throughout (e.g., "Look for a change in the larger objects"). The association between the language cue and the object's size was therefore explicit, whereas for the action prime it was implicit. Furthermore, the action primes did not specify any overt searching behavior. Partially valid (Experiments 2 and 4) or completely valid primes (Experiment 3), simply instructed participants which response device to hold (thereby establishing an action intention for a particular grasp). It is assumed that relative to one another, partially and completely valid cues or primes constituted different weighting strengths (low and high strengths respectively).

Action primes and language cues therefore provided quite different types of top-down bias, and how their effects might integrate (or not) when biasing competition between objects was at the heart of these investigations. **Table 1** summarizes the methods for each experiment in terms of the different action primes and language cues used.

CHANGE DETECTION PARADIGM

Some of the following methodological details have been adapted from Symes et al. (2008). Change-blindness scenes consisted of an array of 12 grayscale photographs of fruit and vegetables (half were small objects congruent with a precision grasp, and half were large objects congruent with a power grasp). One random object in the scene changed back and forth into another object of a similar size (e.g., an apple changed into an orange), and this change coincided temporally with a visually disrupting screen flicker that provided the necessary conditions for change blindness. Participants were told that the identity of 1 of the 12 objects would change back and forth, and their basic task was to follow the screen instructions that appeared at the start of a trial, detect the change, indicate detection with a manual response, and then identify the change using the keyboard. Screen instructions typically included a language cue and instructions for planning a manual response (i.e., the action prime). The specific details of these instructions are described for each experiment. The stimuli described below however, were used in all four experiments (and Symes et al., 2008).

Change detection stimuli

Change detection scenes arose from cyclically presenting a screen "flicker" (F) between an "original" (O) and "modified" (M) picture-pair in the order OFMFOFMF… This sequence cycled until a response was made, and a "change identification" picture was shown to establish that the correct change had been detected. Thus the stimulus set consisted of a flicker stimulus (a blank gray screen), and 60 "original," "modified" and "change identification" grayscale pictures (1,024 × 768 pixels; 32.5 cm × 24.5 cm; visual angle (VA) ≈ 36.0° × 27.5°). As discernable from **Figure 1** (top left panel), each original picture consisted of a 4 × 3 array of six large objects (e.g., an apple) and six small objects (e.g., a strawberry). These had been selected at random from a pool of 24 items of fruit and vegetables, again, half of which were large and half of which were small. The size of each individual object photograph was manipulated such that all small objects were of a similar size (mean VA ≈ 2.3° × 1.6°), and all large objects were of a similar size (mean VA ≈ 4.9° × 4.1°). These objects and their measurements are listed in the Appendix of Symes et al. (2008).

Table 1 | Summary details of Language Cue and Action Prime conditions across experiments.

	Symes et al. (2008; Experiment 1b)	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Language Cues of target size ("large/small")	–	Partially valid	Completely valid	Partially valid	Partially valid
Action Primes of target size (power/precision grasp intention)	Partially valid	–	Partially valid	Completely valid	Partially valid

In each experiment either a Language Cue and/or an Action Prime preceded the onset of a change-detection scene in which participants searched for an unknown changing target. See preceding text for descriptions of types of cues and primes used.

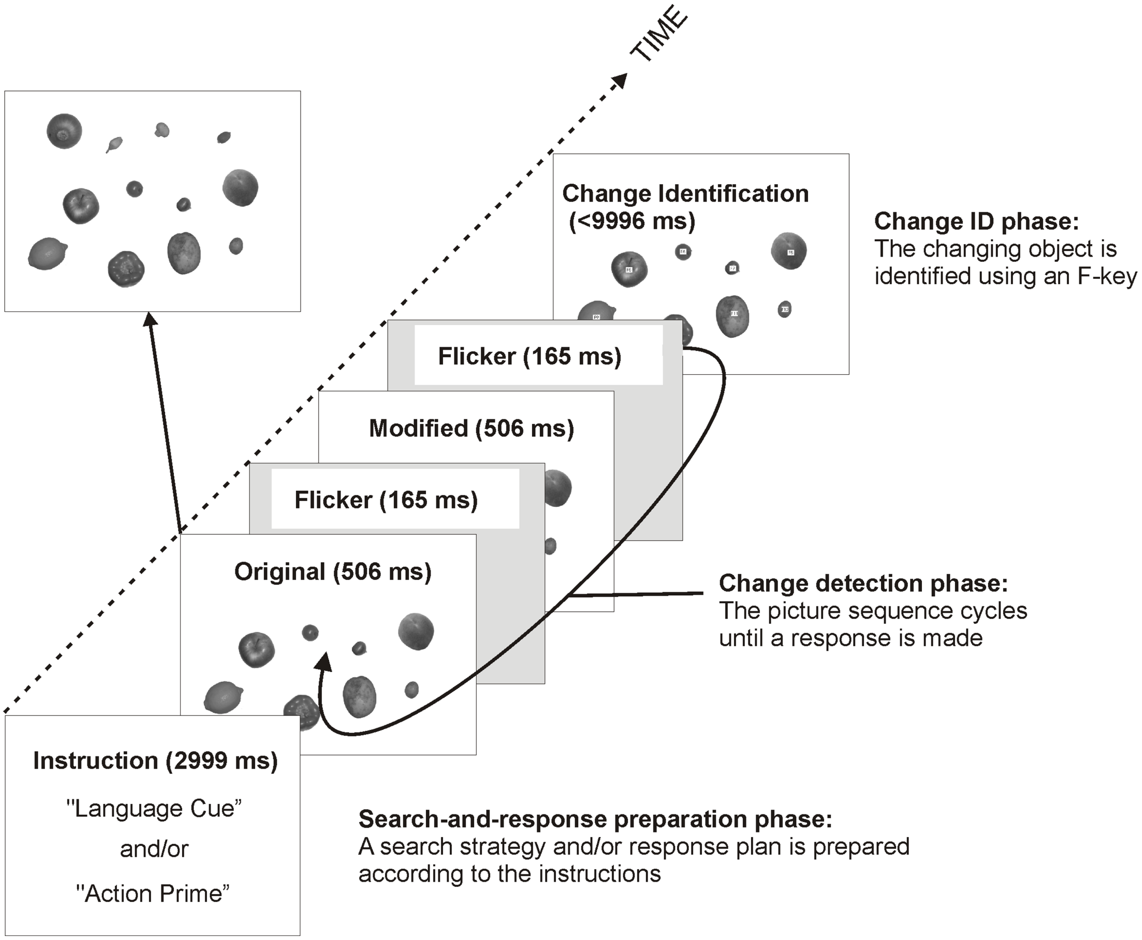

TIME
Change Identification (<9996 ms)
Change ID phase:
The changing object is identified using an F-key
Flicker (165 ms)
Modified (506 ms)
Flicker (165 ms)
Original (506 ms)
Change detection phase:
The picture sequence cycles until a response is made
Instruction (2999 ms)
"Language Cue"
and/or
"Action Prime"
Search-and-response preparation phase:
A search strategy and/or response plan is prepared according to the instructions

FIGURE 1 | Schematic illustration of the sequence and timings of the displays in all four experiments (adapted from Symes et al., 2008). In Experiment 1, the instruction consisted of only a language cue, whereas in the remaining experiments it consisted of a language cue and an action prime.

The order of objects in the array resulted from a random shuffle of the 12 selected objects, and their positions on the screen varied within a loosely defined grid (thereby creating perceptually distinct-looking scenes). An appropriately sized object (30 of each size were required) was selected at random from the 12 to be the changing object. An appropriately sized replacement object was selected at random from the pool (after object selection for the original picture, six large and six small objects remained in the pool). Thus in the modified picture, all objects remained the same as the original picture, except from a single changing object. This was removed and replaced by an object of a similar size (e.g., a strawberry was replaced with a cherry). Each original picture was also reproduced as a change identification picture, whereby each object in the array had an identification "F-number" (F1–F12) superimposed on it. These F-numbers corresponded to the 12 "F-keys" on a keyboard.

EXPERIMENT 1

As mentioned earlier, Symes et al. (2008) demonstrated that partially valid action primes enhanced change detection for prime-congruent sized objects. This first experiment was simply designed to establish a similar effect of partially valid *language cues* using exactly the same stimuli. In terms of the flicker paradigm itself, enhanced detection was expected in principle (explicitly cuing a change with partially valid language cues enhanced detection in a study by Rensink et al., 1997).

Preceding stimulus onset a partially valid language cue appeared on the screen (e.g., "Start looking for a change in the larger objects"). Participants were told to follow the text instructions, and that the identity of 1 of the 12 objects would change back and forth. Their basic task was to press the spacebar as soon they detected which object was changing. The first two predictions set out in the Section "Introduction" are relevant for this experiment, and are summarized below:

1) We expected proxy evidence of language cues biasing competition between objects, with faster detections on trials where cue and target were size-congruent.
2) Since language cues told participants where to look, they should have a strong biasing influence like that of directed attention. Relatedly, the biasing effect of action primes (that arises without directed attention) was expected to be weaker. A cross-experimental comparison of effect sizes tested this prediction.

METHOD

Participants

Twenty volunteers between 39 and 18 years of age [mean (M) = 21.8 years] were paid for their participation in a single session that lasted approximately 20 min. Of these, 17 were females (1 left-handed) and 3 were males (all right-handed). All self-reported normal or corrected-to-normal vision and normal motor control, and all were naïve as to the purpose of the study.

Apparatus and stimuli

Experimental sessions took place in a dimly lit room at a single computer workstation. Situated centrally at the back of the table was a RM Innovator desktop computer that supported a 16-inch RM color monitor (with a screen resolution of 1,024 × 768 pixels and a refresh frequency of 85 Hz). In front of the computer was a keyboard and mouse. The viewing distance was approximately 50 cm, and the hand-to-screen distance was approximately 30 cm. See Section "General Method" for details of the change detection stimuli used.

Design and procedure

Four conditions arose from the orthogonal variation of two within-subjects variables, each with two levels: Language Cue (large or small – specifically, "Start looking for a change in the larger/smaller objects") and Target Size (large or small). At the beginning of the experiment, participants were talked through some written instructions that explained the task. A short practice session of four trials was followed by 120 experimental trials. These consisted of two blocks of 60 trials (4 conditions × 15 replications), with each of the 60 change detection scenes being shown in a random order within each block.

Each trial followed three broad phases: search-and-response preparation, change detection and change identification. Preceding stimulus onset, the language cue appeared center screen and the participant rested the fingertips of both hands on the spacebar of the keyboard (search-and-response preparation phase). The change detection scene then appeared, and as it cycled the participant scrutinized the 12 objects for a change. Upon noticing the change, the participant immediately pressed the spacebar (change detection phase). This response caused the change identification picture to appear, and the participant pressed an F-key on the keyboard corresponding to the F-number of the object they thought they had seen change (change identification phase). Otherwise, it timed-out after 10 s. The sequence and timings for these three phases are illustrated in **Figure 1**.

Response times (RTs) and errors were recorded to a data file for off-line analysis, and the possible source of error related to an F-key response that timed-out or did not correspond to the changing object's F-number.

RESULTS AND DISCUSSION

Errors and RTs more than two standard deviations (SDs) from each participant's condition means were excluded from this analysis and the analyses of all the other experiments reported. 1.8% of trials were removed as change identification errors (i.e., when an F-key identified the wrong object). No further analysis of errors was undertaken; the change identification error data revealed that on the vast majority of trials the correct object had been identified. 4.6% of the remaining trials were removed as outliers, reducing the maximum detection time from 21,065 to 13,113 ms (M = 4,848 ms; SD = 3,028).

The effect of language cues

The condition means of the remaining data were computed for each participant and subjected to a repeated measures analysis of variance (ANOVA) with the within-subjects factors of Language Cue (large or small) and Target Size (large or small). An interaction between Language Cue and Target Size was observed, $F(1, 19) = 120.11, p < 0.001$, that revealed the predicted biasing effect (see prediction 1 above). Mean detection times were faster for large targets following a large (3,622 ms) rather than small (6,131 ms) language cue, and faster for small targets following a small (3,497 ms) rather than large (6,155 ms) language cue.

Comparison with the effect of action primes

In order to evaluate whether the biasing effect of language cues was significantly larger than that of action primes obtained in Experiment 1b of Symes et al. (2008) (see prediction 2 above), cropped correct RT data for each experiment were split by participant and cue/prime–target congruent and incongruent trials. From this a mean effect size for each participant in each experiment was calculated (mean effect size = mean incongruent RTs – mean congruent RTs). These data were compared in a one-tailed independent samples t test. This analysis revealed that the mean effect size associated with language cues was indeed significantly larger (current experiment: language cues = 2,579 ms, Experiment 1b: action primes = 372 ms), $t(40) = 8.874, p < 0.001$.

Distributional analyses

In order to see whether the biasing effect of language cues behaved consistently across different portions of the RT distribution, the Vincentization procedure (Ratcliff, 1979) was used to derive the mean RTs for a new ANOVA (Greenhouse–Geisser corrections for sphericity violations have been used where G is shown). Mean RTs were calculated for five equal bins of rank ordered raw data according to each experimental condition. A statistically significant full interaction, $F(1.249, 23.734) = 8.894, p = 0.004^{G}$, derived from the resulting 2 × 2 × 5 ANOVA (Language Cue – large or small; Target Size – large or small, and Bin – first to fifth). Unpacking this interaction in separate ANOVAS for each bin revealed that all bins had produced significant cue–target compatibility effects ($p < 0.05$). In accounting for the significant full interaction however, it is notable that the effect sizes were smaller in the first and last bins (Bin 1 = 1,805 ms, Bin 5 = 1,151 ms) than in those in-between (Bins 2, 3, and 4 = 2,645, 2,767, and 2,892 ms, respectively).

Overall, these results supported the two predictions generated by the biased competition model, and suggested that the partially valid language cues had successfully biased object representations and ultimately selective attention. This was the case when detection times were averaged across the whole distribution, and when they were divided into individual bins.

The remaining three experiments presented both cues *and* primes in a variety of validity combinations (return to **Table 1** for an overview) to investigate how such different sources of intentional weighting might work together to bias object representations.

EXPERIMENT 2

In this second experiment, language cues were *completely valid* and action primes were only *partially valid*. On any given trial prior to stimulus presentation, a language cue instructed participants how to search for the target (e.g., "Look for a change in the larger objects"), and a separate instruction told participants which grasp-simulating response device to hold and prepare to squeeze (this planned action was the action prime). Participants were told that the identity of 1 of the 12 objects would change back and forth, and their basic task was to search according to the language cue, and to execute their planned grasp as soon as they detected which object was changing.

Prediction 3 set out in the Section "Introduction" is relevant for this experiment, and is summarized below:

3) With two concurrent top-down weighting sources, we expected that their effects on object representations would compete.
   a. We therefore expected to find biasing effects for each source that reflected their different relative strengths (i.e., language cue effects are bigger).
   b. These effects should be *smaller* than when found independently (i.e., they are mutually suppressive).
   c. However, when one weighting source is sufficiently stronger than another, it may *completely* suppress the effect of the weaker source.

Given that the already stronger bias of language cues was maximized in this experiment (i.e., they were *completely valid* cues), it was expected to dominate competition (prediction 3c).

METHOD

Participants

Twenty-one volunteers between 52 and 18 years of age ($M = 24.0$ years) were paid for their participation in a single session that lasted approximately 20 min. Of these, 15 were right-handed females and 6 were males (1 left-handed). All self-reported normal or corrected-to-normal vision and normal motor control, and all were naïve as to the purpose of the study.

Apparatus and stimuli

As Experiment 1. In addition, the keyboard was moved closer to the screen (by 15 cm) to make room for the new response apparatus, which was affixed centrally (from left to right) and set in by 10.5 cm from the table's leading edge. When holding this apparatus, the hand-to-screen distance was approximately 30 cm. The apparatus was fixed to the table top in a vertical position and consisted of two physically connected devices – a cylindrical "power device" (l = 10 cm; diameter = 3 cm), and a square "precision device" ($l = 1.25$ cm, $w = 1.25$ cm, $h = 1.25$ cm). A power grasp was required to hold the power device and a precision grasp was required to hold the precision device. In order to avoid establishing any semantic associations between the devices and their size or required grasp (i.e., ensuring the association between the action prime and the object's size was implicit), the power device was neutrally referred to by the experimenter (and the on-screen instructions) as the "Black device" (it was colored black) and the precision device as the "White device" (it was colored white). Execution of a particular grasp depressed a micro switch embedded in that device, and this response was registered with millisecond accuracy by the computer (micro switches were connected via an input/output box to the parallel interface of the computer).

Design and procedure

Four conditions arose from the orthogonal variation of two within-subjects variables, each with two levels: Action Prime (power or precision) and Language-cued Target Size (large or small). At the beginning of the experiment, participants were talked through some written instructions that explained the task. A short practice session of four trials was followed by 120 experimental trials. These consisted of two blocks of 60 trials (4 conditions × 15 replications), with each of the 60 change detection scenes being shown in a random order within each block.

The trial procedure was similar to Experiment 1 (refer back to **Figure 1**); with three broad phases of search-and-response preparation, change detection and change identification. Preceding stimulus onset this time, the *completely valid* language cue (e.g., "Look for a change in the larger objects") appeared above text instructions for the *partially valid* action prime (which warned participants to prepare a response using either the "Black device" or the "White device"). The participant reached to the instructed device, and held it lightly in their dominant hand (using the device-appropriate hand shape). When the participant detected a change, s/he executed the grasp by squeezing the device. The participant then identified the change as before, by pressing the appropriate F-key.

Response times and errors were recorded to a data file for off-line analysis, and there were two possible sources of error: violations of the response instruction (participants used the wrong device), and change identification errors (an F-key response that timed-out or did not correspond to the changing object's F-number).

RESULTS AND DISCUSSION

1.87% of trials were removed as errors (0.56% response errors, 1.35% change identification errors, 0.04% both errors on same trial). No further analysis of errors was undertaken; response and change identification error data revealed that on the vast majority of trials the response instructions had been adhered to and the correct object had been identified. 3.92% of the remaining trials were removed as outliers, reducing the maximum detection time from 18,943 to 7,986 ms (M = 2,583 ms; SD = 1,152 ms).

The effect of language cues

Mean cropped experimental RTs for each participant were compared for the current experiment and Experiment 1b of Symes et al., (2008) in order to establish whether completely valid language cues enhanced overall detection times (see prediction 3a above). The single methodological difference between the two experiments was the presence of a language cue in the current experiment. If this language cue enhanced detection, we should expect faster overall detection times for the current experiment. Indeed, mean experimental RTs were 2,102 ms faster (current experiment grand mean = 2,584 ms, Experiment 1b grand mean = 4,686 ms), and a one-tailed independent samples t test confirmed that this difference was statistically significant, $t(41) = 10.156$, $p < 0.001$. [In this instance there is no comparison case to test prediction 3b (whether the effect is *smaller* than when completely valid cues are the only weighting source)].

The effect of action primes

The condition means of the cropped data were computed for each participant and subjected to a repeated measures ANOVA with the within-subjects factors of Action Prime (power or precision) and Language-cued Target Size (large or small). The crucial interaction between Action Prime and Language-cued Target Size failed to even approach statistical significance, $F(1, 20) = 0.11$, $p > 0.5$. Thus it seems that any biasing effect of action primes was completely suppressed by the dominant completely valid language cues (thereby supporting prediction 3c above).

In order to establish that this (null) effect of action primes was *smaller* than the biasing effect of action primes in Experiment 1b of Symes et al. (2008) (according to prediction 3b above it should be smaller because it was in competition with another source of bias – a language cue, whereas in Experiment 1b it was not), cropped correct response data for each experiment was split by participant and prime–target congruent and incongruent trials. From this a mean effect size for each participant in each experiment was calculated, and these data were compared in a one-tailed independent samples *t* test. This analysis revealed that the mean effect size of prime–target compatibility was indeed significantly smaller when it was a shared rather than only source of bias (current experiment: shared source = 4 ms, Experiment 1b: single source = 372 ms), $t(41) = 3.360$, $p = 0.001$.

Furthermore, as expected from prediction 3a, the stronger biasing source of Language Cue produced the larger biasing effect of the two (language cue effect = 2,012 ms; action prime effect = 4 ms).

Distributional analyses

In order to see whether the null effect of action primes in this experiment was consistent across different portions of the RT distribution, distributional analyses were performed (see Experiment 1 for procedural details). A statistically significant full interaction, $F(1.232, 24.645) = 4.454$, $p = 0.038^{G}$, derived from the resulting $2 \times 2 \times 5$ ANOVA (Action Prime – power or precision; Language-cued Target Size – large or small, and Bin – first to fifth). Unpacking this interaction in separate ANOVAS for each bin revealed that no significant interactions between Action Prime and Language-cued Target Size were observed under Bins 1, 3, and 4 ($p > 0.05$). Under Bins 2 and 5 however, some interesting patterns emerged. Under the relatively fast RTs of Bin 2, an interaction resembling a reversed compatibility effect was found, $F(1, 20) = 4.271$, $p = 0.052$. Here, mean detection times were actually *slower* for large-cued targets following a power (1,975 ms) rather than precision (1,796 ms) action prime, and marginally slower for small-cued targets following a small (1,886 ms) rather than large (1,894 ms) action prime. Contrastingly, under the relatively slow RTs of Bin 5, an interaction resembling a compatibility effect was found, $F(1, 20) = 3.438$, $p = 0.079$. Here, mean detection times were faster for large-cued targets following a power (5,109 ms) rather than precision (5,299 ms) action prime, and faster for small-cued targets following a small (4,584 ms) rather than large (5,149 ms) action prime. It is plausible that this pattern reflects those longest detection-time cases in which the language cue has lost its potency (i.e., it has not helped participants find the change quickly). In this situation, the language cue weightings no longer dominate suppressive interactions, and hence the action prime is able to exert its influence.

Overall, these results supported the predictions generated by the biased competition model, and suggest that the completely valid language cue had enhanced detection (and dominated competition such that any effect of action primes was completely suppressed in most bins).

EXPERIMENT 3

This third experiment continued to combine language cues and action primes to investigate how such different sources of intentional weighting might work together to bias object representations. In a reversal of the conditions of the previous experiment, now *action primes* were completely valid and *language cues* were only partially valid. On any given trial prior to stimulus presentation, a language cue instructed participants how to commence their search for the target (e.g., "Start looking for a change in the larger objects"), and a separate instruction told participants which response device to hold and prepare to squeeze on target detection (i.e., the action prime).

Prediction 3 set out in the Section "Introduction" is again relevant for this experiment, and is summarized below:

3) With two concurrent top-down weighting sources, we expected that their effects on object representations would compete.
   a. We therefore expected to find biasing effects for each source that reflected their different relative strengths (i.e., language cue effects are bigger).
   b. These effects should be *smaller* than when found independently (i.e., they are mutually suppressive).
   c. However, when one weighting source is sufficiently stronger than another, it may *completely* suppress the effect of the weaker source.

Even though action primes were completely valid here, they were not expected to dominate competition in the same way that completely valid language cues did in the previous experiment. This is because they are an inherently weaker source of bias (as formally established in the cross-experimental analysis of Experiment 1). Thus prediction 3c does not apply here.

METHOD

Participants

Twenty-one volunteers between 51 and 18 years of age ($M = 21.1$ years) were paid for their participation in a single session that lasted approximately 20 min. Of these, 19 were right-handed females and 2 right-handed males. All self-reported normal or corrected-to-normal vision and normal motor control, and all were naïve as to the purpose of the study.

Apparatus and stimuli

As Experiment 2.

Design and procedure

As Experiment 2, differing only in that Action Primes were completely valid, and Language Cues were partially valid. The four conditions that arose from the orthogonal variation of two within-subjects variables, each with two levels were: Language Cue (large or small) and Action-primed Target Size (large or small).

RESULTS AND DISCUSSION

3.29% of trials were removed as errors (0.44% response errors, 2.90% change identification errors, 0.04% both errors on same trial). No further analysis of errors was undertaken. 5.50% of the remaining trials were removed as outliers, reducing the maximum detection time from 27,686 to 13,200 ms (M = 4,087 ms; SD = 2,213).

The effect of action primes

Mean cropped experimental RTs for each participant were compared for the current experiment and Experiment 1, in order to establish whether completely valid action primes enhanced overall detection times (see prediction 3a above). The only methodological difference between the two experiments was the presence of an action prime in the current experiment (this, along with its associated grasp responses). If this action prime enhanced detection, we should expect faster overall detection times for the current experiment. Indeed, mean experimental RTs were 784 ms faster (current experiment grand mean = 4,082 ms, Experiment 1 grand mean = 4,866 ms), and a one-tailed independent samples t test confirmed that this difference was statistically significant, $t(39) = 2.901$, $p < 0.005$. [In this instance there is no comparison case to test prediction 3b (whether the effect is *smaller* than when completely valid primes are the only weighting source)].

The effect of language cues

The condition means of the cropped data were computed for each participant and subjected to a repeated measures ANOVA with the within-subjects factors of Language Cue (large or small) and Action-primed Target Size (large or small).

An unexpected main effect of Language Cue, $F(1, 20) = 7.041$, $p < 0.05$, reflected faster mean detection times following "large" (3,972 ms) rather than "small" (4,191 ms) cues. The crucial interaction between Language Cue and Action-primed Target Size was also observed, $F(1, 20) = 78.609$, $p < 0.001$, revealing the predicted compatibility effect (see prediction 3a above). Mean detection times were faster for large action-primed targets following a large (2,964 ms) rather than small (5,197 ms) language cue, and faster for small action-primed targets following a small (3,186 ms) rather than large (4,980 ms) language cue.

In order to establish whether this biasing effect of language cues was *smaller* than the biasing effect of language cues in Experiment 1 (according to prediction 3b above it should be smaller because it was in competition with another source of bias – an action prime, whereas in Experiment 1 it was not), cropped correct response data for each experiment was split by participant and prime–target congruent and incongruent trials. From this a mean effect size for each participant in each experiment was calculated, and these data were compared in a one-tailed independent samples t test. This analysis revealed that the mean effect size of cue–target compatibility was indeed significantly smaller when it was a shared rather than only source of bias (current experiment: shared source = 2,013 ms, Experiment 1: single source = 2,579 ms), $t(39) = 1.723$, $p < 0.05$.

Furthermore, as expected from prediction 3a, the stronger biasing source of Language Cue produced the larger biasing effect of the two (language cue effect = 2,013 ms; action prime effect = 784 ms).

Distributional analyses

In order to see whether the significant biasing effect of language cues in this experiment was consistent across different portions of the RT distribution, distributional analyses were performed (see Experiment 1 for procedural details). A statistically significant full interaction, $F(1.260, 25.196) = 12.302$, $p < 0.001^{G}$, derived from the resulting 2 × 2 × 5 ANOVA (Language Cue – large or small; Action-primed Target Size – large or small; and Bin – first to fifth). Unpacking this interaction in separate ANOVAS for each bin revealed that the first four bins had all produced highly significant cue–target compatibility effects ($p < 0.001$). The fifth bin revealed a similar, if diminished, pattern of compatibility ($p = 0.103$). Overall then, the effect of language cues seemed highly consistent across bins.

Overall, these results again supported the predictions generated by the biased competition model. The presence of a completely valid action prime enhanced detection, but it did not dominate competition (being an inherently weaker source of bias). Thus language cues exerted a consistent effect across the RT distribution, and as predicted, this was a smaller effect than the one generated in Experiment 1 (where languages cues were the only source of bias).

EXPERIMENT 4

This last experiment combined *partially valid* language cues with *partially valid* action primes. The third and fourth predictions set out in the Section "Introduction" are relevant for this experiment, and are summarized below:

3) With two concurrent top-down weighting sources, we expected that their effects on object representations would compete.
   a. We therefore expected to find biasing effects for each source that reflected their different relative strengths (i.e., language cue effects are bigger).
   b. These effects should be *smaller* than when found independently (i.e., they are mutually suppressive).
   c. However, when one weighting source is sufficiently stronger than another, it may *completely* suppress the effect of the weaker source.
4) Consistent with other sources of additive bias (Reynolds and Desimone, 2003), we expected that the effects of language cues and action primes would be *additive*. Thus on trials when cue, prime and target were all congruent, we expected the best performance.

Language cues were expected to continue to produce a larger effect than action primes. However, because they were only partially valid they were not necessarily expected to dominate competition. Thus prediction 3c does not apply here (although as it turns out, it does help to explain a later unforeseen result that appeared to arise from the additional influence of a bottom-up source of bias).

METHOD

Participants

Twenty volunteers between 51 and 18 years of age ($M = 22.4$ years) were paid for their participation in a single session that lasted approximately 20 min. All were right-handed, with 18 females and 2 males. All self-reported normal or corrected-to-normal vision and normal motor control, and all were naïve as to the purpose of the study.

Apparatus and stimuli

As Experiments 2 and 3.

Design and procedure

As Experiments 2 and 3, except that Action Primes and Language Cues were *both* partially valid. To accommodate this design, the experiment was twice as long, with 240 trials consisting of four blocks of 60 trials, with each of the 60 change detection scenes being shown in a random order within each block (overall, 8 conditions × 30 replications). Eight conditions arose from the orthogonal variation of three within-subjects variables, each with two levels: a) Language Cue (1: large or 2: small), b) Action Prime (1: power or 2: precision), and c) Target Size (1: large or 2: small). These were as follows: 1: a1, b1, c1; 2: a1, b1, c2; 3: a1, b2, c1; 4: a1, b2, c2; 5: a2, b1, c1; 6: a2, b1, c2; 7: a2, b2, c1; 8: a2, b2, c2.

RESULTS AND DISCUSSION

3.6% of trials were removed as errors (0.38% response errors, 3.25% change identification errors, 0.02% both errors on same trial). No further analysis of errors was undertaken. 4.45% of the remaining trials were removed as outliers, reducing the maximum detection time from 56,635 to 16,975 ms ($M = 4{,}203$ ms; SD = 2,305).

Coarse-grained analysis

A coarse-grained analysis was performed as a first look at this more complex data set. Mean RTs were computed from the remaining data for each participant in each of four conditions of target congruence: valid cue + valid prime (e.g., both were target-congruent); valid cue + not valid prime; not valid cue + valid prime; not valid cue + not valid prime. These means were subjected to a repeated measures ANOVA with the within-subjects factors of Cue–Target congruency (congruent or incongruent) and Prime–Target congruency (congruent or incongruent).

Isolating the effects of language cues and action primes. In line with prediction 3a above, separate biasing effects were found for each source as expected, and the stronger biasing source of Language Cue produced the larger effect of the two (by a factor of 15). These biasing effects were reflected by main target-congruency effects of Language Cue, $F(1, 19) = 75.582$, $p < 0.001$; and of Action Prime, $F(1, 19) = 5.207$, $p < 0.05$. Mean change detections were faster for cue-congruent targets (3,229 ms) than for cue-incongruent targets (5,131 ms); and they were faster for prime-congruent targets (4,122 ms) than for prime-incongruent targets (4,239 ms).

According to prediction 3b, each effect should be smaller here, than when it was found alone. In order to establish whether this was the case for language cues, cropped correct response data for this experiment and Experiment 1 (where language cues were the only source of bias) was split by participant and cue–target congruent and incongruent trials. From this a mean effect size for each participant in each experiment was calculated, and these data were compared in a one-tailed independent samples t test. This analysis revealed that the mean effect size of language cues was indeed significantly smaller when it was a shared rather than only source of bias (current experiment: shared source = 1,901 ms, Experiment 1: single source = 2,579 ms), $t(38) = 2.100$, $p < 0.05$. Similarly for action primes, the mean effect size was also significantly smaller when it was a shared rather than only source of bias (current experiment: shared source = 131 ms, Experiment 1b of Symes et al., 2008: single source = 372 ms), $t(40) = 2.105$, $p < 0.05$.

Additive effects of language cues and action primes. Finally, according to prediction 4 above, the biasing effects of language cue and action prime should be additive rather than interactive. The direction of means across the four conditions fully supported an additive model, with detections being driven by valid language cues whilst nevertheless benefiting from concurrently valid action primes (see "All targets" column of **Table 2**). The ANOVA output also supported an additive model, given that there was no significant interaction between Language Cue and Action Prime, $F(1, 19) = 1.074$, $p > 0.10$.

Finer-grained analysis

In keeping with the condition-specific analyses performed for previous experiments, in this finer-grained analysis the condition means of the cropped data were computed for each participant and subjected to a repeated measures ANOVA with the within-subjects factors of Language Cue (large or small); Action Prime (power or precision) and Target Size (large or small).

A main effect of target size. A main effect of Target Size, $F(1, 19) = 5.440$, $p < 0.05$, revealed faster mean change detections for small (4,112 ms) rather than large (4,247 ms) targets. This is a somewhat counter-intuitive finding, since one might expect larger objects to be more salient. Indeed, in testing predictions from the biased competition model, Proulx and Egeth (2008) reported evidence from a singleton paradigm suggesting that similar to increased luminance contrast, increased size contrast also biased competition. Nevertheless, it does seem that smaller objects were genuinely more salient than larger ones *in the specific context of the change detection scenes used here*. Indeed, using the same scenes, Symes et al. (2008) found a robust and reliable advantage for small targets across several experiments – including

Table 2 | Rank ordered RTs (ms) presented with details of their experimental conditions.

Rank	Language Cue	Action Prime	All targets	Small targets	Large targets
1	Valid	Valid	3,143	3,198	3,088
2	Valid	Not valid	3,316	3,243	3,388
3	Not valid	Valid	5,101	4,954	5,246
4	Not valid	Not valid	5,162	5,055	5,266

an eye-tracking experiment that revealed preferential fixating of smaller objects (see Symes et al., 2008, for a possible explanation for this). The crucial question here then, is why these apparently salient smaller objects exerted an influence on detection times in this current experiment (and repeatedly in Symes et al., 2008), and yet they did *not* do so in the preceding three experiments?

What each of the preceding three experiments shared in common were *relatively stronger* sources of top-down bias than were present in this experiment and those of Symes et al. (2008). In Experiment 1, the top-down bias came from partially valid language cues (which were at their most influential, being the only source of bias). In Experiment 2, sources of bias were *completely valid* language cues with partially valid action primes, and in Experiment 3, *completely valid* action primes with partially valid language cues. Relative to these three experiments, top-down sources of bias in the current experiment were at their weakest (*two* partially valid sources). Similarly, top-down sources of bias in Symes et al. (2008) were relatively weak too, always being partially valid action primes. With these cases of relatively weak top-down biases, we argue that *another* source of bias (i.e., a bottom-up bias of small objects) was able to successfully compete for neuronal representation. In the previous three experiments, this relatively weak bottom-up bias had presumably been unable to exert an influence in the context of stronger concurrent top-down biases that dominated competition (see prediction 3c above).

Isolating the effects of language cues and action primes. As was the case with the earlier course-grained analysis, the effects of language cues and action primes supported prediction 3a. Target Size interacted separately with both sources of top-down bias. Language Cue by Target Size, $F(1, 19) = 75.341, p < 0.001$, revealed that mean detection times were faster for large targets following a large (3,238 ms) rather than small (5,256 ms) cue, and faster for small targets following a small (3,220 ms) rather than large (5,005 ms) cue. As already reported above, this effect was significantly smaller as a shared rather than only source of bias (current experiment: shared source = 1,901 ms, Experiment 1: single source = 2,579 ms), $t(38) = 2.100, p < 0.05$.

Action Prime by Target Size, $F(1, 19) = 5.108, p < 0.05$, revealed that mean detection times were faster for large targets following a large (4,167 ms) rather than small (4,327 ms) prime, and faster for small targets following a small (4,076 ms) rather than large (4,149 ms) prime. As already reported above, this effect was also significantly smaller when it was a shared rather than only source of bias (current experiment: shared source = 131 ms, Experiment 1b of Symes et al., 2008: single source = 372 ms), $t(40) = 2.105$, $p < 0.05$.

Additive effects of language cues and action primes. Finally, according to prediction 4 above, the effects of language cue and action prime should be additive rather than interactive. The direction of means across the eight conditions supported an additive model, with detections again being driven by valid language cues whilst nevertheless benefiting from concurrently valid action primes (see "Small and Large targets" columns of **Table 2**).

The ANOVA output revealed a three-way interaction between Language Cue, Action Prime and Target Size that was statistically significant at the 10% level, $F(1, 19) = 3.840, p = 0.065$. In examining this interaction, the biasing effect of action primes appeared stronger under one level of Language Cue – namely "large" cues. Thus with *large language cues* the effect size was 200 ms, with mean detection times that were faster for large targets following a large (3,088 ms) rather than small (3,388 ms) action prime, and faster for small targets following a small (4,954 ms) rather than large (5,055 ms) prime. However, with *small language cues* the effect size was only 33 ms, with mean detection times that were similar for large targets following large and small action primes (5,246 and 5,266 ms respectively), and similar for small targets following large and small action primes (3,243 and 3,198 ms respectively).

Interestingly, this finding makes good sense from the perspective of biased competition, and it does not contradict an additive model. We break down our explanation into two related parts:

1) Recall the main effect of Target Size reported earlier – salient small targets were detected faster overall. It was suggested that this bottom-up bias was able to exert a small influence in this experiment because the concurrent sources of top-down bias were relatively weak. This biasing effect of stimulus salience also appeared to be an additive effect. Indeed, the language cue biasing effect was 269 ms larger when cues were congruent rather than incongruent with this visually preferred small stimulus. Similarly, the prime–target effect was 87 ms larger when primes were congruent rather than incongruent with this visually preferred small stimulus.
2) When the effects of stimulus salience, language cues and action primes concurrently contribute to biasing competition between objects for neuronal representation, certain combinations may result in the effects of one source being heavily suppressed – even to the point where it no longer has a biasing influence of its own (see prediction 3c above). Indeed, previous findings from Experiment 2 (and see also Symes et al., 2009) indicated that dominant biasing signals completely suppressed the weaker effect of action primes. This appears to have been the case here too. In particular, when the biasing influence of salient small objects co-occurred with the biasing influence of a specific language cue ("Small"), their combined influence dominated, such that action primes could no longer exert any real influence. Indeed, the three-way interaction reported above revealed exactly this pattern – the biasing effect of action primes barely arose under "Small" language cues (in fact as the distributional analysis below reveals, it was not significant in any bin).

Distributional analyses. In order to see whether the weaker biasing effect of *action primes* was consistent across different portions of the RT distribution, distributional analyses were performed (see Experiment 1 for procedural details). Separate ANOVAS for each bin were performed under each level of Language Cue. Under "Small" language cues, no significant interactions between Action Prime and Language-cued Target Size were observed in any bins ($p > 0.10$). Under "Large" language cues, signs of an action prime biasing effect began in the first bin, and reached statistical significance in the second and third bins only ($p < 0.05$). By contrast, the more robust biasing effect of *language cues* was statistically significant ($p < 0.001$) in each bin under each level of Action Prime (except for the fifth bin under precision primes, $p = 0.063$).

Overall, these results comprehensively supported the predictions generated by the biased competition model. Firstly, each source of top-down bias produced its own biasing effect, with language cues producing the larger effect (prediction 3a). Through the suppressive interactions of biased competition, each of these effects was significantly *smaller* than when found alone (prediction 3b). Furthermore, the two biasing effects of language cues and action primes seemed to be additive (prediction 4). This was transparently the case in the initial course-grained analysis, and it was also the case following a more careful examination in the finer-grained analysis. Here, it was found that a third source of bias (visually salient *small* objects) had exerted its own bottom-up influence. The combined influence of this bottom-up bias (which also seemed to be an additive effect) and "Small" language cues dominated competition, such that action primes could no longer exert much of an influence (prediction 3c).

GENERAL DISCUSSION

Relatively little is known about how multiple sources of bias might interact. It is known that language cues presented with objects can influence the kinematics of actions directed to those objects (e.g., Gentilucci et al., 2000; Gentilucci, 2003; Lindemann et al., 2006). Superimposing the word "large" on an object, for instance, results in increased maximum grip aperture (Glover and Dixon, 2002). Relatedly, sentence comprehension appears to evoke motor representations – Glenberg and Kaschak (2002) reported that sentence judgments were faster when the required action response (e.g., moving the hand away from or toward the body) matched the actions implied by the sentence. These insights fit well with broader accounts of embodied cognition that suggest that various sources of activation (whether semantic, visual, motoric) may trigger perceptuo-motor simulations (e.g., Barsalou, 2008, 2009).

The biased competition model is an influential theory of attention proposing that objects compete for neuronal representation via mutually suppressive interactions (Desimone and Duncan, 1995). Various top-down and bottom-up factors can bias competition, and some authors have recently suggested that one such top-down factor might be action intentions (e.g., Bekkering and Neggers, 2002). The current study explored the effects of two sources of top-down bias on visual object representation – one explicit (language cues) and the other implicit (action intentions).

OVERVIEW OF RESULTS

Using a change-detection flicker paradigm, participants searched for an unknown identity-changing target amongst 12 graspable objects in a photographed array (half were small objects like cherries, half were larger objects like apples). Prior to the onset of the scene, participants received a language cue that advised them to search for the change in "larger" or "smaller" objects, and an action prime that established an action intention to make a power or precision grip (grips that were congruent with large and small objects respectively). Language cues and action primes were either relatively weak sources of bias (partially valid) or stronger sources of bias (completely valid). Experiment 1b of Symes et al. (2008) has previously established enhanced detections following partially valid action primes, and in the current study, four experiments tested a further variety of cue/prime/validity combinations (Experiment 1: partially valid language cues; Experiment 2: completely valid language cues + partially valid action primes; Experiment 3: partially valid language cues + completely valid action primes; Experiment 4: partially valid language cues + partially valid action primes). The predictions derived from the biased competition model, and their related results, are summarized below (and effect sizes across all experiments are displayed graphically in **Figure 2**).

1) We expected proxy evidence of the single top-down source of language cues biasing competition, with faster detections on trials where cue and target were size-congruent (i.e., valid trials).
   - Experiment 1 (partially valid language cues) found faster detections on trials where cue and target were size-congruent.
2) Since language cues tell participants where to look, they should have a strong biasing influence (like directed attention does). Relatedly, the biasing effect of action primes (that arises without directed attention) is expected to be weaker.
   - A cross-experimental comparison of the effects of partially valid action primes (Experiment 1b of Symes et al., 2008) and the effects of partially valid language cues (Experiment 1) revealed that language cues had a significantly larger biasing effect than action primes.
3) With two concurrent top-down weighting sources, we expected that their effects on object representations would compete.
   a. We therefore expected to find biasing effects for each source that reflected their relative strengths (i.e., stronger language cue effects).
      - Experiments 2–4: Language cues had a larger effect than action primes.

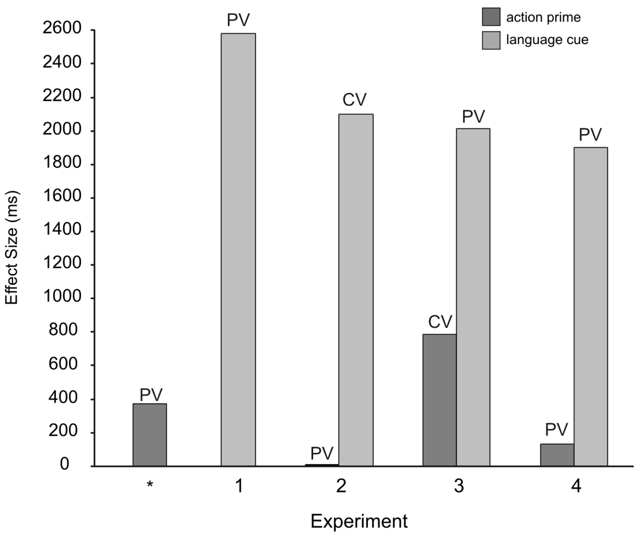


FIGURE 2 | A summary graph of mean effect sizes (mean incongruent RTs – mean congruent RTs) across all experiments. PV, partially valid cue/prime; CV, completely valid cue/prime; * refers to Symes et al. (2008, Experiment 1b).

b. These effects should be *smaller* than when found independently (i.e., than when there is only weighting source).
   - Experiments 2–4: The biasing effects of partially valid cues and primes were significantly smaller than when found alone. [There were no available experimental comparison cases for completely valid cues and primes].

c. However, when one weighting source is sufficiently stronger than another, it may *completely* suppress the effect of the weaker source.
   - Experiments 1–3: The additive effects of language cues and action primes completely suppressed a weaker bottom-up effect of small object saliency.
   - Experiment 2: Completely valid language cues completely suppressed a weaker effect of action primes.
   - Experiment 4: The additive effects of small language cues and small object saliency completely suppressed a weaker effect of action primes.

4) Consistent with other sources of additive bias (Reynolds and Desimone, 2003), we expected that the effects of language cues and action primes would be additive.
   - Experiment 4: Both course-grained and finer-grained analyses supported an additive model (see **Table 2**).

THEORETICAL IMPLICATIONS

As the above summary makes clear, the various tenets of the biased competition model accounted for all degrees of biasing influence derived from action intentions, including when they produced no effect. While some authors have similarly proposed that biased competition may be the mechanism that underlies cases of enhanced visual processing following action intentions (e.g., Bekkering and Neggers, 2002; Hannus et al., 2005; Symes et al., 2008, 2009), other authors have proposed alternative models. Most recently, Wykowska et al. (2009) have suggested combining an intentional weighting mechanism (e.g., Hommel et al., 2001) with the guided search model (e.g., Wolfe, 1994) and the dimensional weighting account (e.g., Müller et al., 1995). In explaining the absence of an action-related biasing effect when selection could be based on bottom-up saliency signals alone (cf. similar results of Symes et al., 2009), Wykowska et al. (2009, p. 1767) suggested that,

"Only if a task-relevance bias occurs, will the action-related weighting also influence perceptual processing. In such a case, bottom-up processing will be modulated by the common weight combining task-relevance and action relevance."

Given the results of the current study, we suggest that the mechanism of biased competition is a sufficient and simpler means of explaining a null-effect of action when stimulus salience is high. We argue this is the case for feature weighting, which our data apply to, although it may also apply to dimension weighting (indeed, Wykowska et al., 2009 suggest that the mechanism underlying dimension weighting may be the same one hypothesized to account for other top-down effects on visual selection). Under our preferred account of this mechanism, there is no common weight that inputs to a master map of activation; rather, ongoing suppressive interactions between objects take place across the various brain regions that represent visual information, (sensory, motor, cortical, and subcortical), Beck and Kastner (2009). Sometimes a particularly strong weighting signal will dominate competition between objects and completely suppress the effects of other weaker signals (e.g., action signals). Nevertheless, the nature of biased competition is such that *any* weighting signal, whether top-down or bottom-up, action-based or language-based, competes to influence perceptual processing (indeed, top-down and bottom-up signals seem to produce very similar neuronal responses, Reynolds and Desimone, 2003; Reynolds and Chelazzi, 2004). Thus action-based sources of bias are not assumed to be "special cases" that only have a modulatory influence that is conditional on higher-order goals (such as a task-relevance bias). To qualify this further, in line with the findings of Reynolds and Desimone (2003) the current results suggested that the various biasing effects were additive, and action-related effects in Experiment 4 for example, were not dependant on task-relevance. Instead they occurred with and without a task-relevance bias (i.e., alongside valid and non-valid language cues).

CONCLUSIONS

Itti (2007, p. 93) captures the essence of the demands that the visual world places on animals (and robots), when he writes;

"Visual processing of complex natural environments requires animals to combine, in a highly dynamic and adaptive manner, sensory signals that originate from the environment (bottom-up) with behavioral goals and priorities dictated by the task at hand (top-down)."

In examining the influences of differently weighted bottom-up and top-down signals, the current series of behavioral experiments revealed a sensitive hierarchy of predicted attentional effects. Such findings serve a "proof-of-principle" role for scientists interested in modeling an embodied neuro-robotic system:

Firstly, the behavioral data suggest that selective perceptual enhancement may be initiated by manual action plans, such as grasping. Although it is perhaps surprising that simply intending to perform an action (even when it is not directed to a known target) might have such diverse influences on an embodied system, complementary neurological evidence does exist. Electrical stimulation of premotor sites within monkey frontal eye fields for example, initiated a bias in the strength of visual signals in corresponding sites of extrastriate visual cortex (Moore and Armstrong, 2003). Recent advances in fMRI methods too, shed further light on the role of different brain areas such as pre-frontal cortex, involved in modulating visual signals (Grill-Spector and Sayres, 2008).

Secondly, the modulatory influence of action planning appeared to integrate with other sources of bias (such as language) through biased competition – a neural mechanism that is sufficiently well-defined for modeling. Indeed, various neural implementations of biased competition have already simulated a wide range of attentional effects that accommodate both top-down and bottom-up influences (e.g., Sun and Fisher, 2003; Lanyon and Denham, 2004a,b; Deco and Rolls, 2005; see also Spratling, 2008a for a review). While sharing similarities with other influential models of visual processing, the physiologically plausible neural architecture of the biased competition model does recommend it (Spratling, 2008a,b). Indeed, it may be more parsimonious than the influential class of saliency map models (e.g., Wolfe, 1994; Itti and Koch, 2001) in three key areas – it does not require a single map for competition to ultimately be resolved, since ongoing competition occurs across a distributed network of interacting brain regions; it does not assume separate preattentive and attentive stages of perceptual processing; and it does not require separate neural pathways for processing saliency and featural information (see Spratling, 2008b for a discussion of these differences). The current behavioral findings therefore recommend future implementations of the biased competition model in robots, and that these consider including action intentions as a form of top-down bias that reflects the behavioral goals of the robot.

ACKNOWLEDGMENTS

This work was supported by a grant from the Economic and Social Research Council (RES-000-23-1497).

REFERENCES

Arbib, M., Metta, G., and van der Smagt, P. (2008). "Neurorobotics: from vision to action," in *Springer Handbook of Robotics*, eds B. Siciliano and O. Khatib (Berlin/Heidelberg: Springer), 1453–1480.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.* 59, 617–645.

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 364, 1281–1289.

Beck, D. M., and Kastner, S. (2009). Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision Res.* 49, 1154–1165.

Bekkering, H., and Neggers, S. F. (2002). Visual search is modulated by action intentions. *Psychol. Sci.* 13, 370–374.

Craighero, L., Fadiga, L., Rizzolatti, G., and Umiltà, C. (1999). Action for perception: a motor–visual attentional effect. *J. Exp. Psychol. Hum. Percept. Perform.* 25, 1673–1692.

Deco, G., and Rolls, E. T. (2005). Neurodynamics of biased-competition and cooperation for attention: a model with spiking neurons. *J. Neurophysiol.* 94, 295–313.

Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 353, 1245–1255.

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222.

Duncan, J. (1998). Converging levels of analysis in the cognitive neuroscience of visual attention. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 353, 1307–1317.

Fagioli, S., Hommel, B., and Schubotz, R. I. (2007). Intentional control of attention: action planning primes action-related stimulus dimensions. *Psychol. Res.* 71, 22–29.

Fischer, M. H., and Hoellen, N. (2004). Space-based and object-based attention depend on motor intention. *J. Gen. Psychol.* 131, 365–377.

Gentilucci, M. (2003). Object motor representation and language. *Exp. Brain Res.* 153, 260–265.

Gentilucci, M., Benuzzi, F., Bertolani, L., Daprati, E., and Gangitano, M. (2000). Language and motor control. *Exp. Brain Res.* 133, 468–490.

Glenberg, A. M., and Kaschak, M. P. (2002). Grounding language in action. *Psychon. Bull. Rev.* 9, 558–565.

Glover, S., and Dixon, P. (2002). Semantics affects the planning but not the control of grasping. *Exp. Brain Res.* 146, 383–387.

Grill-Spector, K., and Sayres, R. (2008). Object recognition: Insights from advances in fMRI methods. *Curr. Dir. Psychol. Sci.* 17, 73–79.

Hannus, A., Cornelissen, F. W., Lindemann, O., and Bekkering, H. (2005). Selection-for-action in visual search. *Acta Psychol.* 118, 171–191.

Hommel, B., Müsseler, J., Aschersleben, G., and Prinz, W. (2001). The theory of event coding (TEC). *Behav. Brain Sci.* 24, 849–937.

Itti, L. (2007). "Computational cognitive neuroscience and its applications," in *Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2007 Symposium* (Washington, DC: The National Academies Press), 87–98.

Itti, L., and Koch, C. (2001). Computational modeling of visual attention. *Nat. Rev. Neurosci.* 2, 194–203.

Kastner, S., De Weerd, P., Desimone, R., and Ungerleider, L. G. (1998). Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. *Science* 282, 108–111.

Lanyon, L. J., and Denham, S. L. (2004a). A model of active visual search with object-based attention guiding scan paths. *Neural Netw.* 17, 873–897.

Lanyon, L. J., and Denham, S. L. (2004b). A biased competition computational model of spatial and object-based attention mediating active visual search. *Neurocomputing* 58–60, 655–662.

Lindemann, O., and Bekkering, H. (2009). Object manipulation and motion perception: evidence of an influence of action planning on visual processing. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 1062–1071.

Lindemann, O., Stenneken, P., van Schie, H. T., and Bekkering, H. (2006). Semantic activation in action planning. *J. Exp. Psychol. Hum. Percept. Perform.* 32, 633–643.

Linnell, K. J. and Humphreys, G. W., McIntyre, D. B., Laitinen, S., and Wing, A. M. (2005). Action modulates object-based selection. *Vision Res.* 45, 2268–2286.

Luck, S. J., Chelazzi, L., Hillyard, S. A., and Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *J. Neurophysiol.* 77, 24–42.

Moore, T., and Armstrong, K. M. (2003). Selective gating of visual signals by microstimulation of frontal cortex. *Nature.* 421, 370–373.

Müller, H. J., Heller, D., and Ziegler, J. (1995). Visual search for singleton feature targets within and across feature dimensions. *Percept. Psychophys.* 57, 1–17.

Proulx, M. J., and Egeth, H. E. (2008). Biased-competition and visual search: the role of luminance and size contrast. *Psychol. Res.* 72, 106–113.

Ratcliff, R. (1979). Group reaction time distribution and an analysis of distribution statistics. *Psychol. Bull.* 86, 446–461.

Rensink, R. A. (2005). "Change blindness," in *Neurobiology of Attention*, eds L. Itti, G. Rees, and J. K. Tsotsos (San Diego, CA: Elsevier), 76–81.

Rensink, R. A., O'Regan, J. K., and Clark, J. J. (1997). To see or not to see: the need for attention to perceive changes in scenes. *Psychol. Sci.* 8, 368–373.

Reynolds J. H., and Chelazzi, L. (2004). Attentional modulation of visual processing. *Annu. Rev. Neurosci.* 27, 611–647.

Reynolds, J. H., Chelazzi, L., and Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *J. Neurosci.* 19, 1736–1753.

Reynolds J. H., and Desimone, R. (2003). Interacting roles of attention and visual salience in V4. *Neuron* 37, 853–863.

Reynolds, J. H., Pasternak, T., and Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron* 26, 703–714.

Schütz-Bosbach, S., Haggard, P., Fadiga, L., and Craighero, L. (2007). "Motor cognition: TMS studies of action generation," in *Oxford Handbook of Transcranial Stimulation*, eds E. Wassermann, C. Epstein, U. Ziemann, V. Walsh, T. Paus, and S. Lisanby (Oxford, UK: Oxford University Press).

Simons, D. J., and Rensink, R. A. (2005). Change blindness: past, present, and future. *Trends Cogn. Sci.* 9, 16–20.

Spratling, M. W. (2008a). Reconciling predictive coding and biased competition models of cortical function. *Front. Comput. Neurosci.* 2, 1–8. doi: 10.3389/neuro.10.004.2008.

Spratling, M. W. (2008b). Predictive coding as a model of biased competition in visual attention. *Vision Res.* 48, 1391–1408.

Sun, Y., and Fisher, R. (2003). Object-based attention for computer vision. *Artif. Intell.* 146, 77–123.

Symes, E., Ottoboni, G., Tucker, M., Ellis, R., and Tessari, A. (2009). When motor attention improves selective attention: the dissociating role of saliency. *Q. J. Exp. Psychol.* doi: 10.1080/17470210903380806.

Symes, E., Tucker, M., Ellis, R., Vainio, L., and Ottoboni, G. (2008). Grasp preparation improves change-detection for congruent objects. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 854–871.

Welsh, T., and Pratt, J. (2008). Actions modulate attention capture. *Q. J. Exp. Psychol.* 61, 968–976.

Wolfe, J. M. (1994). Guided search 2.0: a revised model of visual search. *Psychon. Bull. Rev.* 1, 202–238.

Wykowska, A., Schubö, A., and Hommel, B. (2009). How you move is what you see: action planning biases selection in visual search. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 1755–1769.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 23 December 2009; paper pending published: 13 March 2010; accepted: 21 May 2010; published online: 07 June 2010.

Citation: Symes E, Tucker M and Ottoboni G (2010) Integrating action and language through biased competition. Front. Neurorobot. ***4****:9. doi: 10.3389/fnbot.2010.00009*

frontiers in
NEUROROBOTICS

REVIEW ARTICLE
published: 14 June 2010
doi: 10.3389/fnbot.2010.00003

Sentence comprehension: effectors and goals, self and others. An overview of experiments and implications for robotics

Anna M. Borghi[1,2]*, *Claudia Gianelli*[1] *and Claudia Scorolli*[1]

[1] Department of Psychology, University of Bologna, Bologna, Italy
[2] Institute of Sciences and Technologies of Cognition, National Research Council, Rome, Italy

Edited by:
Angelo Cangelosi,
University of Plymouth, UK

Reviewed by:
Gert Westermann,
Oxford Brookes University, UK
Giovanni Pezzulo, Consiglio Nazionale delle Ricerche, Italy

Correspondence:
Anna M. Borghi, Dipartimento di Psicologia, Viale Berti Pichat 5, 40127 Bologna, Italy.
e-mail: annamaria.borghi@unibo.it

According to theories referring to embodied and grounded cognition (Barsalou, 2008), language comprehension encompasses an embodied simulation of actions. The neural underpinnings of this simulation could be found in wide neural circuits that involve canonical and mirror neurons (Rizzolatti et al., 1996). In keeping with this view, we review behavioral and kinematic studies conducted in our lab which help characterize the relationship existing between language and the motor system. Overall, our results reveal that the simulation evoked during sentence comprehension is fine-grained, primarily in its sensitivity to the different effectors we employ to perform actions. In addition, they suggest that linguistic comprehension also relies on the representation of actions in terms of goals and of the chains of motor acts necessary to accomplish them. Finally, they indicate that these goals are modulated by both the object features the sentence refers to as well as by social aspects such as the characteristics of the agents implied by sentences. We will discuss the implications of these studies for embodied robotics.

Keywords: sentence comprehension, embodied cognition, robotics, action, motor system, language, action goals, social cognition

INTRODUCTION

According to theories of embodied and grounded cognition (from here on EC theories), language is grounded in the sensorimotor system. In this sense, the same sensorimotor and emotional systems are supposed to be involved during perception, action and language comprehension. More specifically, language comprehension would involve an embodied simulation, whose neural underpinnings are to be found in wide neural circuits, crucially involving canonical and mirror neurons (Rizzolatti et al., 1996; Gallese, 2008). In cognitive neuroscience the notion of simulation has been defined differently (for a more detailed analysis of this, see Borghi and Cimatti, 2010; for a review, see Decety and Grezes, 2006). Here we define simulation, with Jeannerod (2007), as the offline recruitment (for instance, during language processing) of the same neural networks involved in perception and action. In addition, we qualify it, as did Gallese (2009), as an embodied and automatic mechanism, which allows us to understand others' behaviors. The automaticity of this process does not imply an intentional strategy to understand intentions and mental states. In keeping with these views, the underlying assumption of our work is that the activation of motor and sensorimotor cortices is not just a side-effect but effectively contributes to language comprehension. In this paper we review behavioral and kinematics studies conducted in our lab which help to characterize the relationship existing between language and the motor system (see also Scorolli et al., 2009). We will focus on studies utilising simple sentences composed for example by a verb and a noun. In the final part of the paper we discuss why we believe these studies have implications for embodied robotics. Further, we will claim that embodied robotics can contribute critically to psychology and neuroscience and can promote more detailed predictions on some critical issues.

THE EXPERIMENTS

SENTENCE COMPREHENSION, SIMULATION AND EFFECTORS

Several recent studies have provided evidence of the involvement of the premotor cortex in reading and hearing action words and action sentences (Aziz-Zadeh and Damasio, 2008). Tettamanti et al. (2005) conducted an fMRI study illustrating that a complex fronto-parietal circuit is activated when presenting sentences describing actions performed with the mouth, the hand or the foot. Within this circuit a critical role seems to be assumed by Broca's area, but in a way that extends the traditional linguistic role of this area. In fact, Broca's area is found to be crucially involved in language processing, as well as in action observation. Pulvermüller et al. (2001) found topographical differences in the brain activity patterns generated by verbs referring to different effectors (mouth, legs, arms: e.g. lick, kick, pick); these differences emerged quite early, starting 250 ms after word onset. This very fast activation, its automaticity and its somatotopic organization render it unlikely that information is first transduced in an abstract format and later influences the motor system, as claimed by critiques of the embodied view. In particular, the early activation of the motor system strongly suggests that this activation is an integrant part of the comprehension process rather than only a by-product of it, or an effect of late motor imagery. Further studies utilising a variety of techniques (fMRI, MEG, etc.) support the hypothesis that action verb processing quickly produces a somatotopic activation of the motor and premotor cortices (e.g. Hauk et al., 2004; Pulvermüller et al., 2005). In line with these results, Buccino et al. (2005) designed a TMS study that showed an amplitude decrease of MEPs recorded from hand muscles when listening to hand-action related sentences, and from foot muscles when listening to foot related sentences. This confirms a somatotopic recruitment of motor areas. As reported in the meta-analysis performed by Jirak, Menz, Borghi and Binkofski (under review), the involvement of motor areas in language processing is consistent over tasks and subjects (for a more critical view, see Willems and Hagoort, 2007). In particular, word and sentence processing involves a variety of brain regions, including parietal, temporal, and frontal, but also cerebellar activity, and, even if the right hemisphere is also activated, there is a clear predominance of activations in the (language and motor areas of the) left hemisphere. In addition, the results of the meta-analysis highlight areas presumably containing mirror neurons in humans, more specifically Broca's region, which may be described as the human homolog of the monkey premotor cortex (Rizzolatti and Craighero, 2004).

We will now describe studies performed in our lab, as they extend the previous behavioral evidence. Here we have illustrated that during language comprehension we are sensitive to the distinction between hand and mouth sentences, and between foot and mouth sentences as well.

In the first study we performed two experiments in which 40 participants read simple sentences from a computer screen that were composed of a verb in the infinitive form followed by an object noun (Scorolli and Borghi, 2007). The sentences referred to either hand, mouth or foot actions. The hand sentences represented the baseline: thus, the same noun was presented after either a foot or hand verb (e.g. "to kick the ball", vs. "to throw the ball") or either after a mouth or hand verb (e.g. "to suck the sweet", vs. "to unwrap the sweet"). Overall, we had 24 object nouns, each preceded by two different verbs, for a total of 48 critical pairs. Presenting the same noun after the verb allowed us to be sure that no frequency effect took place. We did not control for the verb frequency, because the verb was presented before we started recording. However, in a pretest we controlled for the association rate between the verb and the noun, as this might influence performance. Eighteen participants were required to produce the first five nouns they associated with each verb; no difference in production means was present between "mouth sentences" and "hand sentences", $p = 0.65$, and between "foot sentences" and "hand sentences", $p = 1$. The timer started after the noun presentation, and participants were required to respond whether the verb–noun combination made sense or not. Yes responses were recorded either with the microphone or with a pedal. We found a facilitation effect in responses to "mouth sentences" and "foot sentences" compared with "hand sentences" when the effectors – mouth and foot – involved in the motor response and in the sentence were congruent (**Figure 1**). More specifically, participants responding with the microphone were faster with mouth than with hand sentences, $p < 0.01$ (**Figure 1A**), whereas the difference between foot and hand reached significance but was far less marked, $p < 0.05$ (**Figure 1B**). Participants using the pedal responded faster to foot than to hand sentences, $p < 0.0005$ (**Figure 1B**), whereas the difference between hand and mouth sentences was not significant, $p < 0.8$ (**Figure 1A**). These results suggest, in line with the literature, that the simulation activated during sentence comprehension is sensitive to the kind of effector implied by the sentence. In previous behavioral studies only foot and hand sentences were compared; our study extends previous results as we found a difference between mouth and hand sentences as well.

In a further study (Borghi and Scorolli, 2009) we found that the simulation is sensitive not only to the kind of effector (mouth vs. hand, foot vs. hand), but also to the specific effector (right vs. left hand) used to respond. We performed five experiments with the same sentence presentation modality and task used in Scorolli and Borghi (2007); 97 right-handed participants were asked to decide whether verb–noun combinations made sense or not. We analyzed both combinations which made sense (e.g. "to kick the ball") and combinations which did not make sense (e.g. "to melt the chair"). Here we will focus on Experiments 1, 2, and 3, as Experiment 4 was

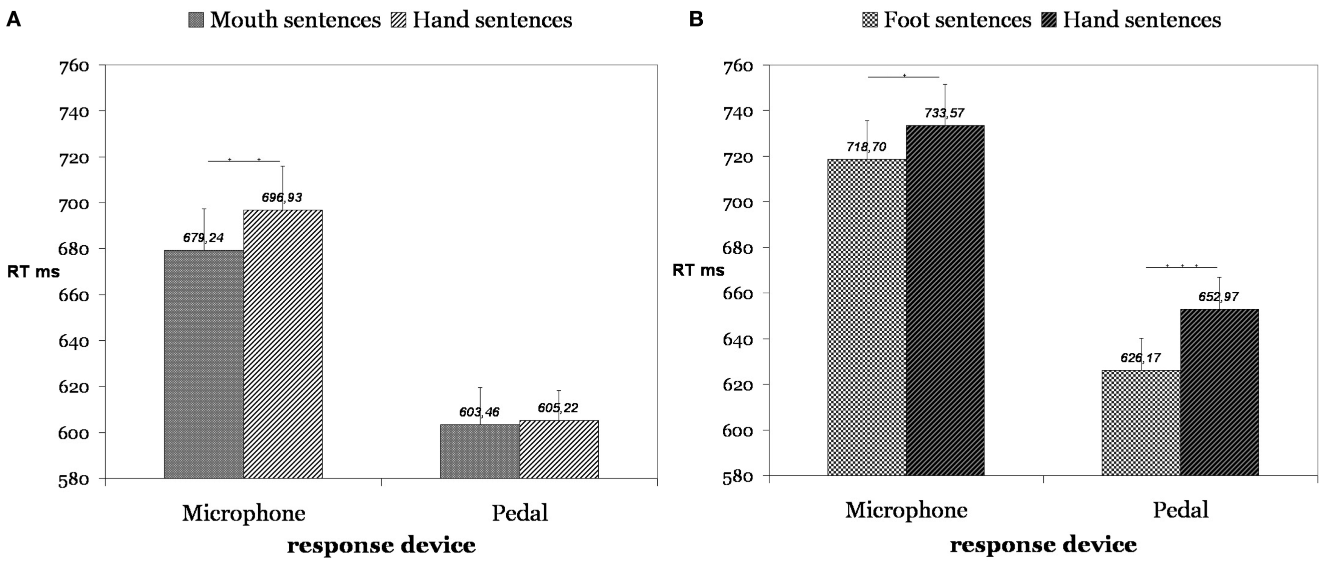


FIGURE 1 | Participants using the microphone responded with greater speed to "mouth sentences" than to "hand sentences" (A), $p < 0.01$. Symmetrically, participants who used the pedal as responding device were significantly faster for "foot sentences" than for "hand sentences" **(B)**, $p < 0.\ 0005$.

a control one. In Experiments 1a,b we used only manual sentences, in Experiment 2 hand and mouth sentences, in Experiment 3 hand and foot sentences. Responses to hand sentences (Experiment 1) were faster than responses to non-sense sentences with the right hand, but not with the left hand (**Figure 2A**), as it appeared in the subject analyses and on materials (we will report the *p*-values for both analyses in sequence): $p < 0.05$; $p < 0.0000001$. Importantly, such an advantage of the right over the left hand was not present when sensible sentences were not action ones: $p = 0.99$; $p = 0.75$. The same advantage of the right over the left hand with sensible sentences was present in Experiment 2 (**Figure 2B**), in which both hand and mouth sentences were presented, even if it reached significance only in the analysis on items, $p < 0.0000001$. This suggests that participants simulated performing the action with the dominant hand. Crucially the advantage of the right hand for sensible sentences was not present with foot sentences, with which, probably due to an inhibitory mechanism, the effect was exactly the opposite, as left hand responses were faster than right hand ones with sensible sentences, $p = 0.055$; $p < 0.0000001$ (**Figure 2C**). These results complement the previous findings as they suggest that the motor simulation formed is not only sensitive to different effectors (mouth, hand, foot), but also to the different action capability of the two hands, the left and the right one. The similarity between the responses with hand and mouth sentences can be due to the fact that different effectors can be involved in single actions, and the similarity of the performance obtained by hand and mouth sentences could be due to the fact that hands and mouth are represented cortically in contiguous areas. However, it may also suggest that not only proximal aspects, such as the kind of effector, modulate the motor responses, but also distal aspects, such as the action goal. Consider an action such as sucking a sweet: it probably also activates manual actions such as the action of grasping the sweet and bringing it to the mouth. In sum: it is possible that the similar modulation of the motor response is due to the common goal evoked by hand and mouth sentences (see also Gentilucci et al., 2008).

Overall, the results of these two studies indicate that language processing activates an action simulation that is sensitive to the effector involved. In addition, they suggest that understanding action sentences implies comprehension of the goals that the actions entail. However, further studies are needed, to deepen the role played by action goals (for a recent study focusing on the importance of goals in action organization in monkeys, see Umiltà et al., 2008. The issue of goals will be discussed later). The results described so far report a facilitation effect in case of congruency between the effector implied by the verb/sentence and the effector used to respond. Even if the evidence we found supports the idea that the different effectors (mouth and foot) are activated during language processing, our behavioral results contrast with the results by Buccino et al. (2005), who found an interference effect between the effector involved in the sentence (hand, foot) and the effector involved in the motor response (hand, foot).

Certainly, in both cases there is clear evidence of a modulation of the motor system during sentence comprehension, thus this evidence is certainly in favor of an embodied cognition perspective. However, knowing more precisely the specific timing of this modulation (Boulenger et al., 2006), as well as the details of this modulation, would be crucial for solving a lot of issues. The first issue is that, even if the somatotopic activation of the motor system suggests that the motor system is involved during language comprehension, we do not yet fully understand if the activation of the motor system is necessary for comprehension or whether it is just a by-product of it (Mahon and Caramazza, 2008). A better understanding of the relationships between the comprehension process and motor system activation, both in terms of time-course and processes, would be crucial as it would allow researchers to formulate clearer predictions.

Many interpretations of the discrepancies between the results have been proposed. One possibility is that these discrepancies are due to timing between linguistic stimulus, motor instructions and motor response. It is possible that, when the motor system is activated both for preparing an action with a given effector and for processing action words referring to the same effector, an interference effect takes place due to the contemporary recruitment of the same resources. Later, a facilitation effect might occur (see Chersi et al., 2010). This explanation is in line with evidence on language and motor resonance that has shown that the compatibility effect between action and sentence (ACE, that is the facilitation effect) was present only when the motor

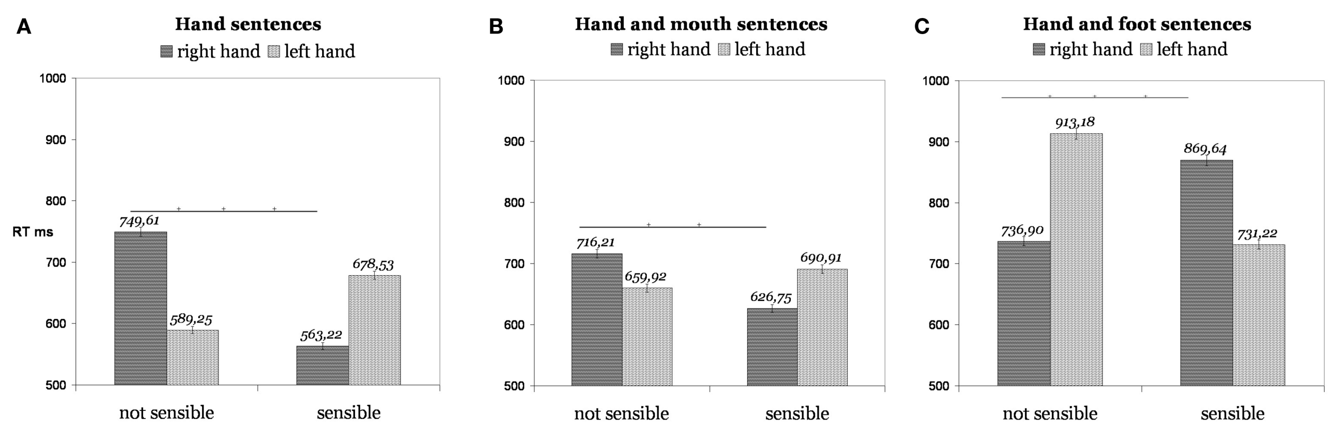


FIGURE 2 | When pairs referred to manual and mouth actions (A, B), participants responded faster with the dominant than with the left hand in case of sensible sentences. When pairs referred to manual and foot actions **(C)** the results were opposite.

instruction was presented simultaneously to the beginning of the sentence rather than after sentence presentation (Borreggine and Kaschak, 2006; Zwaan and Taylor, 2006). In the study by Buccino et al. (2005) participants on presentation of a "go" signal had to respond to the second syllable of a verb preceding a noun; time was measured from this point. Instead in our experiments we didn't use a "go" signal: we first presented a verb, then a noun, and started measuring after the appearance of the noun. Concerning the temporal relationship between language and the motor task, the linguistic stimulus appearance seems to affect not only the movement's speed (reactions times) but also the overt motor behavior, as revealed by detailed analyses of movement kinematics (Boulenger et al. 2006; Dalla Volta et al., 2009). Boulenger et al. (2006) found that when contemporaneously processing language and executing motor tasks, action verbs hinder reaching movements. An interference effect occurred as early as 160–180 ms after word onset when participants started the response movement before word presentation (Experiment 1). On the contrary a priming effect became evident at about 550–580 ms after word onset when the word acted as go-signal for the response movement. Along this line, Dalla Volta et al. (2009) found that there is an early interference effect on the effective movement (kinematics measures) and a late facilitation, detectable through RTs analyses.

Another possibility is that the interference effect is not only aroused by timing but by the interaction between two factors: the temporal overlap and the so called "integrability", that is the degree to which the perceptual input could be integrated into the simulation activated by language. For example, in studies where both sentences and perceptual stimuli were presented, when the perceptual stimuli were abstract and difficult to integrate, an interference effect occurred; otherwise a facilitation effect took place. The difficulty seems to rest on the shared contents between the percept and the simulation of the sentence, and on the temporal overlap (Kaschak et al., 2005; Borreggine and Kaschak, 2006). However, due to the difficulty of integration between perceptual and linguistic stimuli this explanation may be contradicted when accounting for the interference and facilitation effects occurring when using linguistic stimuli.

A further possibility is that these discrepancies are due to the varying paradigms and stimuli used. For example, in some cases tasks requiring superficial processing (e.g. lexical decision tasks) are employed, whereas in other cases tasks requiring deeper semantic processing are used (this position is supported by Sato et al., 2008). More specifically, even in the case of deep semantic processing, results may differ depending on the task at hand. For instance, whether the task requires evaluating the whole sentence (e.g. Scorolli and Borghi, 2007; Borghi and Scorolli, 2009, asked participants to evaluate the sensibility of the verb–noun combination) or the verb (e.g. Buccino et al. required participants to evaluate whether the action verb was abstract or concrete).

A final possibility which should be explored is that the effect emerges differently depending on the type of pronoun used to indicate the agent of the action. In this sense pronouns induce a specific perspective on action, which modulates the motor system. For example, we found that the simulation triggered by the pronouns "I" and "you" have a different effect on kinematics parameters of action. In addition, it is possible that the third person pronoun (see Buccino et al., 2005) may partially activate a simulation, thus relying on more abstract processes.

Overall, further research is necessary to disentangle which mechanisms are underlying interference and facilitation effects. However, we believe that further experimental data are not sufficient. Namely, modeling could help to understand how the process might occur, and might be helpful to propose more detailed and clearer predictions for new experiments. Modeling could help us to understand whether interference and facilitation are two sides of the same coin, or whether they rely on different mechanisms (for an attempt to model interference and facilitation effects, Chersi et al., 2010).

SENTENCE COMPREHENSION, SIMULATION, GOALS AND SOCIAL ASPECTS

In the previous studies we have seen that during language processing we form a simulation sensitive not only to the specific effector, but also to the goal conveyed by the sentence.

Consider for example giving somebody an object: how and to what extent is the action of "giving" represented differently from the action of, say, holding the object? These two actions imply two different goals, and these different goals imply a different chain of motor acts. Namely, in order to hold an object we need to reach and then grasp it, whereas in order to give an object to someone else we need to reach and grasp it, as well as to give it to the other agent involved in the interaction. Thus, in order to pursue the goal it conveys, this "interactive" action implies a longer sequence of chained motor acts.

Goal-relatedness of action has recently received much attention, in particular since Fogassi et al. (2005) demonstrated studying the monkey parietal cortex that motor acts, such as "grasping", are coded according to the specific action (e.g. "grasping for eating" vs. "grasping for placing") in which these acts are embedded. Moreover, this coding is present both when the action is performed and when it is observed, that is a mirror mechanism is involved. The idea that actions have a chained organization has been extended to humans, in particular for what concerns action observation and understanding. Iacoboni et al. (2005) used fMRI to demonstrate the presence of a chained organization that differs depending on the intention of the agent. Other studies have been conducted, showing that impairment of chain organization might be linked to autism spectrum disorder (Cattaneo et al., 2007; Boria et al., 2009; Fabbri-Destro et al., 2009). However, no behavioral task has yet been conducted, demonstrating the importance of chained organization in the normal adult population. Additionally, to our knowledge the only study investigating the extent to which this chained organization is encoded in language was a kinematics study recently performed in our lab, by Gianelli and Borghi (Gianelli and Borghi: I grasp, You give: when language translates actions, submitted), in which we identified different components of verbs (for a similar approach, see Kemmerer, et al., 2008) and distinguished between action verbs (e.g. "to grasp") and interaction verbs (e.g. "to give"). These two kinds of verbs, which differ both for their chained organization and for their goal (acting with an object vs. interacting with another agent), had a different impact on kinematics parameters. That is, participants' response (e.g. reaching–grasping an object) was modulated according to the typical kinematics involved by the actions described by action and interaction verbs. Namely, since interaction verbs describe the interaction with another person, the kinematics in response to interaction verbs is modulated according to an increased requirement for accuracy and precision. That is, the same act of reaching and grasping an object needs to be more accurate when performed in order to give the object to another person, hence performing an additional motor act. Specifically, the deceleration phase is longer. The same effect is found during processing of verbs referring to the same action. This suggests that the chained organization of actions according to more or less interactional goals is translated by language. This chained organization can be reactivated when the motor system is activated, thus similarly contributing to language processing.

The results of this study suggest that sentences referring to actions involving other people (e.g. giving something) are represented differently in comparison to sentences referring to actions involving a relationship between an agent and an object (e.g. holding something). However, this study did not allow us to disentangle whether the difference was due to the different chain of motor events involved in the two actions, or whether it was due to a difference in the social framework the two sentences referred to. To elaborate, do "grasp" and "give" differ at a motor level because of the chain they imply, and the different motor acts used, or do they differ because their "goal", as defined not only by a sequence of motor acts but also by the social dimension in which the action is performed? Namely, in the case of "give" the presence of another person is implied, while in the case of grasp it is not. Hence, their goal and their value differ. Even if the action chain organization characterizes both the canonical and the mirror neuron system, it is possible that, depending on the social framework the sentence describes, there is a different involvement of these two systems. Now consider words referring to objects which differ in valence, to take an example, words such as "nice" or "ugly". Literature on approach/avoidance movements has used a variety of behavioral studies to demonstrate that when we read positive words we are faster in producing a movement with our body; the opposite is true when we read negative words (e.g. Chen and Bargh, 1999; Niedenthal et al., 2005; van Dantzig et al., 2008; Freina et al., 2009).

We conducted three experiments to explore whether the triadic relation between objects, ourselves and other natural and artificial agents modulates the motor system activation during sentence comprehension. We used sentences that referred to nice/ugly objects and to different kinds of recipients (Lugli, Baroni, Gianelli, Borghi and Nicoletti (under review)). Participants were presented with sentences formed by a descriptive part (e.g. the object is attractive/ugly) and by an action part (bring it towards you/give it to another person). Their task consisted of deciding whether the sentence made sense or not by moving the mouse towards or away from their body. In three experiments we manipulated the recipient of the action, which could be "another person", "a table", or "a friend".

Results showed that the direction (away or towards the body) of the movement performed to respond was influenced by the direction of the motion implied by the sentence and the stimuli valence. Crucially, stimulus valence had a different impact depending on the relational context the sentence evoked (action involving another agent or just oneself). We found that, whereas participants tended to move the mouse towards their body when they had to judge actions referring to positive objects, with negative objects the movement varied depending on the action recipient. Namely, when dealing with negative objects participants tended to treat friends as themselves, being equally slow to attract negative objects and to offer them to friends. This was not the case for the recipient "table" and for indistinct "another person". In **Table 1** we compare the effect on RTs of different recipients, "another person", "a table" or "a friend" in the two conditions of giving positive or negative objects.

A further result is worth noting. The paradigm we used in this study allowed us to disentangle information provided by the verb and kinematics information related to the real movement participants were required to produce to respond. Namely, given the experimental design we used, in half of the cases there was a mismatch between the information conveyed by the verb (bring vs. give) and the movement to perform (towards or away from participant's body). Our results showed that the role played by the verb, which defines the action goal, was more important than the role played by the kinematics of the movement. This is in line with the Theory of Event Coding (Hommel et al., 2001), according to which actions are represented in terms of distal aspects, an overall goal, rather than in terms of the proximal ones, and with neurophysiological studies showing that actions are represented in the brain primarily in terms of goals (e.g. Umiltà et al., 2008).

DISCUSSION

Overall, our results suggest that the simulation evoked during sentence comprehension is fine-grained, as it is sensitive both to proximal and to distal information (effectors and goals). Additionally, the results show that actions are represented in terms of goals and of the motor acts necessary to reach them. Finally, they indicate that these goals are modulated by the characteristics of both objects and agents implied by sentences: this is observed due to the difference between actions involving only the self in comparison to those involving others.

Table 1 | Mean response times (RTs, in milliseconds) in the "another person" – "table" – "friend" target/negative object condition and "another person" – "table" – "friend" target/positive object condition.

Experiment	"Another person" – "table" – "friend" target/negative Objects	"Another person" – "table" – " "friend" target/positive objects	difference
Exp. 1 "Another"	1645	1634	11
Exp. 2 – "Table"	1834	1834	0
Exp. 3 – "Friend"	1662	1609	53

We believe that realizing a model of these experiments would be important for understanding the relationships between language and motor system. Namely, modeling could contribute to create a theory of their relationship, which is detailed and advances clear predictions. In this direction, models can help to integrate a variety of different empirical results, obtained with different paradigms and different techniques, within a common framework. However, it is important that models do not only replicate experimental studies, but rather provide general principles and generate predictions to be tested empirically.

One could ask which kinds of models can help to interpret experimental results as the described ones, and help to formulate novel predictions.

Simple feed-forward models are probably not sufficient, as they may not provide an adequate formalization for embodied theories. Namely, feed-forward models are endowed with an input and an output lawyer which strongly resembles the traditional sandwich of dis-embodied theories of cognition. A recurrent network would probably be more suitable to detect the reciprocal influence of perception and action.

On a general level, modeling should respect a variety of constraints (see Caligiore et al., 2009; Caligiore, Borghi, Parisi and Baldassarre (accepted)).

The first kind of constraints are the neurobiological ones. Namely, the model's neural system should be endowed with at least some crucial characteristics of the human neural system. In particular, the neural underpinnings of motor simulations formed during language comprehension are represented by wide neural circuits that – crucially – involve canonical and mirror neurons (Gallese et al., 1996; Rizzolatti and Craighero, 2004). Therefore, the model should be endowed with a simulated neural system which reproduces both canonical and mirror neurons. More specifically, the motor system of the model should be organized in such a way that chains of actions are implemented so that each sequence includes different motor acts, and is organized around goals. One exemplar model, that clearly describes this phenomenon, was presented by Chersi et al. (2005, 2006), who modeled the study by Fogassi et al. (2005) using a chain model. Additionally, the model has been extended to explain how intention understanding and mental simulation take place. We believe that this model could be extended to study whether a chained organization explains the differences between verbs and sentences, for example between action and interaction verbs. Other hierarchical action schemas have been suggested in the literature, for example by Botvinick et al. (2009), adopted a reinforcement learning hierarchical model. Botvinick (2008) reviews how hierarchical models of action are being more frequently referred to. This is probably due to the fact that the way in which how general and abstract action representation emerges from action components, and the role of prefrontal cortex in this process, is becoming an important issue for neuroscientific research. The second kind of constraints are "embodiment constraints". Namely, it would be important to replicate the experiments using embodied models, i.e. models endowed not only with a brain which is similar to that of humans, but also with a body which is similar to ours. In sum: robots should be endowed with a sensorimotor system similar, at least in some respects, to a humans' sensorimotor system. Consider that assuming a strong embodied view would lead to the claim that, given the differences between robots and humans sensorimotor system, embodied robotics models cannot contribute to provide an adequate account of human linguistic comprehension capabilities. We prefer to adopt a weaker embodied view. We propose that robotic models can strongly represent embodied theories of cognition. To elaborate, robotics could be a powerful instrument to explore the extent to which the similarity between the sensorimotor system of different organisms, artificial and natural, constrains the emergence of cognition, and the emergence of language comprehension abilities. In this respect, it might be critical to use robotic models of the sensorimotor system that differ at different degrees from the human one. This could contribute to determine the importance of embodiment theories: namely, it would allow researchers to better understand which aspects of humans' neural and sensorimotor system are critical and determine modifications in humans' behavior.

The third type of constraints, which are referred to as "behavioral constraints", are directly linked with the capability of the model to reproduce and replicate the behaviors produced during the experiments. Having a model which respects the three constraints we outlined would facilitate formulating a synthetic and general theory of the relationships between language and the motor system. Namely, it could contribute to a synthesis effort thus identifying the crucial underpinnings of our behavior.

We believe a model that accounts for the constraints we have illustrated should be able to individuate general principles that combine important characteristics of the relationship between language and the motor system. In the behavioral studies we reported, the critical points which are worth modeling are the following:

The fact that

- during language comprehension the underlying motor and premotor cortices are activated;
- the motor system has a chained organization, and that this organization is encoded in language;
- actions, as well as words and sentences referring to actions, are encoded firstly in terms of distal aspects (overall goal), then of proximal ones (e.g. effectors);
- the different social framework in which the actions are inscribed can change the way in which the action is represented.

On this basis, a model should contribute in detail and explain:

- the time-course as well as the mechanisms underlying the interference and the facilitation effects occurring between effectors implied by action verbs/action sentences and the effectors used to provide a response;
- the mechanisms according to which the different number of motor acts involved in an action chain constrain the comprehension of different action verbs and action sentences;
- the mechanisms according to which, even if the length of a motor chain does not differ, action goals have influences on the comprehension of action verbs/action sentences and how this influences movement;
- the mechanisms according to which, even if the length of a motor chain does not differ the language referring to the presence of objects and/or of other organisms implies the activation of different neural mechanisms (e.g. canonical vs. mirror neurons) which differently affect behavior.

CONCLUSION

In conclusion: we believe that embodied robotics can greatly contribute to a better understanding of the mechanisms underlying the relationship between language and the motor system. We argue that roboticists and modelers should work alongside empirical scientists in order to improve abilities to construe models which do not only account for empirical results, but also formulate predictions that constrain and guide new experimental research.

ACKNOWLEDGMENTS

This work was supported by the FP7 project ROSSI, Emergence of communication in RObots through Sensorimotor and Social Interaction, Grant agreement No: 216125. Thanks to people of the LARAL (http://laral.istc.cnr.it) and LOCEN groups and especially to people of the EMCO group (www.emco.unibo.it) for comments and discussion (and in some cases participation to) on these studies. A special thank to Kate Burke for English revision of the text, and to Fabian Chersi for comments.

REFERENCES

Aziz-Zadeh, L., and Damasio, A. (2008). Embodied semantics for actions: findings from functional brain imaging. *J. Physiol. Paris* 102, 35–39.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.* 59, 617–645.

Borghi, A. M., and Scorolli, C. (2009). Language comprehension and hand motion simulation. *Hum. Mov. Sci.* 28, 12–27.

Borghi, A. M., and Cimatti, F.(2010). Embodied cognition and beyond: acting and sensing the body. *Neuropsychologia*, 48, 763–773.

Boria, S., Fabbri-Destro, M., Cattaneo, L., Sparaci, L., Sinigaglia, C., Santelli, E., Cossu, G., and Rizzolatti, G. (2009). Intention understanding in autism. *PLoS ONE* 4, e5596. doi: 10.1371/journal.pone.0005596.

Borreggine, K. L., and Kaschak, M. (2006). The action-sentence compatibility effect: its all in the timing. *Cogn. Sci.* 30, 1097–1112.

Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends Cogn. Sci.* 12, 201–208.

Botvinick, M. M., Niv, Y., and Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition* 113, 262–280.

Boulenger, V., Roy, A., Paulignan, Y., Deprez, V., Jeannerod, M., and Nazir, T. (2006). Cross-talk between language processes and overt motor behavior in the first 200 msec of processing, *J. Cogn. Neurosci.* 18, 1607–1615.

Buccino, G., Riggio, L., Melli, G., Binkofski, F., Gallese, V., and Rizzolatti, G. (2005). Listening to action-related sentences modulates the activity of the motor system: a combined TMS and behavioral study. *Cogn. Brain Res.* 24, 355–363.

Caligiore, D., Borghi, A. M., Parisi, D., and Baldassarre, G. (2009). "Affordances and compatibility effects: a neural-network computational model," in *Connectionist Models of Behaviour and Cognition II: Proceedings of the 11th Neural Computation and Psychology Workshop*, eds J. Mayor, N. Ruh and K. Plunkett (Singapore: World Scientific), 15–26.

Cattaneo, L., Fabbri-Destro, M., Boria, S., Pieraccini, C., Monti, A., Cossu, G., and Rizzolatti, G. (2007). Impairment of actions chains in autism and its possible role in intention understanding. *Proc. Natl. Acad. Sci. U.S.A.* 104, 17825–17830.

Chen, M., and Bargh, J. A. (1999). Consequences of automatic evaluation: immediate behavioral predispositions to approach or avoid the stimulus. *Pers. Soc. Psychol. Bull.* 25, 215–224.

Chersi, F., Fogassi, L., Rozzi, S., Rizzolatti, G., and Ferrari, P. F. (2005). Neuronal chains for actions in the parietal lobe: a computational model. Soc. Neurosci. Conf., 412.8.

Chersi, F., Mukovskiy, A., Fogassi, L. Ferrari, P.F., and Erlhagen, W. (2006). A model of intention understanding based on learned chains of motor acts in the parietal lobe. In *proceedings of the 15th Annual Computational Neuroscience Meeting* 2006, Edinburgh.

Chersi, F., Ziemke, T., and Borghi, A. M. (2010). Sentence processing: Linking language to motor chains. *Front. Neurorobot*. 4, 1–9.

Dalla Volta, R., Gianelli, C., Campione, G. C., and Gentilucci, M. (2009). Action word under standing and overt motor. *Exp. Brain Res.* 196, 403–412.

Decety, J., and Grèzes, J. (2006). The power of simulation : imagining one's own and other's behavior. *Brain Research* 1079, 4–14.

Fabbri-Destro, M., Cattaneo, L., Boria, S., and Rizzolatti, G. (2009). Planning actions in autism. *Exp. Brain Res.* 192, 521–525.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., and Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. *Science* 308, 662–667.

Freina, L., Baroni, G., Borghi, A. M., and Nicoletti, R. (2009). Emotive conceptnouns and motor responses: attraction or repulsion? *Mem. Cogn.* 37, 493–499.

Gallese, V. (2008). Mirror neurons and the social nature of language: the neural exploitation hypothesis. *Soc. Neurosci.* 3, 317–333.

Gallese, V. (2009). Motor abstraction: a neuroscientic account of how action goals and intentions are mapped and understood. *Psychol. Research* 73, 486–498.

Gallese, V., Craighero, L., Fadiga, L., and Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain* 119, 593–609.

Gentilucci, M., Dalla Volta, R., and Gianelli, C. (2008). When the hands speak. *J. Physiol. Paris* 102, 21–30.

Hauk, O., Johnsrude, I., and Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron* 41, 301–307.

Hommel, B., Müsseler, J., Aschersleben, G., and Prinz, W. (2001). The theory of event coding (TEC): a framework for perception and action planning. *Behav. Brain Sci.* 24, 849–878.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., and Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biol.* 3, e79. doi: 10.1371/journal.pbio.0030079.

Jeannerod, M. (2007). Motor Cognition: what actions tell the self. Oxford: Oxford University Press.

Kaschak, M. P., Madden, C. J., Therriault, D. J., Yaxley, R. H., Aveyard, M., Blanchard, A. A., and Zwaan, R. A. (2005). Perception of motion affects language processing. *Cognition* 94, B79–B89.

Kemmerer, D., Castillo, J. G., Talavage, T., Patterson, S., and Wiley C. (2008). Neuroanatomical distribution of five semantic components of verbs: evidence from fMRI. *Brain Lang.* 107, 16–43.

Mahon, B. Z., and Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *J. Physiol. Paris* 102, 59–70.

Niedenthal, P. M., Barsalou, L. W., Winkielman, P., Krath-Gruber, S., and Ric, F. (2005). Embodiment in attitudes, social perception, and emotion. *Pers. Soc. Psychol. Rev.* 9, 184–211.

Pulvermüller, F., Härle, M., and Hummel, F. (2001). Walking or talking? Behavioral and electrophysiological correlates of action verb processing. *Brain Lang.* 78, 143–168.

Pulvermüller, F., Shtyrov, Y., and Ilmoniemi, R. (2005). Brain signatures of meaning access in action word recognition. *J. Cogn. Neurosci.* 17, 884–892.

Rizzolatti, G., and Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci.* 27, 169–192.

Rizzolatti, G., Fadiga, L., Gallese, V., and Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cogn. Brain Res.* 3, 131–141.

Sato, M., Mengarelli, M., Riggio, L., Gallese, V., and Buccino, G. (2008). Task related modulation of the motor system during language processing. *Brain and Language* 105, 83–90.

Scorolli, C., and Borghi, A. M. (2007). Sentence comprehension and action: effector specific modulation of the motor system. *Brain Res.* 1130, 119–124.

Scorolli, C., Borghi, A. M., and Glenberg, A. M. (2009). Language-induced motor activity in bimanual object lifting. *Exp. Brain Res.* 193, 43–53.

Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., Fazio, F., Rizzolatti, G., Cappa, S. F., and Perani, D. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *J. Cogn. Neurosci.* 17, 273–281.

Umiltà, M. A., Escola, L., Intskirveli, I., Grammont, F., Rochat, M., Caruana, F. Jezzini, A., Gallese, V., and Rizzolatti, G. (2008). When pliers become fingers in the monkey motor system. *Proc. Nat. Acad. Sci. U.S.A.* 105, 2209–2213.

van Dantzig, S., Pecher, D., and Zwaan, R. A. (2008). Approach and avoidance as action effect. *Q. J. Exp. Psychol.* 61, 1298–1306.

Willems, R. M., and Hagoort, P. (2007). Neural evidence for the interplay between language, gesture, and action: a review. *Brain Lang.* 101, 278–289.

Zwaan, R. A., and Taylor, L. J. (2006). Seeing, acting, understanding: motor resonance in language comprehension. *J. Exp. Psychol. Gen.* 135, 1–11.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 December 2009; paper pending published: 12 February 2010; accepted: 27 April 2010; published online: 14 June 2010.

Citation: Borghi AM, Gianelli C and Scorolli C (2010) Sentence comprehension: effectors and goals, self and others. An overview of experiments and implications for robotics. Front. Neurorobot. ***4****:3. doi: 10.3389/fnbot.2010.00003*

frontiers in
NEUROROBOTICS

ORIGINAL RESEARCH ARTICLE
published: 11 November 2010
doi: 10.3389/fnbot.2010.00111

Compositional symbol grounding for motor patterns

Alberto Greco* and Claudio Caneva

Laboratory of Psychology and Cognitive Sciences, Department of Anthropological Sciences, University of Genova, Genova, Italy

Edited by:
Angelo Cangelosi, University of Plymouth, UK

Reviewed by:
Claudia Scorolli, University of Bologna, Italy

****Correspondence:***
Alberto Greco, Department of Anthropological Sciences, University of Genova, Via Balbi 4, 16126 Genova, Italy.
e-mail: greco@unige.it

We developed a new experimental and simulative paradigm to study the establishing of compositional grounded representations for motor patterns. Participants learned to associate non-sense arm motor patterns, performed in three different hand postures, with non-sense words. There were two group conditions: in the first (compositional), each pattern was associated with a two-word (verb–adverb) sentence; in the second (holistic), each same pattern was associated with a unique word. Two experiments were performed. In the first, motor pattern recognition and naming were tested in the two conditions. Results showed that verbal compositionality had no role in recognition and that the main source of confusability in this task came from discriminating hand postures. As the naming task resulted too difficult, some changes in the learning procedure were implemented in the second experiment. In this experiment, the compositional group achieved better results in naming motor patterns especially for patterns where hand postures discrimination was relevant. In order to ascertain the differential effect, upon this result, of memory load and of systematic grounding, neural network simulations were also made. After a basic simulation that worked as a good model of subjects performance, in following simulations the number of stimuli (motor patterns and words) was increased and the systematic association between words and patterns was disrupted, while keeping the same number of words and syntax. Results showed that in both conditions the advantage for the compositional condition significantly increased. These simulations showed that the advantage for this condition may be more related to the systematicity rather than to the mere informational gain. All results are discussed in connection to the possible support of the hypothesis of a compositional motor representation and toward a more precise explanation of the factors that make compositional representations working.

Keywords: motor representation, symbol grounding, compositionality, embodiment

INTRODUCTION

Compositionality and symbol grounding are two fundamental questions that have gained considerable theoretical attention in the last decades. Compositionality consists in the possibility of drawing the meaning of a complex linguistic expression from the systematic combination of meaningful components according to syntactical rules. It is considered one of the key features of human language, differently from animal communication or human ancestor protolanguage, fundamentally holistic and conveying meaning only through single gestaltic expressions (Jeannerod, 1988; Arbib, 2005). Compositionality has been called into play for explaining the ability of producing an indefinite number of linguistic expressions (what is known as productivity), and is relevant in formal languages of mathematics, logic, and computer science. The principle of compositionality, in fact, is a general key concept in all the cognitive sciences, since it has gained interest in philosophy, linguistics, artificial intelligence, robotics, psychology, and neuroscience.

As is well known, compositionality was an essential part of the traditional cognitivist "language of thought" hypothesis (Fodor and Pylyshyn, 1988), positing that human representations acquire their structure by the combination of distinct symbolic parts according to formal rules. This view was first challenged by connectionist theories (Smolensky, 1988; van Gelder, 1990) and more recently by new approaches that accept the idea of non-symbolic representations. These approaches stress the point that cognition cannot be explained solely by abstract symbolic processing, because human beings have a body interacting with environment (*embodiment*: e.g., Glenberg and Kaschak, 2002), and because a sensorimotor ground is needed for symbols. This is the symbol grounding issue (Harnad, 1990; Cangelosi et al., 2000).

Such new stances have influenced also the way of considering language. The question of how actions are internally represented is of general importance because words for action (predicates or verbs) are the essential ingredients of propositions, and actions are also fundamental for understanding, like predicates are essential in logic. In addition, representation of actions and of words could be tightly linked since, according to some theories, linguistic comprehension would be a sort of internal simulation (*re-enactment*) of actions expressed by linguistic symbols (Barsalou, 1999; Pulvermueller, 2005). Many other recent approaches have made similar points, like the "experiential view of language comprehension" (Zwaan, 2004). In the same vein is the finding that motor verbs activate brain regions associated with action (Ruschemeyer et al., 2007). Barsalou comes to considering perceptual non-symbolic representations as a system having the same features of symbolic ones, including *compositionality*. In this sense, Barsalou's approach implies supposing analog representations working compositionally (Wu and Barsalou, 2009).

The motivation for the present study then comes from the fact that, although compositionality has been traditionally considered as concerning the abstract combination of symbols that already must have a grounded meaning, the possibility of an analogical compositionality, and in particular of a motor compositionality, is a still open empirical question.

The hypothesis of a motor compositionality has obtained a substantial interest in current cognitive neuroscience research (Bizzi and Mussa-Ivaldi, 2004, p.415). There are several reasons for hypothesizing compositional motor representations: human motor control has a hierarchical nature, complex motor programs result from motor subroutines, elementary operation of body parts (i.e., joints, muscles, etc.) for action can be identified (Allott, 2003). In robotics, such a system has also obtained significant attention (e.g., Thoroughman and Shadmehr, 2000; Amit and Mataric, 2002; or the "Human Activity Language" primitives for segmenting human motor patterns as a language: Guerra-Filho and Aloimonos, 2006). The theoretical relevance of this issue is clear also since a compositional motor representation would entail that motor primitive elements could be distinguished that keep the same meaning in different contexts, like their possible verbal counterparts.

Some additional clarification seems convenient here about the expression "motor representation." It is obviously possible to consider either symbolic (conceptual, verbal) or analog motor representations; grounding is, of course, just the establishing of an association between these two kinds of representations. But the notion of *analog* motor representations seems to oscillate between psychological and neural senses (Greco, 1995; Peschl, 1997), ambiguously referring to different processes such as: (a) preparing motor action: motor schemata or motor imagery (Jeannerod, 1994; see also the symposium "Mental representations of motor acts" of the European Neurosciences Association: Deecke, 1996); (b) kinesthetic self-perception of motor action during execution; (c) visuospatial perception of motor action executed by others. Such senses evidently refer to different motor tasks that may be related to a more basic distinction between visuospatial and motor aspects (respectively implying perception and execution of motor patterns). The strength of this distinction, however, seems weakened by the celebrated and well-established mirror neuron theory, showing that perception and execution of motor patterns activate the same brain areas (Gallese et al., 1996). The mirror neuron hypothesis is compatible with the assumption that, even if evidence can be found that motor tasks are controlled by different systems at lower levels, at some higher level they should converge into a unique representation. This unique representation is responsible for the uniqueness of meaning, the one that normally is expressed verbally (e.g., when we speak of "walking" we mean the same thing either referring to what we *see* when someone else is walking or what we ourselves *do* when walking).

In any case, whatever the exact nature of analog motor representations is (as a form of imagery, or of mental simulation, or re-enactment), the point is how structured these representations are. Do they include primitive "images" for components of motor performance, or codes for individual features, that are then somehow assembled, or do they work as a whole? The question is relevant also for motor concepts and words that are associated to motor memories.

FRAMEWORK

The present study was aimed at an empirical investigation about the nature, compositional or holistic, of motor representations that provide analog ground for meaningless verbal labels.

The most obvious and ecological way of analyzing the relation between language and motor behavior is considering when a *meaningful* association is established. This is obvious because motor activities are normally goal-directed, and meaningful words are used to describe them. We choose, however, to start from *meaningless* words and motor patterns, a rather extreme situation, because when studying the establishing of symbol grounding the interference of already-known motor patterns and words should be minimized. We needed to study how *new* symbols are associated and eventually combined for representing *new* motor patterns, eventually becoming meaningful. Thus we used non-sense words as arbitrary symbols that would acquire a meaning only (or as much as possible) from grounded sensory experience, namely in connection with perceived visuomotor stimuli. Similarly, we used non-sense motor patterns because if they already had a sense they would also had been already connected with a corresponding linguistic representation and the new word would only consist in a sort of "translation" or a synonym of this existing representation. We actually use the term "motor patterns" and not "gestures" just to stress that we are referring to meaningless motor behavior. We are obviously aware of limits of this perspective, since any stimulus (either verbal or not) is normally put in relation with semantic memory contents; this situation of artificial "semantic vacuum," however, seemed suitable as a starting condition for a study of symbol grounding establishment.

The present work continues a previous one (Greco and Caneva, 2005) where we already associated an artificial language with meaningless motor patterns in holistic and compositional conditions. In the experimental paradigm described in the present paper, there were two conditions. In the first condition one word acquired a grounding for an arm trajectory (irrespective of how it was executed) and a second word was grounded for denoting a particular way of executing it (how to put hands while executing it). In the second condition a single word was grounded for each motor pattern execution taken as a whole.

The main hypothesis tested was that when different verbal labels are learned in association with different aspects of visuomotor patterns in arm motor patterns (namely, in our case, arm trajectory and hand posture), a separate grounding is established for these symbols, based on compositional analog representations, that allows a facilitation in a subsequent naming task for the same patterns.

The rationale is that the ability of correctly naming visuomotor patterns, in our experimental conditions, is a true grounding test (Cangelosi et al., 2000), because this would reveal that labels, that were meaningless at the start, became meaningful symbols for these patterns as a result of an analog grounding. This kind of grounding may be ascribed an analog nature even if it does not necessarily involve really performed motor patterns. This idea is supported by the mirror neuron theory, that strengthens the idea that analogic patterns can be established on observed visuomotor patterns without a direct bodily execution.

If participants in the compositional condition were favored in this task, then, this outcome would show that a separate analog grounding was established for arm trajectory and hand posture, connected with the corresponding two labels. On the contrary, if patterns tended to be better represented by analog holistic codes, a naming task in the condition where each pattern as a whole was learned in association with a single word should be advantaged.

A further account for a possible advantage resulting in compositional condition is that memory load is reduced when the amount of information needed to name stimuli is smaller, as in the case when some words can be reused for recalling the same motor referents. However, not only informational load but also a reliable grounding system must be taken into account in this case: this involves a consistent association between symbols and their analog referents. We shall tackle this question with the help of neural network simulations.

We addressed also the question whether the visuospatial analog coding, on which recognition is based, might be affected by grounding as well. In fact, it is reasonable to suppose that naming implies first some pattern recognition process and after that – if grounding has been established – the retrieval of the corresponding label. We tested this possibility by introducing in our first experiment also a recognition test, in order to assess a possible difference between compositional and holistic groups.

Table 1 | Stimuli.

Basic motor patterns	Hands up	Hands down	Fist
	Nole	*Bote*	*Sove*
Baspi	Terpesova	*PD	Utrimosta
Gispi	Sertamina	Mutiralda	*PD
Respi	Tupifasta	*PD	Mertogala
Tispi	Volsicoda	Feltorana	*PD
Faspi	Patrasina	*PD	Luticanza
Cuspi	Rispaguna	Dortamana	*PD
(mov.#7)	*TD		
(mov.#8)	*TD		
(mov.#9)	*TD		
(mov.#10)	*TD		

**TD indicates trajectory distractors, *PD posture distractors.*

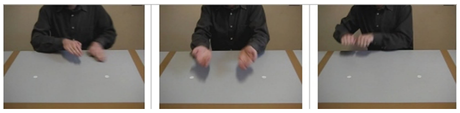


FIGURE 1 | Snapshots from some videoclips of different patterns in the three hand postures.

EXPERIMENT 1

METHOD

The task consisted in associating visuomotor patterns, presented as videoclips, with corresponding words, uttered aloud. There were two conditions: in the compositional condition (group A) motor patterns were associated with *two words*, whereas in the holistic condition (group B) with a *single word*. The two-word sentence presented in the compositional condition can be considered as a "verb–adverb" structure: what motor pattern is performed, how it is performed (i.e., using what posture). In this experiment a recognition test was performed prior to the naming test. The dependent variables were: (a) *recognition* of target motor patterns presented along with distractors; (b) *naming* (retrieving the name corresponding to each target motor pattern).

Stimuli

The structure of stimuli is shown in **Table 1**; some examples are given in **Figure 1**.

Motor stimuli. Consisted in arbitrary arm trajectories (as an example: moving arms toward oneself and then lifting them). Eighteen stimuli were constructed by combining six basic motor patterns, performed in three different hand postures (up, down, fist); four other motor patterns were added, performed in the hand up (called "nole") posture only. All motor patterns were performed by a sitting person, framed half-length, in front of the camera; only the chest and the arms were visible; in the starting position the hands (already in the palm, back or fist posture) rested on two reference circles marked on the table. Only 12 combinations (the ones with a bold name in **Table 1**) were presented during learning. The other 10, indicated with an asterisk, acted as distractors for recognition testing purpose; 4 of them (*TD) were arm trajectory distractors (corresponding to never seen trajectories), the other 6 (*PD) were hand posture distractors (corresponding to seen postures but performed in a different hand posture).

Linguistic stimuli. For group A, a two-word sentence was used to name patterns, resulting from the combination of the word for the trajectory and the word for hand posture (words for group A are in italic in **Table 1**). For group B, a single word (in bold in **Table 1**) was used to define each pattern as a whole. For example, the first pattern was named "baspi nole" for group A and "terpesova" for group B.

As in natural languages syntactical roles are marked by particular morphemes, some constraints were established for pseudowords that had to assume a syntactical role. The six pseudowords denoting verbs were 5-letter and bisyllabic, constructed by adding a consonant-vowel pattern to a fixed ending (–SPI). Pseudowords denoting adverbs were 4-letter and bisyllabic, constructed by the pattern consonant-O-consonant-E. Single pseudowords standing for full motor patterns had 9-letter and 4 syllables (resulting like the sum of the other two words) and all ended in -A.

Participants

Twenty students, volunteers, individually participated in the experiment for course credit. Informed consent was obtained prior to participation in the study. Half of them were randomly assigned to group A, half to group B.

Procedure

Participants seated in front of a 14″ computer monitor, in a different room than experimenter's; in the table in front of the screen a rectangular area measuring cm 77 × 53, including two reference circles identical to ones shown in the videoclips, was traced; this allowed participants to repeat motor patterns when requested. Only a mouse (no keyboard) was available for responses. All instructions and stimuli were presented on the monitor screen. The procedure included the following stages.

Verbal learning. The first stage was aimed at making participants familiar with words. All the words were presented in a panel with 9 (group A) or 12 (group B) buttons, where each single word was printed as a button label. Labels were disposed in alphabetical order. Participants were instructed to click with the mouse on each button to listen to a recorded male human voice that read the corresponding word aloud; the order of presentation was chosen by participants themselves. Only when all words had been listened, a closing button was enabled to proceed to the next step.

Associative learning. This was the main stage of the experiment. Twelve training clips were presented. For each clip, the voice uttering the sentence (gr. A) or word (gr. B) corresponding to the motor pattern was presented at the start, along with a blank screen; the videoclip was then shown immediately. Patterns were presented randomly but paired so that the same pattern was first presented in the "nole" (hands up) posture and then in one of the other two postures, like shown in **Table 1**. Participants were also instructed to repeat each pattern after having seen it while uttering its name aloud, in order to learn it better. It was stressed that the correctness of this performance would have not been assessed in any way. The full set of stimuli was repeated three times.

Integrated test. In the testing phase, recognition test and naming test were integrated. All 22 stimuli clips (12 target and 10 distractors) were presented in random order. For each stimulus, participants were first asked if they had already seen it; if they answered yes, then they were also asked to say the corresponding sentence/name. Motor performance was not requested. A final debriefing was conducted in order to assess possible task difficulties and hints for improvement.

Post-experimental debriefing. After completion of the experiment, a structured interview was conducted in order to assess task difficulty, the use of associations with known words or gestures, and above all to verify whether participants in group A had been able to identify the syntactic role of the two words. Almost all participants found the task difficult or very difficult, but the syntactical roles were identified without uncertainty by participants in group A, with the exception of only two subjects. Associations reported by participants were somewhat subjective and not consistently related to particular stimuli.

RESULTS AND DISCUSSION

Recognition test

Very high recognition scores resulted without any difference in both groups (condition A, $M = 0.81$, SD = 0.39; condition B, $M = 0.82$, SD = 0.38). This outcome shows that motor recognition, at least in our experimental conditions, is not related with the availability of a specific verbal label for components. Motor patterns were presumably not recognized using a verbal code but accessing to a specific visuomotor representation.

We also analyzed recognition scores for distractors only (**Table 2**, PD = posture distractors, TD = trajectory distractors). Recognition was almost fully correct ($M = 0.92$) for MD, i.e., different trajectories, but recognition scores were lower ($M = 0.71$) for PD, i.e., same trajectories with different hand postures. This difference is highly significant ($t = -4.41$, $p < 0.0001$) and depends on the fact that differences between motor patterns resulted very salient, whereas it was more difficult to distinguish hand postures. This result shows that, in a pure recognition test, motor stimuli were not processed at the hand posture detail level, characterized by more confusability, but only at the motor pattern level, more macroscopic, where a more immediate holistic representation seems sufficient for recognition. Retrieval in this case was based on perceptual similarities and not on the symbolic association with arbitrary labels.

Naming test

Naming task results were completely opposite to recognition ones, as very low scores resulted in both groups (condition A, $M = 0.16$, SD = 0.37; condition B, $M = 0.17$, SD = 0.37).

A difference between recognition and naming in our task is not surprising, because it is consistent with the well-established finding that performance is generally better in recognition memory than in retrieval memory, and that these are based on substantially different processes (Yonelinas, 2002). This difference holds in many areas of cognition, from words (Peynircioglu, 1990), to pictures (Langley et al., 2008), to faces (Cleary and Specker, 2007), to melodies (Kostic and Cleary, 2009). This effect was found also with pseudowords and even non-words (Arndt et al., 2008). Our result matches such theoretical premises, and seems to suggest that the recognition-retrieval difference could be extended also to motor memory. The dramatic extent of this difference in our task, however, suggests some caution in reaching this conclusion. Our outcome evidently indicates that name-pattern association was too a difficult learning task in these conditions and this could have amplified the recognition-naming difference. This issue would have deserved a deeper investigation in different learning conditions. We strived, in the course of our study, to remedy such learning difficulties, but, since the recognition-retrieval issue was not the main concern of our current research, this result was not further analyzed and the recognition task was abandoned.

EXPERIMENT 2

The main outstanding question from results of Experiment 1 was the floor effect we found for naming, clearly denoting that learning conditions were inadequate for grounding. This

Table 2 | Mean recognition proportion for distractors and target stimuli in Experiment 1.

	Distractor type							
	PD		TD		Target		Total	
Group	*M*	SD	*M*	SD	*M*	SD	*M*	SD
A	0.74	0.44	0.91	0.29	0.87	0.34	0.81	0.39
B	0.66	0.48	0.94	0.24	0.88	0.32	0.82	0.38
Total	0.71	0.46	0.92	0.27	0.88	0.32	0.82	0.39

motivated a revision of experimental setup in order to make learning easier. We must make clear that our interest is currently focused on differences between compositional and holistic conditions in comparable learning conditions, sufficiently adjusted as to difficulty, and not on learning conditions or mechanisms *per se*.

A new paradigm for Experiment 2 was then planned. In order to make learning easier, method and procedure were simplified. Instructions were improved by introducing an interactive example of task execution. A different stimuli presentation system was also adopted: in the first learning stage, all patterns were presented only in a single hand posture (upwards); in a second learning stage, after having tested that at least four of six stimuli had been learned, the same trajectories were paired with a second posture. As a further change, it was required that verbal stimuli be transformed into an infinitive verb, by adding the (Italian) ending "-are" (e.g., "baspi" into "baspare"). This helps categorizing such words as verbs reducing the cognitive load. An additional reason that motivated this change was that the task resulted rather passive, since names were still in echoic memory when repeated just after having being heard. This change was then aimed also at encouraging an active stimulus processing, so that echoic memory effect be removed or reduced, and participants be less passive and more attentive.

METHOD

The independent variables and the main task (i.e., learning to associate motor patterns with sentences or words) were the same as in Experiment 1. Naming was the only dependent variable.

Stimuli

The conceptual universe was the same as in Experiment 1 (**Table 1**), but only 12 target clips were used (no distractors were needed since no recognition test was performed).

Participants

28 students, volunteers, individually participated in the experiment for course credit; informed consent was obtained prior to participation in the study. Half of them were randomly assigned to group A, half to group B. As in Experiment 1, in group A each motor pattern was associated with a two-word sentence (one for trajectory, one for hand posture), in group B each motor pattern performance as a whole, i.e., regardless of hand posture, was associated with only one word.

Procedure

Instructions and stimuli were presented in the same conditions as in Experiment 1. Learning was split up into two stages. In the first phase (*target pattern learning*) only six patterns in the "hands up" posture were learned. This stage was followed by a first test (*target pattern test*, TPT). In the next learning stage (*posture learning*), each learned pattern was paired with the same pattern in a different hand posture. The procedure included the following stages.

Verbal learning. This stage was exactly as in Experiment 1. The same word panel was used; it included the full set of words for the group (9 A, 12 B), arranged in alphabetical order.

Target pattern learning. The purpose of this stage was to make participants learn motor patterns irrespective of hand postures. As in Experiment 1, clips started with a blank screen while a male human voice uttered the corresponding word, then the motor pattern performance was shown on the screen. Only six clips were presented, in random order, and only one word, referring to the pattern, was used. The only difference between groups A and B was the word used (e.g., "baspi" for gr. A and "terpesova" for gr. B). Subsequently, the word panel was shown, where word labels appeared, transformed into the infinitive form (e.g., "baspare" or "terpesovare"), and the participant was requested to mouseclick the corresponding button. It was possible to correct mistaken choices before confirming. Then, the pattern was shown again without audio and the participant had the opportunity of performing it while uttering the verb aloud. In instructions it had been explained that the purpose of this procedure was to help participants "learn better" motor patterns; it had been also stated clearly the absolute irrelevance of correct performance. The series of six stimuli was repeated three times.

Target pattern test. At this stage, learning of six previously presented names was tested. A minimum learning threshold of 4/6 was required for passing this test, otherwise the first learning stage was repeated (up to two times, after that the protocol was discarded).

Posture learning. After a new warm-up example trial, at this stage all 12 stimuli, in different hand postures, were presented with the same procedure as in Target Pattern Learning. For participants in group A, motor patterns were described using a sentence where the first word was the same word for the trajectory previously learned, and the second was the word for the posture (e.g., "baspi nole," "baspi sove"); for group B the word uniquely denoting the motor pattern was used (e.g., "terpesova," "utrimosta"). Participants in group A could compose the corresponding sentence by clicking on two-word buttons (in group B just one button); all could correct mistakes before confirming. In the word panel all words denoting motor patterns were put into the infinitive Italian form (e.g., "baspare," "terpesovare") and this was the form that participants had to use when repeating aloud the verbal part. The presentation sequence was random, but, to make learning easier, motor patterns were paired so that each randomly selected trajectory was always followed by the same trajectory performed in the other scheduled posture. As in the previous Target Pattern Learning, the full set of stimuli was repeated three times, so that 36 stimuli were presented overall.

Final test. All 12 videoclips showing motor patterns without audio were randomly presented, each followed by the word panel. Participants in group A were requested to click on two words to compose the corresponding sentence; in group B they had just to click on the corresponding word. It was always possible to correct mistakes before confirming.

Post-experimental debriefing. A final debriefing was conducted following the same procedure used for Experiment 1. The task was still perceived as difficult but, as in the previous experiment, the syntactical role of words in group A was easily identified by all participants (only one failed). Very few verbal or visuomotor associations were reported, that were not commonly shared but rather had a personal character. In any case, there is no reason to suppose that particular associations could favor one group over the other.

RESULTS AND DISCUSSION

We first analyzed learning progress in different experimental stages. **Figure 2** shows the learning curve (mean proportion of correct responses) from the first to the final phase. At the first TPT there were no significant differences between the two groups (*A* = 0.47, SD = 0.50; *B* = 0.41, SD = 0.49; *t* = 0.82, *p* = 0.41). This shows that there were no differences between subjects at the start and, importantly, that stimuli used for the two groups were equivalent. Mean values of correct responses at the final test (FT), instead, were significantly different (*A* = 0.60, SD = 0.49; *B* = 0.46, SD = 0.50; *t* = 2.51, *p* = 0.01).

As it is clear from our experimental set up, two kinds of stimuli were tested in the FT phase, i.e., motor patterns presented in both Target Pattern and Posture Learning phases (all with hands up posture) and motor patterns only presented in the latter, differing from previous ones because they had different hand postures. It seems obvious to expect that motor patterns seen in both learning phases (hands up posture) are considerably easier than others; in fact, there is no difference between groups for such stimuli learned during both training phases (see **Table 3**, 0.68 vs.0.62, *t* = 0.83, *p* = 0.41). If we consider other motor patterns, however, the difference between the two groups is dramatic (*A* = 0.51, *B* = 0.29) and statistically highly significant (*t* = 2.83, *p* = 0.005). Since new stimuli differed from previous ones only for hand posture, this supports the hypothesis that the compositional task was easier because a specific word denoting posture was available. We can say, then, that in the present experimental conditions verbal descriptions for motor patterns were better learned when a compositional verbal system was available.

In our experimental setup, in the FT, naming was influenced by having seen a pattern before. The effect of a verbal system could only be revealed by considering trajectories with a new hand posture. In fact, the compositional group (A) had better results than the holistic one (B) in naming new stimuli. If we compare the outcomes of the experiments 1 and 2, we find that there was no compositional representation in recognition (Experiment1) and a sort of compositional representation in naming (Experiment 2). The procedure in Experiment 2 presented two main differences from Experiment 1: (a) having splitted learning of trajectories and of postures; (b) having introduced the addition of the Italian suffix for verbal conjugation ("-are"). The first change may have helped participants identify more easily stimulus features. The second change may have helped group A (where an elementary syntactical system was needed) by giving a hint about the syntactical role of the first word.

Table 3 | Mean correct naming proportion in Experiment 2 (final test).

Group	Stimuli	Correct	
		M	SD
A		0.60	0.49
	Hands up posture	0.68	0.47
	Other posture	0.51	0.50
B		0.46	0.50
	Hands up posture	0.62	0.49
	Other posture	0.29	0.46
Total		0.53	0.50

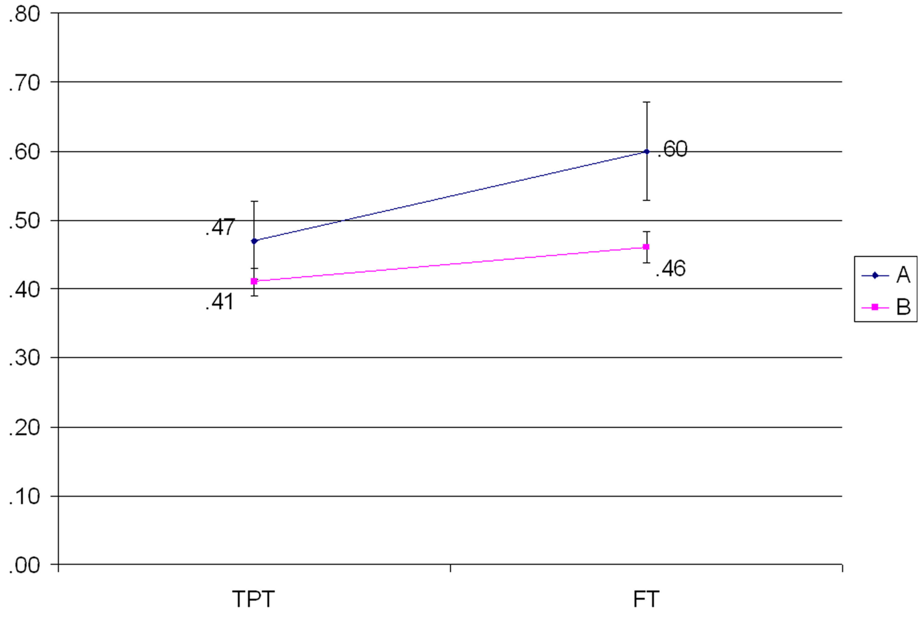


FIGURE 2 | Learning curve in Experiment 2.

NEURAL NETWORK SIMULATIONS

As we have mentioned previously, a possible account for the advantage resulting in the compositional group is that in this condition more motor stimuli can be coded using a fewer number of words. This implies an informational gain, that would be maximally exploited with all theoretically expressible stimuli: using the nine available words in condition A, 18 motor patterns can be named combining six trajectories with three postures (type-token ratio 9/18 = 0.50). Even if in our actual condition only 12 patterns were learnt (type-token ratio = 9/12 = 0.75), anyway a consistent reduced memory load results. In this account, however, two aspects are not clearly distinguished, i.e., the informational-syntactic aspect (i.e., the mere number of alternatives and word positions) and the need for consistent and systematic semantic associations.

In order to test some different possible changes to our paradigm without having to engage a number of new human participants, we reproduced and modified the task using neural network models. Considering that grounding analog information in symbolic codes is tantamount to use more compact representations, we can expect a still greater advantage for the compositional condition respect to a condition where the number of words is equal to the number of stimuli to be distinguished and remembered (type-token ratio = 1.00). To take into account the role of systematic correspondence between words and analog patterns, upon which syntax and grounding are based, we also devised a simulation where such correspondence was disrupted, while maintaining an equivalent informational load.

We then performed three simulations:

- *basic* simulation, that faithfully reproduced the Experiment 2 conditions;
- *extended* simulation, where the basic simulation was augmented using an increased number of inputs;
- *non-systematic* simulation, similar to the extended simulation, where verbal inputs were modified in order to be informationally equivalent to original ones, but without any systematicity.

GENERAL METHOD AND BASIC SIMULATION

General neural network architecture and I/O encoding

For all our simulations we used a set of 50 neural networks that implemented three modules of hidden units, with the function of processing motor and verbal information, and of establishing an associative grounding between the two kinds of data.

The architecture (shown in **Figure 3**) included an input layer, divided into two distinct modules (12 motor units and 9 verbal units), two hidden layers, and only one (verbal) output layer. The first hidden layer was divided into two distinct (not interconnected) modules, each including six units, with the purpose of independently processing motor and verbal inputs. The second hidden layer (including 15 units) had an “associative” function, that is to relate the two kinds of inputs and to generate the output. Each layer was connected in a recurrent way with its lower layer.

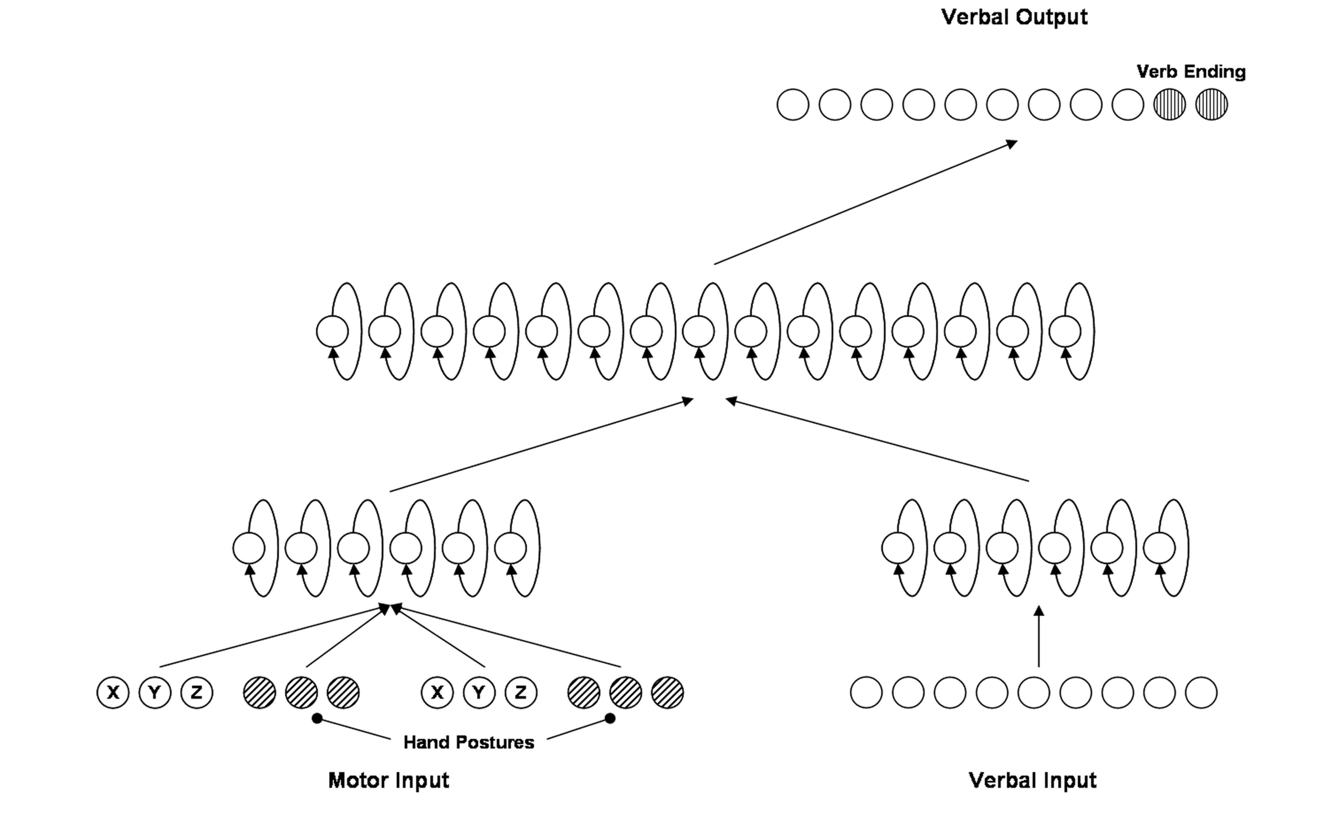


FIGURE 3 | Neural network architecture.

Since motor and verbal data flows had a different length, we introduced a parametric bias to synchronize the data flow. This bias was computed during a pre-training stage by a set of four "timing" units (not shown in figure) supervisioned by a backpropagation algorithm. In this pre-training, networks were given 10 motor pseudo-inputs that had the same streaming structure of real inputs, but that did not represent points fitting on the same curve. During this stage, the supervision algorithm acted uniquely on timing units, while weights of all other connections were not modified. At the simulation start, all timing units were set up the same way in all networks; weights of other units were generated randomly. Parameters of timing units were never modified during simulations. Thus our networks can be considered as discrete-time RNNPB (Recurrent Neural Networks with Parametric Bias) with a dynamic input. During simulation learning was achieved by a supervised Bayesian algorithm using Gibbs sampling.

The basic conceptual universe consisted of a number of inputs equal to the number of stimuli used for the Experiment 2. Verbal and motor input were given as data streamings. Words or sentences were input as a flow of four consecutive strings (corresponding to the 4-syllable verbal inputs, e.g., ba-spi-no-le or ter-pe-so-va). Motor patterns were encoded in a pseudo-analog way, i.e., input as 25 consecutive sets of spatial coordinates of the two hand postures in different time moments (frames) during pattern execution; such coordinates were obtained through the analysis of motor patterns of a virtual dummy (Poser 7.0). Each stream also included information about the hand posture, encoded using three binary units.

General procedure

Each simulation was composed of two sets of 50 nets and they followed the same steps considered in the corresponding experiment. Simulations followed the same steps as in the Experiment 2. In the Basic simulation, the conceptual universe was exactly the same used in Experiment 2 (referred as "standard" in **Figure 5** that summarizes all simulations), including 12 stimuli.

Basic simulation results

Networks learning in both conditions was very close to human participants performance. The learning curve from the TPT and the FT is shown in **Figure 4**. As for human subjects, there were no significant differences between the two conditions at the TPT ($A = 0.41$, SD = 0.26; $B = 0.38$, SD = 0.31; $t = 0.41$, $p = 0.98$), while significant differences were found at the FT ($A = 0.53$, SD = 0.34; $B = 0.40$, SD = 0.31; $t = 1.80$, $p < 0.05$).

ADDITIONAL SIMULATIONS

Extended simulation procedure

This simulation differed from the Basic simulation only because a larger number of input (motor and verbal) stimuli was used. 48 new motor patterns were created using a custom program for generating random 3-D trajectories, using a procedure based on random point generation and spline interpolation; some constraints were included in this procedure to avoid trajectories impossible to be performed by human-like arms; each trajectory was also planned to be performed in 3 + / − 1 s by virtual hands moving at constant velocity. When the final spline system defining a new trajectory was completed, a new set of 25 coordinates was calculated getting points at regular intervals.

Correspondingly, 24 new words were introduced for denoting patterns in condition A and 48 for condition B. New words were also generated using a custom software that reproduced the structure of original words. The constraint was established that each new word be different from previously generated ones, by computing the number of repeated letters (for condition A) or syllables (for condition B). The three original motor and verbal codes were kept for hand postures. The total number of stimuli was then augmented to 60.

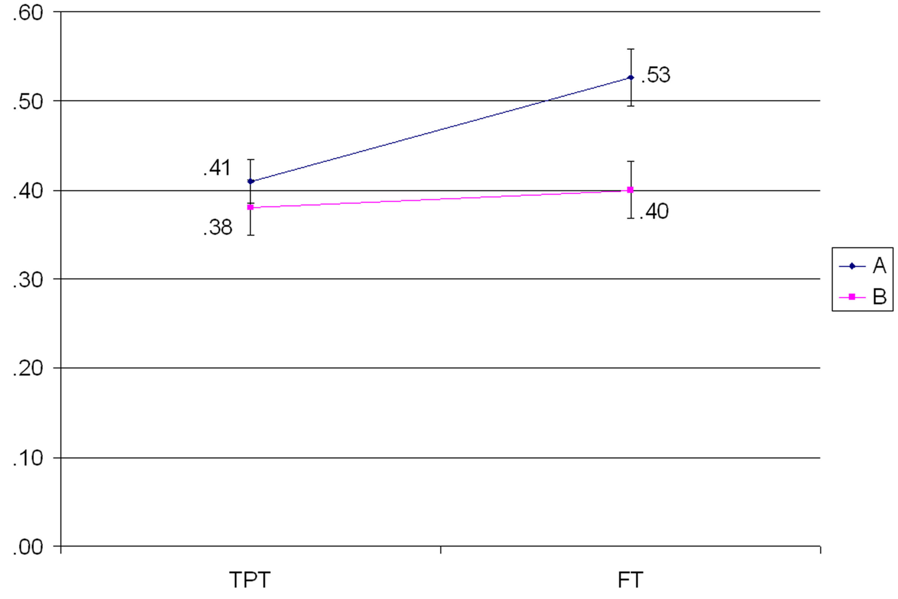


FIGURE 4 | Learning curve in basic simulation.

Extended simulation results

Even if the simulation was run using all 60 stimuli, only 30 can be considered in results since this number is already sufficient for significant differences between conditions A and B. As shown in **Figure 5**, when extending the simulation with an increasing number of stimuli the advantage for condition A persisted and was even more robust. When 30 stimuli were used, correct performance was 0.30 in condition A and 0.07 in condition B ($t = 2.04$; $p < 0.05$).

Non-systematic simulation procedure

In this simulation the same data and procedure of previous ones were replicated, with the exception that a new condition was introduced. In this condition C, the set of verbal inputs included the same bisyllable words used for condition A. The original syntactic structure was kept (5-letter words first and 4-letter words following), but words were associated with randomly selected trajectories and postures. The association was arbitrary when sentences were generated: the first word was chosen randomly in the list of words used for trajectories (e.g., baspi), the second word similarly chosen randomly in the list of words used for hand posture (e.g., nole). For example, "baspi bote" and "baspi nole" in condition A were referred to the same trajectory performed in two different hand postures, while in condition C these word combinations were referred to different trajectories and hand postures. So there was no consistent association between single words and single components of motor patterns: there was only a formal compositional-like structure but without any true compositional meaning.

Once generated, such composed sentences were obviously used consistently throughout the experiment. Since word combinations predicted the same referent as if they were single words, condition C was somewhat similar to condition B, but from the informational quantity aspect (number of different words, type-token ratio) it was equivalent to condition A (0.75).

Non-systematic simulation results

Performance in condition C (**Figure 5**) was always very scarce and smaller than other conditions, and comparable to condition B (even if the type-token ratio in this case was more favorable than in condition B). Already at the standard 12 stimuli level, there was no significant difference between B and C conditions performance ($B = 0.40$, $C = 0.22$; $t = 1.78$, $p = 0.10$) and, as the curve shows, the distance between the two conditions becomes shorter and shorter when the number of stimuli increases.

Additional simulations discussion

Results of additional (Extended and Non-systematic) simulations, taken together, suggest that the better performance in condition A may be explained by an informational advantage only when this is joined with a systematic and consistent association between words and their referents.

As we have seen, the main advantage of symbol grounding is its ability to offer more compact representations than analog ones, but even if representations exhibiting one-to-one correspondences between symbols and referents are still more compact than original

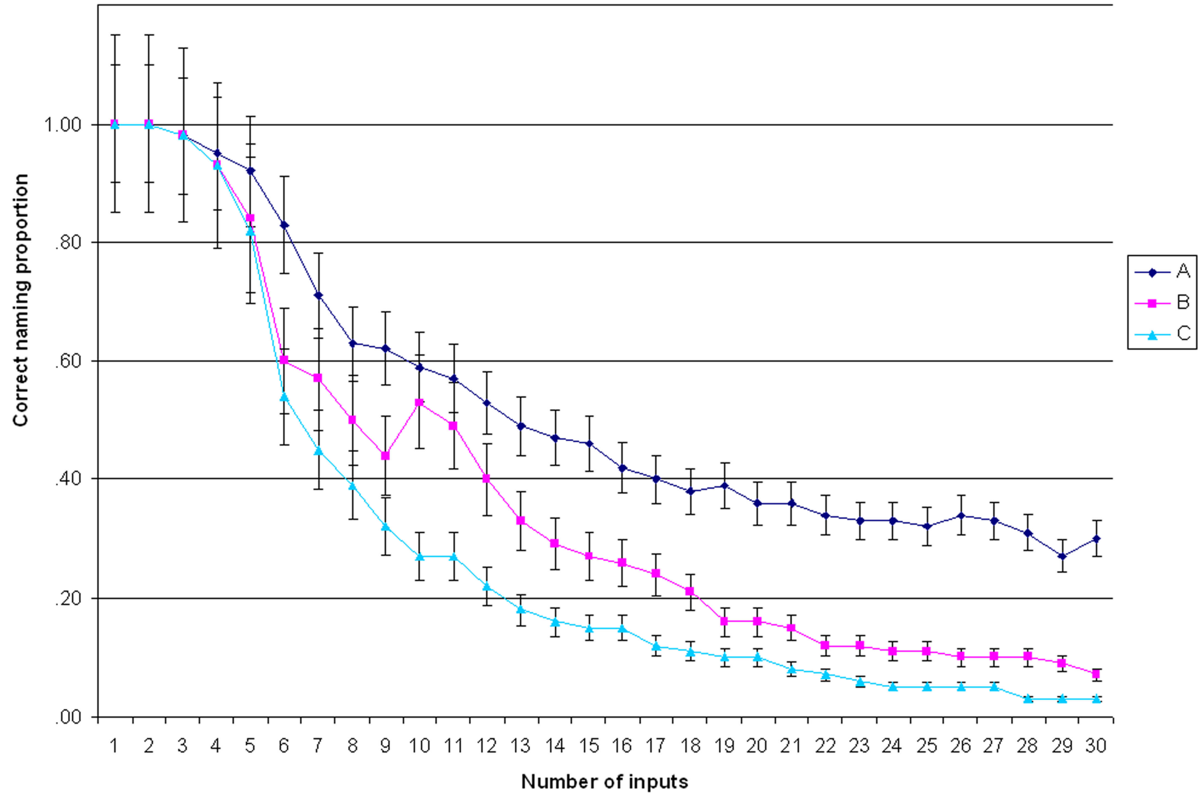


FIGURE 5 | Summary of three simulations.

analog representations, they do not work efficiently in a world where there are regularities and redundancies. We can speculate that the most important reason why compositional systems work better is not their ability of reducing cognitive load but, instead, their ability of making possible a systematic reusing of corresponding grounded analog representations.

GENERAL DISCUSSION

In this paper we described an empirical paradigm aimed at studying the possible compositional nature of grounded analog motor representations. The question asked was whether a compositional internal representation, for arm trajectories that could be executed in different hand postures, can help recognition and naming of such patterns when associated with symbolic (verbal) labels. We performed two experiments and simulations with neural networks, using meaningless stimuli, in two conditions, i.e., when labels were single words, corresponding to motor patterns regardless of hand postures (holistic condition), and when two-word sentences (the first word for the arm trajectory, the second for the hand posture) acted as labels (compositional condition).

In the first experiment, a good performance in pattern recognition was generally achieved regardless of verbal compositionality, but was poorer when distractors differed from targets just in hand postures. This showed that verbal labels did not help reducing the main source of confusability in this task, concerning hand postures, because recognition was only based on perceptual, not symbolic, cues. The mediating representation in this case was a purely analog and holistic code. Nothing could be drawn from this experiment as to the naming task, because of difficulties of the learning procedure.

In the second experiment, as a result of substantial changes to the learning procedure, we obtained acceptable learning performances in the naming task for all participants. In this task, that as we have noted above is a true grounding test, we found a significant difference between the compositional and holistic groups after having introduced a condition where the hand posture was relevant for differentiating between stimuli. Since in the compositional condition the second word had been consistently associated with the hand posture aspect, we can say that a separate grounding representation was established for it, different from the one acting as a ground for the word denoting the arm trajectory. This means that different analog (visuomotor) representations worked compositionally as a ground for the corresponding symbols, similarly to what happens with symbolic composition. The full representation of each new concept that we have tried to construct, on this view, includes both verbal and sensorimotor information corresponding, in different conditions, either to the whole pattern or to aspects of it.

In both our experiments, but notably in the first one, where no grammatical cues were present, almost all (99.9%) responses of participants in the compositional condition were syntactically correct: the first word denoted a trajectory, the second a posture. This happened even when participants, during the post-experimental debriefing, did not show to be aware of the syntactical functions of words. This is not surprising, because the automatic emerging of syntax is a well known fact, evident also from natural language acquisition in infancy. However, we would like to stress here that the correct binding between perceived patterns and appropriate grammatical word categories was established without explicit teaching, showing that syntax and semantics acquisition cannot be clear-cut separated, much like in the experiments of Sugita and Tani (2004) where a robot learned from scratch the compositional meaning of simple sentences from correspondences between sentences and sensory-motor patterns.

Several studies have stressed the role of verbal labels in motor learning. Helstrup (2000) findings support the hypotheses that motor sequences are coded as verbal strings rather than motorically or visuospatially; Frencham et al. (2004) found a better recall of hand movement sequences associated with verbal labels congruent with hand postures, still supporting the hypothesis that motor sequences are coded as verbal strings. These authors explain such results with a greater familiarity of verbal codes, easier to rehearse than actions. In our conditions, where two kinds of unfamiliar stimuli (verbal and motor) were associated, we can hypothesize that, assuming independent symbolic and analog representations, this coupling may rather reinforce a sort of mutual grounding. In fact, our findings support the idea that when an association is established between meaningless analog patterns and verbal symbols, grounding may work in a two-way direction: symbols become meaningful on the sensorimotor grounds, but also analog representations aspects (e.g., in our case, trajectories and postures) become more distinguishable when a specific label is available for them.

These remarks address also a possible issue stating that the use of two linguistic labels (words) in the compositional condition could have been a hint to look for two different components of the shown motor patterns[1]. Even if this turned out to be true, however, it could only be a demonstration that grounding can work bidirectionally, since in this case words had the power of facilitating the perceptual discriminations that in turn must necessarily be considered part of the grounding representations for the same words. This would also be evidence that grounding representations do not depend only on the visuomotor information, but language is fundamental. In any case, this does not cancel the fact that a compositional grounding was established but rather provide a further explanation of how it was obtained.

This somewhat Whorfian hypothesis, obviously, would deserve some deeper investigation, but is compatible with an interpretation of compositionality as a function of cognitive economy. If we take, as a baseline condition, that a single word for each motor pattern (group B, holistic condition) is matched with one fixed corresponding composite sentence (group A, compositional condition), then if group A performs *worst* than B, this indicates the *cost* of compositionality. On the other hand, if group A performs *better* than B, this indicates the *gain* of compositionality. Our results indicate that condition A led to a gain especially for patterns where hand postures *discrimination* was relevant. Some computational studies on language evolution (Kirby, 2002; Vogt, 2005; Smith et al., 2003) have claimed that compositional language has emerged in the cultural evolution as a consequence of the fact that examples actually encountered during verbal learning are necessarily limited (what has been called a *bottleneck* in cultural transmission); in this view, the advantage of compositionality is maximized in more structured

[1] We thank an anonymous referee for pointing out this issue.

environments, and depends on the structure of the meaning space, i.e., the number of distinctions that can be made. When distinct words are available for different aspects, attention can be best focused on such aspects, and this mutual grounding can further explain the cognitive gain of the compositional condition.

Results from our simulations clarify that such cognitive gain is not just the effect of a reduced memory load in a compositional symbol system. In fact, the environmental structure of meaning space is not important just because when the number of stimuli increases more information can be tackled with a smaller number of symbols, but because some symbols, by virtue of their grounding, can be reused as far as they are able to reinstantiate the same analog representations.

Our research can be continued in several directions. Our tasks only required that participants recognized or named visually presented (in videoclips) motor patterns. Interesting additional information about the compositional nature of motor representation could be provided by requiring the inverse performance, i.e., performance of corresponding motor patterns when being told their name. We did not implement this task for practical reasons, because we found it difficult to assess the correctness of motor performance. We are planning to use a robotic arm, programmed both to directly "teach" motor patterns associated with words/sentences and to assess how much the participant's subsequent performance matches with the original motor pattern.

Another significant improvement is trying different forms of composition. In tasks that we have considered, composition comes out from the combination of a motor pattern and a hand posture while executing it. There are theoretical reasons for supposing that other forms of composition could be considered: for example, considering other ways of executing the same pattern (slowly, firmly…), or as a concatenation of different motor patterns (e.g., left push + right push = grab). These two different composition kinds can be considered examples of the two different kinds of compositionality (concatenative and functional) described by van Gelder (1990). They can also be regarded as different ontological "wholes," as Meirav (2003) defined them: either as a sum (juxtaposition) or as a unit (according to a perspective or function).

It would be also of interest manipulating task informational complexity (e.g., the number of different words/features) and motor stimuli complexity. Using the motor coding system adopted in our networks, we can devise new simulations to obtain similarity and complexity scores by the analysis of different motor patterns, scores to be used for assessing the complexity of motor stimuli used in experimental tasks.

ACKNOWLEDGMENTS

The authors wish to thank Art Glenberg and Ivan Vankov for valuable remarks. This study was supported by European Union grant FP7-ICT-2007-1, as part of the research program I-TALK.

REFERENCES

Allott, R. (2003). *Language and Speech as Motor Activities*. Language Origins Society, Nijmegen 4–5 July 2003.

Amit, R., and Mataric, M. J. (2002). "Parametric primitives for motor representation and control," in *Proceedings of International Conference on Robotics and Automation (ICRA), May 11–15, 2002*, Washington, DC.

Arbib, M. A. (2005). From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics. *Behav. Brain Sci.* 28, 105–124.

Arndt, J., Lee, K., and Flora, D. B. (2008). Recognition without identification for words, pseudowords and non-words. *J. Mem. Lang.* 59, 346–360.

Barsalou, L. W. (1999). Perceptual symbols systems. *Behav. Brain Sci.* 22, 577–660.

Bizzi, E., and Mussa-Ivaldi, F. A. (2004). "Toward a neurobiology of coordinate transformations," in *Cognitive Neurosciences*, ed. M. S. Gazzaniga (Cambridge, MA: MIT Press), 413–425.

Cangelosi, A., Greco, A., and Harnad, S. (2000). From robotic toil to symbolic theft: grounding transfer from entry-level to higher-level categories. *Conn. Sci.* 12, 143–162.

Cleary, A. M, and Specker, L. E. (2007). Recognition without face identification. *Mem. Cognit.* 35, 1610–1619.

Deecke, L. (1996). Planning, preparation, execution, and imagery of volitional action (Introduction to the Special Issue "Mental representations of motor acts"). *Cogn. Brain Res.* 3, 59–64.

Fodor, J. A., and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition* 28, 3–71.

Frencham, K. A., Fox, A. M., and Maybery, M. T. (2004). Effects of verbal labeling on memory for hand movements. *J. Int. Neuropsychol. Soc.* 10, 355–361.

Gallese, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain* 119, 593–609.

Glenberg, A. M., and Kaschak, M. P. (2002). Grounding language in action. *Psychon. Bull. Rev.* 9, 558–565.

Greco, A. (1995). The concept of representation in psychology. *Cogn. Syst.* 4-2, 247–256.

Greco, A. and Caneva, C. (2005). "From actions to symbols and back: are there action symbol systems? COGSCI 2005," in *Proceedings of XXVII Annual Conference of the Cognitive Science Society, 2005 July 21–23,* Stresa.

Guerra-Filho, G., and Aloimonos, Y. (2006). Understanding visuo-motor primitives for motion synthesis and analysis. *J. Vis. Comput. Animat.* 17, 207–217.

Harnad, S. (1990). The symbol grounding problem. *Physica D* 42, 335–346.

Helstrup, T. (2000). The effect of strategies and contexts on memory for movement patterns. *Scand. J. Psychol.* 41, 209–215.

Jeannerod, M. (1988). *The Neural and Behavioural Organization of Goal-Directed Movements*, Vol. 15. Oxford: Clarendon Press.

Jeannerod, M. (1994). The representing brain: neural correlates of motor intention and imagery. *Behav. Brain Sci.* 17, 187–245.

Kirby, S. (2002). "Learning, bottlenecks and the evolution of recursive syntax," in *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, ed. E. Briscoe (Cambridge: Cambridge University Press), 173–203.

Kostic, B, and Cleary, A. M. (2009). Song recognition without identification: when people cannot "name that tune" but can recognize it as familiar. *J. Exp. Psychol. Gen.* 138, 146–159.

Langley, M. M., Cleary, A. M., Kostic, B. N., and Woods, J. A. (2008). Picture recognition without picture identification: a method for assessing the role of perceptual information in familiarity-based picture recognition. *Acta Psychol.* 127, 103–113.

Meirav, A. (2003). *Wholes, Sums and Unities*. Dordrecht: Kluwer Academic Publishers.

Peschl, M. (1997). The representational relation between environmental structures and neural systems: autonomy and environmental dependency in neural knowledge representation. *Non-linear Dynamics Psychol. Life Sci.* 1, 99–121.

Peynircioglu, Z. F. (1990). A feeling-of-recognition without identification. *J. Mem. Lang.* 29, 493–500.

Pulvermueller, F. (2005). Brain mechanisms linking language and action. *Nat. Rev. Neurosci.* 6, 576–582.

Ruschemeyer, S. A., Brass, M., and Friederici, A. D. (2007). Comprehending prehending: neural correlates of processing verbs with motor systems. *J. Cogn. Neurosci.* 19, 855–865.

Smith, K., Kirby, S., and Brighton, H. (2003). Iterated learning: a framework for the emergence of language, *Artificial Life*, 9, 371–386.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behav. Brain Sci.* 11, 1–74.

Sugita, Y. and Tani, J. (2004). "A holistic approach to compositional semantics: a connectionist model and robot experiments," in *Advances in Neural Information Processing Systems*, D. S. Touretzky, S. Thrun, L. K. Saul, and

B. Schölkopf (Cambridge, MA: MIT Press),969–976.

Thoroughman, K. A., and Shadmehr, R. (2000). Learning of action through combination of motor primitives. *Nature* 407, 742–747.

van Gelder, T. (1990). Compositionality: a connectionist variation on a classical theme. *Cogn. Sci.* 14, 355–384.

Vogt, P. (2005). The emergence of compositional structures in perceptually grounded language games. *Artif. Intell.* 167, 206–242.

Wu, L.-l., and Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: evidence from property generation. *Acta Psychol. (Amst.)* 132, 173–189.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: a review of 30 years of research. *J. Mem. Lang.* 46, 441–517.

Zwaan, R. A. (2004). "The immersed experiencer: toward an embodied theory of language comprehension," in *The Psychology of Learning and Motivation*, Vol. 44, ed. B. H. Ross (New York: Academic Press), 35–62.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 26 March 2010; paper pending published: 07 September 2010; accepted: 16 October 2010; published online: 11 November 2010.

Citation: Greco A and Caneva C (2010) Compositional symbol grounding for motor patterns. Front. Neurorobot. ***4****:111. doi: 10.3389/fnbot.2010.00111*

HYPOTHESIS AND THEORY ARTICLE
published: 28 May 2010
doi: 10.3389/fnbot.2010.00004

Sentence processing: linking language to motor chains

Fabian Chersi[1]*, Serge Thill[2], Tom Ziemke[2] and Anna M. Borghi[1,3]

[1] *Institute of Sciences and Technologies of Cognition, National Research Council, Rome, Italy*
[2] *School of Humanities and Informatics, University of Skövde, Skövde, Sweden*
[3] *Department of Psychology, University of Bologna, Bologna, Italy*

Edited by:
Angelo Cangelosi,
University of Plymouth, UK

Reviewed by:
Wolfram Erlhagen,
University of Minho, Portugal
Michel Hoen, Institut National de la Santé et de la Recherche Médicale, France
Vadim Tikhanoff,
Italian Institute of Technology, Italy

****Correspondence:***
Fabian Chersi, Istituto di Scienze e Tecnologie della Cognizione, National Research Council, Via S. Martino della Battaglia 44, 00185 Roma, Italy.
e-mail: fabian.chersi@istc.cnr.it

A growing body of evidence in cognitive science and neuroscience points towards the existence of a deep interconnection between cognition, perception and action. According to this embodied perspective language is grounded in the sensorimotor system and language understanding is based on a mental simulation process (Jeannerod, 2007; Gallese, 2008; Barsalou, 2009). This means that during action words and sentence comprehension the same perception, action, and emotion mechanisms implied during interaction with objects are recruited. Among the neural underpinnings of this simulation process an important role is played by a sensorimotor matching system known as the mirror neuron system (Rizzolatti and Craighero, 2004). Despite a growing number of studies, the precise dynamics underlying the relation between language and action are not yet well understood. In fact, experimental studies are not always coherent as some report that language processing interferes with action execution while others find facilitation. In this work we present a detailed neural network model capable of reproducing experimentally observed influences of the processing of action-related sentences on the execution of motor sequences. The proposed model is based on three main points. The first is that the processing of action-related sentences causes the resonance of motor and mirror neurons encoding the corresponding actions. The second is that there exists a varying degree of crosstalk between neuronal populations depending on whether they encode the same motor act, the same effector or the same action-goal. The third is the fact that neuronal populations' internal dynamics, which results from the combination of multiple processes taking place at different time scales, can facilitate or interfere with successive activations of the same or of partially overlapping pools.

Keywords: sentence comprehension, embodied cognition, motor system, neural network, action chains

INTRODUCTION

In recent years an increasing number of studies have adopted an embodied approach. According to embodied cognition theories (Barsalou, 2008), language is grounded in the sensorimotor system and language processing enhances previous sensorimotor experiences with objects or situations language refers to. Within the embodied approach, many studies focused on the role of motor simulation in language comprehension (e.g., Decety and Grèzes, 2006; Gallese, 2008). In particular, it has been highlighted that the comprehension of action verbs and action sentences involves the same sensorimotor and emotional brain circuits that are also activated during the actual interaction with the objects, situations and events the sentences refer to (for reviews, see Barsalou, 2008; Fischer and Zwaan, 2008; Toni et al., 2008). In particular, studies show that the simulation formed during language comprehension is sensitive to the involved effector (e.g., mouth, hand, leg). Although there is thus increasing evidence for a relation between action and language, the precise nature of this relation is still poorly understood. At the same time, an attractive aspect of this area of research is that both behavioral and neuroscientific data is available. In a sense, these are ideal conditions for carrying out computational modeling work that furthers our understanding of observed behavior. It is therefore our intention to use such an approach to elucidate the relationship between the neural mechanisms underlying language and the motor system.

Here we first review the relevant behavioral and brain imaging studies and emphasize the differences in results. We then present a computational model of the underlying neural circuits (based on the Chain model, see Chersi et al., 2006, 2007) which is able to account for the different findings. The model has been chosen as it is strongly motivated by neurophysiological findings which are relevant for the behavioral data discussed below. The fact that a single model can reproduce all results (in particular controversial and apparently conflicting data on timing in sentence comprehension) is a strong indication that they are not intrinsically in conflict but have rather captured different aspects of a single system.

REVIEW: STUDIES ON WORDS AND EFFECTORS

NEUROPHYSIOLOGICAL AND BRAIN IMAGING RESULTS

A number of neurophysiological and brain imaging studies have demonstrated that during action words and sentence comprehension different areas of the brain are activated depending on the effector (arm/hand, mouth, leg/foot) involved. The first study showing this was performed by Pulvermüller et al. (2001), who recorded neurophysiological data (specifically, they calculated event-related current source densities from EEG) pertaining to the processing of verbs referring to actions performed with the face, the arm/hand and the leg/foot. They found topographical differences in the brain activity patterns generated by the different verbs (e.g., to "lick", "pick", "kick") in a lexical decision task, starting at 250 ms after word onset. Hauk et al. (2004) confirmed the result with functional magnetic resonance imaging (fMRI). They found that during a passive reading task, words referring to face, arm, or leg actions differentially activated areas along the motor strip that were contiguous or overlapped with areas where that particular effector is represented. Tettamanti et al. (2005) showed with fMRI that passive listening to sentences expressing actions performed with the mouth, the hand or the foot led to signal increase in regions of the premotor cortex that are related to the effector involved in that sentence.

Overall, these studies thus reveal that during processing of words and sentences part of the brain is activated in a somatotopic way. Importantly, this early activation suggests that the activation of motor and premotor cortices is not simply a by-product. Rather, it appears to play an important functional role in action word comprehension even in tasks which require a rather shallow processing (such as lexical decision or even passive listening tasks). The hypothesis that the motor system is activated in a direct and straightforward way is much more plausible and economical compared to the idea that information is first translated into an abstract format which then influences the motor system (Mahon and Caramazza, 2008).

TRANSCRANIAL MAGNETIC STIMULATION RESULTS

Results like those reported above strongly suggest that the motor system activation is a fundamental part of the word and sentence comprehension process. However, it is still a matter of debate whether or not the activation of the motor system plays a causal role for sentence comprehension (for a position different from the one presented here, see Mahon and Caramazza, 2008). In addition, as we will show in this section, the actual effect the motor system activation can have on the comprehension process is not well understood. Results obtained in studies with Transcranial magnetic stimulation (TMS) are controversial, as some report facilitation while others find interference during the processing and execution of combinations of actions, verbs and action sentences.

Interference

In a recent study, Buccino et al. (2005) found an interference effect when the effector stimulated through TMS and the stimulus were congruent. More specifically, they acoustically presented three kinds of action sentences, referring to either hand action (e.g., he/she sewed the skirt), foot action (e.g., he/she kicked the door) or abstract content (e.g., he/she loved his land) related sentences. Participants were simply required to listen to the sentences. A TMS pulse was delivered at the end of the second syllable of the verb and motor evoked potentials (MEPs) were recorded from hand and foot muscles. Results showed a decrease in amplitude of MEPs recorded from hand muscles while listening to hand-action-related sentences, and from foot muscles when listening to foot-related sentences.

Facilitation

In contrast to the previous results, Pulvermüller et al. (2005) used a lexical decision task in which participants had to respond with a lip movement to arm- and leg-related words (e.g., "pick" vs. "kick"), and to refrain from responding to pseudowords. Transcranial magnetic stimulation pulses were delivered 150 ms after stimulus onset. Arm area TMS led to faster lexical decision times with arm words, whereas leg area TMS led to faster RTs with leg words; no facilitation was found in control conditions. A similar facilitation effect was found by Oliveri et al. (2004), who applied TMS to the left motor cortex when participants produced action-related and non-action verbs (e.g., "pour" vs. "detest") and nouns (e.g., "key" vs. "hill"). The motor cortex activation increased for action words (verbs and nouns) compared to non-action words during paired-pulse TMS at 10 ms ISI, no difference was present at 1 ms ISI. Recently, Papeo et al. (2009) recorded TMS-induced MEPs from right hand muscles. They found an increase of M1-activity only at 500 ms, while no increase was present when they delivered single pulse TMS at 170 and 350 ms after action verbs appearance.

BEHAVIORAL RESULTS

Interference

In a behavioral experiment performed by Buccino et al. (2005), participants were required to respond with either the hand or the foot if a presented verb was concrete and had to refrain from responding if the verb was abstract. Results showed that, if subjects responded with the same effector necessary for executing the action described by the sentence, response times were slower than if participants had to respond with the other effector. Sato et al. (2008) performed three experiments using a go-no go paradigm; participants had to answer with the right hand to verbs referring to hand actions (e.g., to applaud), foot actions (e.g., to walk) or abstract content (e.g., to love). Stimuli were presented both in the acoustic and visual modality. The authors manipulated both the task and the delivery of the go signal. More specifically, they used both a task implying shallower processing (a lexical decision task) and one implying deeper processing (a semantic decision task). In the semantic decision task, response times were slower with hand-related compared to foot-related verbs when the go signal was delivered early (at the isolation point). No effect was found with a late delivery of the go signal. In the lexical decision task no effect was found independently of the delivery of the go signal. This result suggests that the interference effect occurred only with deep semantic processing of sentences, and that it was confined to early delivery of the go-signal. In a kinematics study by Boulenger et al. (2006) participants were required to reach and grasp a cylindrical object. In the first experiment they had to start reaching when a fixation cross appeared, and continue moving when words appeared but stop for pseudowords. Words could either be verbs referring to hand, leg or mouth actions, or nouns representing non-manipulable objects. Results showed a modulation of kinematics parameters: processing action verbs interfered with concurrent early reaching movements.

Facilitation

Scorolli and Borghi (2007) extended the results of Buccino et al. (2005) using combination of nouns and verbs referring to hand, mouth and foot actions. Participants were presented with pairs of nouns and verbs that could refer to either hand and mouth actions (e.g., to unwrap vs. to suck the sweet) or to hand and foot actions (e.g., to throw vs. kick the ball). An equal number of non sensible pairs were presented. The participants' task consisted in deciding whether or not the combination made sense. Half of them were asked to respond by saying *yes* loudly into a microphone whereas the other half responded by pressing a pedal. If the combination did not make sense, they were invited to refrain from responding. The authors found a facilitation in response to "mouth" and "foot" sentences compared to "hand" sentences in case of congruency between the effectors – mouth and foot – involved in the motor response and in the sentence. It should be noted that the task, although different from the one by Buccino et al. (2005) and Sato et al. (2008), required deep semantic processing as well. Importantly, however, the presentation modality of the stimuli differed: the stimuli were presented visually and the noun was presented when the verb was processed. Given that Sato et al. (2008) did not find any difference in the stimulus modality (visual vs. auditory), and that both tasks require deep semantic processing, we have reason to believe that the most influential difference between the two studies is related to different timing.

Borghi and Scorolli (2009) performed experiments where, instead of using a go-no go paradigm, participants used both hands to choose between two possible answers. When pairs of words were presented that referred to manual and mouth actions, participants responded faster with the dominant hand. The advantage of the dominant hand was limited to sensible sentences.

Finally, in a second experiment of the same study by Boulenger et al. (2006) reported above, participants had to start reaching when a string of letters appeared on the screen. It was found that action verbs assisted the reaching movement when processed before movement onset. Despite the interest of this study, the results obtained are only partially relevant for our model, as a rather different paradigm was used, and kinematics measures were recorded, while our model focuses on RTs (see below).

A REASON OF THE DISCREPANCY: TIMING?

The discrepancies in TMS and behavioral results support the hypothesis that the precise task timings play a fundamental role in determining the type of interaction between language processing and action execution. For a similar interpretation see Boulenger et al. (2006), and, although not related to the role played by effectors during sentence comprehension, see Borreggine and Kaschak (2006) and De Vega et al. (2004).

All results support embodied theories as they demonstrate that there is a modulation of the motor system during sentence processing. However, the precise mechanisms underlying the conflicting data presented above are still poorly understood. In this respect, the detailed modeling of the possible processes could help to shed a new light on these phenomena. The model we will describe in the following section addresses this issue and leads to novel predictions.

MATERIALS AND METHODS

THE CHAIN MODEL

Recent neurophysiological experiments (Fogassi et al., 2005; Bonini et al., 2009) have shown that in the parietal and premotor cortices, the great majority of motor and mirror neurons coding a specific motor act (e.g., reaching, grasping, etc.) show markedly different activation patterns according to the final goal of the action sequence in which the act is embedded. More specifically, a neuron that is highly active during the grasping phase in a "grasping to eat" sequence may fire very little during a "grasping to place" sequence. These results have led to the hypothesis that motor and mirror neurons in the parietal and premotor cortices are organized in chains that encode short habitual action sequences (Chersi et al., 2006, 2007). According to this view, for example, the action of taking a piece of food is encoded as the concatenation of neurons that represent the reaching, the grasping and the retrieving motor act (see **Figure 1**). The execution and the comprehension of motor sequences correspond to the propagation of activity within specific chains. This chained organization allows a smooth and automatic execution of action sequences, and can be used to mentally simulate action sequences by "running" chains decoupled from the overt motor output.

Due to the dual property of mirror neurons (i.e., the fact that they are active both during execution and observation of action sequences executed by others) mirror chains can be used to understand others' actions and intentions by mapping the observed acts on one's own motor repertoire.

OUR HYPOTHESIS

Taken together, the reviewed results strongly support the notion that the processing of language stimuli, at least for sentences expressing a motor content, modulates the activity of the motor system and that this modulation specifically concerns those sectors of the motor system where the effector involved in the processed sentence is represented. Interestingly, depending on the temporal relation between language and motor tasks, processing action words can facilitate or interfere with overt motor behavior.

The model we propose to explain these observations is based on three main points. First, the processing of action-related sentences involves the chained activation of specific pools of mirror neurons that encode the motor acts referred to in the sentences (Chersi et al., 2006). This is the same mechanism as the one taking place during the recognition of actions done by other individuals.

Second, as shown by recent experiments (Fogassi et al., 2005; Bonini et al., 2009), part of the neurons representing a motor act (e.g., reaching) embedded in a sequence dedicated to a specific goal (e.g., grasping an object) respond also when the same act is embedded in another sequence (e.g., pressing a button).

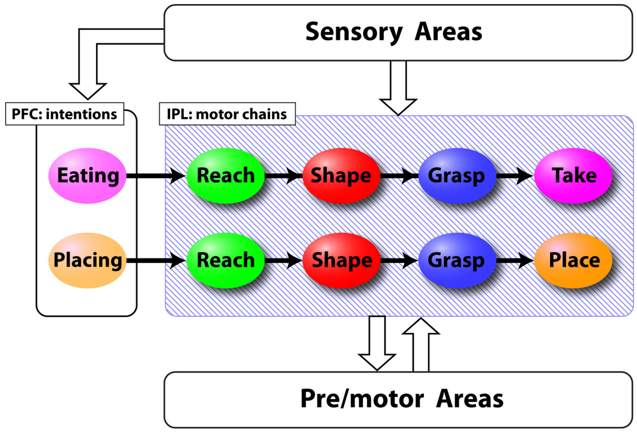


FIGURE 1 | Schematic representation of the chain model derived from Chersi et al. (2007). Colored ellipses represent pools of neurons that encode specific motor acts in the parietal cortex (IPL) or intentions in the prefrontal cortex (PFC). Lines indicate the connections between different pools. Sensory areas provide information about the ongoing action and pre/motor areas interpret abstract commands to generate motor output.

The third point concerns the dynamics of neuronal pools. The detailed analysis of the experiments reported above has revealed that interference occurs between 160 and 500 ms after stimulus presentation, whereas facilitation becomes evident between 550 and 800 ms after sentence appearance (Boulenger et al., 2006). These time scales suggest that short term neural dynamics may be the cause underlying these phenomena. *In vitro* recordings have shown that neuronal responses result from the combination of several dynamic processes occurring at different time scales. In general it is possible to distinguish two main components that determine the neuronal response: (1) an early but brief buildup of ionic currents (typically potassium) that causes an adaptation of the firing rates; (2) a slow but long lasting accumulation of neurotransmitters (NMDA, GABA, AMPA) and other ions (e.g., calcium) that facilitate neuronal firing. More precisely, for high enough spike frequencies a calcium-dependent potassium current (see e.g., McCormick et al., 1985; Sah, 1996) builds up lasting up to a few hundred milliseconds and reducing the firing frequency of neurons. Simultaneously, due to incoming spikes the concentration of neurotransmitters increases rapidly and fades away slowly after the input has ceased (this is especially true for NMDA). Additionally, the accumulation of calcium (Powers et al., 1999) produces a spiking facilitation effect that can last up to more than a second. Taken together these effects produce a time window (up to half a second after stimulation) during which neurons decrease their firing rate and thus reach their maximum activity more slowly, and a facilitation time window (from half a second to about a second) during which pools react more rapidly.

The general mechanism proposed in our study is therefore the following. During the processing of an action-related sentence, pools of mirror neurons that encode the single phases (motor acts) of the expressed action are activated due to a motor resonance mechanism. Neuronal activity propagates along the chain and sequentially activates the motor neurons connected downstream. Although pools fire only for a short interval of time (around 200–300 ms) synaptic currents decay at a much slower rate due to their slower internal dynamics. The firing rate adaptation current is active shortly after the firing of the pool causing a momentary activity slowdown. When a response action has to be produced, the prefrontal cortex (PFC) activates the corresponding neuronal chain. The precise activation profile of each pool in the chain will depend on the degree of overlap it has with any previously activated pools of other chains and on how big the time interval between the activations is. More precisely, the larger the overlap, the stronger the influence. Furthermore, pools will respond faster or more slowly depending on whether their activation falls within the adaptation or the facilitation phase of previous pools.

SIMULATED EXPERIMENT

In order to test our hypothesis, we simulated an experiment by virtually combining those by Buccino et al. (2005) and Scorolli and Borghi (2007) previously discussed. In our experiment, a hypothetical subject has to watch a screen where one of two short sentences can appear. The first sentence is "to grasp the apple", while the second one is "to kick the ball". The subject has to read the sentence and, when the "Go" signal is given, reach and press a button. The delay between the sentence presentation and the "Go" signal varied between 200 and 1200 ms.

We suppose that the first of the two sentences is represented as the concatenation of a "reaching with the hand" and a "grasping" motor act; the second as a "reaching with the foot" followed by a "hitting" motor act. The action that the subject has to perform consists in a "reaching with the hand" and "pressing" motor act. Each action is encoded by a neuronal chain composed of pools that represent the different motor acts. When the subject reads the displayed sentences, neurons that encode the described motor acts start to fire due to a mirror resonance process. More precisely, if the participant reads the first sentence, initially the "reaching (with the hand)", then the "grasping" pools are activated. If the subject reads the second sentence, first the "reaching (with the foot)" then the "kicking" pool is activated. When the subject has to respond by pressing a button, the "reaching-pressing" chain is run, i.e., the "reaching" pool is activated first and this in turn activates the "pressing" pool.

One important characteristics of our model is that neuronal pools encoding the same motor act (involving the same effector) but being part of different chains share a small fraction of neurons and axonal projections. In our case, the common part between the action described in the first sentence and the subject's motor response is the "reaching" motor act. Consequently, the pool encoding "reaching" in the "reaching-pressing" chain is partially activated when the sentence is read. If the "reaching-pressing" chain is then executed shortly afterwards, the previously activated sub-threshold dynamics affect the firing rate of the pool in either a positive or negative way.

In our simulated experiment the elaboration of the sentence is assumed to last around 300 ms, with the peak to peak time interval between two pools being around 150 ms. We would like to emphasize that the motor content of each sentence is independent of the agent (here impersonal) and of the target objects ("the apple", "the ball", "the button"), all of which are not explicitly encoded in the chain but rather considered as parameters of the action. Note that this is possible because mirror neurons do not explicitly encode the agent of an action nor the objects involved.

NETWORK CONFIGURATION

The neural network we used in our simulations was composed of six pools of neurons, each one coding a specific motor act. The pools were arranged in three chains of two pools each (see **Figure 2**).

MODEL DETAILS

The behavior of each neuronal pool is described by a firing rate model with time-dependent synaptic currents (Dayan and Abbott, 2001). This allows us to both compactly represent complex interactions between excitatory and inhibitory neurons within the pools and explicitly take into account the dynamics of ionic currents and neurotransmitters. The set of equations is the following:

$$\begin{cases} \tau_v \dfrac{d\nu_i}{dt} = -\nu_i + g(I_{\text{syn},i} - I_{\text{fra},i}) + \eta \\ \tau_I \dfrac{dI_{\text{syn},i}}{dt} = -I_{\text{syn},i} + I_{\text{ext},i} + \sum_h W_{hi} \cdot \nu_h \\ \dfrac{dI_{\text{fra},i}}{dt} = \sum_h \alpha_{hi} \cdot \nu_h - \beta \end{cases} \qquad (1)$$

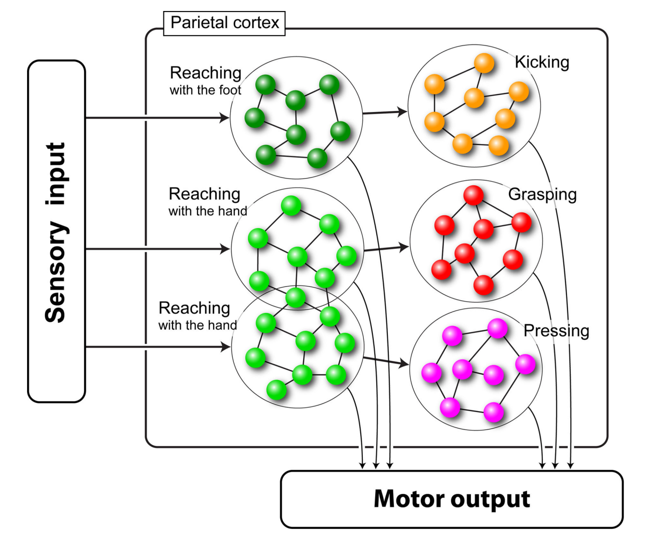


FIGURE 2 | Schematic representation of the chained organization of the network. Each large circle represents a pool of neurons (small spheres) encoding a specific motor act. Lines represent the connections between neurons. Lines on the left represent external inputs that start the chains.

where ν_i is the mean firing rate of the i-th pool and $\tau_\nu = 70$ ms the corresponding time constant, $g()$ is the I–f pools' response function (see below), η is an additional term that simulates spontaneous activity, $I_{\text{syn},i}$ is the total synaptic current and $\tau_I = 260$ ms the corresponding time constant, $I_{\text{fra},i}$ is the firing rate adaptation current, W_{hi} is the connection strength from unit h to unit i, and $I_{\text{ext},i}$ is the external input current arriving from areas that are active while reading the sentence or executing an action. This signal has been modeled as a bell shaped activity peak lasting 200 ms.

In the present implementation a fitting procedure has been used in order to determine the synaptic weights that produce the activation of pools encoding subsequent motor acts in each chain with the correct timing and amplitude (yielding $W_{i,i+1} \approx 0.03$). Furthermore, the connectivity (i.e., the overlap) between the first pool of the "reaching-grasping" chain and the first pool of the "reaching-pressing" chain (both pools encoding "reaching") has been set to a value that produces an activation of 30% of the maximum firing rate when the other chain is activated ($W_{hi} \approx 0.02$). All other connections (including self connections) have been set to zero. Note that the "reaching with the hand" pools have no overlap with the "reaching with the foot" pool because the effectors involved are not the same.

The firing rate adaptation has been modeled as a current, that, when activated, will hyperpolarize the neurons of a pool, slowing down any spiking that may be occurring. We assume that this current is proportional (through α) to the firing rate and relaxes to zero at a rate of β. In our implementation $\alpha \approx 0.09$ nA and $\beta \approx 3$ nA/s.

In order to reproduce more faithfully the behavior of real neurons (in particular the fact that there is a minimum value for the injected current below which no firing takes place) the pools' response function has been modeled in the following way:

$$\begin{cases} g(I) = g_0 \cdot \tanh\left[\gamma(I - I_{\text{thr}})\right] & \text{for } I > I_{\text{thr}} \\ g(I) = 0 & \text{for } I \leq I_{\text{thr}} \end{cases} \qquad (2)$$

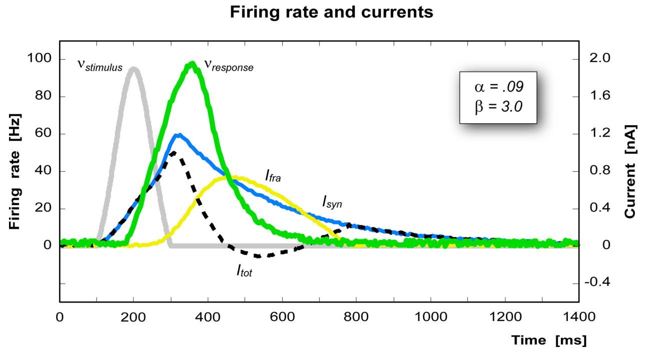


FIGURE 3 | Time course of each variable of a pool after stimulation (gray peak). The green curve represents the response of the pool (scale on the left), the blue curve is the synaptic current, the yellow curve is the firing rate adaptation current, and the dashed line is the resultant total current (current scale on the right).

where g_0 determines the maximum firing rate, γ determines the steepness of the response and I_{thr} is the firing threshold. In this implementation we have chosen $g_0 = 150$ Hz, $\gamma = 1.5$ cm^2/nA, and $I_{\text{thr}} = 0.25$ nA/cm^2. All the parameters in this model have been chosen in order to reproduce as close as possible biological data. **Figure 3** shows the currents and the firing rate of a single pool in response to an external stimulus.

The parameters α and β determine the shape of the firing rate adaptation current curve I_{fra}. Increasing α, for instance, increases the influence of the firing rate on the growth of I_{fra}, which in turn decreases the firing rate (**Figure 4A**). Decreasing β instead causes a slower deactivation of I_{fra} thus shifting the inhibitory phase of I_{tot} further in time (**Figure 4B**).

RESULTS

The results of the simulated experiment are reported in **Figure 5**. **Figure 5A** shows the activity profile of the "reaching" pool (green curve) of the "reaching-grasping" chain activated by the presentation of the sentence "to grasp the apple" (gray curve). In our implementation this input (I_{ext}) is simulated as a bell shaped activation of the duration of 200 ms. Note that both in the experiments and in the model each sentence is considered as a whole. The detailed modeling of single words comprehension is beyond the scope of this paper. The pool reaches its maximum activity 254 ms after stimulus onset. **Figure 5B** shows the response of the "reaching" pool of the "reaching-pressing" chain. The first bump is due to the crosstalk between the first chain and the second chain. The Go signal (gray curve) is given 350 ms after the stimulus presentation. The activity peak is reached 276 ms after the Go signal.

Figure 5C shows the response of the same pool to the presentation of the same sentence and below the response of "reaching" pool of the "reaching-pressing" chain when the Go signal is given 650 ms after the sentence presentation. In order to remove the reaction time component due solely to the physical execution of the action (executed only virtually in our case), we calculate the "facilitation factor" ($\Delta t_D - \Delta t_C$) and the "interference factor" ($\Delta t_B - \Delta t_A$) as the decrease or increase of the reaction time of the specific task compared to the control task.

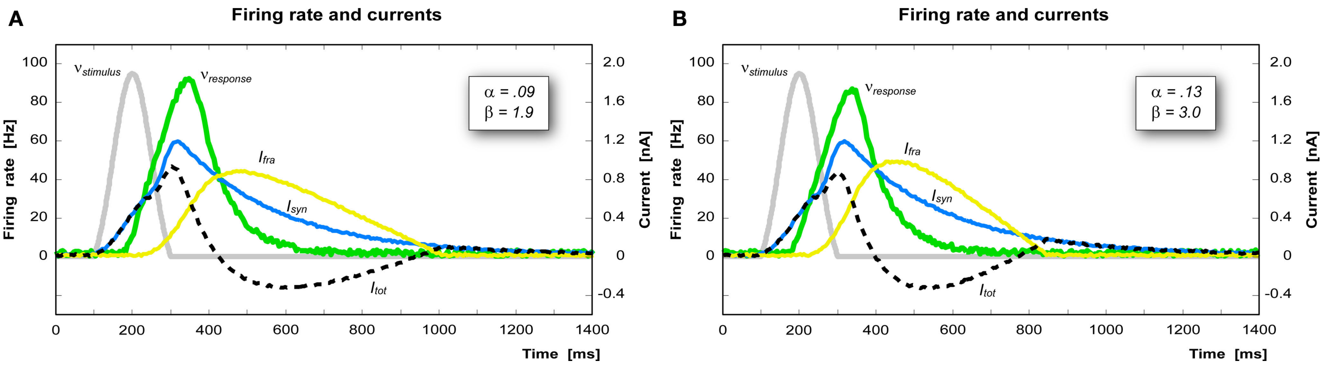


FIGURE 4 | Effect of the parameters α and β on the shape of the firing rate adaptation current. (A) A decrease of β (from 3.0 to 1.9 nA/s) produces a slower deactivation of I_{fra}, thus shifting the inhibitory phase further in time. **(B)** An increase of α (from 0.09 to 0.13 nA) produces a more intense I_{fra} and thus a stronger inhibition in case of re-activation of the same pool within 700 ms.

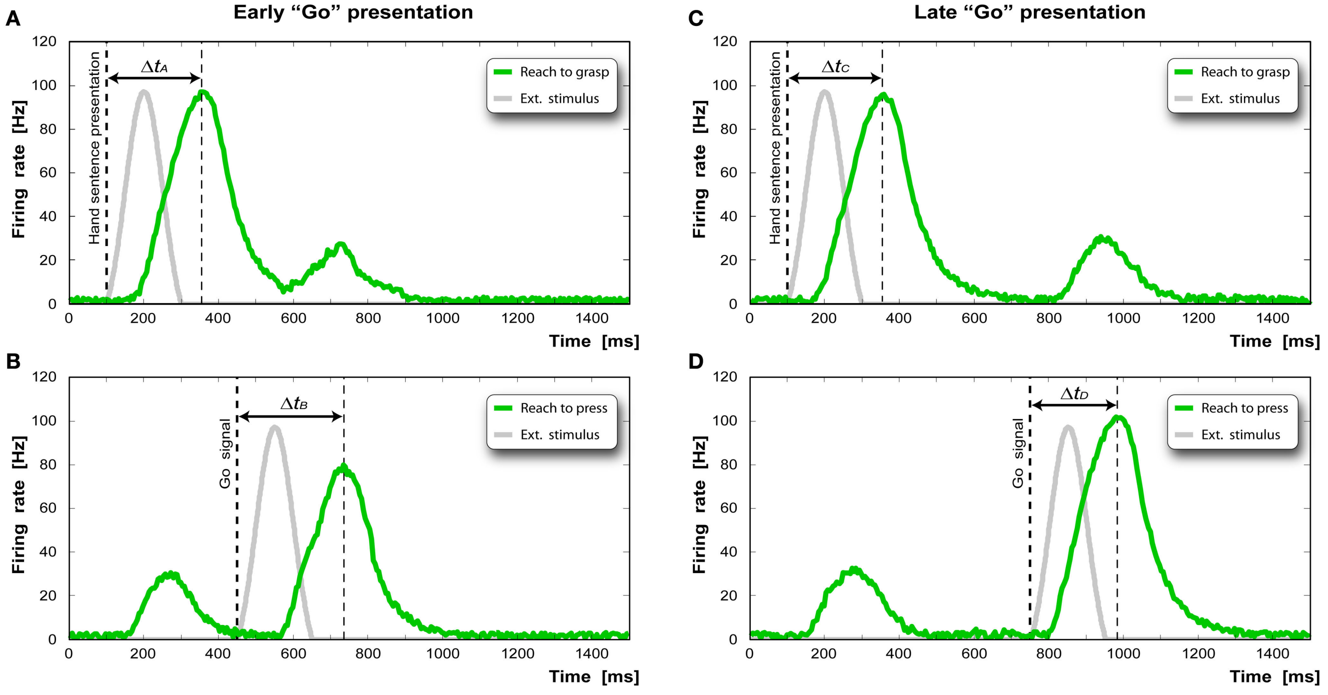


FIGURE 5 | (A,C) Time course of the activation of the "reaching" pool in the "reaching-grasping" chain after the presentation of the hand-related sentence. **(B,D)** Time course of the activation of the "reaching" pool in the "reaching-pressing" chain after the appearance of the "Go" signal.

In our simulations, we obtain a facilitation factor of –25 ms, and an interference factor of 20 ms, which is comparable to the results found by Buccino et al. (2005).

Figure 6 shows the time course of the activation of the "reaching-pressing" chain after the presentation of the two sentences in the late Go signal condition. **Figure 6A** represents the sequential activation of the "reaching" (green curve) and the "grasping" pool (red curve) after the presentation of the sentence "to grasp the apple" and a late Go signal. **Figure 6B** represents the activation of the same pools after the presentation of the sentence "to kick the ball" and a late Go signal. As can be seen reading a sentence that contains a motor act present also in the response motor sequence produces an overall decrease in the reaction time of 25 ms.

Figure 7 shows the modulation of the reaction times of the simulated "reaching" pool as a function of the time interval between the presentation of the hand-related sentence and the "Go" signal. The reaction time data shows that there is a first phase in which interference dominates (up to 500 ms) and a phase in which facilitation dominates. This effect eventually fades to zero. In our model, for time intervals below 200-ms input signals overlap and pools' responses merge thus not allowing a clear interpretation of the results.

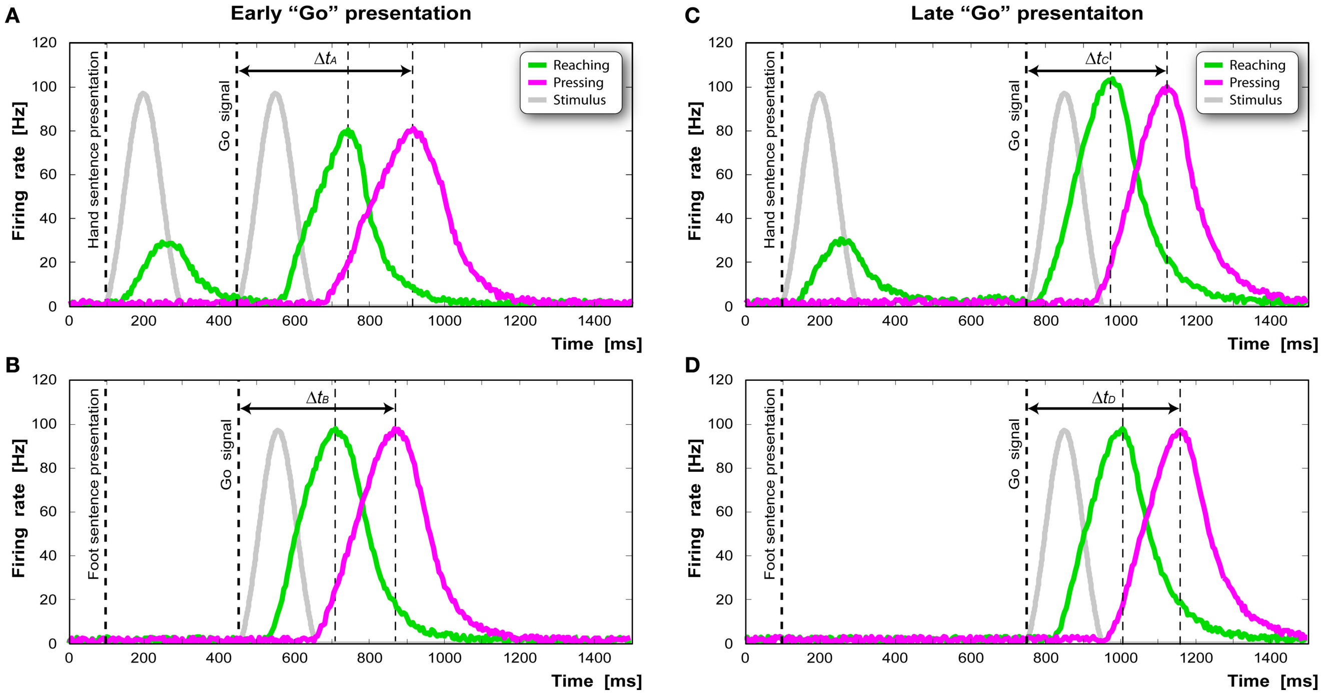


FIGURE 6 | Activation profile of the "reaching-pressing" chain after the presentation of the two sentences in the simulated task. The dashed vertical lines indicate the sentence presentation onset (left), the Go signal presentation (middle), and moment of maximal activity (right). In case of early "Go" presentation, the "reaching-pressing" chain responds slower when the hand sentence was presented **(A)** compared to when the foot sentence was presented **(B)**. In case of late "Go" presentation, the effect is the opposite [**(C)** vs. **(D)**].

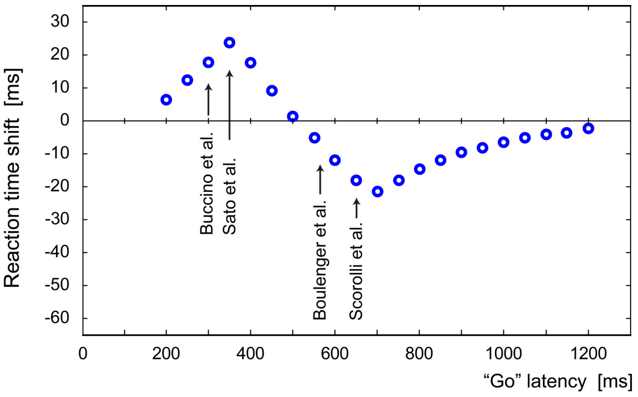


FIGURE 7 | Modulation of the reaction times of the simulated "reaching" pool as a function of the time interval between the presentation of the "hand" sentence and the "Go" signal. Reported below are the timings of the experiments of Buccino et al. (2005), Sato et al. (2008), Boulenger et al. (2006) (Exp. 2) and Scorolli and Borghi (2007).

DISCUSSION AND CONCLUSION

As reviewed in the first part of the paper, both interference and facilitation are widely observed in TMS and behavioral experiments on language comprehension and motor system activation. The underlying mechanisms, however, are a topic of ongoing debate. It is interesting to note that one can find similar facilitation and interference effects also in the action observation literature (e.g., Brass et al., 2001). In the present work, however, we focused on the controversial results related to language processing.

Recently, Sato et al. (2008) postulated that the cause may be the nature and the deepness of the involvement of the motor system determined by the different difficulty of the single tasks. Boulenger et al. (2006) hypothesized that facilitation could result from side or after-effects of linguistic processes while competition for common resources, for instance, could give rise to interference.

In this work we proposed a simple neural mechanism that is capable of explaining both the facilitatory and the inhibitory interactions between language and action. Our model is based on a chain structured organization of the parietal and premotor cortex (Fogassi et al., 2005; Chersi et al., 2007) in which action sequences are encoded as concatenations of neuronal pools representing specific motor acts. Interactions between sensory and motor modalities have been modeled in the present work as a crosstalk between neuronal pools in motor and mirror chains and we have shown that the neural dynamics governing the activation of the pools can qualitatively reproduce the timings observed in behavioral experiments well.

Taken together, these results allow us to draw the following conclusions. First, the fact that our simple model can reproduce different experimental results by exploiting only "low level" properties of neurons supports the idea that these interaction effects might be principally due to neurodynamical factors within the mirror neuron circuit rather than to high-level cognitive processes. Second, this unifying theory suggests that seemingly conflicting behavioral experiments may have observed different time windows of the same mechanism rather than different mechanisms,. This has important theoretical implications because, as previously discussed, it is currently debated in the literature whether the activation of motor and premotor cortices is essential for language understanding or just a by-product of the process. The early activation of the motor system is typically considered a strong point in support for the first thesis. Showing that interference and facilitation are actually two manifestations of the same process greatly strengthens the embodied view according to which the recruitment of the motor system is fundamental for sentence comprehension.

Finally, on the basis of our model we can formulate a variety of predictions that could guide future experimental research.

(1) It should be possible to produce precise interference and facilitation profiles by carefully designing experiments.
(2) If language processing produces a modulation of action execution timings due to the overlap of neural representations, it is reasonable to expect that action execution has the same effect on language processing because overlaps are most probably bidirectional.
(3) Since timing variations are supposed to be caused by the reactivation of neuronal pools, it should be possible to obtain a similar or even greater interaction effect if the tasks were "language following language" or "action following action".
(4) The fact that modeling results support the idea that all the interaction effects between language and action might be due principally to neurodynamical processes taking place within the mirror neuron circuit rather than to high-level cognitive processes, leads us to think this might be a general principle valid for other sensorimotor interactions as well.
(5) If more perceptual modalities exploit the same motor representations it should be possible to observe interactions between these modalities mediated by the common motor substrate.
(6) If the neurodynamical and the embodiment hypotheses are true then we expect to find a mixture of interference and facilitation patterns also in tasks that involve, for example, object affordances (AIP-F5 circuit) and interpersonal interactions (PG-F4 circuit).
(7) Using more sophisticated experimental and/or data analysis techniques, such as for example signal correlation studies, it should be possible to discover weak or very late interactions.

Notwithstanding these interesting results, we are perfectly aware that the mechanisms coming into play during the elaboration of stimuli and decision making are much more complex than depicted here, so our proposal should be considered as a first attempt to model such a complex system. We believe that this computational modeling work may also prove useful in building a biologically inspired robotic model for use in human–humanoid interaction, which is the longer-term goal of this work. From this perspective it is important that embodiment is taken into account at an appropriate level of abstraction that allows computational models of human biological mechanisms to be transferred to a robotic context. Furthermore, from a scientific perspective, it is clear that additional targeted experimental and modeling work is necessary to better understand the mechanisms underlying the relationship between sentence comprehension and motor system activation. As a first step, however, we believe it was important to show in this paper that interference and facilitation may well be two sides of the same coin.

ACKNOWLEDGMENTS

This work was supported by the FP7 project ROSSI, "Emergence of communication in Robots through Sensorimotor and Social Interaction", Grant agreement no. 216125, and by the project IM-CLeVer, "Intrinsically Motivated Cumulative Learning Versatile Robots", Grant agreement no. 231722. A special thank goes to Gianluca Baldassarre for his help and support.

REFERENCES

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.* 59, 617–645.

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 364, 1281–1289.

Bonini, L., Rozzi, S., Ugolotti, F., Maranesi, M., Ferrari, P. F., and Fogassi, L. (2009). Ventral premotor and inferior parietal cortices make distinct contribution to action organization and intention understanding. *Cereb. Cortex.* doi: 10.1093/cercor/bhp200.

Borghi, A. M., and Scorolli, C. (2009). Language comprehension and hand motion simulation. *Hum. Mov. Sci.* 28, 12–27.

Borreggine, K. L., and Kaschak, M. (2006). The action-sentence compatibility effect: its all in the timing. *Cogn. Sci.* 30, 1097–1112.

Boulenger, V., Roy, A. C., Paulignan, Y., Deprez, V., Jeannerod, M., and Nazir, T. A. (2006). Cross-talk between language processes and overt motor behavior in the first 200 msec of processing. *J. Cogn. Neurosci.* 18, 1607–1615.

Brass, M., Bekkering, H., and Prinz, W. (2001). Movement observation affects movement execution in a simple response task. *Acta Psychol. (Amst)* 106, 3–22.

Buccino, G., Riggio, L., Melli, G., Binkofski, F., Gallese, V., and Rizzolatti, G. (2005). Listening to action related sentences modulates the activity of the motor system: a combined TMS and behavioral study. *Cogn. Brain Res.* 24, 355–363.

Chersi, F., Fogassi, L., Bonini, L., Rizzolatti, G., and Ferrari, P. F. (2007). Modeling intentional neuronal chains in parietal and premotor cortex. *Soc. Neurosci. Abstr.* 636.6, DDD23.

Chersi, F., Mukovskiy, A., Fogassi, L., Ferrari, P. F., and Erlhagen, W. (2006). A model of intention understanding based on learned chains of motor acts in the parietal lobe. *Comput. Neurosci.* 69, 48.

Dayan, P., and Abbott, L. F. (2001). *Theoretical Neuroscience. Computational and Mathematical Modelling of Neural Systems*. Cambridge, MA: MIT Press.

De Vega, M., Robertson, D. A., Glenberg, A. M., Kaschak, M. P., and Rinck, M. (2004). On doing two things at once: temporal constraints on actions in language comprehension. *Mem. Cognit.* 32, 1033–1043.

Decety, J., and Grèzes, J. (2006). The power of simulation: imagining one's own and other's behavior. *Brain Res.* 1079, 4–14.

Fischer, M., and Zwaan, R. (2008). Embodied language: a review of the role of the motor system in language comprehension. *Q. J. Exp. Psychol.* 61, 825–850.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., and Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. *Science* 308, 662–667.

Gallese, V. (2008). Mirror neurons and the social nature of language: the neural exploitation hypothesis. *Soc. Neurosci.* 3, 317–333.

Hauk, O., Johnsrude, I., and Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron* 41, 301–307.

Jeannerod, M. (2007). *Motor Cognition: What Actions Tell the Self*. Oxford: Oxford University Press.

Mahon, B. Z., and Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *J. Physiol. (Paris)* 102, 59–70.

McCormick, D. A., Connors, B. W., Lighthall, J. W., and Prince, D. (1985). Comparative electrophysiology of pyramidal and sparsely stellate neurons of the neocortex. *J. Neurophysiol.* 54, 782–806.

Oliveri, M., Finocchiaro, C., Shapiro, K., Gangitano, M., Caramazza, A., and Pascual-Leone, A. (2004). All talk and no action: a transcranial magnetic stimulation study of motor cortex activation during action word production. *J. Cogn. Neurosci.* 16, 374–381.

Papeo, L., Vallesi, A., Isaja, A., and Rumiati, R. I. (2009). Effects of TMS on different stages of motor and non-motor verb-processing in the primary motor cortex. *PLoS ONE* 4, e4508. doi: 10.1371/journal.pone.0004508.

Powers, R. K., Sawczuk, A., Musick, J. R., and Binder, M. D. (1999). Multiple mechanisms of spike-frequency adaptation in motoneurones. *J. Physiol.* 93, 101–114.

Pulvermüller, F., Härle, M., and Hummel, F. (2001). Walking or talking? Behavioral and electrophysiological correlates of action verb processing. *Brain Lang.* 78, 143–168.

Pulvermüller, F., Hauk, O., Nikulin, V. V., and Ilmoniemi, R. J. (2005). Functional links between motor and language systems. *Eur. J. Neurosci.* 21, 793–797.

Sah, P. (1996). Ca-activated K currents in neurons: types, physiological roles and modulation. *Trends Neurosci.* 19, 150–154.

Sato, M., Mengarelli, M., Riggio, L., Gallese, V., and Buccino, G. (2008). Task related modulation of the motor system during language processing. *Brain Lang.* 105, 83–90.

Scorolli, C., and Borghi, A. M. (2007). Sentence comprehension and action: effector specific modulation of the motor system. *Brain Res.* 1130, 119–124.

Rizzolatti, G., and Craighero, L. (2004). The mirror neuron system. *Annu. Rev. Neurosci.* 27, 169–192.

Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., Fazio, F., Rizzolatti, G., Cappa, S. F., and Perani, D. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *J. Cogn. Neurosci.* 17, 273–281.

Toni, I., de Lange, F. P., Noordzij, M. L., and Hagoort, P. (2008). Language beyond action. *J. Physiol. (Paris)* 102, 71–79.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 December 2009; paper pending published: 29 January 2010; accepted: 27 April 2010; published online: 28 May 2010.

Citation: Chersi F, Thill S, Ziemke T and Borghi AM (2010) Sentence processing: linking language to motor chains. Front. Neurorobot. ***4****:4. doi: 10.3389/fnbot.2010.00004*

frontiers in
NEUROROBOTICS

ORIGINAL RESEARCH ARTICLE
published: 03 June 2010
doi: 10.3389/fnbot.2010.00006

Reading as active sensing: a computational model of gaze planning in word recognition

Marcello Ferro[1,2], Dimitri Ognibene[2], Giovanni Pezzulo[1,2]* and Vito Pirrelli[1]

[1] *Istituto di Linguistica Computazionale "Antonio Zampolli" – CNR, Pisa, Italy*
[2] *Istituto di Scienze e Tecnologie della Cognizione – CNR, Rome, Italy*

Edited by:
Angelo Cangelosi, University of Plymouth, UK

Reviewed by:
Frank van der Velde, Leiden University, Netherlands
Marco Mirolli, Istituto di Scienze e Tecnologie della Cognizione, Italy

****Correspondence:***
Giovanni Pezzulo, Istituto di Scienze e Tecnologie della Cognizione - CNR, Via S. Martino della Battaglia, 44 - 00185 Rome, Italy.
e-mail: giovanni.pezzulo@cnr.it

We offer a computational model of gaze planning during reading that consists of two main components: a *lexical representation network*, acquiring lexical representations from input texts (a subset of the Italian CHILDES database), and a *gaze planner,* designed to recognize written words by mapping strings of characters onto lexical representations. The model implements an active sensing strategy that selects which characters of the input string are to be fixated, depending on the predictions dynamically made by the lexical representation network. We analyze the developmental trajectory of the system in performing the word recognition task as a function of both increasing lexical competence, and correspondingly increasing lexical prediction ability. We conclude by discussing how our approach can be scaled up in the context of an active sensing strategy applied to a robotic setting.

Keywords: reading, active sensing, SOM, prediction, serial order encoding, lexical representation network

INTRODUCTION

The human visual system is essentially active, its processing strategies being tightly coupled with the specific demands of an ongoing task (Yarbus, 1967; Ballard, 1991; Johansson et al., 2001; O'Regan and Nöe, 2001). There is ample evidence that in everyday activities, such as driving, walking or reading, gaze shifts are used to gather task-relevant information (Triesch et al., 2003; Hayhoe and Ballard, 2005; Land, 2006). Whenever possible, this is done through efficient, timely selection of the specific information required for a given stage of the task to be carried out, with no need to store information (Ballard et al., 1995). In most tasks, since visual information is required at the very early stages of action planning, the strategy gives rise to anticipatory saccades (e.g., by fixating objects that are manipulated shortly later, or even seconds later).

One visual task that has been the focus of intense investigation is text reading. Somewhat contrary to commonsense, it does not consist in the serial fixation of written words from left-to-right, but it is a truly active task. In reading a text, some words are skipped, and occasionally a gaze regression is made to words that were either already fixated, or skipped. Patterns of eye movements (including, among other things, the time spent on each fixation and the average distance the eyes move along while scanning a text) are complex and depend on a number of factors, including word frequency, lexical predictability and ambiguity, complexity in the syntactic structure of input text etc. (see Rayner, 2009 for a recent review).

In line with this evidence, the present paper intends to investigate the interlocked relationship between *processes of self-organizing lexical storage and learning* on the one hand, and, on the other hand, *active sensing strategies for reading* that exploit expectations on stored lexical representations to drive gaze planning. For this purpose, we shall capitalize on currently emerging views on morphological processing and on the role of anticipatory processes in reading.

Word processing has recently been conceptualized as the outcome of simultaneously activating patterns of cortical connectivity, reflecting (possibly redundant) distributional regularities in the input at the graphemic, morpho-syntactic and morpho-semantic levels (Burzio, 2004; Baayen, 2007; Post et al., 2008). This view argues for a more complex and differentiated neurobiological substrate for human language than both classical dual-route (Pinker and Prince, 1988; Prasada and Pinker, 1993; Pinker and Ullman, 2002; Ullman, 2004) and connectionist one-route models (McClelland and Patterson, 2002; Westermann and Plunkett, 2007) can posit. Brain areas devoted to word processing appear to maximize the opportunity of using both general and specific information simultaneously (Libben, 2006), rather than maximize processing efficiency and economy of storage.

Topological models of lexical self-organization can shed light on such a dynamic view of word processing from a computational perspective (Pirrelli, 2007; Pirrelli et al., in press). In these models, lexical storage and learning is based on the concurrent self-organization of "spatial" word-based information (e.g. segmental or graphemic patterns) and temporal (i.e. sequential) information, accounting for concomitant effects of redundant morphological structure and predictive parsing, as well as for short-term and long-term memory effects in the encoding and processing of symbolic sequences. This makes spatio-temporal self-organizing networks of this kind ideally suitable for investigating anticipatory processes in word recognition and reading.

Experimental studies based on ERP (event-related potentials) and eye-movement evidence show that people use prior (lexical and semantic) contextual knowledge to anticipate upcoming words (Altmann and Kamide, 1999; Federmeier, 2007). DeLong et al. (2005) demonstrate that expected words are pre-activated in subjects' brain in a graded fashion, reflecting their expected probability. This body of evidence provides a solid empirical ground to the probabilistic approach to lexical prediction and gaze planning proposed here. In our model, the probability distribution of stored lexical representations is the main input to the gaze planner, since (parts of) words predicted with high accuracy can be skipped safely during reading (as demonstrated empirically by Ehrlich and Rayner, 1981; Rayner and Well, 1996). Moreover, new information that is (retrospectively) judged as unpredictable and surprising can determine longer fixations, regressions, or revision of lexical representations.

The aforementioned evidence provides the foundations of our modeling approach to gaze planning, in which two components interact: a lexical representation network, and a gaze planner proper. We offer a model of how lexical representations and lexical predictions can be exploited as a basis for an active reading strategy, and analyze the developmental trajectory of the system in a word recognition task as a function of increasing lexical competence and lexical prediction ability. It is worth noting that the interactions between (predictive) learning of task representations and active sensing strategies during task learning and execution are not confined to the linguistic domain, addressed here, but are characteristic of a wide variety of sensorimotor tasks: hence the interest of our approach in developmental robotics studies in general.

MATERIALS AND METHODS

MODEL ARCHITECTURE AND COMPONENTS

Our gaze planning model consists of a *lexical representation network*, and the *gaze planner* proper. The lexical representation network is implemented as a *Temporal Hebbian Self-Organizing Map* (*THSOM*; Koutnik, 2007), an extension of Kohonen's Self-Organizing Maps (*SOMs*; Kohonen, 2001) that, in addition to developing topological patterns of input data, models their temporal sequences and supports prediction.

Based on the input provided by a *THSOM* trained on written words, the *gaze planner* implements an *active sensing* strategy for reading. The model actively selects where the next fixation should be placed, rather than passively scanning all text input, from left-to-right at an even pace. We model the problem of planning gaze sequences in reading as a Bayesian sequential decision process. The eye/gaze controller plans an optimal active sensing strategy (under uncertainty) by weighting up future (lexical) information gain and costs. In particular, our target function is to maximize the (expected) information gain (i.e., how much new lexical information is gained through each gaze), minimize the amount of uncertainty in lexical representations (i.e., disambiguate between competing words, say, "house" and "horse"), and minimize costs (i.e., time spent, effort required for short and long saccades). We tested the *gaze planner* at different stages of lexical acquisition and analyzed the developmental trajectories of eye-movement patterns as a function of (i) the growing lexical complexity of input text, and (ii) the level of reader's lexical competence modeled by a *THSOM*. Our gaze planning algorithm was eventually compared with two (Bayesian) strategies that use complete information on word statistics.

THE LEXICAL NETWORK

(Topological) Temporal Hebbian Self-Organizing Map (T$^{(2)}$HSOM)

SOMs define a class of unsupervised clustering algorithms that mimic the behavior of medium to small aggregations of neurons in the cortical area of the brain, involved in the specialized processing of classes of sensory data. Processing in such neural aggregations (called *brain maps*) consists in the activation (or *firing*) of one or more neurons, each time a particular stimulus is presented. A crucial feature of brain maps is their topological organization (Penfield and Rasmussen, 1950; Penfield and Roberts, 1959): nearby neurons in the map are fired by similar stimuli. Although some brain maps are taken to be genetically pre-programmed, there is evidence that at least some aspects of such global neural organization emerge as a function of the sensory experience accumulated through learning (Kaas et al. 1983; Jenkins et al. 1984). Functionally, brain maps are thus dynamic memory stores, directly involved in input processing, and exhibiting effects of dedicated long-term topological organization.

A *THSOM* is a *SOM* augmented with a *temporal connection layer* (**Figure 1**). Classical components of a *SOM* are parallel processing nodes (or *receptors*) arranged in a grid or *map*. Each node in the map is synaptically connected with all elements of the *input layer*, where input vectors are encoded. Each connection is treated as a communication channel with no time delay, whose synaptic strength is modeled by a weight value. Each receptor is thus associated with one *space weight vector* defined on the *spatial connection layer*. We distinguish here the *input space*, staked out by the defining dimensions of the input layer, from the *map space*, i.e. the (usually two-dimensional) grid where receptors are spatially located.

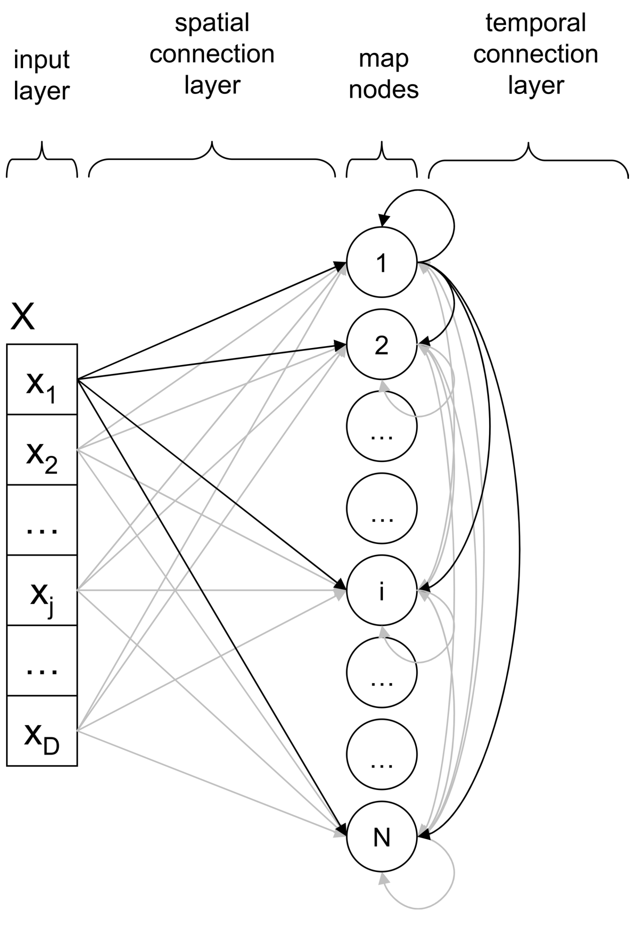


FIGURE 1 | Architecture of a *THSOM*.

In a classical *SOM*, learning is measured in time steps, with each step corresponding to exposure to a single stimulus token. A time step includes three phases: *input encoding*, *input activation* and *input learning*. When a stimulus is encoded on the input layer, all map nodes are activated in parallel as a function of how close their weights are to values of the current input vector. Learning consists in adjusting weights on the spatial connection layer for them to get closer to the corresponding values on the input layer. Weight adjustment does not apply evenly across map nodes and time steps, but depends on *similarity* to the input vector, *learning rate* and *space topology*. At each time step, the most strongly adjusted node is the most highly firing one, or *Best Matching Unit* (*BMU*). All other nodes are adjusted as a function of their distance from *BMU* on the map (or *neighborhood function*). Weights of nodes that lie close to *BMU* are made more similar to input values than weights of nodes lying further away from *BMU*. After adjustment, the time step counter is increased by one tick, the map activation is reset and another input stimulus is encoded. Both learning rate (α) and neighborhood function (ν) vary through time to simulate the behavior of a brain map losing its plasticity.

A *THSOM* models synchronization between two *BMUs* firing at consecutive time steps. This means that a *THSOM* can remember, at time *t*, its state of activation at time *t*–1 and can make an association between the two states. This is possible by augmenting traditional *SOMs* with an additional layer of synaptic connections between each single node and all other nodes on the map (**Figure 1**). For each node, this defines a further association with a *time weight vector*. Connections on this layer (referred to in **Figure 1** as the *temporal connection layer*) are treated as communication channels whose synaptic strength is modeled by a weight value updated with a fixed one-step time delay. Weights on the temporal layer are adjusted with a Hebbian learning strategy (Hebb, 1949) based on activity synchronization of *BMU* at time *t*–1 and *BMU* at time *t*.

During training, the temporal connection between the two *BMUs* is potentiated (**Figure 2A**), while the temporal connections between all other nodes and *BMU* at time *t* are depressed (**Figure 2B**). Logically, this amounts to enforcing the entailment $BMU_t \rightarrow BMU_{t-1}$. Finally, unlike classical *SOMs*, the level of activation of a *THSOM* node at time *t* is determined by the summation of two vector distances: the distance between the current *input vector* and the node's *space weight vector* (as in traditional *SOMs*), and the distance between the node's *time weight vector* and the state of activation of the whole map at time *t*–1.

When trained on time series of input vectors, a *THSOM* develops (i) a topological organization of receptors by their sensitivity to similar input vectors (or spatial similarity) and (ii) a first-order time-bound correlation between *BMUs* activated at two consecutive time steps.

Knowledge of a trained *THSOM* is stored in the synaptic weights of its nodes. We can calibrate the map by assigning a label to each map node. A label is the input symbol which the node is most sensitive to, that is whose input vector matches the node's space vector best. Labeling reveals the topological coherence of the resulting organization (**Figure 4**). Receptors that are fired by similar

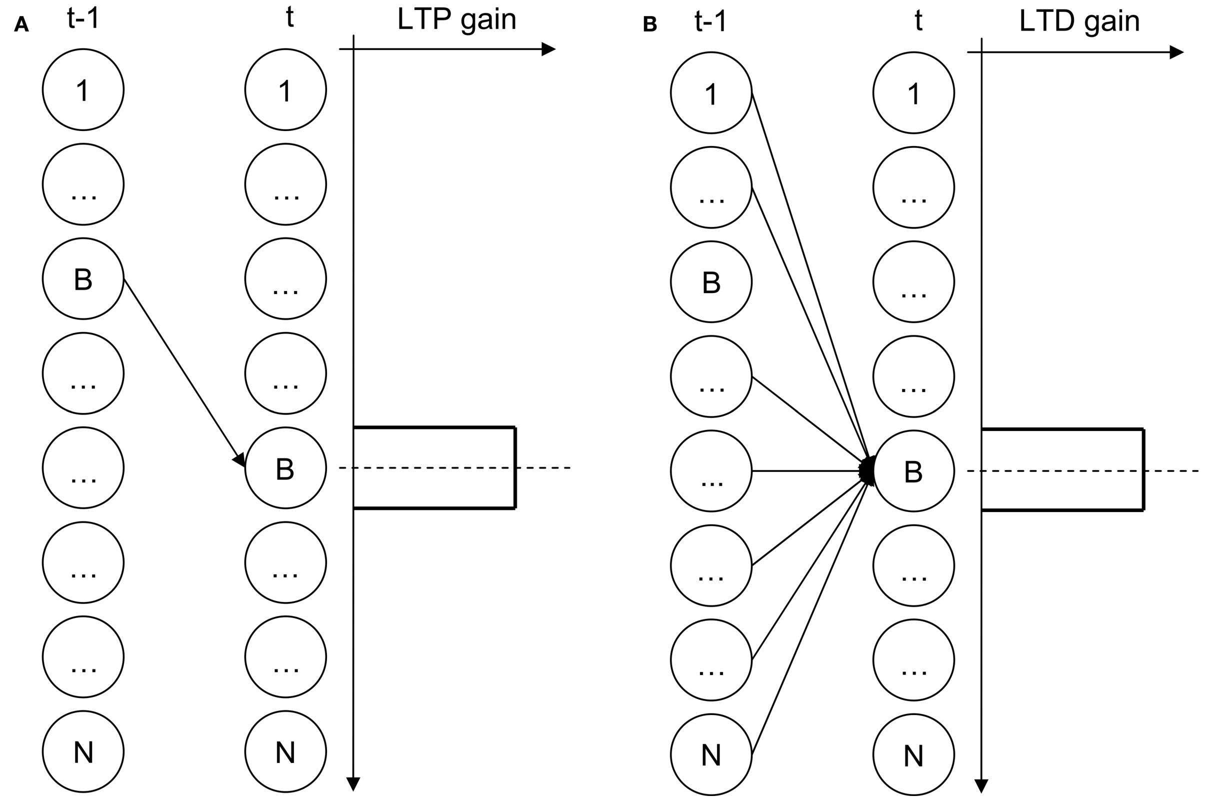


FIGURE 2 | *THSOM* temporal layer plasticity. (A) potentiation; **(B)** depression.

input vectors tend to stick together in the map space. Large areas of receptors are recruited for frequently occurring input vectors. In particular, if the same input vector occurs in different contexts, the map tends to recruit specialized receptors that are sensitive to the specific contexts where the input vector is found. The more varied the distributional behavior of an input vector, the larger the area of dedicated receptors (space allowing).

This dynamics is coherent with a learning strategy that minimizes entropy over inter-node connections. For each map node n_j, we transform connection weights into *transition probabilities* by simply normalizing the weight of a single outgoing (post-synaptic) connection by the summation of the weights over all outgoing connections from n_j. The resulting transition matrix is used to analyze the performance of the model at recall and in particular: (1) the entropy level of each node according to Shannon and Weaver's equation; (2) variation in the entropy of an input sequence as it unfolds its activation over the map; (3) the ability of the map to predict an input sequence, expressed in terms of average (un)certainty in guessing the next transition.

We shall return to a detailed analysis of these aspects later in the paper. Suffice it to say at this juncture that the topological dynamics of a map constrains the degree of freedom to recruit dedicated receptors, as all receptors compete for space on the map. As a result, low-frequency input vectors may lack dedicated receptors after training. By the same token, dedicated receptors may generalize over many instances of the same input vector, gaining in generality but losing in modeling their distributional behavior. The main consequence of a poor modeling of the time-bound distribution of input vectors is an increasing level of entropy, as more *context-free* nodes present more post-synaptic connections. However, topological generalization is essential for a map to learn symbolic sequences whose complexity exceeds the map's memory resources (i.e. the number of available nodes). Moreover, lack of topological organization makes it difficult for a large map to converge on learning simple tasks, as the map has no pressure to treat identical input tokens as instances of the same type.

Pirrelli et al. (in press) originally extend Koutnik's *THSOM* architecture by using the neighborhood function as a principle of organization of connections on the temporal connection layer (**Figures 3A,B**). An additional depressant Hebbian rule penalizes the temporal connections between *BMU* at time *t*–1 and all nodes lying outside the neighborhood of *BMU* at time *t* (**Figure 3C**). This is equivalent to the logical entailment $BMU_{t-1} \rightarrow BMU_t$. Taken together, the temporal connections depicted in **Figure 3** enforce a bi-directional entailment between BMU_{t-1} and BMU_t inducing a bias for biunique first-order Hebbian connections. *THSOMs* that are augmented with this bias are called *Topological Temporal Hebbian Self-Organizing Map (T^2HSOM)*.

In *T^2HSOM*, input vectors can be similar for two independent and potentially conflicting reasons: (i) they have vector representations that are close in the input space; (ii) they distribute similarly, i.e. they tend to be found in similar sequences. Unlike a *THSOM*, which is sensitive to space similarity only, a *T^2HSOM* tries to optimize topological clustering according to both criteria for similarity at the same time. Pirrelli and colleagues show that the dynamic cooperation/competition between the two criteria for similarity is instrumental in capturing paradigmatic effects in the topological organization of the morphological lexicon.

To sum up this long excursus, the overall organization of a *$T^{(2)}HSOM$*[1] after training can be characterized as follows: (1) if space allows, one topologically connected cluster is present for each

[1] Hereafter, we shall use the acronym *$T^{(2)}HSOM$* when we want to say things that apply to both temporal variants of *SOMs* illustrated in the present section.

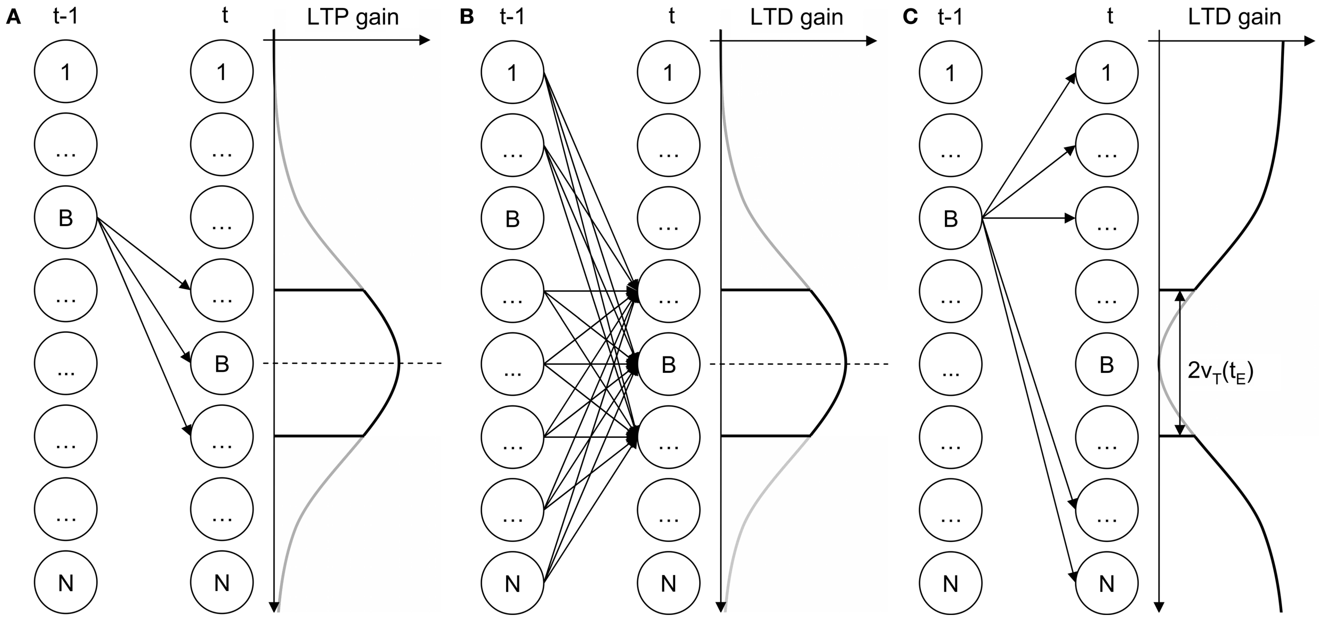


FIGURE 3 | *T^2HSOM* temporal layer plasticity. (A) potentiation; **(B,C)** depression.

symbol; for lack of space, receptors can act as abstract states, fired by a class of similar symbols; (2) receptors that are sensitive to similar symbols are close on the map; (3) the temporal distribution of a symbol may carve out hierarchical sub-clusters within the main cluster for that symbol; (4) the size of a cluster depends on both frequency and the temporal distribution of the corresponding symbol. In the following section we illustrate how *$T^{(2)}HSOM$* can be used to develop lexical representations.

Building a Lexical Network with a $T^{(2)}HSOM$

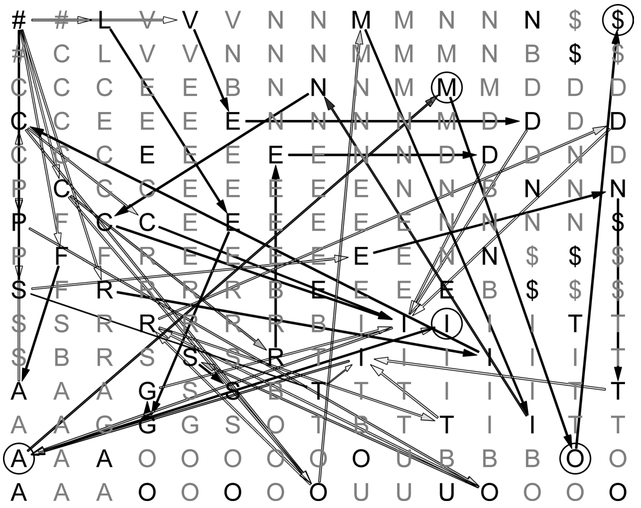
A *$T^{(2)}HSOM$* can learn word forms as time series of alphabetic characters flanked on either side by a start-of-word symbol ('#') and an end-of-word symbol ('$'), as in "#,F,A,C,C,I,O,$".

At each time step, the map is exposed to one single character in its left-to-right order of appearance. Upon exposure to the end-of-word symbol '$', the map resets its Hebbian connections thus losing memory of the correlation between two consecutive word forms. In fact, word forms are repeatedly presented to the map in a random order as a function of their frequency in the training data set. Such a deliberately simplified version of the language learning task helps the map to focus on aspects of word-internal structure, abstracting away from other potentially confounding factors.

By being trained on several lexical sequences of this kind, a *$T^{(2)}HSOM$* (i) develops internal representations of alphabetic characters, (ii) connects them through first-order Hebbian links, (iii) clusters developed representations topologically. The three steps are not taken one after the other but dynamically interact in non trivial ways. From a logical view point, step (i) corresponds to learning individual symbols by recruiting specialized receptors that are increasingly more sensitive to one symbol or class of symbols. Generally speaking, low-frequency symbols are slower in recruiting dedicated receptors than high-frequency symbols are. Step (ii) allows the map to develop selective paths through consecutively activated *BMUs*. This corresponds to learning word forms or recurrent parts of them. Once more, this is a function of the frequency with which symbol sequences are presented to the map. Finally, step (iii) uses either spatial information only (*THSOMs*) or both spatial and temporal information (*T^2HSOMs*) to cluster nodes topologically. Accordingly, nodes that compete for the same symbol stick together on the map. Moreover, they tend to form sub-clusters to reflect distributionally different instances of the same symbol. For example, the symbol A in "#,F,A,C,C,I,O,$" (*faccio*, 'I do') will fire, if space allows, a different node than the same symbol in "#,S,E,M,B,R,A,$" (*sembra*, 'it seems').

An example of a trained lexical map is shown in **Figure 4**. The map is calibrated, with each node being labeled by the alphabetic character that most strongly activates it. Arrows pictorially represent synaptic connections between consecutively activated *BMUs*. In the figure, shades of grey represent different transition probabilities (connection weights), from black (high values) to light grey (low values).

In some cases, it is possible to follow a continuous path of connections going from '#' (start-of-word) to '$' (end-of-word). Only high-frequency word forms, however, are associated with a full path of inter-node connections after training. In the vast majority of cases, only recurring subsequences of activated nodes show strong connection patterns. These may correspond to inflectional endings (such as "I,A,M,O,$" in the figure), verb stems or parts of them.



# L V V V N N M M N N N $ $
# C L V V N N N M M M N B $ $
C C C E E B N N N M M M D D D
C C E E E E N N N N M D D D D
C C C E E E E E N N D D D N D
P C C C E E E E E N N B N N N
P F C C E E E E E E N N N N $
P F F R E E E E E E N N $ $ $
S F R R R R B E E E E B $ $ $
S S R R R R R B I I I I I T T
S B R S S S R T I I I I I I T
A A A G S S B T T T I I I T T
A A G G G S O T B T T I I T T
A A A O O O O O O U B B B O O
A A A O O O O O U U U O O O O

FIGURE 4 | Sample map during learning. Darker edges represent more probable transitions, and lighter edges represent less probable ones.

GAZE PLANNING IN READING: A BAYESIAN IDEAL-OBSERVER PERSPECTIVE

The second component of our model is the *gaze planner*. A gaze planner can be conceptualized as a *Bayesian ideal-observer*, i.e. "a theoretical device that performs a given task in an optimal fashion, given the available information and some specified constraints" (Geisler, 2003, p. 825) spelled out in the framework of Bayesian statistical decision theory. In this framework, one typically assumes that in vision tasks humans behave as (approximate) optimal Bayesian decision makers. Alternatively, one can use the ideal-observer perspective to derive an optimal strategy, without assuming that humans use it, and compare human performance against it with the objective to discover analogies and differences.

In Bayesian analysis, one important aspect of information acquisition is the reduction of uncertainty over the variables that are relevant to the task at hand (e.g., location of objects in space, and/or their orientation, etc.). Reduction of uncertainty is not only valuable *per se*, but also in connection with action execution and behavioral decisions to be taken in the task. This aspect is captured by the notion of *value of information* (Howard, 1966): information has a *value*, which depends on the extent to which it is expected to disambiguate alternative beliefs and (particularly) make behavioral choice effective. That is, new information that could prompt a decision change is more valuable. By estimating the expected value of gazes, a system can select the gaze planning strategy that maximizes the value of acquired information (Sprague and Ballard, 2003; Nelson and Cottrell, 2007 among others).

To design our gaze planning algorithms, we drew inspiration from the Bayesian ideal-observer analysis. Here 'task knowledge' consists in lexical representations, and the task to be performed is recognizing written words in a text by reading a variable number of characters from left-to-right. Note that word recognition is simpler than reading, as only the latter requires a grapheme-to-phoneme mapping function. In word recognition, a Bayesian ideal-observer strategy makes use of lexical predictions to estimate the expected value gain of prospective gazes. This is conducive to gaze plans that aim to maximize such gain under time constraints and in the presence of uncertainty. On the basis on this general idea, we tested three gaze planning algorithms.

Algorithm 1

The first algorithm implements a simplified prediction-based procedure, which consists in skipping all characters that can be predicted reliably (i.e., above a given threshold) by a *$T^{(2)}$HSOM*.

All characters (with the exception of the start-of-word symbol '#') making up a written input word are initially masked by '*'. For example, at the outset, the word "#,F,A,C,C,I,O,$" is shown to the gaze planner as the string '#,*,*,*,*,*,*,*'. The algorithm starts from the first unmasked character '#' and looks into a trained *$T^{(2)}$HSOM* for a set of (probabilistic) predictions over all '#'-ensuing characters. This is done by looking at the most highly activated node (*BMU*) when the input symbol '#' is shown to the map, and by inspecting the set of current *BMU*'s post-synaptic connections (i.e. its outgoing transitions). The gaze planner then decides whether the coming written character(s) should be skipped or not depending on how accurate the *$T^{(2)}$HSOM*'s prediction(s) are. If the highest weight of a *BMU*'s post-synaptic connection (say '#' → 'C') is above a set threshold, then an input character is skipped in reading and the gaze planner takes 'C' as the next input character. If no post-synaptic weight exceeds the threshold, control is returned to reading and the ensuing written character is unmasked. When the system reaches the end-of-word symbol '$', then the sequence of guessed/read symbols is returned and evaluated against the current input word.

Note that the gaze planner is provided with a fovea that fixates only one character at a time (there being no periphery). In other terms, each landing position provides information about one character at a time. Due to the absence of periphery, the system cannot use the strategy that appears to be the most widely used by human readers, i.e., planning the landing positions around the word center (with an additional systematic error, which might derive from Bayesian estimation; see Engbert and Krügel, 2010). For the sake of simplicity, we further assume here that there are no landing errors, and that gazed characters are perfectly recognized. The algorithm, intended to focus on the importance of prediction, is not only (computationally) simpler than minimizing vocabulary entropy (as in Algorithm 3 below), but takes into account at the same time reduction of uncertainty and sequential nature of the reading task, without introducing motor costs for planning saccades of different amplitude (i.e. longer saccades are more costly for the motor system to execute, and more noisy on average).

Algorithms 2 and 3

Like Algorithm 1, Algorithm 2 scans an input word from left-to-right, starting from the first symbol and trying to make predictions about the upcoming characters on the basis of information on their immediate predecessor. Transition probabilities are estimated here through complete statistical information about the distribution of characters in the full training lexicon. If transition probabilities exceed a set threshold, a prediction is made and the corresponding letter in the input word is skipped. If the guessed character is not '$', then a novel belief about another upcoming character is entertained, based on the previously guessed information.

Algorithm 3 makes no full left-to-right scanning of the input text and tries to minimize the number of reading steps required to identify the full word correctly. At each reading step, it places the gaze upon that position in the input string associated with the lowest possible entropy score. Entropy here is defined as a function of the number (and frequency) of outstanding word candidates that remain to be evaluated once the character in the selected position is read off. Suppose, for the sake of concreteness, that the lexicon is made up out of two strings only, say *ABC* and *ABD*. In this case, to establish which of the two words is currently input, reading either the first or the second character would not minimize entropy, as it does not reduce the number of possible candidates. Only the character in third position would reduce uncertainty to zero and thus represents the optimal character to be gazed at. In realistic scenarios, at each reading step new entropic scores are estimated on the basis of a shrinking set of candidate words, until one candidate word only is left.

RESULTS AND DISCUSSION

The three algorithms were tested in two different experiments. For all of them, we used the same set of training data. Training data and testing data were identical in all reported simulations.

EXPERIMENT 1

We tested the Algorithm 1 from Section "Gaze planning in reading: a Bayesian ideal-observer perspective", where gaze planning is based on the capacity of a trained *$T^{(2)}$HSOM* to predict written lexical representations. A *THSOM* and a *T^2HSOM* were independently trained on the same set of Italian written verb forms and results on both trials were compared. Both *SOMs* were bi-dimensional square grids of 25 × 25 nodes.

Training materials

The training data set contained distinct present indicative forms of 10 Italian verbs, for a total of 66 different forms, whose frequency distributions were sampled from the *Calambrone* section of the Italian CHILDES sub-corpus (MacWhinney, 2000), of about 110,000 token words. The average word length was 6.5 characters (see the frequency distribution in **Figure 7A**). Forms were mostly selected from regular, formally transparent morphological paradigms. Nonetheless, some subregular high-frequency forms were introduced in the training set to monitor their representational trajectories during learning.

Written forms were represented as sequences of alphabetic characters between '#' and '$'. To train the lexical network, alphabetic characters were encoded through a distributed, grapheme-based representation consisting of a 20-element vector, with each element encoding a specific feature of the graphical rendering of orthographic symbols cast into the grid of **Figure 5**.

Training protocol

Lexical network. Both maps were trained over 100 epochs. For each epoch, the training data set was treated as an urn containing verb forms. In the urn, the number of (identical) verb forms of the same type reflected the frequency of the verb type in our reference corpus. One verb form at a time was drawn from the urn, and its spelling retrieved. Each character in the spelling was converted into a distributed grapheme-based input vector and was shown to

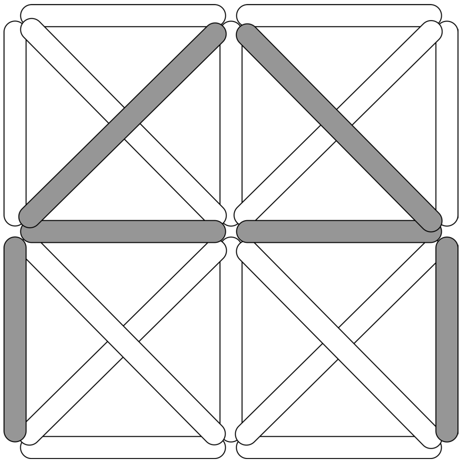


FIGURE 5 | Representation of a capital "A" in the graphical grid.

a $T^{(2)}HSOM$ in its order of appearance. When the '$' symbol of the current input word form was shown, the internal clock of the map was reset and the word discarded. Another word was then drawn from the lexical urn and the whole training process was repeated over again until the urn was emptied.

Gaze planner. The same set of verb forms used for training the *SOMs* was then used for testing the gaze planner. Word forms are presented as dynamically unmasked sequences of characters (see "Gaze planning in reading: a Bayesian ideal-observer perspective").

Figure 6 shows the results of the two networks in the word recognition task, broken down by learning epochs (which is also an indirect evaluation of the topological organization of the trained SOMs, see Pirrelli et al., in press). The values reported in **Figure 6** are averaged over repeated (10) experiments for each network. In particular, we measured the algorithm's *accuracy rate* (the percentage of words that were identified correctly) and *prediction rate* (the percentage of characters that were predicted, not necessarily correctly, and thus skipped in reading) over 100 learning epochs, by plotting them against increasing levels of confidence (*x* axis). Low levels of confidence indicate that the gaze planner has a tendency to skip characters even though they are not strongly predicted by the network connections. Higher confidence thresholds correspond to a more conservative attitude towards reading, whereby only highly predictable ensuing characters are skipped. Clearly, lower thresholds yield less accurate results (the ascending solid line in the panels) and higher percentages of guessed symbols (descending dashed line in the panels).

Careful analysis of the developmental trajectories of both models throws some notable phenomena in relief. Both models increase their overall *accuracy rate* as learning progresses. At the beginning, there are no specialized receptors for each character in the alphabet. Hence, networks are not able to recognize every single character. For instance, it might happen that a 'C' is presented to a network, but the corresponding *BMU* is labeled as a 'G'. This explains the poor performance in the first 20 epochs, even when almost all characters are read. In addition, over the first 30 epochs, transition probabilities are too low to be used effectively, and nearly every character has to be read.

Observe the different developmental stages the two networks go through (**Figure 6**). Both maps converge on full scale accuracy rates (i.e. 100%) and comparable prediction rates, with Koutnik's *THSOM* averaging 44.7% per word prediction at a 0.93 level of confidence, and the T^2HSOM scoring 40.6% per word prediction at 0.89, after 100 learning epochs. Note, however, that Koutnik's *THSOM* converges remarkably more quickly than T^2HSOM. *THSOM* exhibits a tendency to retain longer stretches of input words at a faster pace than T^2HSOM, as shown by the overall number of saccades of varying length in the two models (**Figures 7B,C** respectively). The reason for this behavior lies in the capacity of *THSOM*s to "pack" more nodes that are competing for the same symbol in a comparatively smaller area of the map. Recall that, in T^2HSOMs, competing receptors strongly inhibit each other and can coexist only at a distance. The same constraint does not hold for *THSOM*s, where context-sensitive receptors of the same symbol do not fight for short-range survival. A wider range of context-sensitive receptors minimizes the number of post-synaptic connections, thereby minimizing per node entropy and facilitating memorization of longer symbol chains.

On the other hand, strong competition between symbol tokens in complementary distribution is helpful in learning morphological structure. Tested on the task of identifying morpheme boundaries within inflected forms, the two maps show a reversed accuracy pattern: T^2HSOMs are consistently better at finding morpheme transitions than *THSOMs* are. A 15 × 15 nodes T^2HSOM is able to identify morpheme boundaries with 71% accuracy, while a *THSOM* of the same size has an accuracy of 64% on the same task and test data. Once more, when map size increases, accuracy scores of the two maps level out. **Figure 8** shows transition probabilities at morpheme boundaries in the present indicative forms of the verb CREDERE ('believe'), plotted against learning epochs. In a *THSOM* (**Figure 8A**) lack of inhibition between complementarily distributed endings blots out the difference in frequency distribution among them. On the other hand, a T^2HSOM proves to be sensitive to the uneven distribution of forms in the paradigm (**Figure 8B**). This is shown to have important consequences in learning and access of lexical representations in human speakers (Baayen, 2007) and is demonstrably related to levels of difficulty in reading morphologically complex words by dyslexic and non dyslexic subjects (Burani et al., 2008).

EXPERIMENT 2

In this experiment we tested the results of the two Bayesian models of gaze planning informally described in Section "Gaze planning in reading: a Bayesian ideal-observer perspective". Like our $T^{(2)}$ *HSOM*-based models, Algorithm 2 skips upcoming characters that are predicted reliably, but operates on complete word statistics and uses Bayes rules to update transition probabilities. Results are illustrated in **Figure 9**, plotted against levels of confidence. Unsurprisingly, the performance of the system is better; in particular, with a threshold of 0.85, the system reaches 100% performance and predicts 54% of the characters. In addition, even with lower thresholds the correctness rate is high; this is due to the high prediction accuracy of the system. Therefore, the main lesson learned from this comparison is that the lexical representation network is still limited in its prediction ability, due to its local learning steps and its incrementality. We argue that this is the price we have to pay for modeling human behavior in a more

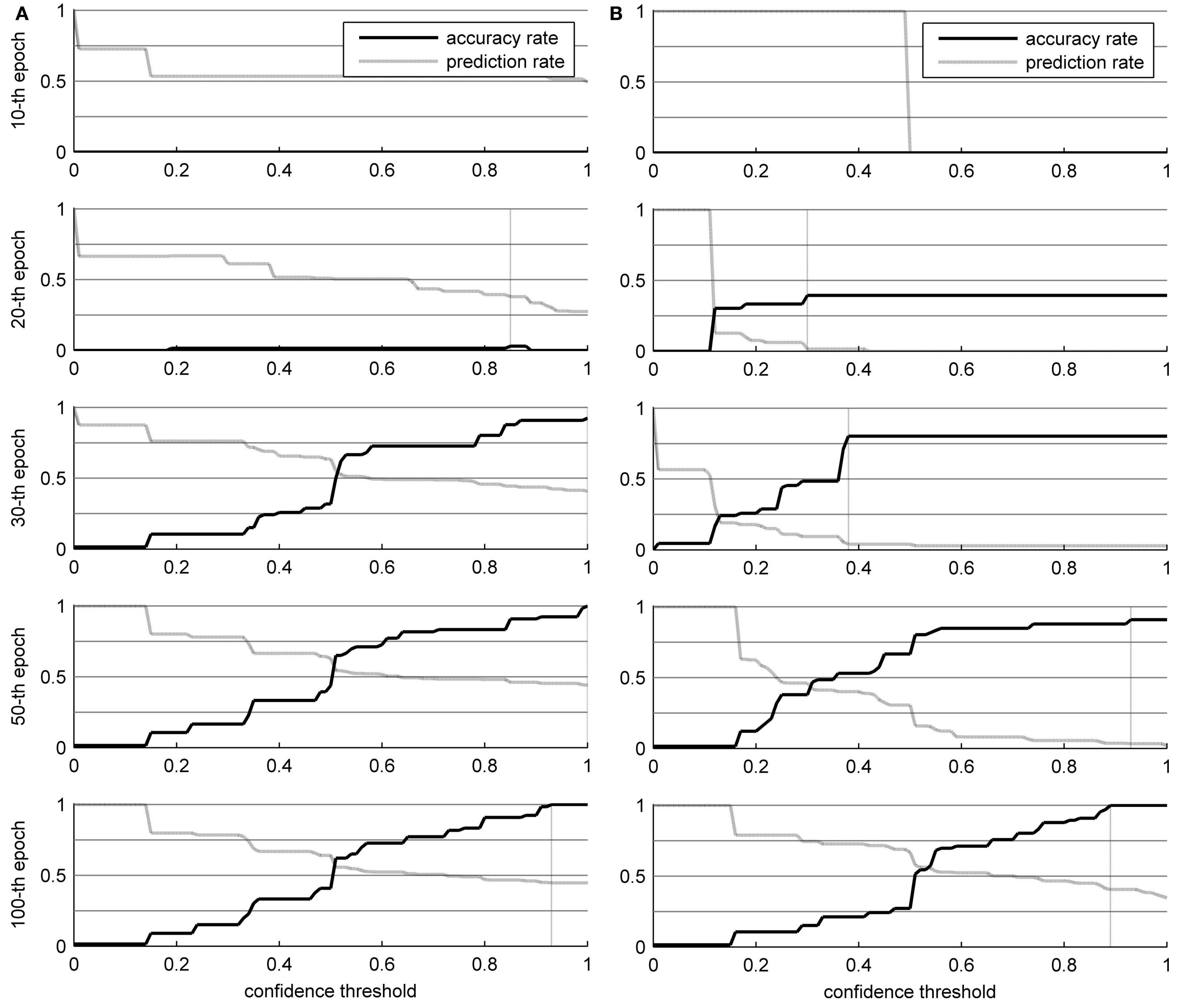


FIGURE 6 | Results of the word identification experiment for 66 words in the corpus. The vertical dotted line indicates the optimal confidence threshold. **(A)** *THSOM* model; **(B)** *T²HSOM* model.

realistic way. In fact, it is dubious that children can supposedly be engaged in a search for global optimization strategies in learning word reading.

Algorithm 3 (also adopted in the design of Mr. Chips, Legge et al., 1997, 2002) implements the Bayesian ideal-observer procedure described above[2]. It calculates the expected information gain (i.e., difference between future and present entropy) of each possible character, and gazes the one with the highest information gain, independently of its position in the word. This is done again until the word is identified with 100% probability. This algorithm is optimal in Bayesian terms, with 2.42 gazes on average per word (from 2 to 4 gazes), corresponding to 30.1% read characters only, with a variance of 0.09. Recognition is 100% accurate. As expected, its performance is significantly better than the other algorithms presented here, at the cost of stronger assumptions (complete knowledge and indifference to the order of characters in words). The comparison sheds light on the difficulty of the task we designed. Indeed, our results show that the number of characters that could be skipped while preserving optimal performance is limited (consider however that in human reading and comprehension, predictions can be done at multiple levels, e.g., lexical, syntactic, semantic; see Pickering and Garrod, 2007).

[2] The algorithms we present here were selected as benchmarks for their simplicity, and many others could be adopted that implement similar ideal-observer strategy, with the addition of extra constraints. First, note that the strategy implemented here is myopic, in that the information gain is calculated only for the next saccade, and not (cumulatively) for whole sequences of saccades. Although the latter strategy is optimal in principle, it is however extremely demanding in computational terms. In addition, one could take into consideration extra factors, such as (motor) costs for the saccades, so that longer saccades are dispreferred, or costs for errors in the word recognition, so that system must find the minimum cumulative loss instead of simply minimizing the number of saccades. Note also that alternative Bayesian strategies have been proposed such as the "optimal ambiguity resolution" procedure of (Chater et al., 1998), which introduces a bias to choose interpretations which make specific predictions, and which might be falsified quickly.

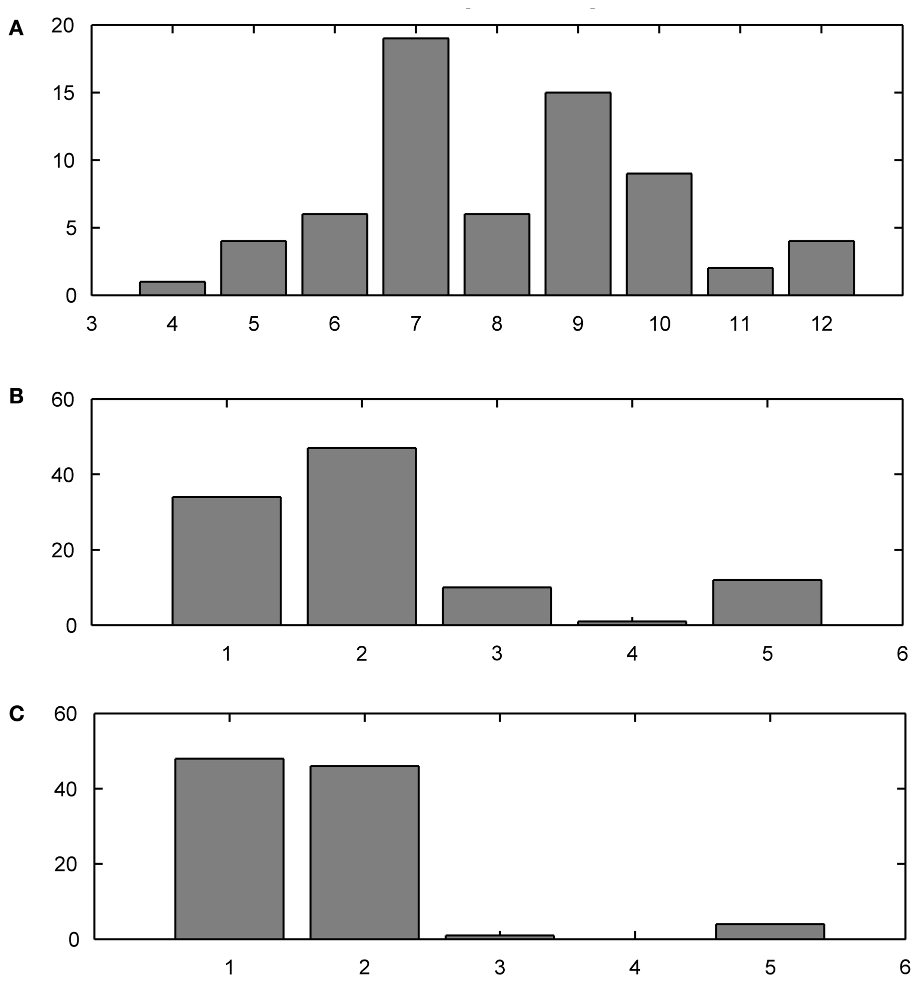


FIGURE 7 | Training corpus word frequency histogram (A) and saccade frequency histogram test results; (B) THSOM model; (C) *T²HSOM* model.

Our experimental results, on the other hand, cannot be compared directly to human reading data. Not only human reading skills are considerably more sophisticated compared to our algorithm, but there are differences in the task requirements too. The human fovea can see about four or five characters around the fixation point with 100% acuity, and up to 10 times more with increasingly less acuity. On the contrary, we used a 'fovea' that only extracts 1 character per time. For this reason, it is reasonable that human saccades are on average 2–3 times longer (7–9 characters) than those obtained in our experiments (2–3 characters on average). In addition, the task we used was simplified compared to reading. For instance, humans 'backtrack' while reading (probably for correcting implausible interpretations). Our system was not allowed to backtrack, instead; wrong interpretations counted as errors.

DISCUSSION AND CONCLUDING REMARKS

We have implemented a computational model of eye movements in language reading that integrates two components: a *lexical representation network* and a *gaze planner*. The *lexical representation network* is a temporal self-organizing map, combining overlaying memory patterns with chains of first-order weighted Hebbian connections. From a cognitive perspective, this novel network architecture has two interesting implications.

A trained temporal map behaves like a first-order stochastic Markov chain, with inter-node connections building expectations about possible word forms on the basis of a global topological organization of already known forms. The model prompts a reappraisal of the traditional melee between one-route and dual-route models of morphology processing and learning, as it *contextually* represents lexical memory patterns *and* rule-like predictions. Furthermore, the architecture has something to say about the representation of serial order information in short-term and long-term memory structures.

Botvinick and Plaut (2006) contrast two general computational approaches to modeling short-term memory for serial order: *weight-based models* and *activation-based models*. In weight-based approaches (see, e.g., Grossberg, 1986; Houghton, 1990; Burgess and Hitch, 1992, 1999; Houghton and Hartley, 1996; Hartley and Houghton, 1996; Henson, 1996, 1998; Brown et al., 2000), serial encoding and recall depend on *transient* associative links between item and context representations, with associative links being

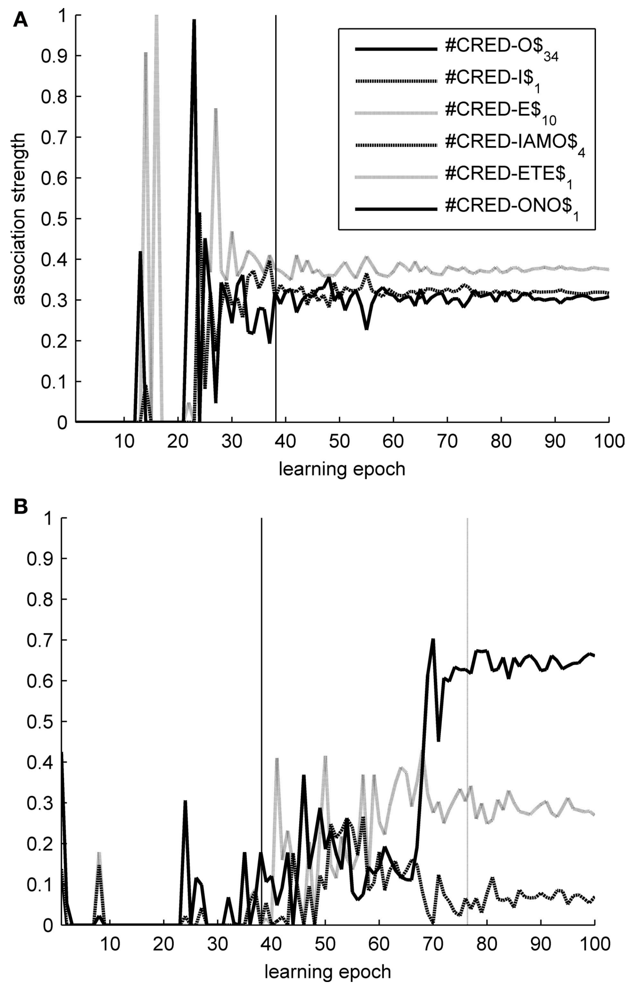


FIGURE 8 | Transition probabilities over morpheme boundaries in CREDERE ('believe'). (A) *THSOM* model; **(B)** T^2*HSOM* model.

established by changing the connection weights between processing units, upon presentation of a sequence to be recalled. Weight-based models may differ in the nature of the context representation they use, but they all agree that serial recall does not involve incremental learning. Thus, although they prove to be able to replicate a wide range of detailed behavioral findings about human subjects, they have so far failed to simulate effects of background long-term knowledge (e.g. Baddeley's so-called *bigram frequency effect*, Baddeley, 1964).

Unlike weight-based approaches, activation-based memory mechanisms (such as *recurrent neural networks* and the $T^{(2)}$ *HSOMs* presented here) adjust weights gradually, over many learning trials, but performance of network recall is evaluated by holding weights constant and using sustained activation patterns. Botvinick and Plaut (2006) show that recurrent neural networks can account for long-term memory effects, while, at the same time, replicating several behavioral facts of human recall. However, this is achieved by accounting for short-term effects of serial recall on the basis of long-term memory effects. This is somewhat questionable. First, it makes short-term memory entirely depend on long-term memory mechanisms. In a developmental perspective, the causal relationship is in fact reversed (although reciprocal effects are also observed). For example, problems with short-term memory processing are known to cause delays in child vocabulary acquisition (Shallice and Vallar, 1990; Papagno et al., 1991; Service, 1992 to mention a few). As observed by Baddeley (2007), children with higher short-term memory capacity are able to hold on to new words for longer, increasing the likelihood of long-term lexical learning. Finally, Botvinick and Plaut's (2006) approach makes the paradoxical suggestion that human performance on immediate serial recall develops through direct practice on the task, rather than using the task to probe short-term memory capacities.

In $T^{(2)}$*HSOMs*, the learning regime is unsupervised and memory effects are not based upon recall performance. Moreover, short-term memory and long-term memory work according to two different dynamics. Serial encoding in a temporal map requires sustained activation of *BMUs* and their one-way associative connections. Sustained activation chains of this kind are triggered upon presentation of an input sequence (see Building a Lexical Network with a $T^{(2)}$*HSOM* above). We further argue here that, by smoothing the decay function over consecutive time steps, activation chains can also simulate effects of immediate serial recall. Serial learning, on the other hand, adjusts connection weights gradually, for them to keep track of the most frequently activated connections. Hence long-term entrenchment of one-way Hebbian connections is the result of repeated exposure to frequent time series of symbols. When long-term entrenchment sets in, it can affect immediate recall through anticipatory activation of the most frequently activated connection chains. In fact, this is the same mechanism we used in this paper to predict upcoming words. Temporal maps thus point to a profound continuity between word prediction, repetition and learning. Nonetheless they assume that short-term memory and long-term memory are based on different temporal dynamics, in line with neurobiological approaches (Pulvermüller, 2003) according to which long-term memory refers to consolidation of associative networks and short-term memory is (transient) activation of the same networks.

The *gaze planner* is motivated by a Bayesian ideal-observer perspective. It bears resemblances to *Mr. Chips* (Legge et al., 1997, 2002), the first computational model based on an ideal-observer analysis, to the *Bayesian reader* (Norris, 2006), and to other Bayesian computational models of reading (Sprague and Ballard, 2003; Nelson and Cottrell, 2007). In all these systems, lexical predictions drive attention in such a way that uncertainty about environmental variables that are task relevant is reduced. This is done either by minimizing entropy, or by minimizing a combination of entropy and movement (i.e. saccade amplitude) costs. Compared to these models, our system adopts the simpler principle of gazing at the next character that cannot be reliably predicted, and works on top of learned (self-organized) lexical representations and lexical predictions.

Since a $T^{(2)}$*HSOM* modifies its lexical representations and predictions during learning, our computational model allows us to analyze how gaze planning varies during reading, depending on the system's lexical knowledge. In particular, it offers a framework to study the interrelated developmental trajectories of (lexical) knowledge acquisition and gaze planning during

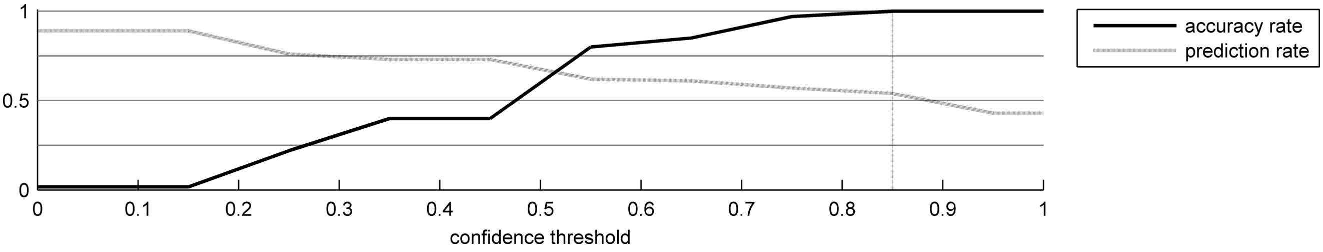


FIGURE 9 | Results of the first algorithm having complete knowledge of the word statistics.

reading. To the best of our knowledge, there is no extensive empirical study of this aspect in reading, whereas relevant data exist related to other tasks. For instance, a recent study has investigated how visual strategies change when the subject learns a novel visuomotor task (Sailer et al., 2005). The authors found that better performance correlated with changes in gaze planning. At a first stage, hit rate was low and gaze was reactive, whereas in the second and third stages hit rate was higher and gaze become increasingly more predictive. In our experiments, we observed the same pattern of behavior, with the development of increasingly reliable predictions that were conducive to planning anticipatory strategies.

Surely, this developmental pattern is not confined to the domain of reading or vision. Several studies in other fields, such as motor development (von Hofsten, 2004), have revealed that the development of predictive abilities determines an increasing reliance on prospective behavior and is a necessary precondition for the rise of more and more complex cognitive abilities (for a discussion of this topic, see Pezzulo and Castelfranchi, 2007; Butz, 2008; Pezzulo, 2008).

RELEVANCE OF OUR STUDY FOR (DEVELOPMENTAL) ROBOTICS

Our approach to reading as an active sensing process is based on representations and predictions that are increasingly refined through learning. This makes our model particular fit for developmental robotic implementations. Through our methodology, lexical representations can be acquired and further exploited to engage in both linguistic and extra-linguistic tasks in human-robot, or robot-robot scenarios. In addition, the model can be extended to study the acquisition of referential capabilities in robots. This could be done, for instance, by coupling many $T^{(2)}$ *HSOMs*, one for each domain (visuomotor, linguistic, etc.), for acquiring a combined lexical representation of a word such as *ball*, a visual representation of balls, and a set of actions to be performed on balls, so that the robot can use language to refer to objects and actions in the world, along the lines of recent computational studies that combine linguistic and sensorimotor processes (Cangelosi and Harnad, 2001; Roy, 2005; Sugita and Tani, 2005; Wermter et al., 2005).

It is worth noting that our active sensing methodology is applicable outside the linguistic domain. In general, the problem of how, during development, task representations are acquired and determine increasingly sophisticated active sensing strategies, is characteristic of any form of sensorimotor learning. In addition, as pointed out above, there is substantial evidence that anticipatory processes drive visual strategies in many visuomotor tasks (Hayhoe and Ballard, 2005). Therefore, by using $T^{(2)}$ *HSOMs* to encode sensorimotor rather than linguistic predictions, our methodology could be adopted for the visual guidance of actions, with attention going where (task) relevant information is expected to be.

FUTURE WORK

We rapidly mention here two aspects of our model that are particularly promising for future work. The predictive nature of our model makes room for *novelty detection* (Bishop, 1994), i.e. identification of novel data from on the basis of marginal density. In particular, the model could classify words or sentences as novel. In turn, novelty detection is a fundamental precondition for active learning based on adaptive curiosity, which consists in focusing learning on novel but still predictable parts of the data, for which the system can actually improve its predictions (Schmidhuber, 1991). In our current model, the two sub-tasks of lexical acquisition and word recognition are carried out independently. However, they could be combined so that the gaze planning mechanism is active during learning and the novelty detection mechanism can affect learning lexical representations in the $T^{(2)}$*HSOM*. In the first learning stages, when lexical representations in the $T^{(2)}$*HSOMs* are not fully developed and reliable, most input text contributes novel information, with few characters being skipped and lexical representations being frequently revised. When lexical representations in the $T^{(2)}$*HSOMs* get more deeply entrenched and dependable, novelties become more rare, more characters are skipped, and lexical representations get revised only occasionally.

Another possible extension of our model is using a cascaded asynchronous $T^{(2)}$*HSOM* architecture, with higher-level maps sampling the activation state of lower-level maps at increasingly larger time intervals. In this architecture, short-range (i.e., phonological and morphological) serial correlations are captured through low-level maps, and long-range serial correlations (i.e., word sequences) are represented on top-level maps. Although a single $T^{(2)}$*HSOM* could in principle capture correlations at all levels (size allowing), with the benefit of the hindsight (Calderone et al., 2007) we conjecture that cascaded architectures of this type can encode correlations more efficiently, avoiding information overload/interference and effectively simulating the interaction of short-term and long-term memory effects in human serial recall.

APPENDIX

THE *T*$^{(2)}$*HSOM* MODEL

Short-term dynamics: activation and filtering

In the topological processing phase, activation of each node is a function of the Euclidean distance in the input space between its weight vector and the input vector. The resulting topological activation of the *i*-th node at time *t* is:

$$y_{S,i}(t)=\sqrt{D}-\sqrt{\sum_{j=1}^{D}[x_j(t)-w_{i,j}(t)]^2}$$

where *D* is the number of components of the input vector $X(t) = [x_1(t),\ldots,x_D(t)]$, and $w_{i,j}(t)$ is the synaptic weight of the topological connection between the *i*-th node and the *j*-th input component.

In the temporal processing phase, activation of each neuron is a function of the correlation between its temporal synaptic connections and the overall activation state at the previous time step. The resulting temporal activation of the *i*-th node at time *t* is:

$$y_{T,i}(t)=\sum_{h=1}^{N}[y_h(t-1)\cdot m_{i,h}(t)]$$

where *N* is the number of node of the map, $Y(t–1) = [y_1(t–1),\ldots, y_N(t–1)]$ is the output of the $T^{(2)}HSOM$ at the previous time step, and $m_{i,h}(t)$ is the synaptic weight of the temporal connection from the *h*-th pre-synaptic neuron to the *i*-th post-synaptic neuron.

The resulting two activation values are summed up, so that the resulting activation value of the *i*-th neuron at time *t* is:

$$y'_i(t)=y_{S,i}(t)+y_{T,i}(t)$$

The filtering module identifies BMU at time *t* by looking for the maximum activation level:

$$y'_{bmu}(t)=\max_i\{y'_i(t)\}$$

The output is subsequently normalized to ensure the network stability over time:

$$Y(t)=\frac{Y'(t)}{y'_{bmu}(t)}$$

Long-term dynamics: learning

In $T^{(2)}HSOM$ learning consists in topological and temporal co-organization.

Topological learning. In classical SOMs, this effect is taken into account by a neighborhood function centered around BMU. Nodes that lie close to BMU on the map will be strengthened as a function of BMU's neighborhood. The distance between BMU and the *i*-th node on the map is calculated through the following Euclidean metrics:

$$d_i(t)=\sqrt{\sum_{c=1}^{n}[i_c-bmu_c(t)]^2}$$

where *n* is 2 when the map is two-dimensional. The topological neighborhood function of the *i*-th neuron is defined as a Gaussian function with a cut-off threshold:

$$c_{S,i}(t)=\begin{cases} e^{-\frac{\mathrm{d}_i^2(t)}{2\sigma_S^2(t_E)}} & \text{if } \mathrm{d}_i(t)\le \nu_S(t_E) \\ 0 & \text{if } \mathrm{d}_i(t)> \nu_S(t_E)\end{cases}$$

where $\sigma_S(\mathrm{t}_E)$ is the topological neighborhood shape coefficient at epoch time t_E, and $\nu_S(\mathrm{t}_E)$ is the topological neighborhood cut-off coefficient at epoch time t_E.

The synaptic weight of the *j*-th topological connection of the *i*-th node at time *t* + 1 and epoch t_E, is finally modified as follows:

$$\Delta w_{i,j}(t)=\alpha_S(t_E)\cdot c_{S,i}(t)\cdot[x_j(t)-w_{i,j}(t)]$$

$$w_{i,j}(t+1)=w_{i,j}(t)+\Delta w_{i,j}(t)$$

where $\alpha_S(\mathrm{t}_E)$ is the topological learning rate at t_E.

Temporal learning

On the basis of BMU at time *t*–1 and BMU at time *t*, three learning steps are taken:

- temporal connections from BMU at time *t*–1 (the *j*-th neuron) to the neighborhood of BMU at time *t* (the *i*-th neurons) are strengthened:

$$m_{i,j}(t+1)=m_{i,j}(t)+\alpha_T(t_E)\cdot c_{T,i}(t)\cdot[1-m_{i,j}(t)+\beta_T(t_E)]$$

$$c_{T,i}(t)=\begin{cases} e^{-\frac{\mathrm{d}_i^2(t)}{2\sigma_T^2(t_E)}} & \text{in} T^2HSOM \\ 1 & \text{in} THSOM\end{cases}$$

- temporal connections from all neurons except BMU at time *t*–1 (the *j*-th neurons) to the neighborhood of BMU at time *t* (the *i*-th neurons) are depressed as well:

$$m_{i,j}(t+1)=m_{i,j}(t)-\alpha_T(t_E)\cdot[1-c_{T,i}(t)]\cdot[m_{i,j}(t)+\beta_T(t_E)]$$

$$c_{T,i}(t)=\begin{cases} e^{-\frac{\mathrm{d}_i^2(t)}{2\sigma_T^2(t_E)}} & \text{in} T^2HSOM \\ 0 & \text{in} THSOM\end{cases}$$

- temporal connections from BMU at time *t*–1 (the *j*-th neuron) to outside the neighborhood of BMU at time *t* (the *i*-th neurons) are depressed as well:

$$m_{i,j}(t+1)=m_{i,j}(t)-\alpha_T(t_E)\cdot c_{T,i}(t)\cdot[m_{i,j}(t)+\beta_T(t_E)]$$

$$c_{T,i}(t)=\begin{cases} e^{-\frac{\mathrm{d}_i^2(t)}{2\sigma_T^2(t_E)}} & \text{in} T^2HSOM \\ 0 & \text{in} THSOM\end{cases}$$

Learning decay. As an epoch ends, an exponential decay process applies to each learning parameter so that the generic parameter *p* at t_E is calculated according to the following equation:

$$p(t_E) = p(0)\cdot e^{-\frac{t_E}{\tau_p}}$$

A complete list of the learning parameters is shown below:

- α_S: learning rate of the topological learning process
- σ_S: shape parameter of the neighborhood Gaussian function for the topological learning process
- ν_S: cut-off distance of the neighborhood Gaussian function for the topological learning process
- α_T: learning rate of the temporal learning process
- σ_T: shape parameter of the neighborhood Gaussian function for the temporal learning process
- ν_T: cut-off distance of the neighborhood Gaussian function for the temporal learning process
- β_T: offset of the Hebbian rule within the temporal learning process

Post processing. At a given epoch t_E, the transition matrix is extracted from the temporal connection weights $m_{i,j}(t_E)$, so that $P_{i,j}(t_E)$ is the probability to have a transition from the *i*-th node to the *j*-th node of the network (i.e., the *j*-th node will be the *BMU* at time $t + 1$, given the *i*-th node is the *BMU* at time *t*):

$$P_{i,j} = m_{j,i}\cdot\frac{1}{\sum_{h=1}^{N} m_{h,i}}$$

At the same time the labeling procedure is applied. A label L_i (i.e., an input symbol) is assigned to each node, so that the grapheme-base coding of the *c*-th symbol matches the *i*-th node's space vector best:

$$L_i = \arg\min_c \sqrt{\sum_{j=1}^{D}[x_{c,j}(t)-w_{i,j}(t)]^2} \qquad (i = 1\ldots N)$$

Parameter configuration

The experiments shown in the present work were performed using the following parameter configuration:

- 25 × 25 map nodes
- 20 elements in the input vector (grapheme-based orthographic character coding)
- 100 learning epochs
- learning rates starting from maximum value (i.e. 1.0), exponentially decaying over epochs with a time-constant equal to 25 epochs
- shape parameters starting from a value so that the Gaussian function has a gain equal to 30% at the maximum cut-off distance, with no decay over epochs
- spatial cut-off distance starting from the maximum distance between two nodes in the map, exponentially decaying over epochs with a time-constant equal to 12.5 epochs
- temporal cut-off distance starting from the maximum distance between two nodes in the map, exponentially decaying over epochs with a time-constant equal to 25 epochs
- offset of the Hebbian rule within the temporal learning process starting from 0.01), exponentially decaying over epochs with a time-constant equal to 25 epochs

The *THSOM* version of the model was tested by using $\nu_T = 0$ and $\sigma_T = \infty$.

ALGORITHM 1

The performance of the $T^{(2)}$*HSOM* model is evaluated in terms of accuracy and prediction rate during the execution of the reading task of single words. During this stage the learning algorithm of the model is turned off. The algorithm takes into account all the words contained in the dictionary, and all the symbols contained in each word. With the aim to identify the optimal confidence threshold θ, the corresponding domain ($0 \le \theta \le 1$) is sampled in 100 steps and the performance rates are evaluated at each step.

For each word in dictionary, assuming $s_{i,j}$ represents the *j*-th symbol of the *i*-th word, the algorithm starts from the left-most symbol (i.e. $j \leftarrow 1$) and performs the following steps:

(1) the *j*-th symbol of the *i*-th input word is collected:

$$c \leftarrow \mathrm{s}_{i,j}$$

(2) the symbol *c* is queued in the output word:

$$s'_{i,j} \leftarrow \mathrm{c}$$

(3) a look-up table provides the D-element vector *V* representing the grapheme-based coding belonging to the symbol *c*:

$$V \leftarrow (x_{c,1}, x_{c,2}, \ldots, x_{c,D})$$

(4) the input vector *V* is propagated into the model and, as a result, a new BMU gets activated:

$$k \leftarrow BMU$$

(5) the algorithm looks for the highest transition probability among all the outgoing (post-synaptic) connections from the *k*-th node of the network:

$$q \leftarrow \arg\max_h (P_{k,h}) \qquad (h = 1\ldots N)$$

(6) if $P_{k,q}$ is above the confidence threshold θ, then the next symbol can be directly obtained (i.e., predicted) as the label of the *q*-th node of the network:

$$c \leftarrow \mathrm{L}_q$$

(7) if this the case, the algorithm returns to step (2). Otherwise, the next symbol must be collected (i.e., read) from the input word, returning to step (1). In both cases, the algorithm continues with the next symbol ($j \leftarrow \mathrm{j} + 1$) of the current word. If the end-of-word is reached, the next word is processed ($j \leftarrow 1$; $i \leftarrow i + 1$) until the end-of-dictionary is reached.

During the previous steps, the algorithm evaluates the following scores:

- for each word, the ratio between the number of predicted symbols and the number of total symbols of the word (the start-of-word symbol is excluded)
- the *prediction rate*, which is obtained averaging the above mentioned ratio over all the words
- for each word, the Boolean comparison between the input word s_i and the output word s'_i
- the *accuracy rate*, which is obtained as the ratio between the number of words predicted correctly (i.e., there is no difference between the input and output word) and the total number of words in the dictionary

ALGORITHM 2

The first algorithm described in Section "Experiment 2" operates with complete knowledge (of the order/probability of the characters in the words) and skips predictable characters. Given the current belief state [i.e. a vector $b_t(w_i)$ that describes the probability that the already gazed characters belong to one of the words in the dictionary (w_i)] and the current position a_t, the algorithm selects the character o_m that has the maximum probability P_m to be the next character (at position a_{t+1}) in the word being read.

$$P_m = \max_{o \in O} \sum_{w_i : w_i(a_t+1)=o} b_t(w_i)$$

$$o_m = \arg\max_{o \in O} \sum_{w_i : w_i(a_t+1)=o} b_t(w_i)$$

If the maximum probability is more than a threshold θ, the algorithm assumes that o_m has been read (or can be skipped), otherwise it reads the character at position $a_t + 1$. Then, it updates the belief state $b_{t+1}(w_i)$ and sets the new initial position ($a_{t+1} \leftarrow \mathrm{a}_t + 1$). This procedure continues until the end of the word.

ALGORITHM 3

The second algorithm described in Section "Experiment 2" uses the probability distribution of the words in the dictionary, given the characters already read and the priors (of which it has complete knowledge). The aim of the algorithm is selecting the action (i.e., gaze position) that results in an observation (i.e. a read character), which, in turn, minimizes (on average) the expected entropy, or the entropy of the resulting probability distribution of the words in the entire dictionary, given the current belief state (i.e. word probability)[3]. Note that this approach is myopic, since it minimizes entropy of the next step only (not of the entire sequence of gazes). In general, there is no guarantee that a sequence of myopic actions achieves the same decrease of entropy as an optimal non-myopic sequence.

The initial probability of word w_i is $b_0(w_i)$, and corresponds to the frequency of the word in the corpus. The vector b_0 is the belief state of the agent. The following formulas describes how beliefs (b_{t+1}) are updated based on (i) the previous belief state (b_t), (ii) the new observation (o_{t+1}), and (iii) the executed action (a_t).

$$b_{t+1}(w_i) = P[w_i \,|\, b_t(w_i), a_t, o_{t+1}] = \begin{cases} \dfrac{b_t(w_i)}{\sum\limits_{w_j : w_j(a_t)=o_{t+1}} b_t(w_j)} & \mathit{if}\ w_i(a_t) = o_{t+1} \\ 0 & \mathit{if}\ w_i(a_t) \neq o_{t+1} \end{cases}$$

When the algorithm gets the character o_{t+1} at position a_t, the probability distribution of words is updated as follows: (i) it becomes zero for all words that have a different character in that position, (ii) for all the other words, the previous probability is divided by the sum of the previous probability of all the words that have the character in the right position. Expected entropy (*EH*), given the current belief and the position gazed (a_t), is calculated as indicated by the next formula:

$$\begin{aligned} EH(b_t, a_t) &= \sum_{b' \in \{b = SE(b_t, a_t, o), o \in O\}} \left[\tau(b_t, a_t, b') \cdot H(b') \right] = \sum_{o}^{O} \left\{ H[SE(b_t, a_t, o)] \cdot g(b_t, a_t, o) \right\} \\ &= \sum_{o}^{O} \left[H[\mathrm{SE}(b_t, a_t, o)] \cdot \sum_{w_j : w_j(a_t)=o} b_t(w_j) \right] = \sum_{o}^{O} \left\{ \sum_{w_j : w_j(a_t)=o} \left[\frac{b_t(w_i)}{\sum\limits_{w_j : w_j(a_t)=o} b_t(w_j)} \cdot \log \left(\frac{b_t(w_i)}{\sum\limits_{w_j : w_j(a_t)=o} b_t(w_j)} \right) \right] \cdot \sum_{w_j : w_j(a_t)=o} b_t(w_j) \right\} \end{aligned}$$

Function $\tau(b_t, a_t, b')$ gives the probability of obtaining the belief state *b'* given current belief state b_t and gazing at position a_t, while *H(b')* is the entropy of the belief state *b'* corresponding to the distribution of probability over the dictionary $\{w_i\}$. $SE(b_t, a_t, o)$ is the belief state that, starting from belief state b_t is obtained after the execution of action a_t resulting in the observation *o*. $g(b_t, a_t, o)$ is the probability of getting observation *o* by executing action a_t in belief state b_t (i.e., the sum of probabilities of all words matching all read characters and with character *o* at position a_t).

It is worth noting that the use of this computational approach in realistic reading tasks is hindered by its computational cost (which grows quadratically with the length of the word/text to be read), and by its huge demands in terms of knowledge (it implicitly assumes that all the possible words/texts are already known, and the current task consist in recognizing which word/text one is currently reading). For text reading, a more feasible computational approach could be adopted that uses this method at two or more levels in parallel, for instance at the level of single words and at the same time at the level of whole sentences (using words and not characters as observations, and changing the priors on words). Another limit of this algorithm

[3] The notation used, (action *a*, belief *b* and observation *o*) is typical of POMDP, which is a formalization of the problem of choosing sequences of actions under uncertainty in order to achieve an optimal total reward.

is that it doesn't model noise in action (e.g., one can believe to be reading the 5th character, but actually read the 6th) and observation (e.g., one can mistake an "l" for a "i"). Modeling noise would result in more complex algorithms like those for planning in POMDP.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European's Community Seventh Framework Programme under grant agreement no FP7-231453 (HUMANOBS, Humanoids That Learn Socio-Communicative Skills Through Observation).

REFERENCES

Altmann, G. T. M., and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* 73, 247–264.

Baayen, H. (2007). "Storage and computation in the mental lexicon," In *The Mental Lexicon: Core Perspectives*, eds G. Jarema and G. Libben (Amsterdam: Elsevier), 81–104.

Baddeley, A. D. (1964). Immediate memory and the "perception" of letter sequences. *Q. J. Exp. Psychol.* 16, 364–367.

Baddeley, A. D. (2007). *Working Memory, Thought, and Action*. Oxford: Oxford University Press.

Ballard, D. H. (1991). Animate vision. *Artif. Intell.* 48, 1–27.

Ballard, D. H., Hayhoe, M. M., and Pelz, J. B. (1995). Memory representations in natural tasks. *J. Cogn. Neurosci.* 7, 66–80.

Bishop, C. M. (1994). Novelty detection and neural network validation. *IEE Proc., Vis. Image Process* 141, 217–222.

Botvinick, M. M., and Plaut, D. C. (2006). Short-term memory for serial order: a recurrent neural network model. *Psychol. Rev.* 113, 201–233.

Brown, G., Preece, T., and Hulme, C. (2000). Oscillator-based memory for serial order. *Psychol. Rev.* 107, 127–181.

Burani, C., Marcolini, S., De Luca, M., and Zoccolotti, P. (2008). Morpheme-based reading aloud: evidence from dyslexic and skilled Italian readers. *Cognition* 108, 1, 243–262.

Burgess, N., and Hitch, G. J. (1992). Toward a network model of the articulatory loop. *J. Mem. Lang.* 21, 429–460.

Burgess, N., and Hitch, G. J. (1999). Memory for serial order: a network model of the phonological loop and its timing. *Psychol. Rev.* 106, 551–581.

Burzio, L. (2004). "Paradigmatic and syntagmatic relations in Italian verbal inflection," in *Contemporary Approaches to Romance Linguistics*, eds J. Auger, J. C. Clements and B. Vance (Amsterdam: John Benjamins), 17–44.

Butz, M. V. (2008). How and why the brain lays the foundations for a conscious self. *Constructivist Found.* 4, 1–42.

Calderone, B., Herreros, I., and Pirrelli, V. (2007). Learning Inflection: the importance of starting big. *Lingue e Linguaggio* 2, 175–200.

Cangelosi, A., and Harnad, S. (2001). The adaptive advantage of symbolic theft over sensorimotor toil: grounding language in perceptual categories. *Evol. Commun.* 4, 117–142.

Chater, N., Crocker, M. J., and Pickering, M. J. (1998). "The rational analysis of inquiry: the case of parsing" in *Rational Models of Cognition,* eds M. Oaksford and N. Chater (Oxford: Oxford University Press), 441–469.

DeLong, K. A., Urbach, T. P., and Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat. Neurosci.* 8, 1117–1121.

Ehrlich, S. E, and Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *J. Verbal Learn. Verbal Behav.* 20, 641–655.

Engbert, R., and Krügel, A. (2010). Readers use Bayesian estimation for eye-movement control. *Psychol. Sci.* 21, 366–371.

Federmeier, K. D. (2007). Thinking ahead: the role and roots of prediction in language comprehension. *Psychophysiology* 44, 491–505.

Geisler, W. S. (2003). "Ideal observer analysis," in *The Visual Neurosciences,* eds L. Chalupa and J. Werner (Boston: MIT press) 825–837.

Grossberg, S. (1986). "The adaptive self-organization of serial order in behavior: speech, language, and motor control," in *Pattern Recognition by Humans and Machines Vol. 1: Speech Perception*, eds E. C. Schwab and H. C. Nusbaum. (New York: Academic Press), 187–294.

Hartley, T., and Houghton, G. (1996). A linguistically constrained model of short-term memory for nonwords. *J. Mem. Lang.* 35, 1–31.

Hayhoe, M., and Ballard, D. H. (2005). Eye movements in natural behavior. *Trends Cogn. Sci. (Regul. Ed.)* 9, 188–193.

Hebb, D. O. (1949). *The Organisation of Behaviour*. New York: Wiley.

Henson, R. N. A. (1996). *Short-term memory for serial order. Unpublished doctoral dissertation, MRC Applied Psychology Unit.* Cambridge: University of Cambridge.

Henson, R. N. A. (1998). Short-term memory for serial order: The start-end model. *Cogn. Psychol.* 36, 73–137.

Houghton, G. (1990). "The problem of serial order: a neural network model of sequence learning and recall," in *Current Research in Natural Language Generation*, eds R. Dale, C. Nellish and M. Zock (San Diego: Academic Press), 287–318.

Houghton, G., and Hartley, T. (1996). Parallel models of serial behaviour: Lashley revisited. *Psyche* 2, 2–25.

Howard, R. A. (1966). Information value theory. *IEEE Trans. Syst. Sci. Cybern.* 2, 22–26.

Jenkins, W., Merzenich, M. M., and Ochs, M. (1984). Behaviorally controlled differential use of restricted hand surfaces induces changes in the cortical representation of the hand in area 3b of adult owl monkeys. *Abstr. - Soc. Neurosci.* 10, 665.

Johansson, R. S., Westling, G., Bäckström, A., and Flanagan, J. R. (2001). Eye-hand coordination in object manipulation. *J. Neurosci.* 21, 6917–6932.

Kaas, J. H., Merzenich, M. M., and Killackey, H. (1983). The reorganization of somatosensory cortex following peripheral nerve damage in adult and developing mammals. *Annu. Rev. Neurosci.* 6, 325–356.

Kohonen, T. (2001). *Self-Organizing Maps*. Heidelberg: Springer-Verlag.

Koutnik, J. (2007). "Inductive modelling of temporal sequences by means of self-organization," in *Proceeding of International Workshop on Inductive Modelling (IWIM 2007)*, Prague, 269–277.

Land, M. F. (2006). Eye movements and the control of actions in everyday life. *Prog Retin Eye Res* 25, 296–324.

Legge, G. E., Hooven, T. A., Klitz, T. S., Mansfield, J. S., and Tjan, B. S. (2002). Mr. Chips 2002: New insights from an ideal-observer model of reading. *Vision Res.* 42, 2219–2234.

Legge, G. E., Klitz, T. S., and Tjan, B. S. (1997). Mr. Chips: An ideal-observer model of reading. *Psychol. Rev.* 104, 524–553.

Libben, G. (2006). "Why studying compound processing? An overview of the issues," in *The Representation and Processing of Compound Words*, eds G. Libben and G. Jarema (Oxford: Oxford University Press), 1–22.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk, volume 2: The Database*. Hillsdale, NJ: Lawrence Erlbaum.

McClelland, J., and Patterson, K. (2002). Rules or connections in past-tense inflections: what does the evidence rule out? *Trends Cogn. Sci.* 6, 465–472.

Nelson, J. D., and Cottrell, G. W. (2007). A probabilistic model of eye movements in concept formation. *Neurocomputing* 70, 2256–2272.

Norris, D. (2006). The Bayesian reader: explaining word recognition as an optimal Bayesian decision process. *Psychol. Rev.* 113, 327–357.

O'Regan, J., and Nöe, A. (2001). A sensorimotor account of vision and visual consciousness. *Behav. Brain Sci.* 24, 883–917.

Papagno, C., Valentine, T., and Baddeley, A. (1991). Phonological short-term memory and foreign-language learning. *J. Mem. Lang.* 30, 331–347.

Penfield, W., and Rasmussen, T. (1950). *The Cerebral Cortex of Man*. New York: Macmillan.

Penfield, W., and Roberts, L. (1959). *Speech and Brain Mechanisms*. Princeton: Princeton University Press.

Pezzulo, G. (2008). Coordinating with the future: the anticipatory nature of representation. *Minds Machine* 18, 179–225.

Pezzulo, G., and Castelfranchi, C. (2007). The symbol detachment problem. *Cogn. Process.* 8, 115–131.

Pickering, M. J., and Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends Cogn. Sci. (Regul. Ed.)* 11, 105–110.

Pinker, S., and Prince, A. (1988). On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition* 29, 195–247.

Pinker, S., and Ullman, M. T. (2002). The past and future of the past tense. *Trends Cogn. Sci.* 6, 456–463.

Pirrelli, V. (2007). Psychocomputational issues in morphology learning and processing: an ouverture. *Lingue Linguaggio* 2, 131–138.

Pirrelli, V., Ferro, M., and Calderone, B. (in press). "Learning paradigms in time and space. Computational evidence from Romance languages," in *Morphological Autonomy: Perspectives from Romance Inflectional Morphology*, eds M. Goldbach, M. O. Hinzelin, M. Maiden and J. C. Smith (Oxford: Oxford University Press).

Post, B., Marslen-Wilson, W., Randall, B., and Tyler, L. K. (2008). The processing of English regular inflections: Phonological cues to morphological structure. *Cognition* 109, 1–17.

Prasada, S., and Pinker, S. (1993). Generalization of regular and irregular morphological patterns. *Lang. Cogn. Process.* 8, 1–56.

Pulvermüller, F. (2003). Sequence detectors as a basis of grammar in the brain. *Theory Biosci.* 122, 87–103.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Q. J. Exp. Psychol.* 62, 1457–1506.

Rayner, K., and Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychon. Bull. Rev.* 3, 504–509.

Roy, D. (2005). Semiotic schemas: a framework for grounding language in action and perception. *Artif. Intell,* 167, 170–205.

Sailer, U., Flanagan, J. R., and Johansson, R. S. (2005). Eye-hand coordination during learning of a novel visuomotor task. *J. Neurosci.* 25, 8833–8842.

Schmidhuber, J. (1991). *Adaptive Confidence and Adaptive Curiosity*. Institut für Informatik, Technische Universitat, Munchen.

Service, L. (1992). Phonology, working memory and foreign-language learning. *Q. J. Exp. Psychol. A.* 45, 21–50.

Shallice, T., and Vallar, G. (1990). "The impairment of auditory-verbal short-term storage," in *Neuropsychological Impairments of Short-Term Memory*, eds G. Vallar and T. Shallice (Cambridge: Cambridge University Press), 121–141.

Sprague, N., and Ballard, D. H. (2003). "Eye movements for reward maximization," in *Proceedings of Advances in Neural Information Processing Systems 16 (NIPS'03)*, eds S. Thrun, L. Saul and B. Schölkopf (Cambridge: MIT Press), 1467–1474.

Sugita, Y., and Tani, J. (2005). Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. *Adapt. Behav.* 13, 33–52.

Triesch, J. J., Ballard, D. H., Hayhoe, M., and Sullivan, B. (2003). What you see is what you need. *J. Vis.* 3, 86–94.

Ullman, M. T. (2004). Contributions of memory circuits to language: the declarative/procedural model. *Cognition* 92, 231–270.

von Hofsten, C. (2004). An action perspective on motor development. *Trends Cogn. Sci.* 8, 266–272.

Wermter, S., Weber, C., and Elshaw, M. (2005). "Associative neural models for biomimetic multi-modal learning in a mirror neuron-based robot," in *Modeling Language, Cognition and Action*, eds A. Cangelosi, G. Bugmann and R. Borisyuk (Singapore: World Scientific), 31–46.

Westermann, G., and Plunkett, K. (2007). Connectionist models of inflection processing. *Lingue Linguaggio* 2, 291–311.

Yarbus, A. (1967). Eye *Movements and Vision*. New York: Plenum Press.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 18 December 2009; paper pending published: 28 January 2010; accepted: 28 April 2010; published online: 03 June 2010.

*Citation: Ferro M, Ognibene D, Pezzulo G and Pirrelli V (2010) Reading as active sensing: a computational model of gaze planning in word recognition. Front. Neurorobot. **4**:6. doi: 10.3389/fnbot.2010.00006*

frontiers in
NEUROROBOTICS

PERSPECTIVE ARTICLE
published: 19 November 2010
doi: 10.3389/fnbot.2010.00010

Robots with language

Domenico Parisi*

Institute of Cognitive Sciences and Technologies, National Research Council, Rome, Italy

Edited by:
Angelo Cangelosi, University of Plymouth, UK

Reviewed by:
Dimitar Kazakov, University of York, UK
Davide Marocco, University of Plymouth, UK

****Correspondence:***
Domenico Parisi, Institute of Cognitive Science and Technologies, National Research Council, Rome, Italy.
email: domenico.parisi@istc.cnr.it

Trying to understand human language by constructing robots that have language necessarily implies an embodied view of language, where the meaning of linguistic expressions is derived from the physical interactions of the organism with the environment. The paper describes a neural model of language according to which the robot's behaviour is controlled by a neural network composed of two sub-networks, one dedicated to the non-linguistic interactions of the robot with the environment and the other one to processing linguistic input and producing linguistic output. We present the results of a number of simulations using the model and we suggest how the model can be used to account for various language-related phenomena such as disambiguation, the metaphorical use of words, the pervasive idiomaticity of multi-word expressions, and mental life as talking to oneself. The model implies a view of the meaning of words and multi-word expressions as a temporal process that takes place in the entire brain and has no clearly defined boundaries. The model can also be extended to emotional words if we assume that an embodied view of language includes not only the interactions of the robot's brain with the external environment but also the interactions of the brain with what is inside the body.

Keywords: emotional words, language, robots

STUDYING LANGUAGE BY CONSTRUCTING ROBOTS THAT HAVE LANGUAGE

If we want to construct *human* robots rather than just *humanoid* robots, that is, if we want to construct robots which actually behave like human beings rather than robots which only resemble human beings in their external morphology, it will be necessary for our robots to possess language because language is such a prominent feature of human beings. Some robots give us the impression of being able to use language but they do not actually understand the language they hear or produce. They are programmed to respond with specific actions to specific acoustic inputs and to generate specific sounds in specific circumstances but human language is much more than that. Of course, human language is a very complicated behavior and it will not be easy to construct robots that can be said to really possess language. But we can make some steps in that direction.

Studying language by constructing robots that have language implies a specific conceptual framework with which to look at human language. Robots are real or simulated physical artifacts. They have a body, they have sensors and effectors with which they interact with the physical environment, their behavior is controlled by a simulated "brain" (an artificial neural network), and their body contains (or should contain; cf. Parisi, 2004) not only a "brain" but also other internal organs and systems. Therefore, robots necessarily imply an "embodied" conception of cognition according to which cognition depends on, and is shaped by, the possession of a body and the movements of the body's different parts. Cognitive representations have been traditionally thought of as based on perception or as abstract representations that do not contain sensory-motor information. However, recent empirical findings and theoretical developments favor a different conception of cognitive representations according to which the body of the organism and the movements of the body's effectors play a critical role in shaping the organism's cognitive representations (Gibson, 1979; Clark, 1999; Barsalou, 2008a,b). Furthermore, if the robot's behavior is controlled by a neural network (as it should be since the brain is part of the body and, to be consistent, robots should be neurorobots), cognitive representations become neural representations, that is, patterns of activation or successions of patterns of activation in a set of units that simulate the brain's neurons. This usefully operationalizes the rather vague concept of cognitive or mental representation since, unlike cognitive or mental representations, artificial neural representations can be observed, measured, and compared with empirical data on the brain.

Constructing robots that have language necessarily extends the embodied conception of cognition to language. Empirical evidence in favor of an embodied conception of cognition has accumulated not only in experiments in which participants respond to the sight of objects but also when they respond to words that refer to those objects (for a recent review, see Fischer and Zwaan, 2008). In both cases, the sensory input activates the neural representation of the action with which participants usually respond to the object. However, a well-developed embodied theory of language should be able to answer many questions that remain still open, and constructing robots that have language should help us to answer these questions. Here are some examples. Are there differences in what happens in the brain when a participant responds to the sight of an object and when he/she responds to a word which refers to the object? Do nouns evoke what are called the stable affordances of an object, i.e., the action with which one responds to the object and which is represented in the brain independently of the actual movements with which the action is physically realized in different circumstances, while seeing an object in a particular orientation also evokes the variable affordances of the object which specify at least some aspects of the movements one has to produce to physically realize the action (Borghi and Riggio, 2009)? Do verbs that refer to actions only or mainly activate the neural representation of the state of the environment which is produced by the movements of the effectors, i.e., the effect of the action, while the sight of an action activates a neural representation of both this state and the movements of the effectors that will produce it? Does possession of a language change how the brain responds to perceived objects and actions in that an individual tend to internally label those objects and actions and therefore he or she responds to both the perceived object and action and the self-produced linguistic signal? Are there differences between the neural representations evoked by verbs and by nouns, or by nouns that refer to tools (e.g., hammer) and nouns that refer to natural objects (tree) (Cangelosi and Parisi, 2001)? How can an embodied theory of language account for abstract words? Do abstract words imply going back and forth between the part of the brain that processes words as sounds and the part which constructs a meaning for the sound, while this is less true for concrete words? How is an embodied meaning for combinations of words (phrases and sentences) constructed? How can an embodied theory account for emotional words and for the emotional component of non-emotional words?

In this paper we describe a simple neural network architecture for language-using simulated robots living in simulated environments and we try to show how this architecture may explain a (very limited) number of linguistic behaviors, where to explain a behavior is to construct a robot that reproduces the behavior. We will refer to robots that have been actually constructed and we will indicate how these robots could be modified to account for other linguistic phenomena.

OBJECTS ARE INTERNALLY REPRESENTED IN TERMS OF THE ACTIONS WITH WHICH WE RESPOND TO THEM

Neurorobots develop internal (neural) representations of perceived objects which are based on the motor actions with which they respond to the objects rather than on the objects' perceptual properties. Imagine a robot whose neural network has sensory units encoding the visual properties of objects, motor units encoding the movements of the robot's arm, and an intermediate layer of internal units. The robot lives in an environment in which it may be exposed to one of four possible objects possessing two properties, color and shape, both with two values, blue and red, square and circle. The robot sees one object at a time and it has to respond by reaching with its arm one of two buttons, one on the right and the other on the left (Borghi et al., 2003, 2005; Di Ferdinando and Parisi, 2004). (The connection weights of the neural network of all the robots described in this paper are evolved using a genetic algorithm. See Mitchell, 1998). If the button on the right has to be reached when the robot sees a square object, independently of the object's color, and the button on the left when the robot sees a circular object, again independently of the object's color, we find that the four objects are represented in the internal units of the robot's neural network in terms of the action that the robot has to do in response to the objects (go to the button on the right, go to the button on the left) rather than in terms of the perceptual properties of the objects as such. In fact we observe only two activation patterns in the neural network's internal units, one which controls the action of reaching the button on the right and the other one which controls the action of reaching the button on the left. The object's perceptual property which is critical to decide which action to do, in our case shape, determines the internal representation of the object, while the other property, color, is ignored. Notice that the internal representation of an action is abstract in the sense that it needs to be translated into a succession of specific movements of the robot's arm which vary as a function of the starting position of the arm. In fact, the robot's neural network includes an additional set of proprioceptive input units encoding the current position of the robot's arm which project directly to the motor units. The two activation patterns that constitute the internal representations of the two actions interact with this proprioceptive information from the arm so that, for any starting position of the arm, each of the two abstractly represented action can be translated in the appropriate succession of movements.

A NEURAL NETWORK'S ARCHITECTURE FOR LANGUAGE-USING ROBOTS

The robots we have described in the preceding Section do not have language but they only respond to objects with the appropriate non-linguistic action. We now ask: What is the basic architecture of the neural network that controls the behavior of a language-using robot? The robot's overall neural network is made up of two sub-networks, the non-linguistic sub-network (NL) and the linguistic sub-network (L) (Mirolli and Parisi, 2005). Both sub-networks include three layers of units: a sensory layer, a motor layer, and an intermediate layer of internal units. The sensory units of NL encode perceived objects and its motor units encode movements of the robot's effectors such as the robot's arm. The sensory units of L encode heard linguistic sounds and its motor units encode movements of the robot's phono-articulatory organs that result in the production of linguistic sounds. NL maps non-linguistic sensory input into non-linguistic actions. In fact, NL is identical to the neural network of our robots that had to reach with their arm one of two different buttons in response to a visually perceived object. L maps heard linguistic sounds into phono-articulatory movements. This is the network that controls the behavior of a robot which is able to imitate (repeat) heard sounds without associating any meaning to them. The robot hears a linguistic sound and it responds with movements of its phono-articulatory organs that reproduce the sound.

The two sub-networks remain functionally or perhaps even anatomically separate during an initial period of the robot's existence which corresponds to children's first year of life. During this period the robot learns to respond to non-linguistic sensory input (say, perceived objects) with movements of its non-linguistic effectors (arm, hands, legs, eyes) using its NL sub-network. In addition the robot uses its L sub-network to produce linguistic sounds with its phono-articulatory organs, either spontaneously or in response of its own heard sounds (babbling) and, later on, by imitating the linguistic sounds produced by already speaking robots.

At the end of this period the two sub-networks begin to be connected together by two-way connections that go from the internal units of NL to the internal units of L, and vice versa, and the synaptic weights of these two-way connections are learned based on the co-variation of specific linguistic sounds with specific objects and actions in the robot's experience. From this point on our robot becomes a language-using robot. The robot still uses NL to produce non-linguistic actions in response to non-linguistic sensory input but, in addition, it begins to understand and to produce language. Language understanding consists in responding to heard linguistic sounds with the appropriate movements of the non-linguistic effectors while language production consists in responding to non-linguistic input with phono-articulatory movements that produce the appropriate linguistic sounds. In language understanding neural activation spreads from the sensory layer of L (heard words) to the internal units of L and from there to the internal units and to the motor layer of NL (non-linguistic actions). In language production activation spreads from the sensory layer of NL (perceived objects and actions) to the internal units of NL and from there to the internal units and to the motor layer of L (phono-articulatory movements that produce words). (We are talking here of overt responses to sensory input but activation can stop at the internal layer of the two sub-networks, where non-linguistic and linguistic actions are neurally represented, without producing overt responses, that is, without translating these actions into actual physical movements of either the non-linguistic or linguistic effectors).

INFLUENCE OF LANGUAGE ON THE ROBOT'S CATEGORIES

The network architecture described in the preceding Section allows us to (begin to) answer the question of what are the consequences of possessing a language for a robot's cognition, that is, for the functioning of the robot's NL sub-network. More specifically, in this Section we will see what are the consequences of possessing a language for the robot's categories.

As we have seen in Section "Objects are Internally Represented in Terms of the Actions with Which We Respond to Them", perceived objects that have to be responded to with the same action elicit an identical activation pattern in the network's internal units even if they are perceptually different, while perceived objects that have to be responded to with different actions elicit different activation patterns in the internal units. This is the basis for defining an action-based notion of categories. A category is an internal activation pattern elicited by different objects that have to be responded to with the same action. For the robots of Section "Objects are Internally Represented in Terms of the Actions with Which We Respond to Them", two objects with different color elicit the same activation pattern in the robot's internal units if they have to be responded to with the same action. Hence, for those robots square objects form one category and circular object form another category.

In the robots we have described categories correspond to a single activation pattern in the robot's internal units. As we consider more complex environments, however, we have to qualify this claim. Imagine a mobile (and armless) robot living in an environment that contains a large variety of perceptually different objects that have to be responded to with the same action (say, approaching and reaching perceptually different edible mushrooms) and a variety of different objects which have to be responded to with another action (avoiding perceptually different poisonous mushrooms) (Cangelosi and Parisi, 1998). In these circumstances, we cannot expect that all the mushrooms that have to be responded to with the same action will evoke an identical activation pattern in the internal units of the robot's neural network, completely eliminating the differences among the individual mushrooms that have to be responded to with the same action. In fact, if we evolve a population of robots in this new environment and we examine the activation patterns elicited by the mushrooms in the internal units of the robot's neural network, we find that even perceptually different mushrooms that have to be responded to with the same action tend to elicit somewhat different activation patterns in the internal units. However, we can still maintain the basic assumption of the action-based conception of cognition, i.e., that different objects tend to be internally represented on the basis of the action with which they have to be responded to, rather than in terms of their purely perceptual characteristics. When the robots have (evolutionarily) learned to respond appropriately to the mushrooms, i.e., they eat the edible ones and avoid the poisonous ones, we discover that the mushrooms of one category do not evoke exactly the same internal activation pattern in the internal units of the robots' neural network. However, the internal activation patterns evoked by the mushrooms belonging to one action-defined category resemble each other and are very different from the activation patterns evoked by the mushrooms belonging to the other category. The two categories of mushrooms can be formally represented as two "clouds" of points in the abstract hyperspace of the internal units, where each point represents the internal activation pattern of an individual mushroom, each dimension of the space corresponds to one internal unit, and the position of the point on that dimension corresponds to the activation level of the unit. The points that correspond to one category of mushrooms form one cloud, and they are close to one another, and the points corresponding to the other category form another, separate, cloud.

What if when the robot encounters a mushroom, it does not only visually perceive the mushroom but it also hears the word that describes the category of the mushroom, for example the words "edible" and "poisonous"? Now the neural network that controls the robot's behavior includes both a NL sub-network and L sub-network. When the robot encounters an edible mushroom and it visually perceives the mushroom with its NL sub-network, it also hears the sound "edible" with its L sub-network, while when it encounters a poisonous mushroom it hears the sound "poisonous". (Notice that these two sounds can be produced by another robot or they can be self-produced by the robot. For the self-production of language as an important component of what we can call a robot's mental life, see Section "Robot That Talk to Themselves"). If we examine the clouds of points representing the two categories of mushrooms in the internal units of a robot's NL sub-network, we find that, compared to the robots without language, the two clouds have a smaller size and there in a greater distance between the centers of the two clouds. As a consequence, we find that the robots are better able to distinguish between the two categories of mushrooms and to avoid making errors by eating a poisonous mushroom or avoiding an edible mushroom (Mirolli and Parisi, 2005). The linguistic labeling of categories of objects makes these categories better able to support effective behavior. This appears to be an important consequence of possessing a language and may have had a crucial role in its evolutionary emergence.

Notice that in many neural network models of "semantic knowledge" (e.g., Rogers and McClelland, 2004), word meanings are identified with categories or concepts and therefore it is in principle impossible to ask the question of what might be the influence of language on categories. The neural network of our language-using robots is made up of two parts, one which is non-linguistic and the other one which is linguistic, and categories emerge in the non-linguistic part as a consequence of the non-linguistic interactions of the robot with the non-linguistic environment. Only when the linguistic part becomes operational (in children at around 1 year of age) one may pose the question of what are the consequences of possessing a language for the non-linguistic functioning of the organism.

THE EMERGENCE OF DIFFERENT TYPES OF WORDS IN THE ROBOTS' LANGUAGE

In the robots described so far the NL sub-network has a single layer of internal units that receive activation from the sensory units and send activation to the motor units, and we have seen that the pattern of activation appearing in the NL internal units encodes the action with which the robot has to respond to the sensory input and ignores the properties of the sensory input which are not relevant to decide the action. The edible mushrooms are all different from each other but this variation tends to be ignored (or minimized) by the internal units because all edible mushrooms, independently of their differences, have to be responded to with the same action: approaching and reaching the mushroom. What if NL has not one but two successive layers of internal units, with the sensory units sending their activation to the first layer, this layer sending its activation to the second layer, and the second layer sending its activation to the motor units? If we construct a robot with this type of NL, we find that, while the second layer of internal units specifies the action to be executed and ignores the properties of the perceived object which are irrelevant for the action (like in our robots with a single layer of units), the first layer of internal units preserve more of the properties of the perceived object, even if they appear not to be relevant for the action (Borghi et al., 2003).

Why might it be useful for our robots to have two layers of internal units and not only one? The robots described so far have only to approach and reach the edible mushrooms in order to eat them. But imagine another robot which has an arm and a hand and which to eat an edible mushroom has first to grasp the mushroom with its hand. The robots lives in an environment in which the edible mushrooms are of two sizes: small and large. The robot has to approach and reach both small and large mushrooms but then it has to grasp the mushrooms with its hand in order to bring them to the mouth. There are two actions of grasping. To grasp small mushrooms the robot has to produce a precision grip by using the thumb and index finger of its hand while to grasp large mushrooms the action has to be a power grip which uses all the fingers of the hand. This robot would find it useful to have two layers of internal units, one which specifies the action of approaching and reaching edible mushrooms or the action of avoiding poisonous mushrooms and the other one which specifies the action of grasping small edible mushrooms with a precision grip or the action of grasping large edible mushrooms with a power grip.

These robots would make it possible to ask some interesting questions about the neural representation of language. As we know, the internal units of L are bi-directionally connected with the internal units of NL. But now NL has two layers of internal units, one specifying the actions of approaching and reaching the edible mushrooms and avoiding the poisonous ones, and the other one specifying the actions of producing a precision grip of the hand for small edible mushrooms and a power grip for large edible mushrooms. With which of these two different internal layers of NL will the internal layer of L be bi-directionally connected? Would the internal layer of L be bi-directionally connected with both the first and the second internal layers of NL, or would it be preferentially connected with the more perceptually abstract second layer? Notice that the answer to this question might depend on the robots' language. Small and large edible mushrooms might co-vary with two different sounds, i.e., two different nouns, that is, there might be one sound (name) for small edible mushrooms and a different sound (a different name) for large edible mushrooms. Or the robots' language might include a sound which co-varies with both small and large edible mushrooms and two other sounds which co-vary, respectively, with small and large mushrooms (and probably with other things that require a precision or a power grip). This would make it possible to being to recognize different types of words in our robots' language (say, nouns and adjectives).

A MORE SOPHISTICATED MODEL OF THE INTERNAL LAYER OF BOTH THE NL AND N SUB-NETWORKS

We have assumed so far that external input, from within the network or from the outside environment, evokes one single pattern of activation in the internal units of NL and L. But let us change the model and imagine that the internal units of both NL and L have internal (horizontal) connections that allow one activation pattern to elicit another activation pattern in the same set of units. In this manner, an external input will evoke not a single (static) activation pattern but a succession of activation patterns in the internal units of both NL and L. This is just one particular instance of a general property of brain activity which is not well captured in most neural network models: brain activity is made up of continuous processes, not states. Time is a crucial property of brain activity but it is not well captured by neural network models that conceive network activity as a succession of discrete time cycles. Objects and words should elicit processes in a neural network which at some point or another cause a response in the network's motor units, and this response will in turn cause other processes in the neural network. (We do not address here how this could be implemented in our language-using robots).

This new type of neural network for our language-using robots may help us explain one type of word associations, i.e., word-word associations. The sound of a word, represented as an action of the phono-articulatory effectors of the robot in the internal units of L, will evoke the sound of another word in the same internal units of L. Another type of word associations, based on the meaning of words and not just on their sound, can be reproduced with our robots if an activation pattern in the L internal units evokes an activation pattern in the NL internal units which evokes a second activation pattern in the same NL internal units which in turn evokes an activation pattern in the L internal units. The first type of word associations requires a succession of activation patterns within the L internal units while this second type requires going back and forth between the internal layers of L and NL.

This more sophisticated (and, we believe, more realistic) neural network model may lead to a number of interesting conclusions concerning the nature of language, with particular reference to three issues: (a) how the meaning of words is represented in the brain; (b) the ambiguity of all words (and not only of ambiguous words); and (c) the idiomatic character of all multi-word expressions (and not only idiomatic expressions). Let us consider these three issues.

(a) There is no such thing as the meaning of a word in the brain
If an activation pattern in the L internal units evokes an activation pattern in the NL internal units which in turn evokes another activation pattern in the same NL internal units, and so on, one is led to the conclusion that there is nothing like the meaning of a word in the brain. Heard words are specific entry points to the NL sub-network but they elicit in the NL sub-network not a single activation pattern but a succession of activation patterns in many possible directions as a function of the current context and many other factors, and it is difficult to say where the process ends. The existence of a "semantic module" is assumed in many symbolic models and in many traditional, i.e., non-embodied, neural network models of language. But in the neural network that controls the behavior of our language-using robots there is no special "semantic module" which contains the "meanings" of words. The NL sub-network is the "rest of the brain" which is activated in many possible directions when one hears a word. Words do not have well-defined meanings but they are just entry points for activating the entire brain.

(b) All words are ambiguous
The more sophisticated neural network of our language-using robots should help us to explain the role of context in language understanding. We define context as any additional input, linguistic or non-linguistic, arriving from outside the brain or self-generated inside the brain, that may influence what activation patterns are sequentially elicited in the internal units of the NL sub-network. Among other things, context explains how the brain disambiguates ambiguous words. The context is an additional input that directs the activation process in the NL internal units in one direction or another. The ambiguous word "club" activates one activation pattern in the internal units of the NL sub-network in the context of golf and another activation pattern in the context of the social behavior of some people.
What is more interesting is that our model can explain the less well-recognized fact that all words are to some extent ambiguous and there is no clear dividing line between ambiguous and non-ambiguous words. For all words, the context in which a word is used directs the understanding of the word, i.e., the succession of activation patterns in the internal units of the NL sub-network, in one or another direction. For example the word "water", which is not normally considered to be an ambiguous word, may elicit different activation patterns in the NL sub-network as a function of the particular context in which the word is used. This may be extended to the metaphorical use of words and to words that have both a literal meaning and a metaphorical meaning, such as the verb "to grasp".

(c) All multi-word expressions are idiomatic
Not only there is no clear separation between ambiguous words and non-ambiguous words but there is no clear separation between idiomatic expressions and non-idiomatic expressions. Idioms are defined as multi-word expressions whose meaning cannot be derived from the meanings of the component words (Cacciari and Tabossi, 1993). Idiomatic expressions are considered as different from non-idiomatic expressions in that non-idiomatic expressions are sequences of words which elicit an overall pattern of activation in the NL sub-network which is made up of the activation patterns elicited by the words that make up the sequence (phrase or sentence) according to some general rules (syntax). We claim that all multiple-word expressions, when they are actually used, are to some extent idiomatic, that is, they elicit an activation pattern in the NL internal units is something more than, and more specific, than the sum of the component words. (This applies even to what appear to be the simplest multi-word expressions, i.e., verb-noun expressions). Our model can provide an explanation both for idioms and for the fact that all multi-word expressions possess some degree of idiomaticity (Wray, 2002). The model should explain these different degrees (and types; cf. Wray, 2002) of idiomaticity because the overall activation pattern which is activated in the internal units of the NL sub-network when the robot arrives to the end of the sequence of heard words may be related to the activation patterns elicited by the single words of the sequence in a variety of different and unique ways.

ROBOT THAT TALK TO THEMSELVES

The neural network of our language-using robots allows us to (begin to) explain an important aspect of mental life, that is, mental life as talking to oneself (Parisi, 2007). The simple network architecture described in Section "A Neural Network's Architecture for Language-Using Robots" appears to be generally appropriate to capture how language can influence cognition, providing the basis for a Vygotskyan robotics (Mirolli and Parisi, 2009, 2010). But if we assume that the internal units of the robots' neural network have horizontal connections and these connections can produce a succession of activation patterns in both the NL and L internal units, we can see how the reciprocal connections linking the NL and L internal units can explain mental life as talking to oneself. We have seen the role of these reciprocal connections in explaining language understanding and language production. But when a non-linguistic input arrives to the sensory units of NL, for example the robot sees a cat, and the activation spreads to the internal units of NL and then to the internal units of L, two different things can happen. One is that the activation reaches the motor units of L and the robots pronounces the word "cat". The other is that the activation pattern in the internal units of L elicits another activation pattern in the same set of units, for example the activation pattern corresponding to word "dog". This activation pattern in turn will elicit in the internal units of NL the activation pattern (or rather the succession of activation patterns) that gives a meaning to the word "dog". This is already talking to oneself. The robots hears the self-produced word "dog" and understands the word. But what is interesting is that the process can go back and forth between NL and L. An activation pattern in the internal units of NL can elicit another activation pattern in the same units and this other activation pattern can elicit an activation pattern in the internal units of L. The process can go on an indefinite number of times, as when one is immersed in his or her thoughts.

Talking to oneself really takes off when the process of going back and forth between the L and NL internal units interacts with the process of generating a succession of activation patterns in many possible directions in the NL internal units. The result of the interaction between these two processes is that the activation patterns evoked in the L internal units (words) influence and control the succession of activation pattern evoked in the NL units. This is an important component of talking to oneself as thinking.

LANGUAGE PRODUCTION DETERIORATES MORE THAN LANGUAGE UNDERSTANDING WITH AGE

A robot that has language must also be able to exhibit the rich phenomenology of pathological linguistic behaviors. By lesioning in different places and in different ways the neural network that controls both the non-linguistic and the linguistic behavior of the robot, we should be able to reproduce a variety of linguistic disorders. We will only mention here a phenomenon which is not considered as really pathological but still involves some malfunctioning of language. With old age many people find it difficult to find the word that expresses something they appear to have in their mind. This difficulty in producing the word is not normally accompanied by a parallel difficulty in understanding the word. Our model of language might be able to reproduce this asymmetry if we assume that in old age there is a gradual but diffuse loss of neurons or of connections between neurons. The model can explain both facts if we make the very reasonable assumption that the internal layer of the NL sub-network contains many more units (neurons) than the internal layer of the L sub-network. If there is a diffuse loss of units or of connections between units (including the two-way connections between the two layers of units) it might be easier for the network to go from a pattern of activation in the L sub-network (internal representation of the heard word) to the appropriate pattern of activation in the NL sub-network (understanding the word) than to go from a pattern of activation in the NL sub-network to the appropriate pattern of activation in the L sub-network (finding the word to express something one has in mind) simply because the larger sub-network is more robust than the smaller sub-network (Mirolli et al., 2007).

THE EMOTIONAL MEANING OF WORDS

Words do not only have a cognitive or informational meaning but they also have an emotional meaning. Can the neural network of our language-using robots be modified so that our robots might be able to appreciate the emotional meaning of words?

As we have said at the beginning of the paper, robots indicate the importance of the body and its movements in determining cognition so that robotics naturally converges with theories of cognition as embodied and as action-based. But both current robots and these theories have two related limitations that need to be overcome if we want to construct a more complete theory of the human mind. The first limitation is due to the fact that an organism's body does not only have an external morphology and sensory and motor organs but it also includes internal organs and systems which exist inside the body beyond the brain. The second limitation is that the mind is not only cognition but also motivation and emotion. The two limitations are related because while cognition mainly results from the interactions of the brain with the external environment, motivations and emotions mainly result from the interactions of the brain with the other organs and systems that exist inside the body.

An embodied conception of the entire mind (not just cognition) assumes two levels of functioning of the behavioral system of an animal, a strategic or motivational level and a tactical or cognitive level. All animals have many different motivations that are generally impossible to satisfy at the same time. Therefore these different motivations necessarily compete with one another for the control of the animal's behavior and at any given time the strategic level of functioning of the animal has to decide which motivation the animal should pursue with its behavior. The decision is taken on the basis of the current intensity of the different motivations, which is determined by many different factors, both intrinsic (the overall adaptive pattern of the animal and the specific environment in which the animal lives) and contextual (sensory input from the body and from the external environment). Once a decision is taken at the strategic level, the cognitive level executes the activity which will hopefully satisfy the motivation decided at the strategic level. Emotions operate at the strategic or motivational level by increasing the current intensity of one or another motivation so that the strategic level may function more effectively (fewer errors, faster decisions, increasing the persistence of important motivations, etc.). The tactical level is mostly implemented through the interactions of the animal's brain with the external environment, while the strategic level is mostly implemented through the interactions of the brain with what is inside the body. If robots should help us to develop a complete embodied theory of the mind, what is needed is an internal robotics, that is, the construction of robots that do not have only the external morphology of an animal's body and a "brain" which interacts only with the external environment but also have internal (artificial) organs and systems and a "brain" which interacts with these internal organs and systems (Parisi, 2004).

How are this more complete conception of the mind and this more complete robotics related to the construction of robots that have language? Not only so-called emotional words but all words have an emotional component that plays a role in their use and, unless we are able to endow the words used by a robot with this emotional component, we are not authorized to say that we have constructed robots that have language.

How should we proceed? The first step is to construct robots that have many different motivations and have to choose which one of these different motivations will control the robot's behavior at any given time. Current robots tend to have just one single motivation, and this motivation is not chosen by them but by their users, that is, by us. The second step is to endow our robots with emotions (not just with the capacity to express emotions that they do not actually have, as in most current "emotional" robots). This can be done by adding an "emotional circuit" to the neural network that controls the robot's behavior, where the function of this emotional circuit is to enable the robot to make more effective and more efficient motivational choices. The emotional (neural) circuit can be activated by input from the body (e.g., hunger or thirst) or from the external environment (e.g., a predator or a possible mate) and it sends activation to the rest of the robot's neural network, influencing the motivational decision taken by the neural network and therefore the actual behavior exhibited by the robot. The emotional circuit also interacts with the rest of the robot's body, sending and receiving activation to and from internal organs (e.g., heart and gut) and systems (e.g., endocrine and immunological systems).

The first steps in this direction have already been made by constructing robots that to survive and reproduce have to both eat and drink, or to both eat and avoid being killed by a predator, or to both eat and approach a mate. The results indicate that in all cases adding an emotional circuit to the neural network that controls the robot's behavior leads to more effective behaviors and, therefore, to longer lives and more offspring. (For a detailed description of these robots, see Parisi and Petrosino, in press).

How can we extend our model of language-using robots so that our robots can understand and produce words that have an emotional component? The answer is to add an emotional circuit to the NL sub-network of our robots so that the emotional circuit can also be activated when the robot understands or produces a word. The use (understanding and production) of some words will more directly and more extensively involve the activation of the emotional circuit of the NL sub-network but all words will in one manner or other and to a greater or smaller degree activate the circuit. Adding an emotional circuit to the NL sub-network of our language-using robots will be necessary if the motivational/emotional level of behavior of our robots should be influenced by hearing both words produced by other robots and self-produced words. If we further assume that exercising our emotions in safe conditions such as those implied in exposing oneself to artistic artifacts leads to a more sophisticated motivational/emotional functioning, our emotional language-using robots might also be able to understand and enjoy poetry and other forms of verbal art as humans do. Poems and novels are verbal stimuli that to be really understood and enjoyed should activate the emotional circuit of the NL sub-network of our language-using robots.

CONCLUSION

A crucial step toward the construction of really human, and not simply humanoid, robots is to construct robots that have language. In this paper we have described a simple neural network architecture that controls the behavior of a language-using robot and we have illustrated a number of language-related phenomena that can be explained (reproduced) with our language-using robot. However, most of the work to construct robots that can be said to have language has still to be done since human language is such a complex and multi-faceted phenomenon. Language has emerged from animal-like non-linguistic communication systems, is culturally transmitted, and it changes historically. Language is learned through a succession of specific stages. Linguistic expressions are made up of simpler expressions, from morphemes to words, from phrases to sentences. Language is a crucial ingredient of human social life and it is used to accomplish a large number of different social goals. We think that all these aspects of language which are studied by a variety of scientific disciplines might be illuminated by a well-developed linguistic robotics. (For a description of the different goals of such a linguistic robotics, see Parisi and Cangelosi, 2002).

REFERENCES

Barsalou, L. W. (2008a). Grounded cognition. *Annu. Rev. Psychol.* 50, 617–645.

Barsalou, L. W. (2008b). Grounded cognition: past, present, and future. *Top. Cogn. Sci.* 2, 716–724.

Borghi, A., Di Ferdinando, A., and Parisi, D. (2003). "The role of perception and action in object representation," in *Connectionist Models of Cognition and Perception*, eds J. A. Bullinaria and W. Lowe (Singapore: World Scientific), 40–50.

Borghi, A., Parisi, D., and Di Ferdinando, A. (2005). Action and hierarchical levels of categories: A connectionist perspective. *Cogn. Syst. Res.* 6, 99–110.

Borghi, A. M., and Riggio, L. (2009). Sentence comprehension and simulations of objects' temporary, canonical, and stable affordances. *Brain Res.* 1253, 117–128.

Cacciari, C., and Tabossi, P. (eds) (1993). *Idioms: Processing, Structure, and Interpretation*. New York, Psychology Press.

Cangelosi, A., and Parisi, D. (1998). The emergence of language in an evolving population of neural networks. *Connection Sci.* 10, 83–97.

Cangelosi, A., and Parisi, D. (2001). "How nouns and verbs differentially affect the behaviour of artificial organisms," in *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, eds J. D. Moore and K. Stenning (London: Erlbaum), 170–175.

Clark, A. (1999). An embodied cognitive science? *Trends Cogn. Sci.* 3, 345–351.

Di Ferdinando, A., and Parisi, D. (2004). "Internal representations of sensory input reflect the motor output with which organisms responds to the input," in *Seeing and Thinking*, ed. A. Carsetti (Dordrecht: Kluwer), 115–141.

Fischer, M. H., and Zwaan, R. A. (2008). Embodied language. A review of the role of the motor system in language comprehension. *Q. J. Exp. Psychol.* 61, 825–850.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Mirolli, M., Cecconi, F., and Parisi, D. (2007). A neural network model for explaining the asymmetries between linguistic production and linguistic comprehension. in *Proceedings of the 2007 European Cognitive Society Conference*, eds S. Vosniadou, D. Kayser, and A. Protopapas. Hove: Lawrence Erlbaum, 670–675.

Mirolli, M., and Parisi, D. (2005). "Language as an aid to categorization: A neural network model of early language acquisition," in *Modeling Language, Cognition, and Action. Proceedings of the 9th Neural Computation and Psychology Workshop*, eds A. Cangelosi, G. Bugmann, and R. Borysyuk (Singapore: World Scientific), 97–106.

Mirolli, M., and Parisi, D. (2009). Language as a cognitive tool. *Minds Mach.* 19, 517–528.

Mirolli, M., and Parisi, D. (2010). Towards a Vygotskyan cognitive robotics. The role of language as a cognitive tool. *New Ideas Psychol.* 1–14.

Mitchell. M. (1998). *An Introduction to Genetic Algorithms*. Cambridge, MA: MIT Press.

Parisi, D. (2004). Internal robotics. *Connection Sci.* 16, 325–338.

Parisi, D. (2007). "Mental robotics," in *Artificial Consciousness*, eds A. Chella and R. Manzotti (New York: Imprint-Academic).

Parisi, D., and Cangelosi, A. (2002). "A unified simulation scenario for language development, evolution, and historical change," in *Simulating the Evolution of Language*, eds A. Cangelosi and D. Parisi (London: Springer), 255–276.

Parisi, D., and Petrosino, G. (in press). Robots that have emotions. *Adapt. Behav.*

Rogers, T. T., and McClelland, J. L. (2004). *Semantic Cognition: A Parallel Processing Approach*. Cambridge, MA: MIT Press.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 16 December 2009; paper pending published: 22 March 2010; accepted: 15 July 2010; published online: 19 November 2010.

Citation: Parisi D (2010) Robots with language. Front. Neurorobot. ***4****:10. doi: 10.3389/fnbot.2010.00010*

ORIGINAL RESEARCH ARTICLE
published: 03 June 2010
doi: 10.3389/fnbot.2010.00008

Linking language with embodied and teleological representations of action for humanoid cognition

Stephane Lallee, Carol Madden, Michel Hoen and Peter Ford Dominey*

Robot Cognition Laboratory, Integrative Neuroscience, Stem Cell and Brain Research Institute, Institut National de la Santé et de la Recherche Médicale U846, Bron, France

Edited by:
Angelo Cangelosi,
University of Plymouth, UK

Reviewed by:
José F. Fontanari,
University of Sao Paulo, Brazil
Claudia Scorolli,
University of Bologna, Italy

****Correspondence:***
Peter Ford Dominey, Robot Cognition Laboratory, Stem Cell and Brain Research Institute, Institut National de la Santé et de la Recherche Médicale, 18, avenue du doyen Jean Lépine, 69675 Bron Cedex, France.
e-mail: peter.dominey@inserm.fr

The current research extends our framework for embodied language and action comprehension to include a teleological representation that allows goal-based reasoning for novel actions. The objective of this work is to implement and demonstrate the advantages of a hybrid, embodied-teleological approach to action–language interaction, both from a theoretical perspective, and via results from human–robot interaction experiments with the iCub robot. We first demonstrate how a framework for embodied language comprehension allows the system to develop a baseline set of representations for processing goal-directed actions such as "take," "cover," and "give." Spoken language and visual perception are input modes for these representations, and the generation of spoken language is the output mode. Moving toward a teleological (goal-based reasoning) approach, a crucial component of the new system is the representation of the subcomponents of these actions, which includes relations between initial enabling states, and final resulting states for these actions. We demonstrate how grammatical categories including causal connectives (e.g., because, if–then) can allow spoken language to enrich the learned set of state-action-state (SAS) representations. We then examine how this enriched SAS inventory enhances the robot's ability to represent perceived actions in which the environment inhibits goal achievement. The paper addresses how language comes to reflect the structure of action, and how it can subsequently be used as an input and output vector for embodied and teleological aspects of action.

Keywords: human-robot interaction, action perception, language, cooperation

INTRODUCTION – A FRAMEWORK FOR LANGUAGE AND ACTION

One of the central functions of language is to coordinate cooperative activity (Tomasello, 2008). In this sense, much of language is about coordinating action. Indeed, language constructions themselves become linked to useful actions in our experience, as emphasized by Goldberg (1995, p. 5) "constructions involving basic argument structure are shown to be associated with dynamic scenes: experientially grounded gestalts, such as that of someone volitionally transferring something to someone else, someone causing something to move or change state…" Interestingly, this characterization is highly compatible with the embodied language comprehension framework, which holds that understanding language involves activation of experiential sensorimotor representations (Barsalou, 1999; Bergen and Chang, 2005; Zwaan and Madden, 2005; Fischer and Zwaan, 2008; Pulvermüller et al., 2009). We have pursued this approach in developing neurally inspired systems that make this link between language and action. We introduce this approach in the remainder of this section, describing the path we have taken to arrive at our present work.

In this context of linking language and action, we first developed an action recognition system that extracted simple perceptual primitives from the visual scene, including contact or collision (Kotovsky and Baillargeon, 1998), and composed these primitives into templates for recognizing events like give, take, touch and push. Siskind and colleagues (Fern et al., 2002) developed a related action learning capability in the context of force dynamics. A premise of this approach is that it is not so much the details of spatial trajectories of actions, but more their resulting states which characterize action in the context of perception and recognition (Bekkering et al., 2000). The resulting system provided predicate–argument representations of visually perceived events, which could then be used in order to learn the mapping between sentences and meaning. We demonstrated that naïve humans could narrate their actions which were perceived by the event recognition system, thus providing sentence-meaning inputs to the grammatical construction model, which was able to learn a set of grammatical constructions that could then be used to describe new instances of the same types of events (Dominey and Boucher, 2005).

We subsequently extended the grammatical construction framework to robot action control. We demonstrated that the robot could learn new behaviors (e.g., Give me the *object*, where *object* could be any one of a number of objects that the robot could see) by exploiting grammatical constructions that define the mapping from sentences to predicate–argument representations of action commands. This work also began to extend the language–action framework to multiple-action sequences, corresponding to more complex behaviors involved in cooperative activity (Dominey et al., 2009b). Cooperation – a hallmark of human cognition (see Tomasello et al., 2005) – crucially involves the construction of action plans that specify the respective contribution of both agents, and the representation of this shared plan by both agents. Dominey and Warneken (in press) provided the Cooperator – a 6DOF arm and monocular vision robot – with this capability, and demonstrated that the resulting system could engage in cooperative activity, help the human, and perform role reversal, indicating indeed that it had a "bird's eye view" of the cooperative activity. More recently, Lallee et al. (2009) extended this work so that the robot could acquire shared plans by observing two humans perform a cooperative activity.

An important aspect of this area of research is that the source of meaning in language is derived directly from sensory-motor experience, consistent with embodied language processing theories (Barsalou, 1999; Bergen and Chang, 2005; Zwaan and Madden, 2005). For instance, Fontanari et al. (2009) have demonstrated that artificial systems can learn to map word names to objects in a visual scene in a manner that is consistent with embodied theories. However, we also postulated that some aspects of language comprehension must rely on a form of "hybrid" system in which meaning might not be expanded completely into its sensory-motor manifestation (Madden et al., 2010). This would be particularly useful when performing goal-based inferencing and reasoning. Indeed, Hauser and Wood (2010) argue that understanding action likely involves goal-based teleological reasoning processes that are distinct from the embodied simulation mechanisms for action perception. These authors state that, "Integrating insights from both motor-rich (simulation, embodiment) and motor-poor (teleological) theories of action comprehension is attractive as they provide different angles on the same problem, a set of different predictions about the psychological components of action comprehension, and enable a broad comparative approach to understanding how organisms interpret and predict the actions of others" (Hauser and Wood, 2010, p. 4). This is consistent with a hybrid approach to action understanding that we have recently proposed (Madden et al., 2010; for other dual-representation approaches see: Barsalou et al., 2008; Dove, 2009). In that model, action perception and execution take place in an embodied sensorimotor context, while certain aspects of planning of cooperative activities are implemented in an amodal system that does not rely on embodied simulation.

A fundamental limitation of this approach to date is that the system has no sense of the underlying goals for the individual or joint actions. This is related to the emphasis that we have placed on recognition and performance of actions, and shared action sequences, without deeply addressing the enabling and resulting states linked to these actions. In the current research, we extend our hybrid comprehension model to address aspects of goal-based reasoning, thus taking a first step toward the type of teleological reasoning advocated by Hauser and Wood (2010). The following section describes how this new framework addresses the limitations of the current approach.

A NEW FRAMEWORK FOR ACTION AND LANGUAGE – COMBINING TELEOLOGICAL AND EMBODIED MECHANISMS

In Lallee et al. (2009) the iCub robot could observe two human agents perform a cooperative task, and then create a cooperative plan, which includes the interleaved temporal sequence of coordinated actions. It could then use that plan to take the role of either of the two agents in the learned cooperative task. This is illustrated in **Figure 1**. A limitation of this work is that the task is represented as a sequence of actions, but without explicit knowledge of the results of those actions, and the link between them. In the current work, this limitation is addressed by allowing the robot to learn for each action, what is the enabling state of the world which must hold for that action to be possible, and what is the resulting state that holds once the action has been performed. We will refer to this as the $S_E AS_R$ state-action-state (SAS) representation of action. This is consistent with our knowledge that humans tend to represent actions in terms of goals – states that result from performance of the action (Woodward, 1998). Furthermore, neurophysiological evidence of such a goal specific encoding of actions has been observed in monkeys (Fogassi et al., 2005) whereby the same action (grasping) can be encoded in different manners according to intentions or goals (grasping for eating/grasping for placing).

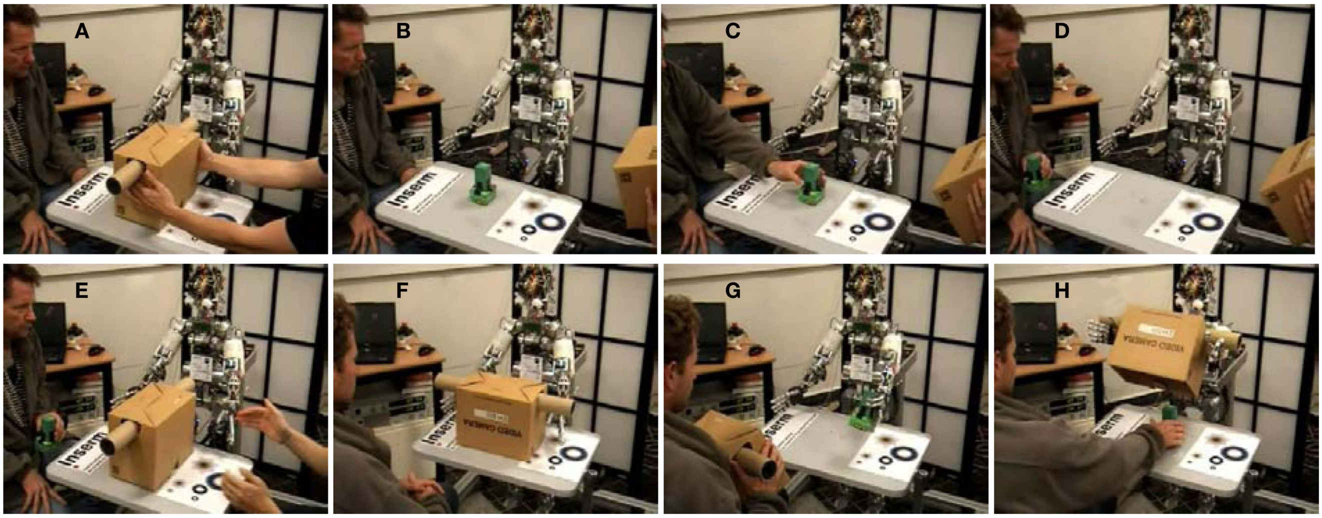


FIGURE 1 | On-line learning of a cooperative task. (A,B) Larry (left of robot) lifts the box that covers the toy. **(C,D)** This allows Robert (right of robot) to take the toy. **(E)** Larry replaces the box. **(F)** Robot now participates. **(G)** Human takes box, so Robot can take the toy. **(H)** Robot takes box so human can take the toy.

Interestingly, we quickly encountered limitations of the perceptual system, in the sense that when an action causes an object to be occluded, the visual disappearance of that object is quite different from the physical disappearance of the object, yet both result in a visual disappearance. The ability to keep track of objects when they are hidden during a perceived action, and the more general notion of object constancy is one of the signatures of core object cognition (Spelke, 1990; see Carey, 2009). This introduces the notion that human cognition is built around a limited set of "core systems" for representing objects, actions, number and space (Spelke and Kinzler, 2007). Robot cognition clearly provides a testing ground for debates in this domain, and the current study uses this platform to investigate the nature of the core system for agency. Embodied theories hold that actions are interpreted by mental simulation of the observed action, while teleological theories hold that this is not sufficient, and that a generative, rationality-based inferential process is also at work in action understanding (Gergely and Csibra, 2003). In our work, we employ both embodied learning of actions as well as a higher-level symbolic processing of these actions to yield a better understanding of the causes and consequences of events in the world. There are several research teams conducting very important and interesting work in scaling up from the primary perceptual layers (e.g., Fontanari et al., 2009; Tikhanoff et al., 2009). Our aim is to use the output of these layers in a more abstract and symbolic reasoning mainly driven by language, combining two approaches that are not antagonistic but rather complementary. This dual approach is consistent with Mandler's (2008) ideas of developmental concepts, as well as the role of amodal lexical associations in embodied language theories (e.g., Glaser, 1992; Kan et al., 2003), and several representational theories of meaning (Borghi and Cimatti, 2009; Dove, 2009).

As event understanding often involves inferences of links between intentions, actions, and outcomes, language can play an important role in helping children learn about relations between actions and their consequences (Bonawitz et al., 2009). The following section provides an overview of how language is used to enrich perceptual representations of action, and some of the corresponding neurophysiological mechanisms that provide some of these capabilities, based largely on data from humans. It is our belief that understanding these behavioral and neurophysiological mechanisms can provide strong guidelines in constructing a system for robot event cognition in the context of human–robot cooperation.

ASPECTS OF LANGUAGE AND CAUSALITY

One of the hallmarks of human cognition is the ability to understand goal-directed events. This ability surely entails the representation of events in terms of their causes and effects or goals (Bekkering et al., 2000; Sommerville and Woodward, 2005), but how does it work? Although some theorists have postulated that causality itself is a conceptual primitive, it has become evident that causality can be decomposed into constituent elements (see Carey, 2009 for discussion). According to physicalist models of causality, causes and effects are understood in terms of transfer or exchange of physical quantities in the world, such as energy, momentum, impact forces, chemical and electrical forces (Talmy, 1988; Wolff, 2007). Furthermore, nonphysical causation (e.g., forcing someone to decide) is understood by analogy to these physical forces. In this sense, physicalist models necessitate the ability to perceive kinematics, and dynamic forces, in order to represent causal relationships between entities. That is, to understand causality, one must have a body, and thus any implementation model of causal understanding necessitates an embodied system, to sense physical forces.

Dynamic forces are often invisible, such as the difference in the feeling of contact when an object is moving fast or slow, and how a pan feels when it is hot or cold. Because invisible dynamic forces map so well onto our experience of *kinematic* forces, or visual experience of forces (shape, size, position, direction, velocity, accelerations), humans often rely solely on visual information when attributing causal relationships in the world. In the same vein, causal understanding in non-human systems can be implemented through the use of kinematics as perceived via vision (e.g. Michotte, 1963). Thus, in our current work, we capitalize on this aspect of visual perception and restrict our representation of events to perceptual primitives that fall out of the visual input, leaving other perceptual modalities as well as motor actions for future implementation. Fern et al. (2002) and Siskind et al. (2001, 2003) have exploited the mapping of force dynamic properties into the visual domain, for primitives including contact, support and attachment. This results in robust systems in which event definitions are prespecified or learned, and then used for real-time event classification. Dominey and Boucher (2005) employed a related method for the recognition of events including give, take, push, touch in the context of grounded language acquisition.

In the context of development, once a toddler is able to sense and understand physical forces in the environment, he has the tools to understand causal relationships. Pioneering studies have shown that this understanding of causality and causal language is acquired very early in development, as infants may already perceive cause-effect relationships at only 27 weeks (Leslie and Keeble, 1987), and toddlers can already express many types of causal language by the age of 2–3 years (Bowerman, 1974; Hood et al., 1979). At this stage, exposure to language may help to accelerate the development of causal understanding. One study has shown that when toddlers are exposed to a causal relationship between two events accompanied by a causal description, they are more likely to initiate the first event to generate the second, and expect that the predictive relations will involve physical contact, compared to when they are exposed to the causal situation in the absence of causal language (Bonawitz et al., 2009). That is, though the toddler associates the two events in either case, this association might not be recognized as a causal link, and causal language, such as "the block makes the light turn on," can help to explicitly establish this link.

In this way, language is used as a tool to further conceptual understanding of goal-directed events and actions by helping toddlers more quickly integrate information about prediction, intervention, and contact causality. Thus, we can exploit language in our current system as a vector for establishing causal links between actions and their resulting states. In particular we are interested in the states that result from the "cover" and "give" actions which involve states related to the covered object being present, but invisible in the first case, and notions of change of possession in the second.

CORTICAL NETWORKS FOR LANGUAGE COMPREHENSION

In our effort to develop a system that can represent events and the state-transition relations between events, we can exploit knowledge of how language and event comprehension are implemented in the human nervous system. Language comprehension involves a cascade of computational operations starting from the decoding of speech in sensory areas to the emergence of embodied representations of the meaning of events corresponding to sensory-motor simulations (Barsalou, 1999; Bergen and Chang, 2005; Zwaan and Madden, 2005; see Rizzolatti and Fabbri-Destro, 2008 for review). These representations are triggered via: observation of others engaged in sensory-motor events; imagination of events and the evocation of these experiences through language. Therefore, we consider the existence of two parallel but interacting systems: one system for language processing, ultimately feeding information processes into a second system, dedicated to the processing of sensory-motor events. These systems are highly interconnected and their parallel and cooperative work can ultimately bootstrap meaning representations. The second system will also accommodate the representation of elaborated events that implicates processes derived from a system sometimes referred to as a "social perception" network (Decety and Grèzes, 2006; see Wible et al., 2009 for review). This second network is directly involved in teleological aspects of reasoning, including agency judgments, attributing goals and intentions to agents, inferring rationality about ongoing events and predicting outcomes of the ongoing simulation (Hauser and Wood, 2010). We will present these two systems and show how they interact to form complex meaning representations through language comprehension.

One central view in the recent models of the cortical processing of language is that it occurs along two main pathways, mostly lateralized to the left cortical hemisphere (Ullman, 2004; Hickok and Poeppel, 2007; see also Saur et al., 2008). The first route is referred to as the ventral-stream. It is dedicated to the recognition of complex auditory (or visual) objects involving different locations along the temporal lobe and the ventralmost part of the prefrontal cortex (BA 45/46). The second one is named the dorsal-stream and is dedicated to the connection between the language system and the sensory-motor system, that is both implicated in the transformation of phonetic codes into speech gestures for speech production, but also in the temporal and structural decoding of complex sentences (Hoen et al., 2006; Meltzer et al., in press). It implicates regions in the posterior part of the temporo-parietal junction (TPJ), parietal and premotor regions and reaches the dorsal part of the prefrontal cortex (BA 44).

In the ventral pathway, speech sounds are decoded in or nearby primary auditory regions of the dorsal superior temporal gyrus (BA 41/42), before phonological codes can be retrieved from the middle posterior superior temporal sulcus (mp-STS – BA 22), and words recognized in regions located in the posterior middle temporal gyrus (pMTG – BA 22/37; see Hickok and Poeppel, 2007 for review; Scott et al., 2006; Obleser et al., 2007). Then, these lexical symbols can trigger the reactivation of long-term stored sensory-motor experiences, either via implications of long-term autobiographic memory systems in the middle temporal gyrus or in long-term sensory-motor memories, with a widespread storage inside the sensory-motor system. Therefore, complex meaning representation can actually engage locations from the ventral pathway but also memories stored inside the dorsal pathway (Hauk et al., 2008; e.g., Tettamanti et al., 2005). This primary network feeds representation into a secondary-extended cortical network, whenever language leads to complex mental representations of complex events. Our initial computational models predicted dual structure-content pathway distinction (Dominey et al., 2003), which was subsequently confirmed in neuroimaging studies demonstrating the existence and functional implication of these two systems (Hoen et al., 2006), leading to further specification of the model (Dominey et al., 2006, 2009a).

TOWARD A NEUROPHYSIOLOGICAL MODEL OF EMBODIED AND TELEOLOGICAL EVENT COMPREHENSION

More recently, we extended this to a hybrid system in which sentence processing interacts both with a widespread embodied sensory-motor system, and with a more amodal system to account for complex event representation and scenario constructions operating on symbolic information (Madden et al., 2010). This second network, seems to engage bilateral parietal–prefrontal connections including bilateral activations in the parietal lobule for the perception and monitoring of event boundaries (Speer et al., 2007) as well as dorsal prefrontal regions seemingly implicated in the global coherence monitoring of the ongoing mental representation elaboration (Mason and Just, 2006). The monitoring of complex event representation includes the ability of deciding if ongoing linguistic information can be inserted in the current representation and how it modifies the global meaning of this representation. These aspects rely on information and knowledge that are not primary characteristics of the language system *per se* but rather include general knowledge about causal relations between events, intentionality and agency judgments etc. These properties are sometimes called teleological reasoning and different authors have now shown that processes involving teleological reasoning are sustained by a distributed neural network, referred to as a "social perception" cognitive network that is closely related to the language system (Wible et al., 2009).

This social perception network is implicated in teleological reasoning as determining agency or intentionality relations and involves regions as the right inferior parietal lobule (IP), the superior temporal sulcus (STS) and ventral premotor regions. All these regions are part of the well-known mirror system (Decety and Grèzes, 2006). The TPJ or IP and STS regions, in addition to being part of the mirror system, are also heavily involved in other social cognition functions. Decety and Grèzes (2006), in an extensive review, have designated the right TPJ as the "social" brain region. Theory of mind is the ability to attribute and represent other's mental states or beliefs and intentions or to "read their mind" ("predict the goal of the observed action and, thus, to "read" the intention of the acting individual" – from Decety and Grèzes, 2006, p. 6). Therefore, it seems that regions that are implicated in social-cognition, that is to say regions implicated in agency, intentionality judgments on others are also implicated in the same judgments on a simulation/representation of mental simulations triggered by language.

Figure 2 illustrates a summary representation of the cortical areas involved in the hybrid, embodied-teleological model of language and event processing. The language circuit involves the frontal language system including BA 44 and 45 with a link to embodied representations in the premotor areas, and in the more posterior parietal areas – both of which include mirror neuron

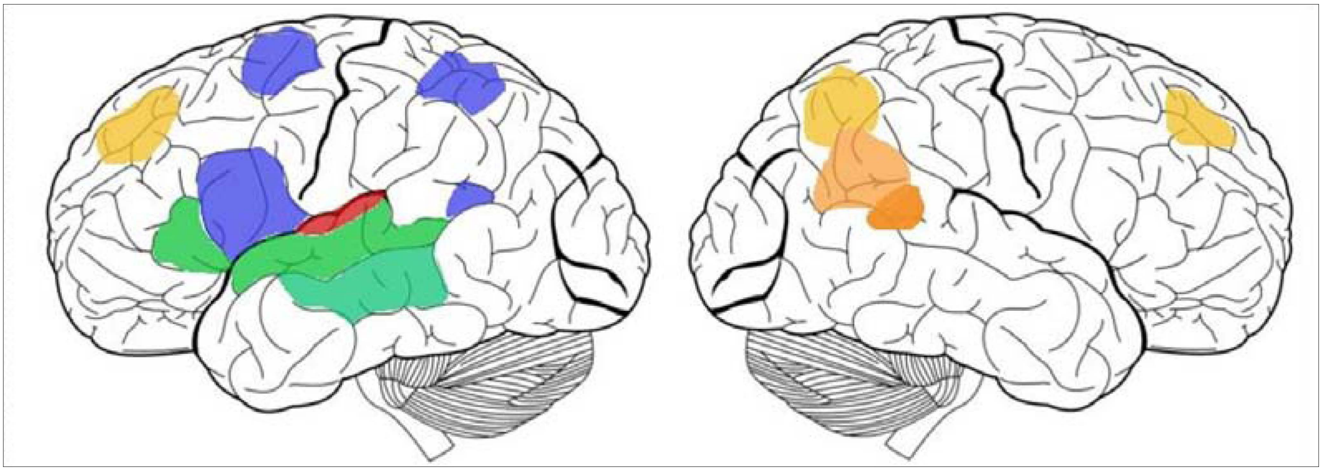


FIGURE 2 | Cortical networks for language processing (simplified). Ventral stream areas (green) are part of a first network dedicated to speech decoding and phonological/lexical processing along the superior temporal sulcus (STS), middle temporal gyrus (MTG) and ventral prefrontal cortex (Pfc). Dorsal stream areas (blue) constitute a sensory-motor interface implicated both in the transcription of phonological codes into articulatory codes (adapted from Hickok and Poeppel, 2007) but also in the temporal/structural organization of complex sentence comprehension, and engage the left temporo-parietal junction, the parietal lobule and dorsal prefrontal regions (Hoen et al., 2006; Meltzer et al., in press). The social perception or teleological cognition network (oranges) is implicated in complex event representation and the attribution of agency, theory of mind in the right TPG (orange, from Decety and Lamm, 2007), causality and intentionality in the posterior STS (dark orange, from Saxe et al., 2004; Brass et al., 2007), and also comprises areas implicated in the global monitoring of the coherence of event representation (light orange, from Mason and Just, 2006). Networks are shown in their specialized hemispheres but most contributions are bilateral.

activity in the context of action representation. This corresponds to the embodied component of the hybrid system. The teleological reasoning functions are implemented in a complimentary network that includes STS and TPJ/IP. In the current research, while we do not model this hybrid system directly in terms of neural networks, we directly incorporate this hybrid architecture into the cognitive system for the robot.

MATERIALS AND METHODS

This section will present in three parts the physical platform, the behavioral scenarios, and the system architecture.

THE iCub HUMANOID AND SYSTEM INFRASTRUCTURE

The current research is performed with the iCub, a humanoid robot developed as part of the RobotCub project (Tsagarakis et al., 2007). The iCub is approximately 1 m tall, roughly the size and shape of a 3-year-old child, and its kinematic structure has a total of 53 degrees of freedom controlled by electric motors, primarily located in the upper torso. The robot hands are extremely dexterous and allow manipulation of objects thanks to their 18 degrees of freedom in total. The robot head is equipped with cameras, microphones, gyroscopes and linear accelerometers. The iCub is illustrated in **Figures 1 and 4**.

Our research focuses on cognitive functions that operate on refined sensory data. We use off the shelf systems for both visual object and word recognition because they handle this raw sensory information quite well. Spoken language processing and overall system coordination is implemented in the CSLU Rad toolkit. The system is provided with an "innate" recognition vocabulary including a set of action names (give, take, touch, cover, uncover), derived predicates (on, has), object names (block, star, sign), and causal language connectives (if–then, because). That is, a list of words is given to the system in advance, so it will be able to recognize them from speech. It is possible to have the speech recognition behaviors emerge (e.g., Fontanari et al., 2009), but as mentioned above, that was not the goal of this work. The ability to recognize this innate vocabulary and use it in recognition grammars is provided by the CSLU toolkit which deals with HHM processing of the sound signal. The grammars that parse the speech signal both for input and output are hard coded into the system, however the system learns to associate a parsed sentence (verb, subject, object) with the visual perception of the corresponding action. Vision is provided by a template-matching system (Spikenet™) based on large spiking neurons networks, here again we use this tool to make a bridge between the raw sensory images, and the symbols of recognized objects. We developed state and action management in C#. Interprocess communication is realized via the YARP protocol.

EXPERIMENTAL SCENARIOS

In this section we describe the experimental human–robot interaction scenarios that define the functional requirements for the system. The current scenarios concentrate on action representation in the embodied and teleological frameworks. They demonstrate how language can be used (1) to enrich the representation of action and its consequences, and (2) to provide access to the structured representation of action definitions, and current knowledge of the robot. An embodied artificial system should incorporate both perceptual and motor representations in action comprehension, and current work is underway on this issue. However, in the current demonstrations we focus solely on perceptual (visual) representations of actions.

First we put the emphasis on the robot's ability to learn from the human when the human performs physical actions with a set of visible objects in the robot's field of view. Typical actions include covering (and uncovering) one object with another, putting one object next to another, and briefly touching one object with another. For actions that the robot has not seen before, the robot should ask the human to describe the action. The robot should learn the action description (e.g., "The block covered the star"), and be capable of generalizing this knowledge to examples of the same action performed on different objects. For learned actions, the robot should be able to report on what it has seen. This should take place in a real-time, on-line manner. Knowledge thus acquired should be available for future use, and in future work, the robot will also be able to learn its own motor representations of actions.

Another element that has to be learned is the causal relation between an action and the resulting state, which is not always trivial. When one object covers another, the second object "disappears" but is still physically present, beneath the covering object. In this scenario actions are performed that cause state changes, in terms of the appearance and disappearance of objects. The robot should detect these changes and attempt to determine their cause. The cause may be known, based on prior experience. If not, then the robot should ask the human, who will use speech for clarification about this causal relation.

The links between actions and their enabling and resulting states correspond directly to grammatical expressions with the if–then construction. The sentence "If you want to take the block then the block must be visible" expresses an enabling relation, where the state "block visible" enables the action "take the block." In contrast, the sentence "If you cover the star with the block, then the star is under the block," or "If you cover the star with the block then the star is not visible" expresses a causal relation. This scenario should demonstrate how by using these forms of grammatical constructions, we can interrogate the system related to these enabling and causal relations.

Once the robot has learned about new actions in one context, we want it to use this knowledge in another context. Concretely, in the cooperative task where Larry uncovers the toy so that Robot can pick it up, the robot should be able to begin to make the link between the resulting state of the "uncover" action as the enabling state of the subsequent "take" action. In this experiment, through a process of interrogation we will demonstrate that the robot has the knowledge necessary to form a plan for getting access to a covered object, by linking goals with resulting states of actions, and then establishing the enabling state as a new goal. After each learning session, the robot knowledge is stored in a long-term memory which we call Knowledge Base. It stores all the action definitions and their causes and consequences in term of states in an XML file that can be loaded on the robot.

We monitor the evolution of the Knowledge Base in order to analyze the performance of the recognition capabilities of the system under extended use. We start with a naïve system (i.e., an empty Knowledge Base), and then for the five actions *cover*, *uncover*, *give*, *take*, and *touch*, we expose the robot to each action with the block and the sign, and then in the transfer condition test the ability to recognize these actions with a new configuration (i.e., with the block and the star). We repeat this exhaustive exposure five times (one for each action). The dependant measure will be the number of presentations required for the five actions to be recognized in the training configuration, and transfer configuration, in each of the five phases. This experiment is detailed in Section "Usage Study" and in **Figure 5**.

COGNITIVE SYSTEM ARCHITECTURE

We developed a cognitive system architecture to respond to the requirements implied in Section "Experimental Scenarios," guided by knowledge of the cognitive linguistic mechanism in humans and their functional neurophysiology, and by our previous work in this area (See **Figure 3**). The resulting system is not neuro-mimetic, but its architecture is consistent with and inspired by our knowledge of the corresponding human system and on neural correlates found in the monkey (Fogassi et al., 2005). We describe the architecture in the context of processing a new action, and illustrated in **Figure 4**.

The human picks up the block and places it on the sign. Vision provides the front end of the perceptual system. Video data from the eyes of the iCub are processed by the Spikenet vision software which provides robust recognition for pretrained templates that recognize all objects in the scene. Each template is associated with a name and the camera coordinates of the recognized location. One to four templates were required per object.

Based on our previous work, inspired by human developmental studies, we identified three perceptual primitives to be extracted from the object recognition, which would form the basis for generic action recognition – these are *visible(object, true/false)*, *moving(object, true/false)*, and *contact(obj1,obj2, true/false)*. These primitives are easily extracted from the Spikenet output based on position and its first derivative, and are provided as input to Temporal Segmentation. The temporal segmentation function returns the most recent set of segmented primitives that occurred within specified time window. This corresponds to our hypothesis that a given complex action will be constituted by a pattern of primitives that occur in a limited time window, separated in time by periods with no action. The resulting pattern of primitives for contact is illustrated in **Figure 4C**.

When the robot detects changes in the visual scene, the above processing is initiated. The Action Management function matches the resulting segmented perceptual primitives with currently defined action in the Knowledge Base. Each action in the Knowledge Base is defined by its pattern of action primitives, its name, the arguments it takes, any preconditions (i.e., the enabling state S_E in the S_EAS_R representation), and the resulting state. Thus, during action recognition, the Action Management function compares this set of segmented primitives with existing action patterns in the Knowledge Base. If no match is found then the system prompts the human to specify the action and its arguments, e.g., "I cover the sign with the block."

The State Management determines that as a result of the action, the World State has changed, and interrogates the user about this. The user then has the opportunity to describe any new relations that result from this action but that are not directly perceptible. When the block covers the sign, the sign is no longer visible, but still present. The State Management asks "Why is the sign no longer visible?" Thus the human can explain this loss of vision by saying "Because the block is on the sign." The action manager binds this relation in a generic way (i.e., it generalizes to new objects when the event "cover" is perceived) to the definition of "cover" (see **Figure 4D**).

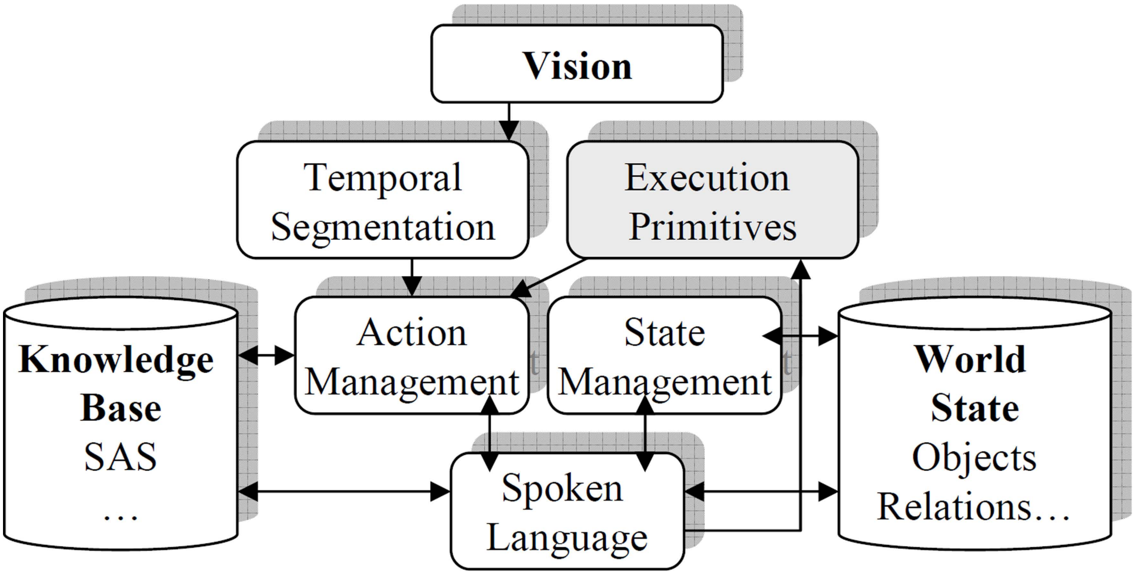


FIGURE 3 | Cognitive system architecture. See text for description.

If a match is found, then the system maps the concrete arguments in the current action segment with the abstract arguments in the action pattern. It can then describe what happened. For a recognized action, State Management updates the World State with any resulting states associated with that action. In the case of cover, this includes encoding of the derived predicate on (block, star).

RESULTS

LEARNING NEW ACTIONS AND THEIR DERIVED CONSEQUENCES

Here we present results from an interaction scenario in which the user teaches the robot four new actions: cover, uncover, give and take. In order to explain the system level functionality, details for learning are illustrated in **Figure 4** for the action "cover." The corresponding dialog is presented in **Table 1**.

For new actions (that have not yet been defined in the Knowledge Base) the system uses the set of observed primitives from Temporal Segmentation to generate a generic pattern of primitives to define the action (**Figure 4C**). If any unexpected perceptual changes occur, the system asks the human why this is the case, and the human can respond by describing any new relation that holds. For example, when the block covers the sign, the sign becomes not visible. The system asks the human why, and the human responds that this is "because the block is on the sign." This new relation on (block, sign) is added as part of the generic definition of the cover action, illustrated in **Figure 4D**.

Table 1 provides a record of the interaction in which the robot learns the meaning of "cover" and then displays this knowledge by recognizing cover in a new example. We observed that executing a given action like cover may sometimes lead to a different ordering of the segmented primitive events, e.g., detecting of the end of the block's movement may occur before or after the sign being visually obstructed. This is accommodated by encoding multiple patterns for a given action in the database. This redundant coding captures the physical redundancy that is expressed in the observations made by the system. The result is that when any of the appropriate patterns for an action are recognized, the action is recognized.

A total of five distinct actions were learned and validated in this manner. The resulting definitions are summarized in **Table 2**. **Figure 5** provides some performance statistics for learning these actions and then using the learned definitions to recognize new actions.

USE OF CAUSAL CONSTRUCTIONS TO INTERROGATE S_EAS_R REPRESENTATIONS

This experiment demonstrates how the "if–then" construction can be used to extract the link between actions, the required enabling states, and the resulting states. Results are presented in **Table 3**.

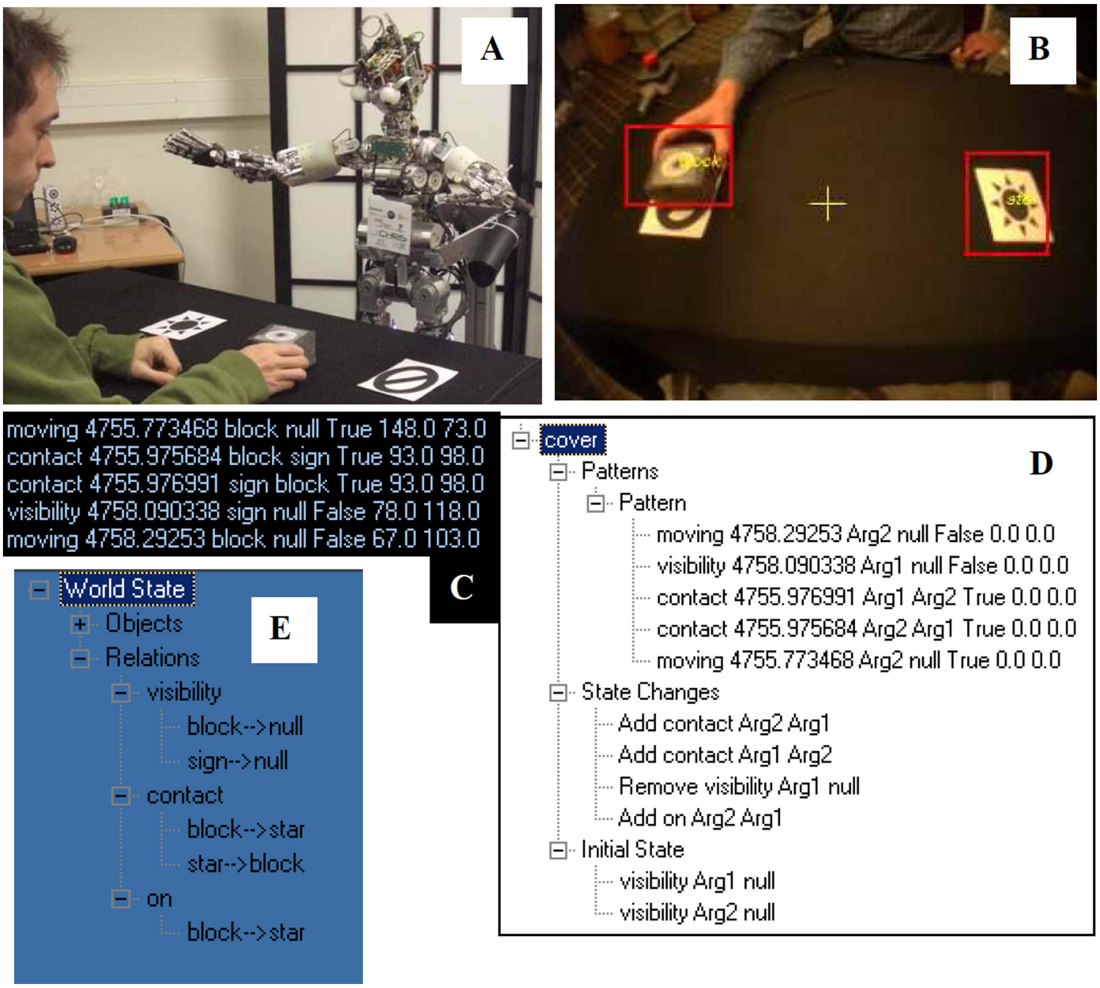
USE OF CAUSAL KNOWLEDGE IN TELEOLOGICAL REASONING

Here we consider a scenario similar to "uncover the block" scenario introduced in Section "Introduction – A Framework for Language and Action," and **Figure 1**. In this context, an object is covered by another, and the user's goal is to use the first object in a new task. The goal then is to find out how to gain access to the first object that is currently covered. The robot observes one human put the toy on the table, and another human cover the toy with the box. The objective is to begin to perform teleological reasoning about action sequences that have never been observed. Results are presented in Table 4.

This experiment demonstrates how the SAS (S_EAS_R) representation provides the required information for goal-based reasoning.

USAGE STUDY

We performed six additional experiments, which involved processing of 111 separate actions, to begin to evaluate the robustness of the system. Experiments 1–4 each started with an empty Knowledge Base, and examined the ability to learn the five actions, and then transfer this knowledge to new object configurations.



```
moving 4755.773468 block null True 148.0 73.0
contact 4755.975684 block sign True 93.0 98.0
contact 4755.976991 sign block True 93.0 98.0
visibility 4758.090338 sign null False 78.0 118.0
moving 4758.29253 block null False 67.0 103.0
```

```
⊟ cover
   ⊟ Patterns
      ⊟ Pattern
           moving 4758.29253 Arg2 null False 0.0 0.0
           visibility 4758.090338 Arg1 null False 0.0 0.0
           contact 4755.976991 Arg1 Arg2 True 0.0 0.0
           contact 4755.975684 Arg2 Arg1 True 0.0 0.0
           moving 4755.773468 Arg2 null True 0.0 0.0
   ⊟ State Changes
        Add contact Arg2 Arg1
        Add contact Arg1 Arg2
        Remove visibility Arg1 null
        Add on Arg2 Arg1
   ⊟ Initial State
        visibility Arg1 null
        visibility Arg2 null
```

```
⊟ World State
   ⊞ Objects
   ⊟ Relations
      ⊟ visibility
           block-->null
           sign-->null
      ⊟ contact
           block-->star
           star-->block
      ⊟ on
           block-->star
```

FIGURE 4 | Learning and generalizing "cover Arg1 with Arg2." (A) Robot setup and visual scene before the action. **(B)** Vision: Robot's view of scene after the block is put on the sign. **(C)** Temporal Segmentation: Time ordered sequence of perceptual events observed during the action. **(D)** Knowledge Base: Abstract pattern template for cover, along with the resulting state changes, and required initial state. **(E)** Relations encoded in the World State after the "cover" is recognized in a new action "covers the star with the block."

The key performance indices are (1) how many trials are required to learn an action with one set of objects, and (2) how well does this learning transfer to recognition of the same actions with different objects? Over the four experiments, a given action required 1.35 demonstrations to be learned accurately. This learning then transferred to new conditions on 70% of the new trials. Thirty percent of the trials required additional learning before the actions were recognized under new conditions. To quantify learning we performed an ANOVA on this data with three independent variables corresponding to Experiments (1–4), Experience (Naive, Experienced), and Condition (Training, Transfer). The dependant variable was the number of trials performed. The main observation was that once the system had learned the actions with one set of objects, it could transfer this knowledge to recognize the same actions with new object with little or no additional training. This was revealed by the significant interaction between Experience and Condition, independent of Experiment. Thus there is a significant effect of training on the ability to recognize new instances of learned action patterns ($p < 0.05$). Closer investigation revealed that in Experiment 2 the vision system was generating false movement recognition which lead to a number of irrelevant patterns being learned. When only Experiments 1, 3 and 4 are considered, an average of 1.13 trials are required for learning, and the knowledge transfers to 100% of the new trials with no additional learning.

In Experiments 5 and 6 we retained the Knowledge Base from Experiment 4, and then tested it with a new user, and examined the evolution over two complete tests with the five actions and the two object configurations. In Experiment 5, a total of six additional demonstrations were required to recognize the five actions in the two different object configurations. In Experiment 6 only one additional demonstration was required during the recognition of the 10 distinct actions. Overall these tests indicate that when the vision system is properly calibrated, the system is quite robust in the ability to learn generalized action recognition patterns.

Table 1 | Interaction fragment as robot learns meaning of "cover."
Completion time (seconds).

LEARNING "COVER" (0:42)

```
12:22:22 User : watch <user then places the block on the sign>
12:22:31 Robot : Need to Learn this one
12:22:58 User : cover the sign with the block
12:23:00 Robot : Did you say cover sign block ?
12:23:02 User : yes
12:23:04 Robot : Adding cover sign block
```

LEARNING RESULTING PREDICATE "ON" (0:17)

```
12:23:08 Robot : Why are the sign and the block close one each other ?
12:23:12 User : because the block is on the sign
12:23:15 Robot : Did you say because block on sign ?
12:23:17 User : yes
12:23:25 Robot : What should I do now ?
```

TRANSFER "COVER" TO NEW SITUATION (0:16)

```
12:24:07 Robot : What should I do now ?
12:24:09 User : watch
12:24:21 Robot : Recognized cover star block
12:24:23 Robot : What should I do now ?
```

USER CAN INTERROGATE THE WORLD STATE:

```
12:24:26 User : Tell about star
12:24:28 Robot : I will query info about the star
12:24:30 Robot : star contact block
12:24:33 Robot : block on star
```

Table 2 | Enabling State – Action – Resulting State (SEASR) representation for the learned actions.

Enabling state	Action	Resulting state
Visible Arg1 Visible Arg2	Cover Arg1 with Arg2	Contact Arg1 Arg2 Visible Arg2 On Arg2 Arg1
Visible Arg2 True On Arg2 Arg1	Uncover Arg1 with Arg2	Visible Arg1 Visible Arg2
Visible Arg1 Visible Arg2	Give Arg1 to Arg2	Contact Arg1 Arg2 Visible Arg1 Visible Arg2 Has Arg2 Arg1
Contact Arg1 Arg2 Visible Arg1 Visible Arg2 On Arg2 Arg1	Take Arg1 from Arg2	Visible Arg1 Visible Arg2

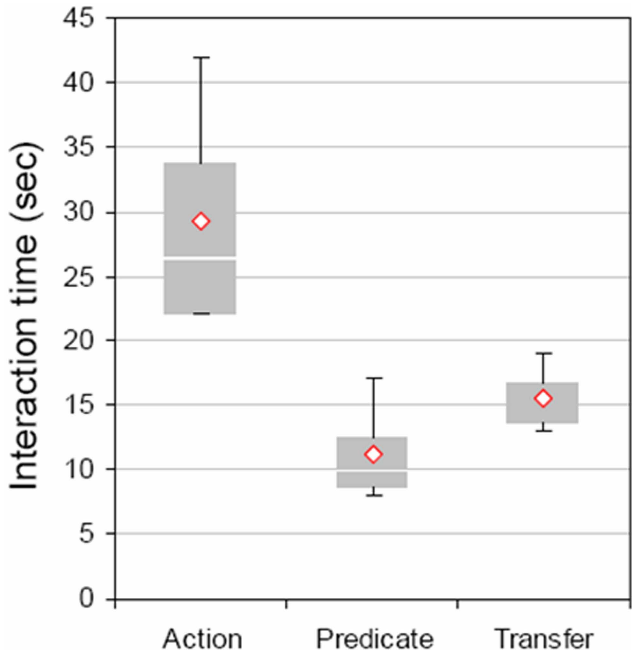


FIGURE 5 | Performance effects for learning and transfer. Interaction times were recorded for the learning of four novel actions, learning the derived predicates for those actions, and then using the resulting definitions to recognize new instances of the five actions (Action, Predicate, and Transfer, respectively). We compared interaction time for these three conditions in a single factor ANOVA. Recognition in transfer to new actions (15.5 s) was significantly faster than initial action learning time (29.25), $p < 0.05$.

Table 3 | Interaction fragment as robot explains the preconditions and results of the cover action.

```
13:43:10 User : What are the conditions for cover the sign with the block?
13:43:12 Robot : If you want to cover the block with the sign then
13:43:14 Robot : sign need to be visible
13:43:16 Robot : block need to be visible
13:43:23 User : What are the consequences of cover the sign with the block?
13:43:26 Robot : If you cover the sign with the block then
13:43:28 Robot : block will be contact with sign
13:43:30 Robot : sign will be contact with block
13:43:32 Robot : sign will no longer be visible
13:43:34 Robot : block will be on sign
```

THE BLOCK IS CURRENTLY ON THE SIGN.

```
13:43:40 User : How can I make the sign visible?
13:43:43 Robot : If you want to make the sign visible then
13:43:45 Robot : You can uncover the sign
```

DISCUSSION

Part of the stated objective of this work has been to implement, and demonstrate the advantages of, a hybrid embodied-teleological approach to action–language interaction, both from a theoretical perspective and via results from human–robot interaction experiments with the iCub robot. This objective was motivated by our observation that true cooperation requires not only that the robot can learn shared action sequences, but that it represents how those actions are linked in a chain of state changes that lead to the goal. This means that the robot must be able to represent actions in terms of the states that allow them to be performed, the states that result from their performance including the "unseen" predicates, for example, related to object permanence.

We developed a perceptual system that extracts patterns of spatio-temporal visual properties in order to encode actions in terms of these patterns. We re-discovered that action and meaning are not purely perceptual (Carey, 2009), and that additional properties related to object permanence and physical possession also form part of the meaning of action. Based on studies indicating that language can be used by toddlers to accelerate the acquisition of such knowledge (Bonawitz et al., 2009), when our cognitive system encounters unexpected results from an action, it interrogates the user, much like a developing child (Hood et al., 1979). This allows the user to explain, for example, that when the block covers the star, the star is not visible (but still there) *because* the block is *on* the star. We refer to these additional predicates (*on*, *has*) as derived predicates. This demonstrates that language can play an essential role in refining the representation of the meaning of action which is first approximated purely from the perceptual stream, by introducing derived predicates that become part of the meaning of the action. These predicates are encoded in the state changes that are to be introduced whenever the action is recognized. Thus, when the *give* and *take* actions are recognized, the derived predicate *has* (indicating possession) will be appropriately updated.

We believe that this is a fundamental development in the link between language and action, because it goes beyond a pure identity mapping between sentences and meaning, and instead uses language to change and enrich forever the meaning of action as part of a developmental/learning process. In this way, the use of language by the iCub to transfer knowledge to new trials is similar to the causal learning of toddlers observed by Bonawitz et al. (2009), as in both cases language is a symbolic processing tool for memory and cognition. This is consistent with theories of language in which words are not only considered as markers for referents in the world, but also as tools that allow us to reason and operate in the world (Borghi and Cimatti, 2009; Mirolli and Parisi, in press) as well as current ideas of how language evolved in humans through sensory-motor and social interaction, as well as possible simulations of these ideas in artificial systems (see Parisi, 2006; Parisi and Mirolli, 2006). These theories explain that language is used not only within the individual for reasoning and memory but also within a broader social network for communicative purposes. Therefore, our ongoing research on cooperative action (Dominey et al., 2009b; Dominey and Warneken, in press) is an important step in better understanding how language acts as a tool to facilitate goal-directed action between two or more agents.

A crucial component of the new system is the representation of actions which includes the link to initial enabling states, and final resulting states. The resulting system produces a Knowledge Base that encodes the representation of action meanings, and a World State that encodes the current state of the world. As mentioned above, we demonstrate how grammatical constructions that exploit causal connectives (e.g., because) can allow spoken language to enrich the learned set of SAS representations, by inserting derived predicates into the action definition. We also demonstrated how the causal connective "if–then" can be employed by the robot to inform the user about the links between enabling states and actions, and between actions and resulting states. Again, this extends the language–action interface beyond veridical action descriptions (or commands) to transmit more subtle knowledge about enabling and resulting states of actions, how to reach goals etc.

Indeed, in the context of the "hybrid" embodied and teleological system, we demonstrated how representations of enabling and resulting states provides the system with the knowledge necessary to make the link between goals as the resulting states of actions, and the intervening actions that are required. This is part of the basis of a teleological reasoning capability (Csibra, 2003). In the current system, we have not implemented a full blown reasoning capability, that can perform forward and backward chaining on the states and action representations. This is part of our ongoing research.

In Foundations of Language, Jackendoff (2002) indicates that while languages may vary in their surface structure, the organization of the conceptual structure that they express appears more universal (see Fadiga et al., 2009 for a review of common syntactic structure in language, action and music). We extended this notion to consider that indeed, the compositional structure of syntax is derived from that of the conceptual system (Dominey, 2003), and Jackendoff (2003) agreed. In this context, one of the most promising results of the current research is the continued observation that language reflects the structure of conceptual representations. We have previously demonstrated this in situations where multiple actions are linked by shared states, resulting in descriptions such as "Larry took the toy that Robert uncovered with the box" (Dominey and Boucher, 2005). The current work extends this to include functional and causal links between elements in the SAS representations (e.g., the if–then constructions in **Tables 3 and 4**).

The experiments described here have focused on perception of events. We realize that action is crucial to event understanding, and we are currently working to integrate this S_EAS_R framework into our existing cooperative action framework (Dominey et al., 2009b; Dominey and Warneken, in press). We will first demonstrate that the mechanism presented here for learning

Table 4 | Interaction fragment where the robot performs basic goal-based reasoning about how to solve the task of getting the star when it is covered by the block, and then uses it to cover the sign.

```
18:03:50 Robot : Recognized cover star with block
18:04:05 User : What are the conditions to cover the sign with star
18:04:07 Robot : If you want to cover the sign with the star then
18:04:09 Robot : sign needs to be visible
18:04:11 Robot : star needs to be visible
The sign is visible, but the star is covered with the block, so the user asks the robot how to make the sign visible:
18:04:18 User : How can I see the star
18:04:23 Robot : You can uncover the star
The user can now uncover the star by removing the block, and then achieve the final goal which is to cover the sign with the star
```

the perceptual patterns associated with perceived actions can be applied to learning motor patterns associated with executed actions. This will result in further enriched action representations that include the enabling and resulting states, the perceptual primitive patterns, and the action primitive patterns. We will then use these representations in the context of learning cooperative tasks by observation. This will yield a situation in which the robot can represent the trajectory from initial state to final goal state via coordinated action sequence, and will thus provide the basis for intentional reasoning, and the extension of the teleological reasoning to cooperative activity.

ACKNOWLEDGMENTS

This research is supported by the European Commission FP7 ICT Projects CHRIS and Organic, and French ANR Projects Amorces and Comprendre.

REFERENCES

Barsalou, L. W. (1999). Perceptual symbol systems. *Behav. Brain Sci.* 22, 577–660.

Barsalou, L. W., Santos, A., Simmons, W. K., and Wilson, C. D. (2008). "Language and simulation in conceptual processing," in *Symbols, Embodiment, and Meaning*, eds M. De Vega, A. M. Glenberg and A. C. Graesser (Oxford, UK: Oxford University Press), 245–284.

Bekkering, H., Wohlschlager, A., and Gattis, M. (2000). Imitation of gestures in children is goal-directed. *Q. J. Exp. Psychol.* 53, 153–164.

Bergen, B., and Chang, N. (2005). "Embodied construction grammar in simulation-based language understanding," in *Construction Grammar(s): Cognitive Grounding and Theoretical Extensions*, eds J.-O. Östman and M. Fried (Amsterdam: John Benjamins), 147–156.

Bonawitz, E. B., Horowitz, A., Ferranti, D., and Schulz, L. (2009). "The block makes it go: causal language helps toddlers integrate prediction, action, and expectations about contact relations," in *Proceedings of the Thirty-first Cognitive Science Society*, eds N. Taatgen and H. van Rijn (Amsterdam: Cognitive Science Society), 81–86.

Borghi, A. M., and Cimatti, F. (2009). "Words as tools and the problem of abstract word meanings," in *Proceedings of the Thirty-first Cognitive Science Society*, eds N. Taatgen and H. van Rijn (Amsterdam: Cognitive Science Society).

Bowerman, M. (1974). Learning the structure of causative verbs: a study in the relationship of cognitive, semantic, and syntactic development. *Proc. Res. Child Lang. Dev.* 8, 142–178.

Brass, M., Schmitt, R. M., Spengler, S., and Gergely, G. (2007). Investigating action understanding: inferential processes versus action simulation. *Curr. Biol.* 17, 2117–2121.

Carey, S. (2009). *The Origin of Concepts*. New York: Oxford University Press.

Csibra, G. (2003). Teleological and referential understanding of action in infancy. *Phil. Trans. R. Soc. Lond. B.* 358, 447–458.

Decety, J., and Grèzes, J. (2006). The power of simulation: imagining one's own and other's behavior. *Brain Res.* 1079, 4–14.

Decety, J., and Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition. *Neuroscientist* 13, 580–593.

Dominey, P. F. (2003). A conceptuocentric shift in the characterization of language. *Comment on Jackendoff, BBS*, 674–675.

Dominey, P. F., and Boucher, J. D. (2005). Learning to talk about events from narrated video in a construction grammar framework. *Artif. Intell.* 167, 31–61.

Dominey, P. F., Hoen, M., Blanc, J. M., and Lelekov-Boissard, T. (2003). Neurological basis of language and sequential cognition: evidence from simulation, aphasia, and ERP studies. *Brain Lang.* 86, 207–225.

Dominey, P. F., Inui, T., and Hoen, M. (2009a). Neural network processing of natural language: II. Towards a unified model of corticostriatal function in learning sentence comprehension and non-linguistic sequencing. *Brain Lang.* 109, 80–92.

Dominey, P. F., Mallet, A., and Yoshida, E. (2009b). Real-time spoken-language programming for cooperative interaction with a humanoid apprentice. *Int. J. HR* 6, 147–171.

Dominey, P. F., and Warneken, F. (in press). The origin of shared intentions in human–robot cooperation. *New Ideas Psychol.*

Dove, G. (2009). Beyond conceptual symbols. A call for representational pluralism. *Cognition* 110, 412–431.

Fadiga, L., Craighero, L., and D'Ausilio, A. (2009). Broca's area in language, action, and music. The neurosciences and music III – disorders and plasticity. *Ann. N. Y. Acad. Sci.* 1169, 448–458.

Fern, A., Givan, R., and Siskind, J. M. (2002). Specific-to-general learning for temporal events with application to learning event definitions from video. *J. Artif. Intell. Res.* 17, 379–449.

Fischer, M. H., and Zwaan, R. A. (2008). Embodied language: a review of the role of the motor system in language comprehension. *Q. J. Exp. Psychol.* 61, 825–850.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., and Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. *Science* 308, 662–667.

Fontanari, J. F., Tikhanoff, V., Cangelosi, A., Ilin, R., and Perlovsky, L. I. (2009). Cross-situational learning of object–word mapping using Neural Modeling Fields. *Neural Netw.* 22, 579–585.

Gergely, G., and Csibra, G. (2003). Teleological reasoning in infancy: the naïve theory of rational action. *Trends Cogn. Sci.* 7, 287–292.

Glaser, W. R. (1992). Picture naming. *Cognition* 42, 61–105.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Hauk, O., Shtyrov, Y., and Pulvermüller, F. (2008). The time course of action and action-word comprehension in the human brain as revealed by neurophysiology. *J. Physiol. Paris* 102, 50–58.

Hauser, M., and Wood, J. (2010). Evolving the capacity to understand actions, intentions, and goals. *Annu. Rev. Psychol.* 61, 303–324.

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393–402.

Hoen, M., Pachot-Clouard, M., Segebarth, C., and Dominey, P. F. (2006). When Broca experiences the Janus syndrome: an ER-fMRI study comparing sentence comprehension and cognitive sequence processing. *Cortex* 42, 605–623.

Hood, L., Bloom, L., and Brainerd, C. J. (1979). What, when, and how about why: a longitudinal study of early expressions of causality. *Monogr. Soc. Res. Child Dev.* 44, 1–47.

Jackendoff, R. (2002). Foundations of Language: brain, Meaning, Grammar, Evolution. Oxford University Press.

Jackendoff, R. (2003). Précis of Foundations of Language: brain, Meaning, Grammar, Evolution, BBS 26, 651–707.

Kan, I. P., Barsalou, L. W., Solomon, K. O., Minor, J. K., and Thompson-Schill, S. L. (2003). Role of mental imagery in a property verification task: fMRI evidence for perceptual representations of conceptual knowledge. *Cogn. Neuropsychol.* 20, 525–540.

Kotovsky, L., and Baillargeon, R. (1998). The development of calibration-based reasoning about collision events in young infants. *Cognition* 67, 311–351.

Lallee, S., Warkeken, F., and Dominey, P. F. (2009). "Learning to collaborate by observation and spoken language," in *Ninth International Conference on Epigenetic Robotics*, Venice.

Leslie, A. M., and Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition* 25, 265–288.

Madden, C. J., Hoen, M., and Dominey, P. F. (2010). A cognitive neuroscience perspective on embodied language for human–robot cooperation. *Brain Lang.* 112, 180–188.

Mandler, J. M. (2008). On the birth and growth of concepts. *Philos. Psychol.* 21, 207–230.

Mason, R. A., and Just, M. A. (2006). "Neuroimaging contributions to the understanding of discourse processes," in *Handbook of Psycholinguistics*, eds M. Traxler and M. A. Gernsbacher (Amsterdam: Elsevier), 765–799.

Meltzer, J. A., McArdle, J. J., Schafer, R. J., and Braun, A. R. (in press). Neural aspects of sentence comprehension: syntactic complexity, reversibility, and reanalysis. *Cereb. Cortex.*

Michotte, A. (1963). *The Perception of Causality*. New York: Basic Books.

Mirolli, M., and Parisi, D. (in press). Towards a Vygotskyan cognitive robotics: the role of language as a cognitive tool. *New Ideas Psychol.*

Obleser, J., Zimmermann, J., Van Meter, J., and Rauschecker, J. P. (2007). Multiple stages of auditory speech perception reflected in event-related FMRI. *Cereb. Cortex* 17, 2251–2257.

Parisi, D. (2006). "Simulating the evolutionary emergence of language: a research agenda," in *The Evolution of Language*, eds A. Cangelosi, A. D. M.

Smith and K. Smith (Singapore: World Scientific), 230–238.

Parisi D., and Mirolli M. (2006). "The emergence of language: how to simulate it," in *Emergence of Communication and Language*, eds C. Lyon, C. Nehaniv and A. Cangelosi (London: Springer), 269–285.

Pulvermüller, F., Shtyrov, Y., and Hauk, O. (2009). Understanding in an instant: neurophysiological evidence for mechanistic language circuits in the brain. *Brain Lang.* 110, 81–94.

Rizzolatti, G., and Fabbri-Destro, M. (2008). The mirror system and its role in social cognition. *Curr. Opin. Neurobiol.* 18, 179–184.

Saur, D., Kreher, B. W., and Schnell, S., et. (2008). Ventral and dorsal pathways for language. *Proc. Natl. Acad. Sci. U. S. A.* 105, 18035–18040.

Saxe, R., Xiao, D. K., Kovacs, G., Perrett, D. I., and Kanwisher, N. (2004). A region of right posterior superior temporal sulcus responds to observed intentional actions. *Neuropsychologia* 42, 1435–1446.

Scott, S. K., Rosen, S., Lang, H., and Wise, R. J. (2006). Neural correlates of intelligibility in speech investigated with noise vocoded speech – a positron emission tomography study. *J. Acoust. Soc. Am.* 120, 1075–1083.

Siskind, J. M. (2001). Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *J. Artif. Intell. Res.* 15, 31–90.

Siskind, J. M. (2003). Reconstructing force-dynamic models from video sequences. *Artif. Intell.* 151, 91–154.

Sommerville, A., and Woodward, A. L. (2005). Pulling out the intentional structure of action: the relation between action processing and action production in infancy. *Cognition* 95, 1–30.

Speer, N. K., Zacks, J. M., and Reynolds, J. R. (2007). Human brain activity time-locked to narrative event boundaries. *Psychol. Sci.* 18, 449–455.

Spelke, E. S. (1990). Principles of object perception. *Cogn. Sci.* 14, 29–56.

Spelke, E. S., and Kinzler, K. D. (2007). Core knowledge. *Dev. Sci.* 10, 89–96.

Talmy, L. (1988). Force dynamics in language and cognition. *Cogn. Sci.* 12, 49–100.

Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., Fazio, F., Rizzolatti, G., Cappa, S. F., and Perani, D. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *J. Cogn. Neurosci.* 17, 273–281.

Tomasello, M. (2008). *Origins of Human Communication*. Cambridge: MIT Press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: the origins of cultural cognition. *Behav. Brain Sci.* 28, 675–691.

Tikhanoff, V., Cangelosi, A., and Metta, G. (2009). "Language understanding in humanoid robots: simulation experiments with iCub platform", in *Proceedings of 2009 International Conference on Integration of Knowledge Intensive Multi-Agent Systems (KIMAS'09)*, St. Louis, MA.

Tsagarakis, N. G., Metta, G., Sandini, G., Vernon, D., Beira, R., Becchi, F., Righetti, L., Victor, J. S., Ijspeert, A. J., Carrozza, M. C., and Caldwell, D. G. (2007). iCub – the design and realization of an open humanoid platform for cognitive and neuroscience research. *Adv. Robot.* 21, 1151–1175.

Ullman, M. T. (2004). Contributions of memory circuits to language: the declarative/procedural model. *Cognition* 92, 231–270.

Wible, C. G., Preus, A. P., and Hashimoto, R. (2009). A cognitive neuroscience view of schizophrenic symptoms: abnormal activation of a system for social perception and communication. *Brain Imaging Behav.* 3, 85–110.

Wolff, P. (2007). Representing causation. *J. Exp. Psychol. Gen.* 136, 82–111.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition* 69, 1–34.

Zwaan, R. A., and Madden, C. J. (2005). "Embodied sentence comprehension," in *Grounding cognition: The Role of Perception and Action in Memory, Language, and Thinking*, eds D. Pecher and R. A. Zwaan (Cambridge, UK: Cambridge University Press), 224–245.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 December 2009; paper pending published: 20 February 2010; accepted: 11 May 2010; published online: 03 June 2010.

*Citation: Lallee S, Madden C, Hoen M and Dominey PF (2010) Linking language with embodied and teleological representations of action for humanoid cognition. Front. Neurorobot. **4**:8. doi: 10.3389/fnbot.2010.00008*

frontiers in
NEUROROBOTICS

ORIGINAL RESEARCH ARTICLE
published: 21 May 2010
doi: 10.3389/fnbot.2010.00005

Integrating verbal and nonverbal communication in a dynamic neural field architecture for human–robot interaction

Estela Bicho[1], *Luís Louro*[1] *and Wolfram Erlhagen*[2]*

[1] Department of Industrial Electronics, University of Minho, Guimarães, Portugal
[2] Department of Mathematics and Applications, University of Minho, Guimarães, Portugal

Edited by:
Angelo Cangelosi, University of Plymouth, UK

Reviewed by:
Domenico Parisi, The National Research Council, Italy

***Correspondence:**
Wolfram Erlhagen, Department of Mathematics and Applications, University of Minho, 4800-058 Guimaraes, Portugal.
e-mail: wolfram.erlhagen@mct.uminho.pt

How do humans coordinate their intentions, goals and motor behaviors when performing joint action tasks? Recent experimental evidence suggests that resonance processes in the observer's motor system are crucially involved in our ability to understand actions of others', to infer their goals and even to comprehend their action-related language. In this paper, we present a control architecture for human–robot collaboration that exploits this close perception-action linkage as a means to achieve more natural and efficient communication grounded in sensorimotor experiences. The architecture is formalized by a coupled system of dynamic neural fields representing a distributed network of neural populations that encode in their activation patterns goals, actions and shared task knowledge. We validate the verbal and nonverbal communication skills of the robot in a joint assembly task in which the human–robot team has to construct toy objects from their components. The experiments focus on the robot's capacity to anticipate the user's needs and to detect and communicate unexpected events that may occur during joint task execution.

Keywords: joint action, neural fields, goal inference, natural communication, mirror system

INTRODUCTION

New generations of robotic systems are starting to share the same workspace with humans. They are supposed to play a beneficial role in the life of ordinary people by directly collaborating with them on common tasks. The role as co-worker and assistant in human environments leads to new challenges in the design process of robot behaviors (Fong et al., 2003). In order to guarantee user acceptance, the robot should be endowed with social and cognitive skills that makes the communication and interaction with the robot natural and efficient. Humans are experts in coordinating their actions with others to reach a shared goal (Sebanz et al., 2006). In collaborative tasks we continuously monitor the actions of our partners, interpret them effortlessly in terms of their outcomes and use these predictions to select an adequate complementary behavior. Think for instance about two people assembling a piece of furniture from its components. One person reaches toward a screw. The co-actor immediately grasps a screw-driver to hand it over and subsequently holds the components that are to be attached with the screw. In familiar tasks, such fluent team performance is very often achieved with little or no direct communication. Humans are very good in combining motion and contextual information to anticipate the ultimate goal of others' actions (Sebanz et al., 2006). Referring to objects or events through the use of language and communicative gestures is essential, however, whenever the observed behavior is ambiguous or a conflict in the alignment of intentions between partners has been detected. Ideally, not only the fact that something might go wrong in the joint action but also the reason for the conflict should be communicated to the co-actor.

The last decade has seen enormous progress in designing human-centered robots that are able to perceive, understand and use different modalities like speech, communicative gestures, facial expressions and/or eye gaze for more natural interactions with human users (for a recent overview see Schaal, 2007). Different control architectures for multi-modal communication have been proposed that address specific research topics in the domain of human-centered robotics. It has been shown for instance that integrating multiple information channels supports a more intuitive teaching within the learning by demonstration framework (McGuire et al., 2002; Steil et al., 2004; Pardowitz et al., 2007; Calinon and Billard, 2008), allows the robot to establish and maintain a face-to-face interaction in crowded environments (Spexard et al., 2007; Koenig et al., 2008), or can be exploited to guarantee a more intelligent and robust robot behavior in cooperative human–robot tasks (Breazeal et al., 2004; Alami et al., 2005; Foster et al., 2008; Gast et al., 2009). Although the proposed multi-modal architectures differ significantly in the type of control scheme applied (e.g., hybrid or deliberative) and theoretical frameworks used (e.g., neural networks, graphical or probabilistic models) they also have an important aspect in common. Typically, the integration of verbal and nonverbal information and the coordination of actions and decisions between robot and human are performed in dedicated fusion and planning modules that do not contain sensorimotor representations for the control of the robot actuators. A representative example are control architectures for HRI based on the theoretical framework of joint intention theory (e.g., Breazeal et al., 2004; Alami et al., 2005) that has been originally proposed for cooperative problem solving in distributed artificial intelligence systems (Cohen and Levesque, 1990). In these architectures a joint intention interpreter and a reasoner about beliefs and communicative acts can feed a central executive that is responsible for joint action planning and coordination on a symbolic level. A different approach to more natural and efficient HRI followed by our and other groups is inspired by fundamental findings in behavioral and neurophysiological experiments analyzing perception and action in a social context (Wermter et al., 2004; Erlhagen et al., 2006b; Bicho et al., 2009; Breazeal et al., 2009). These findings suggest that automatic resonance processes in the observer's motor system are crucially involved in the ability to recognize and understand actions and communicative acts of others', to infer their goals and even to comprehend their action-related utterances. The basic idea is that people gain an embodied understanding of the observed person's behavior by internally simulating action consequences through the covert use of their own action repertoire (Barsalou et al., 2003). In joint action, the predicted sensory consequences of observed actions together with prior task knowledge may then directly drive the motor representation of an adequate complementary behavior. Such shared representations for perception, action and language are believed to constitute a neural substrate for the remarkable fluency of human joint action in familiar tasks (Sebanz et al., 2006).

Many of the experiments on action observation were inspired by the discovery of mirror neurons (MNs) first in premotor cortex and later in the parietal cortex of macaque monkey (di Pellegrino et al., 1992, for a review see Rizzolatti and Craighero, 2004). Mirror neurons fire both when the monkey executes an object-directed motor act like grasping and when it observes or hears a similar motor act performed by another individual. They constitute a neural substrate of an abstract concept of grasping, holding or placing that generalizes over agents and the modality of action-related sensory input. Many MNs require the observation of exactly the same action that they encode motorically in order to be triggered. The majority of MNs however falls in the broadly congruent category for which the match between observed and executed actions is not strict (e.g., independent of the kinematic parameters or the effector). Important for HRI, broadly congruent MNs may support an action understanding capacity across agents with very different embodiment and motor skills like human and robot. The fact that the full vision of an action is not necessary for eliciting a MN response whenever additional contextual cues may explain the meaning of the action has been interpreted as evidence for the important role of MNs in action understanding. It has been shown for instance that grasping MNs respond to a hand disappearing behind a screen when the monkey knew that there is an object behind the occluding surface (Umiltà et al., 2001). A grasping behavior is normally executed with an ultimate goal in mind. By training monkeys to perform different action sequences Fogassi et al. (2005) have recently tested whether MNs are not only involved in the coding of a proximate goal (the grasping) but also in the coding of the ultimate goal or motor intention (what to do with the object). The fundamental finding was that specific neural populations represent the identical grasping act in dependence of the outcome of the whole action sequence in which the grasping is embedded (e.g., grasping for placing versus grasping for eating). This finding has been interpreted as supporting the hypothesis that neural representations of motor primitives are organized in chains (e.g., reaching–grasping–placing) generating specific perceptual outcomes (Chersi et al., 2007, see also Erlhagen et al., 2007). On this view, the activation of a particular chain during action observation is a means to anticipate the associated outcomes of others' actions.

More recently, brain imaging studies of joint action revealed compelling evidence that the mirror system is also crucially involved in complementary action selection. People performing identical or complementary motor behaviors as those they had observed showed a stronger activation of the human mirror system in the complementary condition compared to the condition when the participants imitated the observed action (Newman-Norlund et al., 2007). This finding can be explained if one assumes a central role of the mirror system in linking two different but logically related actions that together constitute a goal-directed sequence involving two actors (e.g. receiving an object from a co-actor).

It has been suggested that the abstract semantic equivalence of actions encoded by MNs is related to aspects of linguistic communication (Rizzolatti and Arbib, 1998). Although the exact role of the mirror mechanism for the evolution of a full-blown syntax and computational semantics is still matter of debate (Arbib, 2005), there is now ample experimental evidence for motor resonance during verbal descriptions of actions. Language studies have shown that action words or action sentences automatically activate corresponding action representations in the motor system of the listener (Hauk et al., 2004; Aziz-Zadeh et al., 2006; Zwann and Taylor, 2006). Following the general idea of embodied simulation (Barsalou et al., 2003) this suggests that the comprehension of speech acts related to object-directed actions does not involve abstract mental representations but rather the activation of memorized sensorimotor experiences. The association between a grasping behavior or a communicative gesture like pointing and an arbitrary linguistic symbol may be learned when during practice the utterance and the matching hand movement occur correlated in time (Billard, 2002; Cangelosi, 2004; Sugita and Tani, 2005).

In this paper we present and validate a dynamic control architecture that exploits the idea of a close perception–action linkage as a means to endow a robot with nonverbal and verbal communication skills for natural and efficient HRI. Ultimately, the architecture implements a flexible mapping from an observed or simulated action of the co-actor onto a to-be-executed complementary behavior which consist of speech output and/or a goal-directed action. The mapping takes into account the inferred goal of the partner, shared task knowledge and contextual cues. In addition, an action monitoring system may detect a mismatch between predicted and perceived action outcomes. Its direct link to the motor representations of complementary behaviors guarantees the alignment of actions and decisions between the co-actors also in trials in which the human shows unexpected behavior.

The architecture is formalized by a coupled system of dynamic neural fields (DNFs) representing a distributed network of local neural populations that encode in their activation patterns task-relevant information (Erlhagen and Bicho, 2006). Due to strong recurrent interactions within the local populations the patterns may become self-stabilized. Such attractor states of the field dynamics allow one to model cognitive capacities like decision making and working memory necessary to implement complex joint action behavior that goes beyond a simple input–output mapping. To validate the architecture we have used a joint assembly task in which the robot has to construct together with a user different toy objects from their components. Different to our previous study in a symmetric construction task (Bicho et al., 2008, 2009), the robot does not directly participate in the construction work. The focus of the present study is on anticipating the needs of the user (e.g., handing over pieces the user will need next) and on the detection and communication of unexpected events that may occur on the plan and the execution level. The robot reasons aloud to indicate in conjunction with hand gestures the outcome of its action simulation or action monitoring to the user. The robot is able to react to speech input confirming or not the prediction of the internal simulation process. It also understands object-directed speech commands (e.g., *Give me object X*) through motor simulation. The results show that the integration of verbal and nonverbal communication greatly improves the fluency and success of the team performance.

JOINT CONSTRUCTION TASK

For the human–robot experiments we modified a joint construction scenario introduced in our previous work (Bicho et al., 2009). The goal of the team is to assemble different toy objects from a set of components (**Figure 1**). Since these components are initially distributed in the separate working areas of the two teammates, the coordination of their actions in space and time is necessary in order to successfully achieve the task. The human performs the assembly steps following a given plan which explains the way how different pieces have to be attached to each other. He or she can directly request from the robot a specific component by using speech commands (e.g., *Give me component X*) and/or communicative hand gestures (e.g., pointing, requesting). The role of the robot is to hand over pieces in response to such requests or in anticipation of the user's needs, to monitor the user's actions and to communicate potential conflicts and unexpected behaviors during task execution to the user. Conflicts may result from a mismatch between expected and perceived goal-directed actions either because the action should have been performed later (sequence error) or the action is not compatible with any of the available construction plans defining possible target objects (wrong component).

The fact that the robot does not perform assembly steps itself simplifies the task representation that the robot needs to serve the user (for a symmetric construction scenario see Bicho et al., 2009). What the robot has to memorize is the serial order of the use of the different components rather than a sequence of subgoals (e.g., attach components A and B in a specific way) that have to be achieved

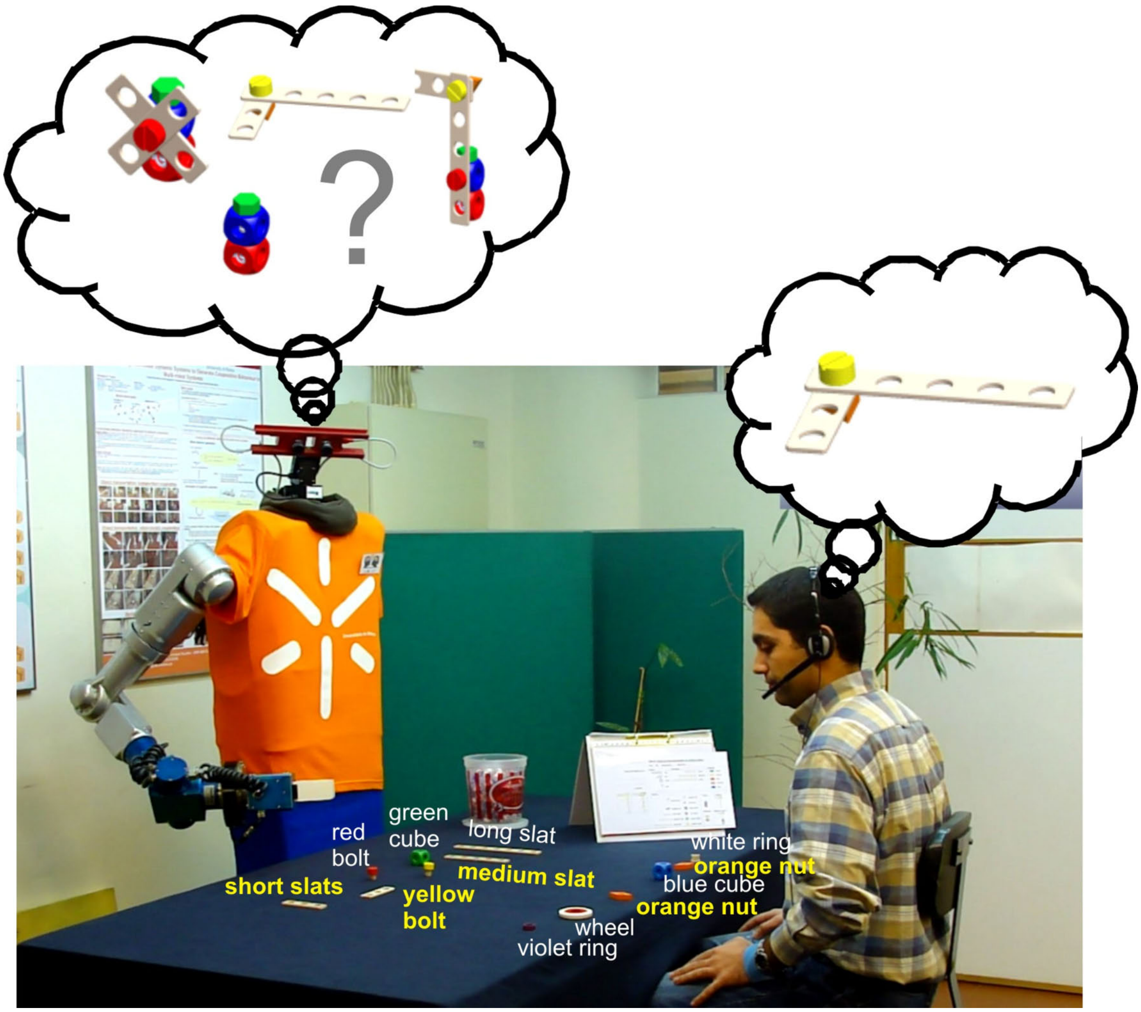


FIGURE 1 | Human–robot joint construction of different toy objects. The robot has first to infer what toy object the human partner intends to build. Subsequently, the team constructs the target object from its components following an assembly plan.

during the course of the assembly work. Importantly, since for each of the target objects the serial order of task execution is not unique, the robot has to simultaneously memorize several sequences of component-directed grasping actions in order to cope with different user preferences. To facilitate the coordination of actions and plans between the teammates, the robot speaks aloud and uses gestures to communicate the outcome of its goal inference and action monitoring processes to the user. For instance, the robot may respond to a request by saying *You have it there* and simultaneously points to the specific piece in the user's workspace. Although the integration of language and communicative gestures in the human–robot interactions will normally promote a more fluent task performance, this integration may also give rise to new types of conflict that the team has to resolve. From studies with humans it is well known for instance that if the verbally expressed meaning of an action or gesture does not match the accompanying hand movement (e.g., pointing to an object other than the object referred to) decision processes in the observer/listener appear to be delayed compared to a matching situation. This finding has been taken as direct evidence for the important role of motor representations in the comprehension of action-related language (Glenbach and Kaschak, 2002).

For the experiments we used the robot **ARoS** built in our lab. It consists of a stationary torus on which a 7 DOFs AMTEC arm (Schunk GmbH) with a two finger gripper and a stereo camera head are mounted. A speech synthesizer/recognizer (Microsoft Speech SDK 5.1) allows the robot to verbally communicate with the user. The information about object type, position and pose is provided by the camera system. The object recognition combines color-based segmentation with template matching derived from earlier learning examples (Westphal et al., 2008). The same technique is also used for the classification of object-directed, static hand postures such as grasping and communicative gestures such as pointing or demanding an object. For the control of the arm-hand system we applied a global planning method in posture space that allows us to generate smooth and natural movements by integrating optimization principles obtained from experiments with humans (Costa e Silva et al., submitted).

ROBOT CONTROL ARCHITECTURE

The multistage control architecture reflects empirical findings accumulated in cognitive and neurophysiological research suggesting a joint hierarchical model of action execution and action observation (van Schie et al., 2006; Hamilton and Grafton, 2008, see also Wolpert et al., 2003 for a modeling approach). The basic idea is that motor resonance mechanism may support social interactions on different but closely coupled levels: an intention level, a level describing the immediate goals necessary to realize the intention, and the kinematics level defining the movements of actions in space and time (**Figure 2A**).

Efficient action coordination between individuals in cooperative tasks requires that each individual is able to anticipate goals and motor intentions underlying the partner's unfolding behavior. As discussed in the introduction, most MNs represent actions on an abstract level sensitive to goals and intentions. For a human–robot team this is of particular importance since it allows us to exploit the motor resonance mechanism across teammates with very different embodiment.

In the following we briefly describe the main functionalities of the layered control architecture for joint action. It is implemented as a distributed network of DNFs representing different reciprocally connected neural populations. In their activation patterns the pools encode action means, action goals and intentions (or their associated perceptual states), contextual cues and shared task information (c.f. 'Model Details' for details on DNFs). In the joint construction task the robot has first to realize which target object the user intends to build. When observing the user reaching toward a particular piece, the automatic simulation of a reach-to-grasp action allows the robot to predict future perceptual states linked to the reaching act. The immediate prediction that the user will hold the piece in his/her hand is associated with the representation of one or more target objects that contain this particular part. In case that there is a one-to-one match, the respective representation of the target object becomes fully activated. Otherwise the robot may ask for clarification (*Are you going to assemble object A or object B?*) or may wait until another goal-directed action of the user and the internal simulation of action effects disambiguate the situation.

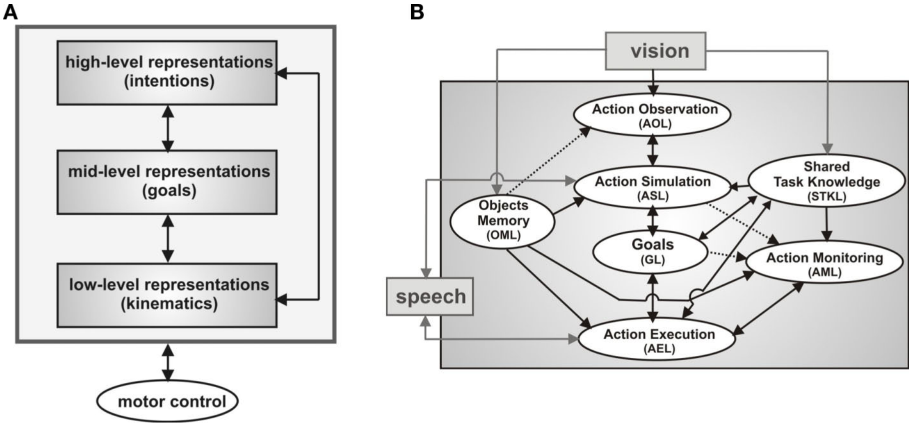


FIGURE 2 | Multistage robot control architecture. (A) Joint hierarchical model of action execution and action observation. **(B)** Mapping from observed actions (layer AOL) onto complementary actions (layer AEL) taking into account the inferred action goal of the partner (layer GL), detected errors (layer AML), contextual cues (layer OML) and shared task knowledge (layer STKL). The goal inference capacity is based on motor simulation (layer ASL).

Once the team has agreed on a specific target object, the alignment of goals and associated goal-directed actions between the teammates have to be controlled during joint task execution. **Figure 2B** presents a sketch of the highly context-sensitive mapping of observed onto executed actions implemented by the DNF-architecture. The three-layered architecture extends a previous model of the STS-PF-F5 mirror circuit of monkey (Erlhagen et al., 2006a) that is believed to represent the neural basis for a matching between the visual description of an action in area STS and its motor representation in area F5 (Rizzolatti and Craighero, 2004). This circuit supports a direct and automatic imitation of the observed action. Importantly for joint action, however, the model allows also for a flexible perception–action coupling by exploiting the existence of action chains in the middle layer PF that are linked to goal representations in prefrontal cortex. The automatic activation of a particular chain during action observation (e.g., reaching–grasping–placing) drives the connected representation of the co-actor's goal which in turn may bias the decision processes in layer F5 towards the selection of a complementary rather than an imitative action. Consistent with this model prediction, a specific class of MNs has been reported in F5 for which the effective observed and effective executed actions are logically related (e.g., implementing a matching between placing an object on the table and bringing the object to the mouth, di Pellegrino et al., 1992). For the robotics work we refer to the three layers of the matching system as the action observation (AOL), action simulation (ASL) and action execution layer (AEL), respectively. The integration of verbal communication in the architecture is represented by the fact that the internal simulation process in ASL may not only be activated by observed object-directed actions but also by action related speech input. Moreover, the set of complementary behaviors represented in AEL consists of goal-directed action sequences like holding out an object for the user but also contains communicative gestures (e.g., pointing) and speech output.

For an efficient team behavior, the selection of the most adequate complementary action should take into account not only the inferred goal of the partner (represented in GL) but also the working memory about the location of relevant parts in the separate working areas of the teammates (represented in OML), and shared knowledge about the sequential execution of the assembly task (represented in STKL). To guarantee proactive behavior of the robot, layer STKL is organized in two connected DNFs with representation of all relevant parts for the assembly work. Feedback from the vision system about the state of the construction and the observed or predicted current goal of the user will activate the population encoding the respective part in the first layer. Through synaptic links this activation pattern automatically drives the representations of one or more future components as possible goals in the second layer. Based on this information and in anticipation of the user's future needs the robot may already prepare the transfer of a part that is currently in its workspace.

In line with the reported findings in cognitive neuroscience the dynamic field architecture stresses that the perception of a co-actor's action may immediately and effortlessly guide behavior. However, even in familiar joint action tasks there are situations that require some level of cognitive control to override prepotent responses. For instance, even if the user would directly request verbally or by pointing a valid part located in the robot's workspace, the robot should not automatically start a handing over procedure. The user may have for instance overlooked that he has an identical object in his own working area. In this case, a more efficient complementary behavior for the team performance would be to use a pointing gesture to attract the user's attention to this fact. Different populations in the action monitoring layer (AML) are sensitive to a mismatch on the goal level (e.g., requesting a wrong part) or on the level of action means (e.g., handing over versus grasping directly). In the example, input from OML (representing the part in the user's workspace) and from ASL (representing the simulated action means) activate a specific neural population in AML that is in turn directly connected to the motor representation in AEL controlling the pointing gesture. As a result, two possible complementary actions, handing over and pointing, compete for expression in overt behavior. Normally, the pointing population has a computational advantage since the neural representations in AML evolve with a slightly faster time scale compared to the representations driving the handing over population. In the next section we explain in some more detail the mechanisms underlying decision making in DNFs. It is important to stress that the direct link between action monitoring and action execution avoids the problem of a coordination of reactive and deliberative components that in hybrid control architectures for HRI typically requires an intermediate layer (e.g., Spexard et al., 2007; Foster et al., 2008).

MODEL DETAILS

Dynamic neural fields provide a theoretical framework to endow artificial agents with cognitive capacities like memory, decision making or prediction based on sub-symbolic dynamic representations that are consistent with fundamental principles of cortical information processing. The basic units in DNF-models are local neural populations with strong recurrent interactions that cause non-trivial dynamic behavior of the population activity. Most importantly, population activity which is initiated by time-dependent external signals may become self-sustained in the absence of any external input. Such attractor states of the population dynamics are thought to be essential for organizing goal-directed behavior in complex dynamic situations since they allow the nervous system to compensate for temporally missing sensory information or to anticipate future environmental inputs.

The DNF-architecture for joint action thus constitutes a complex dynamical system in which activation patterns of neural populations in the various layers appear and disappear continuously in time as a consequence of input from connected populations and sources external to the network (e.g., vision, speech).

For the modeling we employed a particular form of a DNF first analyzed by Amari (1977). In each model layer i, the activity $u_i(x,t)$ at time t of a neuron at field location x is described by the following integro-differential equation (for mathematical details see Erlhagen and Bicho, 2006):

$$\tau_i \frac{\delta u_i(x,t)}{\delta t} = -u_i(x,t) + S_i(x,t) + \int w_i(x-x')f_i(u_i(x',t))dx' - h_i \quad (1)$$

where the parameters $\tau_i > 0$ and $h_i > 0$ define the time scale and the resting level of the field dynamics, respectively. The integral term describes the intra-field interactions which are chosen of lateral-inhibition type:

$$w_i(x) = A_i \exp\left(\frac{-x^2}{2\sigma_i^2}\right) - w_{\text{inhib},i} \tag{2}$$

where $A_i > 0$ and $\sigma_i > 0$ describe the amplitude and the standard deviation of a Gaussian, respectively. For simplicity, the inhibition is assumed to be constant, $w_{\text{inhib},i} > 0$. Only sufficiently activated neurons contribute to interaction. The threshold function $f_i(u)$ is chosen of sigmoidal shape with slope parameter β and threshold u_0:

$$f_i(u_i) = \frac{1}{1 + \exp[-\beta(u_i - u_0)]}. \tag{3}$$

The model parameters are adjusted to guarantee that the field dynamics is bi-stable (Amari, 1977), that is, the attractor state of a self-stabilized activation pattern coexists with a stable homogenous activation distribution that represents the absence of specific information (resting level). If the summed input, $S_i(x,t)$, to a local population is sufficiently strong, the homogeneous state loses stability and a localized pattern in the dynamic field evolves. Weaker external signals lead to a subthreshold, input-driven activation pattern in which the contribution of the interactions is negligible. This preshaping by weak input brings populations closer to the threshold for triggering the self-sustaining interactions and thus biases the decision processes linked to behavior. Much like prior distributions in the Bayesian sense, multi-modal patterns of subthreshold activation may for instance model user preferences (e.g., preferred target object) or the probability of different complementary actions (Erlhagen and Bicho, 2006).

The existence of self-stabilized activation pattern allows us to implement a working memory function. Since multiple potential goals may exist and should be represented at the same time and all relevant components for the construction have to be memorized simultaneously, the field dynamics in the respective layers (STKL and ML) must support multi-peak solutions. Their existence can be ensured by choosing weight functions (Eq. 2) with limited spatial ranges. The principle of lateral inhibition can be exploited on the other hand to force and stabilize decisions whenever multiple hypothesis about the user's goal (ASL, GL) or adequate complementary actions (AEL) are supported by sensory or other evidence. The inhibitory interaction causes the suppression of activity below resting level in competing neural pools whenever a certain subpopulation becomes activated above threshold. The summed input from connected fields u_l is given as $S_i(x,t) = k\Sigma_l S_l(x,t)$. The parameter k scales the total input to a certain population relative to the threshold for triggering a self-sustained pattern. This guarantees that the inter-field couplings are weak compared to the recurrent interactions that dominate the field dynamics (for details see Erlhagen and Bicho, 2006). The scaling also ensures that missing or delayed input from one or more connected populations will lead to a subthreshold activity distribution only. The input from each connected field u_l is modeled by Gaussian functions:

$$S_l(x,t) = \sum_m \sum_j a_{mj} c_l(t) \exp\left(\frac{-(x - x_m)^2}{2\sigma^2}\right) \tag{4}$$

where $c_l(t)$ is a function that signals the presence or absence of a self-stabilized activation peak in u_l, and a_{mj} is the inter-field synaptic connection between subpopulation j in u_l to subpopulation m in u_i. Inputs from external sources (speech, vision) are also modeled as Gaussians for simplicity.

RESULTS

In the following we discuss results of real-time human–robot interactions in the joint construction scenario. The snapshots of video sequences shall illustrate the processing mechanisms underlying the robot's capacity to anticipate the user's need and to deal with unexpected events. To allow for a direct comparison between different joint action situations, the examples all show the team performance during the construction of a single target object called L-shape (**Figure 3**). Details on the connection scheme for the neural pools in the layered architecture and numerical values for the DNF parameters and inter-field synaptic weights may be found in the Supplementary Material.

The initial communication between the teammates that lead to the alignment of their intentions and plans is included in the videos. They can be found at http://dei-s1.dei.uminho.pt/pessoas/estela/JASTVideosFneurorobotics.htm. The plan describing how and in which serial order to assemble the different components is given to the user at the beginning of the trials. We focus the discussion of results on the ASL and AEL. **Figures 4, 5 and 7** illustrate the experimental results. In each Figure, panel A shows a sequence of video snapshots, panel B and C refer to the ASL and AEL, respectively. For both layers, the total input (top) and the field activation (bottom) are compared for the whole duration of the joint assembly work. **Tables 1 and 2** summarize the component-directed actions and communicative gestures that are represented by different populations in each of the two layers. Since the robot does not perform assembly steps itself, AEL only contains two types of overt motor behavior: pointing towards a specific component in the user's workspace or grasping a piece for holding it out for the user.

It is important to stress that the dynamic decision making process in AEL also works in more complex situations with a larger number of possible complementary action sequences linked to each component (Erlhagen and Bicho, 2006).

Figure 4 shows the first example in which the humans starts the assembly work by asking for a medium slat (S1). The initial distribution of components in the two workplaces can be seen in **Figure 1**. The fact that the user simultaneously points towards a short slat creates a conflict that is represented in the bi-modal input pattern to ASL centered over A6 and A7 at time T0. As can be seen in the bottom layer of **Figure 4B**, the field dynamics of ASL resolves this conflict by evolving a self-sustained activation pattern. It represents a simulated pointing act towards the short slat. The decision is the result of a slight difference in input strength which favors communicative gestures over verbal statements. This bias can be seen as reflecting an interaction history with different users. Our human–robot experiments revealed that naive users are usually better in pointing than verbally referring to (unfamiliar) objects. The robot directly communicates the inferred goal to the

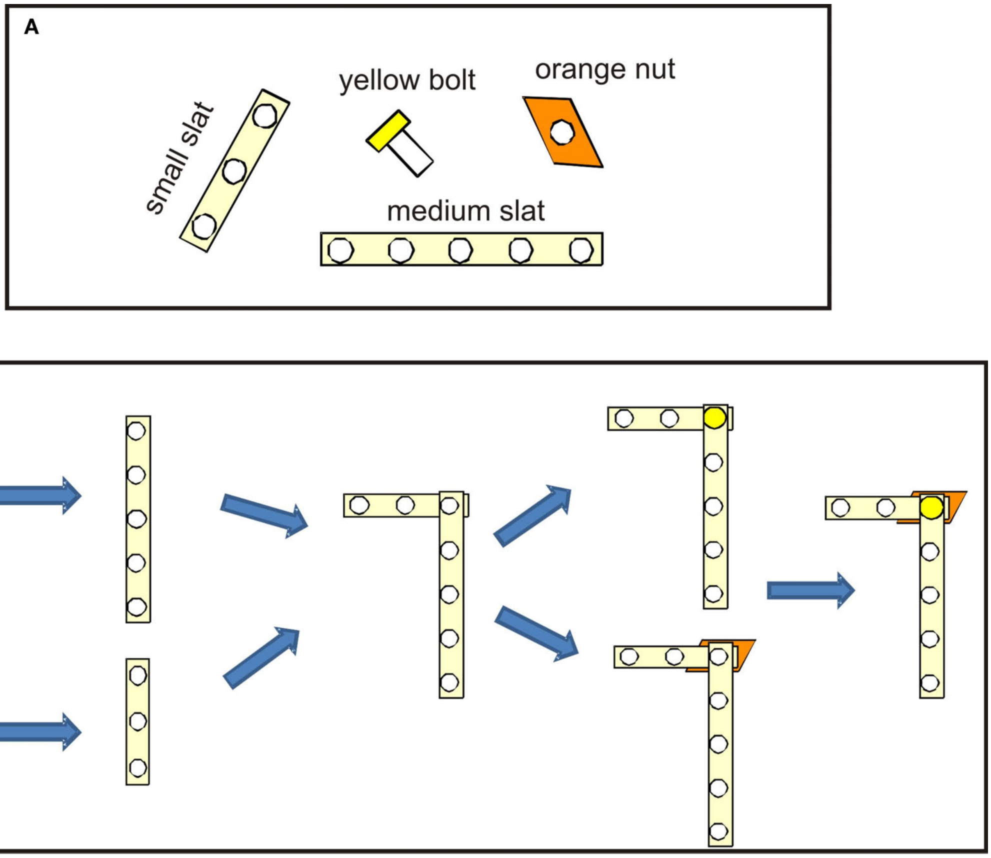


FIGURE 3 | Toy object L-shape. (A) Pieces used to build the L-shape. **(B)** Different serial orders to assemble the L-shape.

user (S2). **Figure 4C** shows that the input to AEL supports two different complementary actions, A1 and A2. However, since the total input from connected layers is stronger for alternative A1, the robot decides to hand over the short slat (S3). Subsequently, the robot interprets the user's request gesture (empty hand, S4) as demanding a medium slat (S5). The observed unspecific gesture activates to some extent all motor representations in ASL linked to components of the L-shape in the robot's workspace (compare the input layer). Goal inference is nevertheless possible due to the input from STKL that contains populations encoding the sequential order of task execution. The field activation of AEL (**Figure 4C**) shows at time T1 the evolution of an activation peak representing the decision to give the medium slat to the user (S6). At time T2 the robot observes the human reaching towards an orange nut (S7). The visual input from AOL activates the motor representation A4 in ASL which enables the robot to predict that the human is going to grasp the nut (S7). Since according to the plan the nut is followed by a yellow bolt and the bolt is in its workspace, the robot immediately starts to prepare the handing over procedure and communicates the anticipated need to the user (S8–S9). Note that the activation patterns representing the inferred current goal of the user (A4 in ASL) and the complementary action (A3 in AEL) evolve nearly simultaneously in time. An additional observation is worth mentioning. The input supporting the complementary behavior A3 starts to increase shortly after the decision to hand over the medium slat, that is, well ahead of the time when the robot predicts the nut as the user's next goal. This early preparation reflects the fact that handing over the medium slat automatically activates the representations of all possible future goals in STKL that are compatible with stored sequential orders. Since a yellow bolt and an orange nut represent both possible next assembly steps, the combined input from STKL and OML (bolt in robot's workspace) explains this early onset of subthreshold motor preparation in AEL.

In the second example (**Figure 5**) the initial distribution of components in the two working areas is identical to the situation in the first example. However, this time the meaning of the verbal request and the pointing act are congruent. Consequently, the input converges on the motor representation in ASL representing the pointing (A6) and a suprathreshold activity pattern quickly evolves. This in turn activates the population encoding the complementary behavior of handing over the short slat in AEL. Compared to the dynamics of the input and the field activity in the previous case (**Figure 4C**) one can clearly see that in the congruent condition the input arrives earlier in time and the decision process is faster. Note that in both cases the alternative complementary behavior representing the transfer of a medium slat (A3) appears to be activated below threshold at time T0. This pre-activation is caused by the input from STKL that supports both the short and the medium slat as possible goals at the beginning of the assembly work.

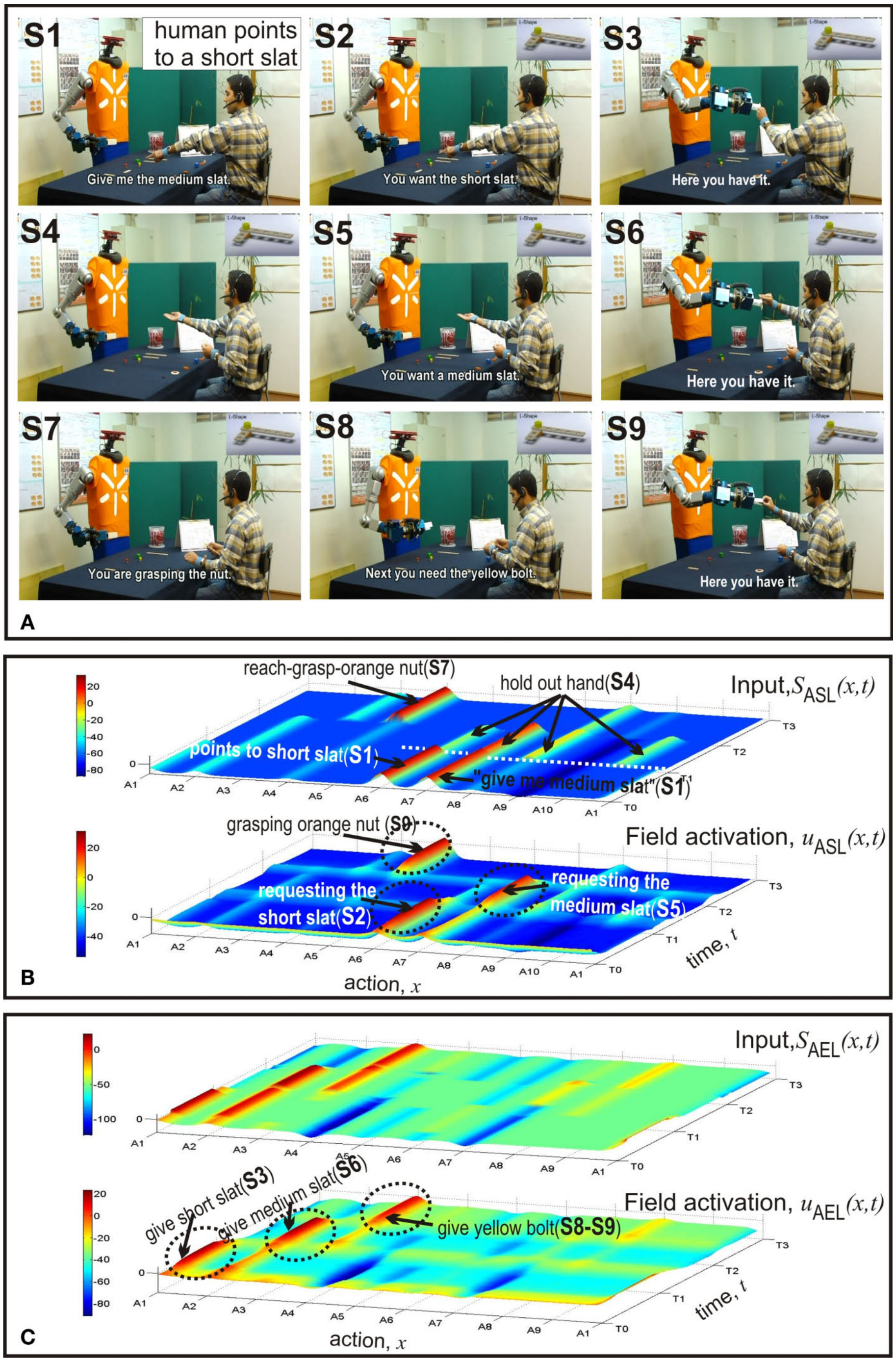


FIGURE 4 | First example: (1) goal inference when gesture and speech contain incongruent information (ASL), and (2) anticipatory action selection (AEL). (A) Video snapshots. **(B)** Temporal evolutions of input to ASL (top) and activity in ASL (bottom). **(C)** Temporal evolutions of input to AEL (top) and activity in AEL (bottom).

In the third example (**Figures 6 and 7**) the robot's action monitoring system detects a sequence error and the robot reacts in an appropriate manner before the failure becomes manifested. The robot observes a reaching towards the short slat (S1) and communicates to the user that it infers the short slat as the user's goal (S2). The input to the AEL (**Figure 7C**) triggers at time T0 the evolution of an activation pattern at A6 representing the preparation of a pointing to the medium slat in the user's workspace. However, this pattern does not become suprathreshold since at time T1 the user request the yellow bolt in the robot's workspace (S3). By internally simulating a pointing gesture the robot understands the request (S4) which in turn causes an activity burst of the population in AEL representing the corresponding complementary behavior (A3). However, also this pattern does not reach the decision level due to

Table 1 | Goal-directed sequences and communicative gestures in ASL.

Action	Sequence of motor primitives	Short description
A_1	Reach short slat → grasp	Use short slat
A_2	Reach medium slat → grasp	Use medium slat
A_3	Reach yellow bolt → grasp	Use yellow bolt
A_4	Reach orange nut → grasp	Use orange nut
A_5	Reach other piece → grasp	Use other part
A_6	Point to short slat	Request short slat
A_7	Point to medium slat	Request medium slat
A_8	Point to yellow bolt	Request yellow bolt
A_9	Point to orange nut	Request orange nut
A_{10}	Point to other part	Request other part

Table 2 | Goal-directed sequences and communicative gestures in AEL.

Action	Sequence of motor primitives	Short description
A_1	Reach short slat → grasp	Give short slat
A_2	Reach medium slat → grasp	Give medium slat
A_3	Reach yellow bolt → grasp	Give yellow bolt
A_4	Reach orange nut → grasp	Give orange nut
A_5	Point to short slat	Attend to short slat
A_6	Point to medium slat	Attend to medium slat
A_7	Point to yellow bolt	Attend to yellow bolt
A_8	Point to orange nut	Attend to orange nut
A_9	Point to other part	Attend to other part

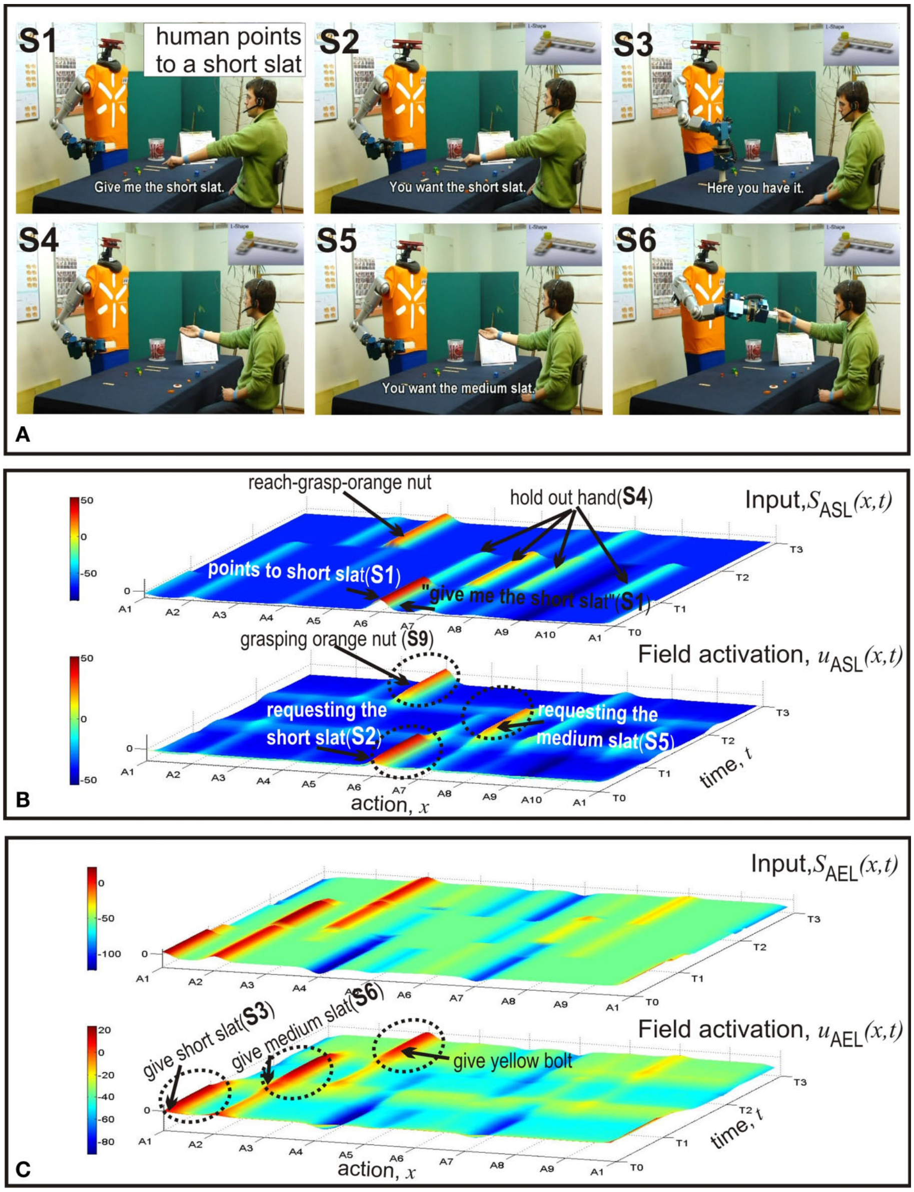


FIGURE 5 | Second example: faster goal inference and speeded decision making due to congruent information from gesture and speech. (A) Video snapshots. **(B)** Temporal evolutions of input to ASL (top) and activity in ASL (bottom). **(C)** Temporal evolutions of input to AEL (top) and activity in AEL (bottom).

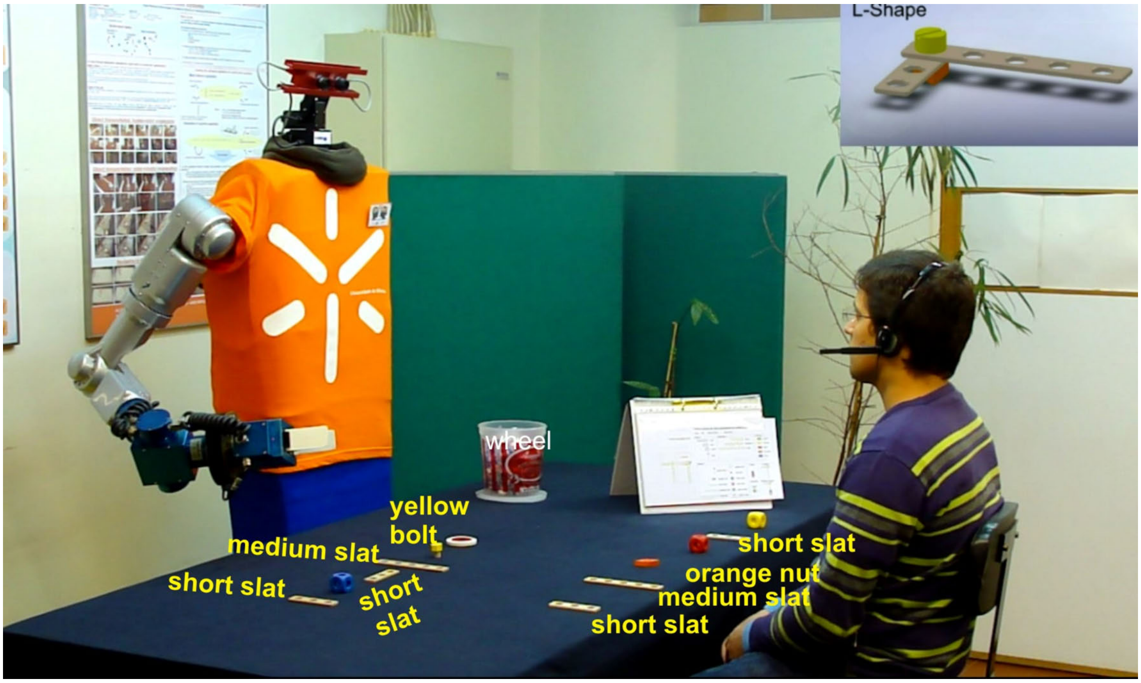


FIGURE 6 | Third example: initial distribution of components in the two working areas.

inhibitory input from a population in the AML. This population integrates the conflicting information from STKL (possible goals) and the input from the action simulation (yellow bolt). The robot informs the user about the sequence error (S5) and suggests the correction by pointing towards the medium slat and speaking to the user (S6). The pointing gesture is triggered by converging input from STKL, OML and the population in AML representing the conflict. The user reacts by reaching towards the correct piece (S7). The internal simulation of this action triggers the updating of the goals in STKL which allows the robot to anticipate what component the user will need next. As shown by the suprathreshold activation pattern of population A3 in AEL, the robot immediately prepares the transfer of the yellow bolt (S8–S9).

DISCUSSION AND SUMMARY

The main aim of the present study was to experimentally test the hypothesis that shared circuits for the processing of perception, action and action-related language may lead to more efficient and natural human–robot interaction. Humans are remarkably skilled in coordinating their own behavior with the behavior of others to achieve common goals. In known tasks, fluent action coordination and alignment of goals may occur in the absence of a full-blown human conscious awareness (Hassin et al., 2005). The proposed DNF-architecture for HRI is deeply inspired by converging evidence from a large number of cognitive and neurophysiological studies suggesting an automatic but highly context-sensitive mapping from observed on to-be-executed actions as underlying mechanism (Sebanz et al., 2006). Our low-level sensorimotor approach is in contrast with most HRI research that employ symbolic manipulation and high-level planning techniques (e.g., Breazeal et al., 2004; Alami et al., 2005; Spexard et al., 2007; Gast et al., 2009). Although it is certainly possible to encode the rules for the team performance in a logic-based framework, the logical manipulations will reduce the effectiveness that a direct decoding of others' goals and intentions through sensorimotor knowledge offers. At first glance, the motor resonance mechanism for nonverbal communication seems to be incompatible with the classical view of language as an intentional exchange of symbolic, amodal information between sender and receiver. However, assuming that like the gestural description of another person's action also a verbal description of that action has direct access to the same sensorimotor circuits allows one to bridge the two domains. In the robot ARoS, a verbal command like *Give me the short slat* first activates the representation of a corresponding motor act in ASL (e.g., pointing towards that slat) and subsequently the representation of a complementary behavior in AEL (e.g., transferring the short slat). We have introduced this direct language–action link into the control architecture not only to ground the understanding of simple commands or actions in sensorimotor experience but also to allow the robot to transmit information about its cognitive skills to the user. Verbally communicating the results of its internal action simulation and monitoring processes greatly facilitates the interaction with naive users since it helps a human to quickly adjust his/her expectations about the capacities the robot might have (Fong et al., 2003).

Our approach to more natural HRI differs not only on the level of the control architecture from more traditional approaches but also on the level of the theoretical framework used. Compared with for instance probabilistic models of cognition that have been employed in the past in similar joint construction tasks (Cuijpers et al., 2006; Hoffman and Breazeal, 2007), a dynamic approach to cognition (Schöner, 2008)

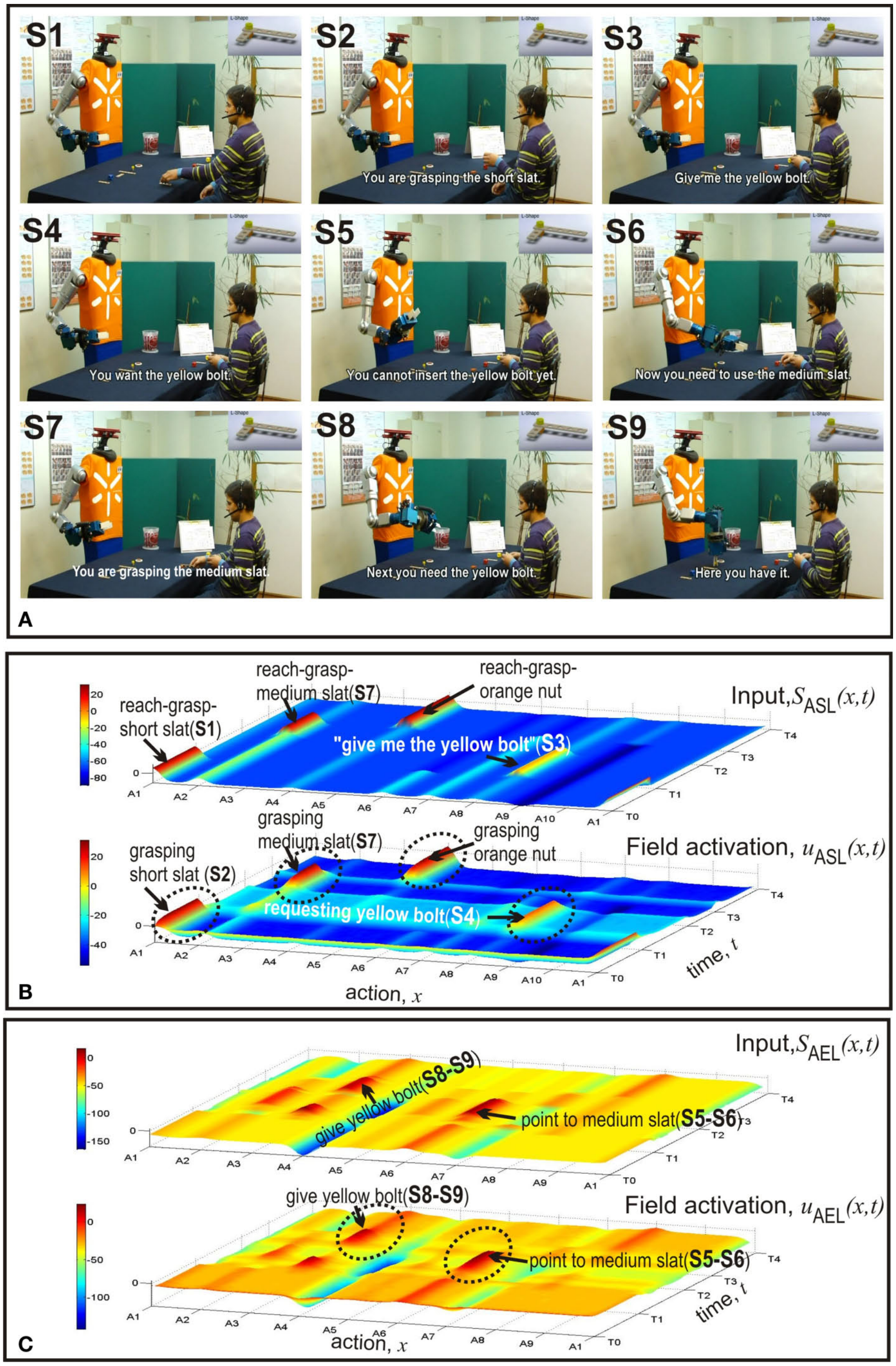


FIGURE 7 | Third example: Error detection and correction. (A) Video snapshots. **(B)** Temporal evolutions of input to ASL (top) and activity in ASL (bottom). **(C)** Temporal evolutions of input to AEL (top) and activity in AEL (bottom).

represented by the dynamic field framework allows one to directly address the important temporal aspects of action coordination (Sebanz et al., 2006). As all activity patterns in the interconnected network of neural populations evolve continuously in time with a proper time scale, a change in the time course of population activity in any layer may cause a change in the robot's behavior. For instance, converging input from vision and speech will speed up decision processes in ASL and AEL compared to the situation when only one input signal is available. Conflicting signals to ASL on the other hand will slow down the processing due to intra-field competition (compare **Figures 4 and 5**). This in turn opens a time window in which input from the AML may override a prepotent complementary behavior (**Figure 7**). We are currently exploring adaptation mechanisms of model parameters that will allow the robot to adapt to the preferences of different users.

A simple change in input strength from STKL to AEL will affect for instance whether the robot will wait for the user's explicit commands or will act in anticipation of the user's needs.

Learning and adaptation has not been a topic of the present study for which all inter-field connections were hand-coded. It is important to stress, however, that the DNF-approach is highly compatible with a Hebbian perspective on how social cognition may evolve (Keysers and Perrett, 2004). In our previous work we have applied a competitive, correlation-based learning rule to explain for instance how intention-related action chains may evolve during learning and practice (Erlhagen et al., 2006a, 2007). The interaction of the field and learning dynamics causes the emergence of new grasping populations that are linked to specific perceptual outcomes (e.g., grasping for handing over versus grasping for placing, compare Fogassi et al., 2005). Evidence from learning studies also support the plausibility of the direct action–language link implemented in the control architecture. Several groups have applied and tested in robots different neural network models to explain the evolution of neural representations that serve the dual role of processing action-related linguistic phrases and controlling the executing of these actions (Billard, 2002; Cangelosi, 2004; Wermter et al., 2004; Sugita and Tani, 2005). The results show that not only simple word–action pairs may evolve but also simple forms of syntax. A promising learning technique seems to be a covert or overt imitation of a teacher who is simultaneously providing the linguistic description. The tight coupling between learner and teacher helps to reduce the temporal uncertainty of the associations (Billard, 2002). The role of brain mechanisms that have been originally evolved for sensorimotor integration in the development of a human language faculty remains to a large extent unexplored (Arbib, 2005). We believe that combining concepts from dynamical systems theory and the idea of embodied communication constitutes a very promising line of research towards more natural and efficient HRI.

ACKNOWLEDGMENTS

The present research was conducted in the context of the fp6-IST2 EU-IP Project JAST (proj. nr. 003747) and partly financed by the FCT grants POCI/V.5/A0119/2005 and CONC-REEQ/17/2001. We would like to thank Rui Silva, Eliana Costa e Silva, Nzoji Hipolito, Toni Machado, Flora Ferreira and Emanuel Sousa for their help during the robotic experiments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/neuroscience/neurorobotics/paper/10.3389/fnbot.2010.00005/

REFERENCES

Alami, R., Clodic, A., Montreuil, V., Sisbot, E. A., and Chatila, R. (2005). "Task planning for human–robot interaction," in *Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence. ACM International Conference Proceeding Series*, Vol. 121, Grenoble, 81–85.

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibitory type neural fields. *Biol. Cybern.* 27, 77–87.

Arbib, M. A. (2005). From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics. *Behav. Brain Sci.* 28, 105–168.

Aziz-Zadeh, L., Wilson, S. M., Rizzolatti, G., and Iacoboni, M. (2006). Congruent embodied representations for visually presented actions and linguistic phrases describing actions. *Curr. Biol.* 16, 1818–1823.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., and Wilson, C. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends Cogn. Sci. (Regul. Ed.)* 7, 84–91.

Bicho, E., Louro, L., Hipolito, N., and Erlhagen, W. (2008). "A dynamic neural field architecture for flexible and fluent human–robot interaction," in *Proceedings of the 2008 International Conference on Cognitive Systems* (Germany: University of Karlsruhe), 179–185.

Bicho, E., Louro, L., Hipolito, N., and Erlhagen, W. (2009). "A dynamic field approach to goal inference and error monitoring for human–robot interaction," in *Proceedings of the 2009 International Symposium on New Frontiers in Human–Robot Interaction*, ed. K. Dautenhahn (Edinburgh: AISB 2009 Convention, Heriot-Watt University), 31–37.

Billard, A. (2002). "Imitation: a means to enhance learning of a synthetic proto-language in autonomous robots," in *Imitation in Animals and Artifacts*, eds K. Dautenhahn and C. L. Nehaniv (Cambridge, MA: MIT Press), 281–311.

Breazeal, C., Gray, J., and Berlin, M. (2009). An embodied cognition approach to mindreading skills for socially intelligent robots. *Int. J. Rob. Res.* 28, 656–680.

Breazeal, C., Hoffman, G., and Lockerd, A. (2004). "Teaching and working with robots as a collaboration," in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems.* New York, USA, 1030–1037.

Calinon, S., and Billard, A. (2008). "A framework integrating statistical and social cues to teach a humanoid robot new skills," in *Proceedings of the ICRA 2008 Workshop on Social Interaction with Intelligent Indoor Robots*, Pasadena, CA, USA, 27–34.

Cangelosi, A. (2004). "The sensorimotor basis of linguistic structure: experiments with grounded adaptive agents," in *From Animals to Animats 8*, eds S. Schaal, A. J. Ijspeert, A. Billard, S. Vijayakumar, J. Hallam, and J.-A. Meyer (Cambridge, MA: MIT Press), 487–496.

Chersi, F., Fogassi, L., Bonini, L., Erlhagen, W., Bicho, E., and Rizzolatti, G. (2007). "Modeling intentional neural chains in parietal and premotor cortex," in *Proceedings of the 37th Annual Meeting of the Society for Neuroscience, Soc. Neuroscience Abs. 636.6* San Diego.

Cohen, P., and Levesque, H. J. (1990). Intention is choice with commitment. *Artif. Intell.* 42, 213–261.

Cuijpers, R. H., van Schie, H. T., Koppen, M., Erlhagen, W., and Bekkering, H. (2006). Goals and means in action observation: a computational approach. *Neural Netw.* 19, 311–322.

di Pellegrino, G., Fadiga, I., Fogassi, L., Gallese, V., and Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Exp. Brain Res.* 91, 176–180.

Erlhagen, W., and Bicho, E. (2006). The dynamic neural field approach to cognitive robotics. *J. Neural Eng.* 3, R36–R54.

Erlhagen, W., Mukovskiy, A., and Bicho, E. (2006a). A dynamic model for action understanding and goal-directed imitation. *Brain Res.* 1083, 174–188.

Erlhagen, W., Mukovskiy, A., Bicho, E., Panin, G., Kiss, C., Knoll, A., van Schie, H., and Bekkering, H. (2006b). Goal-directed imitation for robots: a bio-inspired approach to action understanding and skill learning. *Rob. Auton. Syst.* 54, 353–360.

Erlhagen, W., Mukovskiy, A., Chersi, F., and Bicho, E. (2007). "On the development of intention understanding for joint action tasks," in *6th IEEE International Conference on Development and Learning* (London: Imperial College), 140–145.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., and Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. *Science* 308, 662–667.

Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots. *Rob. Auton. Syst.* 42, 143–166.

Foster, M. E., Giuliani, M., Mueller, T., Rickert, M., Knoll, A., Erlhagen, W., Bicho, E., Hipolito, N., and Louro, L. (2008). "Combining goal inference and natural language dialog for human–robot joint action," in *Proceedings of the 1st International Workshop on Combinations of Intelligent Methods and Applications, CIAM 2008*, Patras, Greece, 25–30.

Gast, J., Bannat, A., Rehrl, T., Wallhoff, F., Rigoll, G., Wendt, C., Schmidt, S., Popp, M., and Farber, B. (2009). "Real-time framework for multimodal human–robot interaction," in *Proceedings of the 2nd Conference on Human System Interactions 2009, HSI '09. IEEE Computer Society*, Catanid, Italy, 276–283.

Glenbach, A. M., and Kaschak, M. P. (2002). Grounding language in action. *Psychon. Bull. Rev.* 9, 558–565.

Hamilton, A. F., and Grafton, S. T. (2008). Action outcomes are represented in human inferior frontoparietal cortex. *Cereb. Cortex* 18, 1160–1168.

Hassin, R. R., Aarts, H., and Ferguson, M. J. (2005). Automatic goal inferences. *J. Exp. Soc. Psychol.* 41, 129–140.

Hauk, O., Johnsrude, I., and Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron* 41, 301–307.

Hoffman, G., and Breazeal, C. (2007). Cost-based anticipatory action selection for human–robot fluency. *IEEE Trans. Robot.* 23, 952–961.

Keysers, C., and Perrett, D. (2004). Demystifying social cognition: a Hebbian perspective. *Trends Cogn. Sci. (Regul. Ed.)* 8, 501–507.

Koenig, N., Chernova, S., Jones, C., Loper, M., and Jenkins, O. C. (2008). "Hands-free interaction for human–robot teams," in *Proceedings of the ICRA 2008 Workshop on Social Interaction with Intelligent Indoor Robots*, Pasadena, CA, USA, 35–41.

McGuire, P., Fritsch, J., Ritter, H., Steil, J. J., Röthling, F., Fink, G. A., Wachsmuth, S., and Sagerer, G. (2002). "Multi-modal human-machine communication for instructing robot grasping tasks," in *2002 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2002. IEEE Computer Society*, Lausanne, Switzerland, 1082–1089.

Newman-Norlund, R. D., van Schie, H. T., van Zuijlen, A. M. J., and Bekkering, H. (2007). The mirror neuron system is more active during complementary compared with imitative action. *Nat. Neurosci.* 10, 817–818.

Pardowitz, M., Knopp, S., Dillmann, R., and D, Z. R. (2007). Incremental learning of tasks from user demonstrations, past experiences and vocal comments. *IEEE Trans. Syst. Man Cybern. B Cybern.* 37, 322–332.

Rizzolatti, G., and Arbib, M. A. (1998). Language within our grasp. *Trends Neurosci.* 21, 188–194.

Rizzolatti, G., and Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci.* 27, 169–192.

Schaal, S. (2007). The new robotics: towards human-centered machines. *HFSP J* 1, 115–126.

Schöner, G. (2008). "Dynamical systems approaches to cognition," in *The Cambridge Handbook of Computational Psychology*, ed. R. Sun (Cambridge University Press, New York), 101–125.

Sebanz, N., Bekkering, H., and Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends Cogn. Sci. (Regul. Ed.)* 10, 70–76.

Spexard, T. P., Hanheide, M., and Sagerer, G. (2007). Human-oriented interaction with an anthropomorphic robot. *IEEE Trans. Robot.* 23, 852–862.

Steil, J. J., Röthling, F., Haschke, R., and Ritter, H. (2004). Situated robot earning for multi-modal instruction and imitation of grasping. *Rob. Auton. Syst.* 47, 129–141.

Sugita, Y., and Tani, J. (2005). Learning semantic combinatoriality from interaction between linguistic and behavioral processes. *Adapt. Behav.* 13, 33–52.

Umiltà, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., and Rizzolatti, G. (2001). I know what you are doing: a neurophysiological study. *Neuron* 31, 155–165.

van Schie, H. T., Toni, I., and Bekkering, H. (2006). Comparable mechanisms for action and language: neural systems behind intentions, goals, and means. *Cortex* 42, 495–498.

Wermter, S., Weber, C., Elshaw, M., Panchev, C., Erwin, H., and Pulvermüller, F. (2004). Towards multimodal neural robot learning. *Rob. Auton. Syst.* 47, 171–175.

Westphal, G., von der Malsburg, C., and Würtz, R. P. (2008). "Feature-driven emergence of model graphs for object recognition and categorization," in *Applied Pattern Recognition, Sudies in Computational Intelligence, Vol. 91*, eds H. Bunke, A. Kandel and M. Last (Berlin/Heidelberg: Springer Verlag), 155–199.

Wolpert, D. M., Doya, K., and Kawato, M. (2003). A unifying computational framework for motor control and social interactions. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 358, 593–602.

Zwann, R. A., and Taylor, L. J. (2006). Seeing, acting, understanding: motor resonance in language comprehension. *J. Exp. Psychol. Gen.* 135, 1–11.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 December 2009; paper pending published: 29 January 2010; accepted: 27 April 2010; published online: 21 May 2010.

Citation: Bicho E, Louro L and Erlhagen W (2010) Integrating verbal and nonverbal communication in a dynamic neural field architecture for human–robot interaction. Front. Neurorobot. ***4****:5. doi: 10.3389/fnbot.2010.00005*

frontiers in
NEUROROBOTICS

ORIGINAL RESEARCH ARTICLE
published: 31 May 2010
doi: 10.3389/fnbot.2010.00007

Grounding action words in the sensorimotor interaction with the world: experiments with a simulated iCub humanoid robot

Davide Marocco[1]*, Angelo Cangelosi[1], Kerstin Fischer[2] and Tony Belpaeme[1]

[1] Centre for Robotics and Neural Systems, School of Computing and Mathematics, University of Plymouth, Plymouth, UK
[2] University of Southern Denmark, Odense, Denmark

Edited by:
Frederic Kaplan, Ecole Polytechnique Federale De Lausanne, Switzerland

Reviewed by:
Pierre Andry, University of cergy pontoise, France
Peter F. Dominey, Centre National de la Recherche Scientifique, France

****Correspondence:***
Davide Marocco, School of Computing and Mathematics, University of Plymouth, Portland Square A322, Plymouth PL4 8AA, UK.
e-mail: davide.marocco@plymouth.ac.uk

This paper presents a cognitive robotics model for the study of the embodied representation of action words. The present research will present how an iCub humanoid robot can learn the meaning of action words (i.e. words that represent dynamical events that happen in time) by physically interacting with the environment and linking the effects of its own actions with the behavior observed on the objects before and after the action. The control system of the robot is an artificial neural network trained to manipulate an object through a Back-Propagation-Through-Time algorithm. We will show that in the presented model the grounding of action words relies directly to the way in which an agent interacts with the environment and manipulates it.

Keywords: grounding problem, cognitive robotics, embodiment

INTRODUCTION

Human language is a formidable communication system. It allows us to describe the world around us and exchange our thoughts. Nevertheless, despite many decades of studies and research, a complete description of its functions and operations is still missing. In particular, the fundamental mechanisms that allow humans to associate meanings to words are still a matter of ongoing debate among scientists.

For instance, Siskind (2001) suggests three major language functions allowing humans: (i) to describe what they perceive, (ii) to ask others to perform a certain action and (iii) to engage in conversation. At the core of all three functions there is our ability to understand the meanings that words represent. Especially the first two language functions require that language be grounded in perception and action processes. Especially in the description of dynamic processes and specific relations between objects and object properties, the process of grounding language in perception and actions means that, when we describe a given scene or we ask someone to perform a certain action, the words used must be linked with physical entities in the scene or in actions that can be either observed or desired.

In order to understand the link by means of which words are connected with objects and actions, it may be useful to look into studies on child language acquisition.

Children acquire word meanings in direct interaction with the environment. Before they begin to learn words, they go through a long phase of perceptual (visual, haptical, motor, interactional, etc.) exploration of objects in their environment. Interactions with preverbal infants and young children are furthermore anchored in the immediate context; that is, interactions are highly situated in the here and now and allow the child to make direct connections between perceptually available objects and events and linguistic utterances (e.g. Snow, 1977; Hatch, 1983; Sachs, 1983; Karmiloff and Karmiloff-Smith, 2001; Veneziano, 2001). Word meanings are therefore directly related to the child's experience, and the amount of situationally detached information presented to children by their caregivers only gradually increases over time (Veneziano, 2001).

While word meanings are partly acquired based on salient perceptual properties (cf. Clark, 1973; Smith et al. 1996), other word meanings are rather based on the role of functional affordances of objects in interaction (Nelson, 1973; Mandler, 1992). Nelson (1973), for instance, shows that 3-years-old children use their sensorimotor experiences about the function of a given object for categorization. But also linguistic information is taken into account in word meaning learning (cf. Gelman and Heyman, 1999; Bowerman and Choi, 2001; Bowerman, 2005), That is, children understand objects and events that share a linguistic label to share underlying characteristics as well (cf. Gelman, 2009).

But also for adult speakers, word meanings are grounded in embodied experience to a considerable extent (Bergen, 2005; Glenberg, 2007). For instance, distinctions between verbs of grasping are motivated by different hand postures and subtle differences in motor control involved in the actions denoted by a particular motion verb (Bailey et al., 1997). Different motor patterns associated with different action verbs were also found to be reflected in differences in location in the motor cortex (Pulvermüller et al., 2001). Furthermore, language understanding was found to interfere with motor actions if the meaning of the respective sentence evokes a motion in the opposite direction than necessary to carry out the action (Glenberg and Kaschak, 2002).

Further evidence for the embodiment of word meanings in adult language comes from the study of cognitive metaphor and image schemata (e.g. Lakoff and Johnson, 1999) and lexical semantics (Wierzbicka, 1985). These studies draw attention to meanings that are shaped by an implicit understanding of dimensions and functions of the human body. To address word learning from a grounded language learning perspective is thus supported by research from both child language acquisition and human language understanding.

Several computational models have been proposed to study communication and language in cognitive systems, such as robots and simulated agents (Cangelosi and Parisi, 2002; Lyon et al., 2007). On the one hand there are models of language focusing on the internal characteristics of the individual agent in which the lexicon is constructed based on a self-referential symbolic system. The cognitive agents only possess a series of abstract symbols used for both communication and for representing meanings (e.g. Kirby, 2001). These models are subject to the symbol grounding problem (Harnad, 1990). That is, symbols are self-referential entities that require the interpretation of an external experimenter to identify the referential meaning of the lexical items.

On the other hand, there are grounded approaches to modeling language, in which linguistic abilities are developed through the direct interaction between the cognitive agents and the physical world they interact with. In these models, the external world plays an essential role in shaping the language used by these cognitive systems. Language is therefore grounded in the cognitive and sensorimotor knowledge of the agents (Steels, 2003). As pointed out by Cangelosi and Riga (2006), the grounding of language in autonomous cognitive systems requires a direct grounding of the agent's basic lexicon. This assumes the ability to link perceptual (and internal) representations to symbols.

In this modeling paradigm, artificial agents are usually asked to associate features of objects to words, where this association is self-organized by the agents itself. An agent discovers autonomously certain features that are peculiar to a given object and learns from a model, which is usually another agent's, to associate the feature to an arbitrary word. Some of these models aim to study the emergence of shared lexicons through biological and cultural evolution mechanisms (Cangelosi and Parisi, 2002). In these models, a population of cognitive agents that are able to interact with the physical entities in the environment and to construct a sensorimotor representation of it, is initialized to use random languages. Within this population, agents converge toward the use of a shared lexicon after an iterative process of communication and language games.

The paradigm of language games for language evolution and acquisition has been used extensively by Luc Steels (Steels, 2001). For example, Steels and collaborators (Steels et al., 2002; Steels, 2003) use hybrid population of robots, internet agents and humans engaged in language games. Agents are in turn embodied into two "talking head" robots to play language games. In this experiment it has been demonstrated that a shared lexicon gradually emerges to describe a world made of colored shapes. This model has been also extended to study the emergence of communication between humans and robots using the SONY AIBO robot (Steels and Kaplan, 2000). Steels's approach is characterized by his focus on the naming of perceptual categories and by his emphasis on the importance of social mechanisms in the grounding and emergence of language.

Other models focus on the developmental factors that favor the acquisition of language by investigating the role of internal motivation and active exploratory behavior. Oudeyer and Kaplan (2006) show that an intrinsic motivation toward the experience of novel situations (i.e. situations that increase the chance of an agent to learn new environmental and communicational features) lead the agent to autonomously focus the attention toward vocal communicative and language features (see also Oudeyer et al. (2007), on a related topic, and Kaplan et al. (2008) for a compelling review and discussion of computational models of language acquisition).

From a different perspective, Marocco et al. (2003) use evolutionary robotics for the self-organization of simple lexicons in a group of simulated robots. Agents first acquire an ability to manipulate objects (e.g. to touch spheres or to avoid cubes). Subsequently, they are allowed to communicate with each other. Populations of agents are able to evolve a shared lexicon to name the objects and the actions being performed on them.

In other robotics models of language grounding, robotic agents acquire a lexicon through interaction with human users. For example, Roy et al. (2003) have developed an architecture that provides perceptual, procedural and affordance representations for grounding the meaning of words in conversational robots. Sugita and Tani (2005) use a mobile robot that follows human instructions based on combinations of five basic commands. Yu (2005) focuses on the combination of word learning and category acquisition to show improvements in both word-to-world mapping and perceptual categorization. This suggests a unified view of lexical and category learning in an integrative framework. Another experiment on human-robot communication has been carried out by Dominey (2005). This particular study provides insight into a developmental and evolutionary transition from idiom-like holophrases to progressively more abstract grammatical constructions.

All of the models presented before adopt the general and widespread assumption that tends to define *nouns* as words associated to physical (or even abstract) entities, and *verbs* as words that represent actions (or, in general, events that happen in time). This practice reflects findings in cognitive (e.g. Langacker, 2008) and functional (e.g. Halliday, 1985) linguistics that nouns are prototypically associated with objects and verbs prototypically correspond to events and actions. Grounded computational models so far mainly focus on grounding nouns on sensorimotor object representations and verbs on actions that are directly performed by the agent (e.g. Sugita and Tani, 2005). In Marocco et al. (2003), for example, a simulated robotic arm was evolved for the ability to discriminate between a sphere and a cube and then to associate different words (nouns) to the two objects. The discrimination was based on a physical exploration of the characteristics of the two shapes. Therefore, the meaning of the nouns was entirely grounded in the sensorimotor dynamic that allowed the discrimination of the two objects. The same procedure has been applied to evolve two different words associated to two different actions performed by the agents. The actions were "avoid" the cube and "touch" the sphere. In this case, the agents were asked to discriminate the objects and perform one of the two actions with respect to the shape. Given the action-word association, these types of words were defined as verbs.

In a different experiment, Cangelosi and Riga (2006) developed a robot able to imitate the actions of another robot (the teacher). The robot was also able to learn from the model an association between actions and words, such as close or open arms. Actions were related to motor patterns performed by the robot and words were directly associated to those motor patterns. Also in this case, therefore, the grounding of a verb is strictly related to an action that is entirely under the motor control of the agent.

Actions, however, are not restricted to agents. Actions can also be produced by physical objects in the environment, for example. Only few studies focus on the acquisition of actions words that are connected to properties of objects, such as *rolling* for a ball, or on the acquisition of words that express a dynamic and force-varied interaction with objects, such as *hit* or *move*. The application range of these two words can be extremely complex and may vary considerably depending on the physical properties of the object. In the case of the rolling ball, moving the ball can be "similar" to hitting the ball, because the ball has the property to roll after being hit; therefore it will move by itself. On the contrary, hitting a solid cube can produce a different effect from moving the cube by sliding it on the surface of a desk.

Other research in this area has mostly focused on disembodied models that aim to ground the meaning of action words by the elaboration of a visual scene acquired by a fixed camera that observes that scene. In Siskind (2001) the computational model is a computer program called *LEONARD* that analyzes a visual scene and is able to recognize different events, such as *pick-up, put-down, move* or *assemble*. As pointed out before, these actions are not directly performed by an agent, but it is the computational system that observes a visual scene through a fixed camera that is able to reconstruct the meaning. The visual scene typically includes a human hand that perform actions on objects of different colors. For instance, the pick-up scene is represented by a hand that picks up a red cube originally positioned on top of a green cube. The model is based on the principle of force dynamics (Talmy, 1988) and on a specifically designed event logic system, to which in later work (Fern et al., 2002) a learning system has been added. This enabled the computational model to learn and describe events based on a general temporal logic.

A similar approach to the identification of dynamic events has been taken by Steels and Baillie (2003). In their models, two artificial agents embedded in two movable cameras negotiate a self-organized lexicon based on dynamical events observed through the cameras. However, the ability of the agents to recognize and communicate about dynamic events is provided by the interaction between two different ways of using information: A bottom-up and a top-down direction of information flow. The bottom-up system, based on vision, provides information about the actual visual scene and a set of layered software detectors that allow to detect changes in the scene, such as movements of contacts between objects at a lower level, and series of changes at a upper level in order to identify complex dynamics. The top-down system, on the other hand, provides a sort of internal guidance for the vision system that allows, for example, to focus on particular aspects of the visual scene. Thus, the agent's representation of the world is constituted by the interaction between the two processes, encoded in a kind of lisp-type logic, and by sharing its communicative interaction with the other agent.

Cannon and Cohen (2010) and Cohen et al. (2005) ground the meaning of action words on the physical interactions between two bodies. In their system, verbs like *push, hit* and *chase* are represented as pathways through a metric space, defined as *maps for verbs*, that represent distances between two interacting physical entities, their velocity, and the observed transfer of energy after the interaction. Bodies are represented as circles of different colors that interact in various ways. Also in this case, as before, the system is purely based on the passive elaboration of a visual scene.

Following the same path, the aim of the present research is to study how a humanoid robot can learn to understand the meaning of action words (i.e. words that represent dynamical events that happen in time) by physically acting on the environment and linking the effects of its own actions with the behavior observed on the objects before and after the action. This will allow the agent to give an interpretation of a given scene that develops in time, and is grounded on its own bodily actions and sensorimotor coordination. Object manipulation, therefore, is the central concept behind the research. We believe that an active manipulation of the object is an opportunity to test the reaction of that object. Imagine the robot hits a ball. As an effect of the hit, the ball will move. Therefore, the dynamics of this event can be characterized by the action performed by the robot and by the sequence of the activation of its sensors during the movement and the physical interaction with the ball. The movement of the ball can be viewed as an instantaneous contact between the ball and the hand of the robot followed by a displacement of the same ball in the space, away from the hand. In such a case, the integration of the contact sensors with vision information can easily characterize this situation as different from another situation in which, e.g., the robot moves a cube by sliding it over the surface of the desk. In this case, although there is movement, i.e. the object displaces in space, the event is characterized by a continuous contact of the hand with the cube. On the other hand, the fact that different objects react differently to the same movement can also characterize a particular property of the object itself. Therefore, *rolling* and *sliding* are action words that pertain to objects that can be understood by the agent on the basis of the same sensorimotor information used to characterize its own actions.

Such types of interactions can be easily regarded as affordances of the objects for the robot. In fact, the robot learns the effects of its own movement on a given object. Several studies have already addressed affordances on robotics models in a similar way, where a robot learns a specific type of affordances using information provided by sensory states. These models have been mainly used in relation with imitation tasks. For example, in Fitzpatrick et al. (2003) a robot learns the motion dynamics of different objects after having pushed them. Subsequently, it uses the sensorimotor information to recognize actions performed by others and to replicate the observed motion. Similarly, Kozima et al. (2002) created a system that enables a robot to imitate actions driven by the effects of that actions (a more general solution on learning affordances is presented in Stoytchev, 2005; Fritz et al., 2006; Dogar et al., 2007). Montesano et al. (2008) created a humanoid robot controller that uses a Bayesian network for learning object affordances and showed the benefit of the model in imitation games. The model presented here, although inspired by a similar approach, does not have an explicit interest in imitation, and also the actions repertoire is simplified in comparison to those models. However, we believe that this simplification helps to better highlight and understand the sensorimotor grounding of action words, which is the primarily scientific question behind this work. This consideration, of course, does not prevent possible extensions of the model towards more applied scenarios that involve imitation tasks, as well as tasks that involve a form of linguistic instruction provided by another agent, which might be a human or another robot.

To approach the research issue related to the grounding of action words in sensorimotor coordination, we present a simulated robotic model equipped with a neural control system. By manipulating the environment, the robot can learn the association between certain objects, located on a desk in front of him, and some physical property of such objects. In the next section, a description of the robot used in the experiment, the environment and the neural control system will be described. Subsequently, the results of the experiment will be present and discussed.

MATERIALS AND METHODS

The robotic model used for the experiments is a simulation of the iCub humanoid robot (Tikhanoff et al., 2008) controlled by a recurrent artificial neural network. The robot can interact with objects located on a desk in front of it, and its neural control system is trained through a supervised learning algorithm, namely the "Back-Propagation-Through-Time" algorithm (Rumelhart and McClelland, 1986). In the following sections we provide details on the robotic platform utilized, the environment and on the robot-object interaction. Moreover, a description of the neural network that acts as a control system and of the training procedure will be presented.

THE SIMULATED HUMANOID ROBOTS AND THE ENVIRONMENT

For the experiments a simulated model of the iCub (**Figure 1**) has been used, a small-size humanoid robot, designed and produced by the European project "Robotcub" (robotcub.org; Metta et al. 2008). The iCub dimensions are similar to that of 2.5-year-old child and the robot has been specifically designed to act in a cognitive robotics domain, where the robotic platform is a physical entity that allows researchers to test hypothetic cognitive models in the real world. The robot is 90-cm tall and has a weight of 23 kg. iCub has 53 degrees of freedom distributed as follow: seven for each arm, six for each leg, three on the waist, three dedicated to eyes movements and three for the neck. In addition, it has two complex hands with 9 degrees of freedom each. For its size, the iCub is the most complete humanoid robot currently being designed, in terms of kinematic complexity. In contrast to similarly sized humanoid platforms, the eyes can also move. All motors and sensors are accessible through a centralized control system that provides an interface between the robot the and the external world. The interface is implemented on a PC104 board located in the head of the robot. For vision, the robot is equipped with two cameras with VGA resolution and 30-fps speed that provide color images (for additional technical details about the robot body and head see Beira et al., 2006; Tsagarakis et al., 2007). Every communication with the robots uses an Ethernet network protocol. The integrated software platform to control all the sensors and actuators is called YARP (Metta et al., 2006).

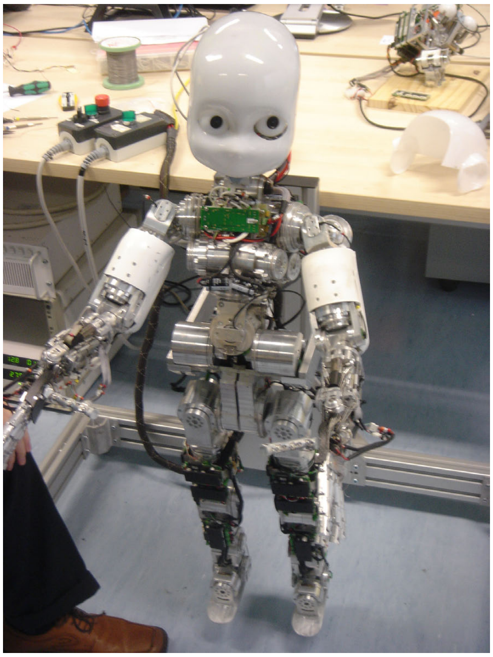


FIGURE 1 | The humanoid robot iCub.

In our experiment we used a carefully designed software simulation of the iCub robot that uses ODE (Open Dynamic Engine) to simulate the dynamics of the physical interactions (for details about the simulator see Tikhanoff et al., 2008). The YARP platform is used as the main communication tool for both the simulator and the real robot. The simulator has been designed to test the robot's software application in a safe, yet realistic, environment. In particular, the simulator can be used to safely test potentially dangerous motor commands that might damage the physical structure of the robot. Moreover, for the specific requirement of the model, we had to use tactile sensors in the hand that are currently not implemented on the real robot available to us.

For the present study we used a sub-set of all the degrees of freedom and only one of the two cameras. In particular, for manipulation purpose we only use a single joint on the shoulder that allows the robot to reach and move an object placed on a desk in front of it. The encoder value of this joint is also used as proprioceptive sensory feedback. When the hand gets in contact with an object, a binary tactile sensor placed on the hand is activated. Its activation value provides a coarse tactile sensory feedback. This tactile sensor is activated whatever part of the hand gets in contact with the object. The vision of the robot is provided by a vision system that acts on the left camera of the robot that automatically fixate the object in the environment, regardless of the action currently performed by the robot. The encoder values of two neck joints, that represent the position of the head, express the position of the object in the visual field relative to the robot. The position of the head is then treated as visual input of the system (Yamashita and Tani, 2008). The vision system, in addition to the object relative position in the visual field, also provides coarse information about the shape of the object. A parameter of the shape, which we call *roundness*, is calculated from the image of the object acquired by the robot and its value is added as input to the neural network controller. The robot automatically generates a movement when it receives a target joint angle as input. The movement corresponds to the target angle and is generated by means a pre-programmed proportional-integral-derivative (PID) controller. The sensorimotor state of the robot is updated every 500 ms.

The environment of the experiment consists in a desk placed in front of the robot. On the desk, one out of three objects is positioned on a given location. These objects are a sphere, a box, and a cylinder placed vertically on the desk. The sphere has a diameter of 12 cm. The three dimensions of the cube are 12 cm on a side and 7 cm on the other two sides. The cylinder has a diameter of 4 cm and 25 cm tall. Roundness values calculated for the three object are ~0.87, ~0.71, and ~0.43 for the sphere, the cube, and the cylinder respectively. Each of these objects has different physical properties associated to the shape and the physical connection to the desk. The sphere, when touched by the robot hand, will roll away on a direction that directly depends on the hand direction and on the applied force. The cube, when touched with the same force and direction as the ball, will slide on the desk while in contact with the robot hand. The cylinder, which was tightly attached to the desk, will not move and will prevent the robot to accomplish its desired movement. Therefore, the three objects represent three different properties, namely, the property to roll, to slide and to resist.

STRUCTURE AND TRAINING OF THE NEURAL CONTROL SYSTEM

The neural system that controls the robot is a fully connected recurrent neural network with 10 hidden units (**Figure 2**), eight input units and eight output units. Activations of input units are divided into five sensory units and three linguistic units. Three of the five sensory units are set to the encoder values of the three corresponding joints (shoulder, pan-neck and tilt-neck), scaled between 0 and 1. Those input units provide information about joints current angles. The fourth sensory unit encodes the value of the binary tactile sensor. This is set to either 0 or 1, depending on the contact of the hand with the object. The fifth sensory unit encodes the value of the roundness. The three linguistic input units represent a local binary encoding of the three objects. The activation value of those units can vary with respect to the experimental phase, that is, training or testing phase, respectively. Activations of hidden and output units y_i are calculated at a discrete time, by passing the net input u_i to the logistic function, as it is described in Eqs 1 and 2:

$$u_i = \sum_{j}^{n} y_j \cdot w_{ij} - k_i \tag{1}$$

$$y_i = \frac{1}{1 - e^{-u_i}} \tag{2}$$

where w_{ij} is the synaptic weight that connects unit *i* with unit *j* and k_i is the bias of unit *i*. The output units encode the values of the input at the time step $t + 1$. That is, the output state corresponds to the next input state of the network. The network is trained to predict its own input. As we will see in the next section, during the testing phase, the predicted input state is also used to provide target angles for the actuators.

The structure of the experiment is divided into two phases: in the first phase the network is trained to predict its own subsequent sensorimotor state. In the second phase the network is tested on the robot, in interaction with the environment.

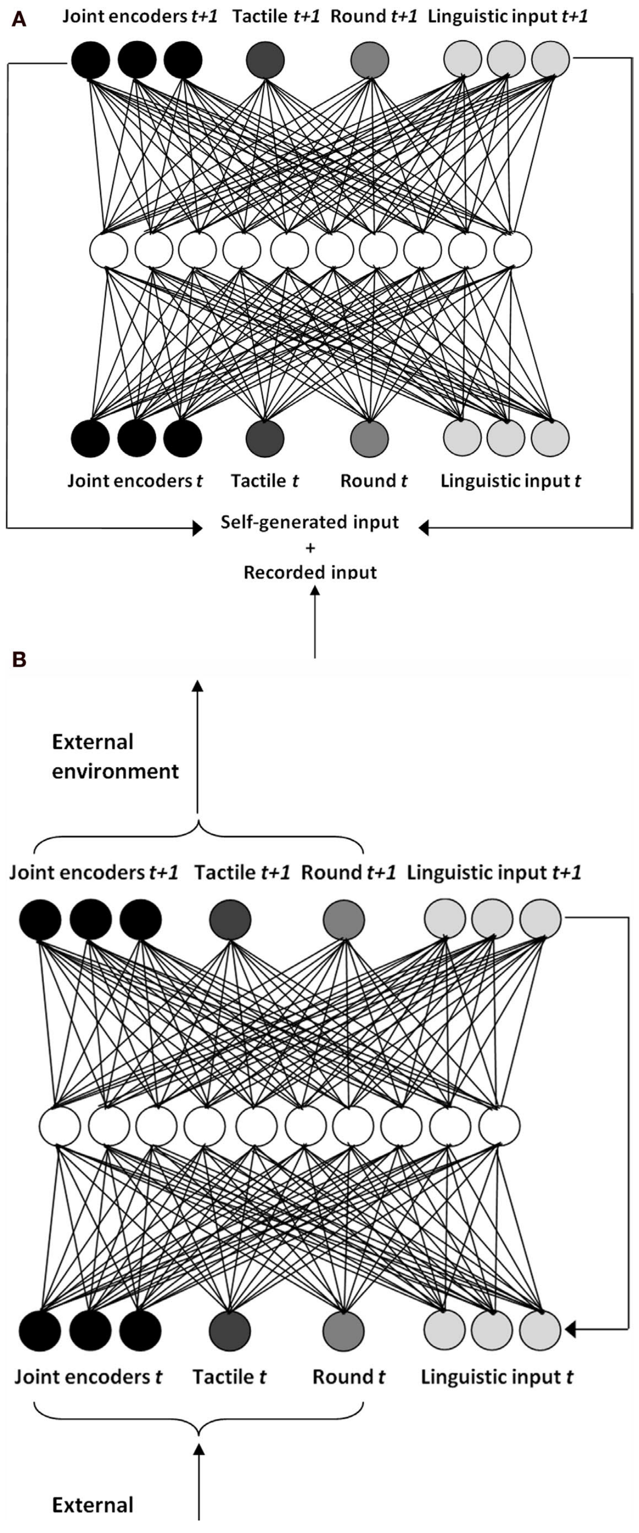


FIGURE 2 | The neural network that acts as a control system for the robot. **(A)** Network activation structure during the training phase (*closed-loop* condition). The output at time *t* with a small portion of the recorded input is used for setting the input at time *t* + 1. See the text for details. **(B)** Network activation structure during the testing phase (*open-loop* condition). The input is taken by the state of the sensors and the output is used to set the target angle of the actuators.

Training phase

For training the neural network we used the Back-Propagation-Through-Time-algorithm (BPTT), which is typically used to train neural network with recurrent nodes (Rumelhart and McClelland, 1986). This algorithm allows a neural network to learn the dynamical sequences of input-output patterns as they develop in time. Since we are interested in the dynamic and time dependent processes of the robot-object interaction, an algorithm that allows to take into account dynamic events is more suitable than the standard Back-Propagation algorithm (Rumelhart and McClelland, 1986). For a detailed description of the BPTT algorithm, in addition to Rumelhart and McClelland (1986) see also Werbos (1990). The main difference between a standard Back-Propagation algorithm and the BPTT is that, in the latter case the training set consists in a series of input-output sequences, rather than in a single input-output pattern. The BPTT allows the robot to learn sequences of actions. The goal of the learning process is to find optimal values of synaptic weights that minimize the error E, defined as the error between the teaching sequences and the output sequences produced by the network. The error function E is calculated as follows:

$$E=\sum_{S}\sum_{t}\sum_{i\in\text{output}}((y_{i,t,s*}-y_{i,t,s})(y_{i,t,s}-(1-y_{i,t,s})))^2 \qquad (3)$$

where $y_{i,t,s*}$ is the desired activation value of the output unit i at time t for the sequence s and $y_{i,t,s}$ is the actual activation of the same unit produced by the neural network, calculated using Eqs 1 and 2.

During the training phase, synaptic weights at learning step $n+1$ are updated using the error δ_{i} (Rumelhart and McClelland, 1986) calculated at the previous learning step (n), that in turn depend on the error E, according to the following equation:

$$\Delta w_{ij}(n+1)=\eta\delta_i y_j+\alpha\Delta w_{ij}(n) \qquad (4)$$

where w_{ij} is the synaptic weight that connects unit i with unit j, y_i is the activation of unit j, η is the learning rate and α is the momentum.

The sequences to learn, in our case, are the sensorimotor contingencies produced by the robot's manipulation of the object present in the environment. In order to produce those sequences, the robot is placed in front of the desk together with one of the three objects placed on the desk, at a given position. At this point, the shoulder joint of the robot is activated so as to move the right arm from the side of the robot to the front. By performing this movement, the hand of the robot moves towards the object and gets in contact with it. At the same time, the automatic vision routing turns the head in the direction of the object and keeps the object in the visual field by moving the neck joints (**Figure 3**). During this activity, we recorded the values of shoulder and neck joint encoders, as well as the state of the tactile sensors and the *roundness* value calculated by the image processing system. Each sequence consists of 30 recorded patterns that represent 15 s of activity by the robot. The graphs in **Figure 4** show the activations of the sensory units when the robot is interacting with the three objects. The information provided to the robot, although extremely simple, is sufficient to allow the neural controller to correctly separate the three conditions.

The input pattern of every sequence is completed by adding the linguistic input in the following form: *[1 0 0]* when the robot is interacting with the sphere, *[0 1 0]* when the robot is interacting with the cube, and *[0 0 1]* when the robot is interacting with the cylinder. The linguistic input is explicitly presented only at the beginning of the sequence. For the rest of the 30 patterns that form the training sequence, the linguistic input is self-generated by the network. It should be noted that at this time we deliberately avoid to give a semantic interpretation of the linguistic input and output. So far the "words" chosen as input and, consequently, as output simply correlate with the interaction with different objects.

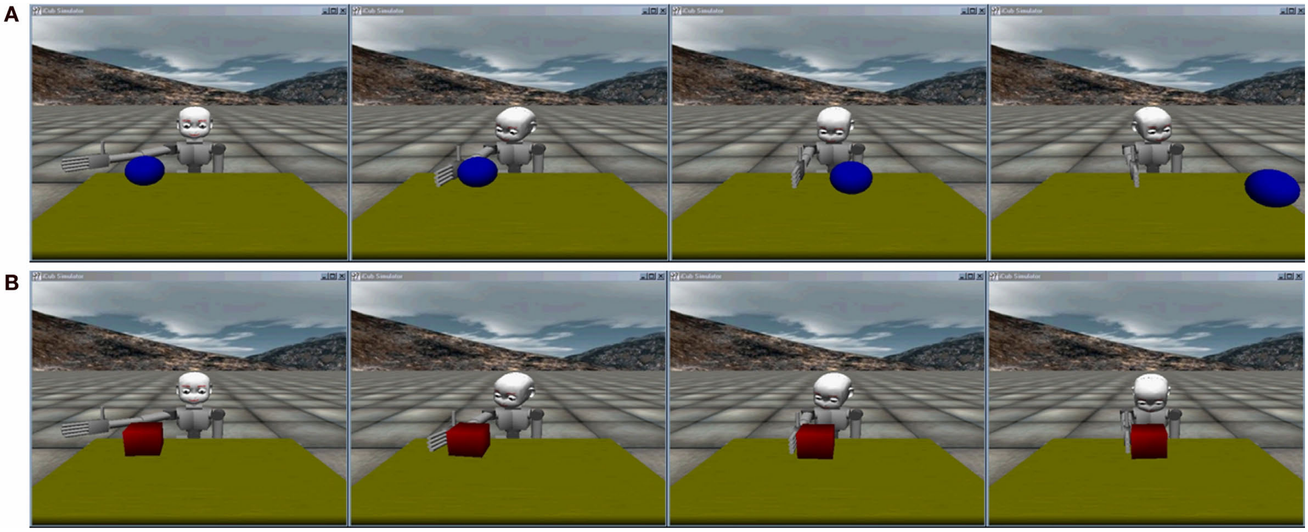


FIGURE 3 | From left to right, a small sample of the 30 step sequences produced for training the network. (A) An example sequence produced by the manipulation of a sphere and **(B)** the same movement towards a cube. The two objects produce different interactions because of their different physical properties. The sequence produced by the interaction with the fixed cylinder is not shown, given the fact that the robot, after the contact with the cylinder does not move anymore. The tracking behavior is due to the automatic visual routine embedded in the control system.

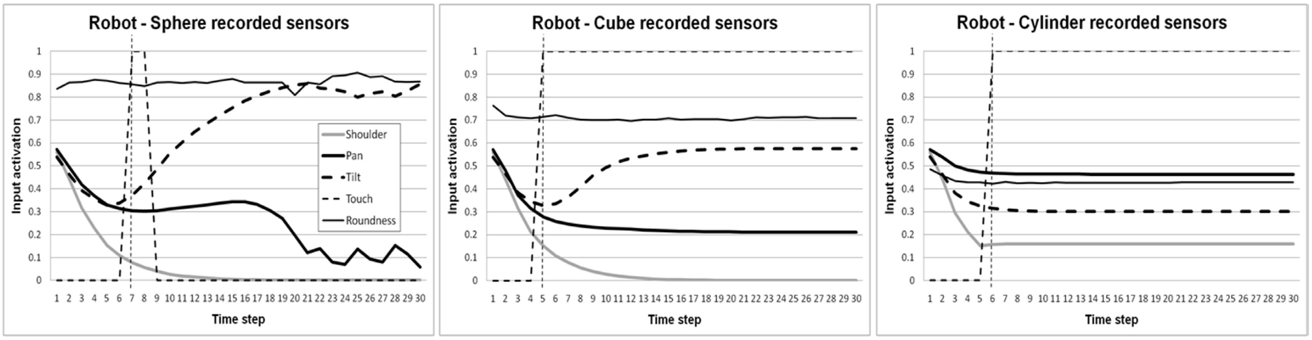


FIGURE 4 | Sensor activations recorded during the 15-s interactions between the robot and the object on the desk. Curves represent activations of the *shoulder* joint, the *pan* and *tilt* movements of the neck, the *touch* sensors and the *roundness* of the object calculated by the vision system over time. Refer to the legend on the graph. The thin dashed vertical line in the three graphs represents the moment in which the robot's arm touches the object.

Their semantic referent, i.e., whether they refer to objects (sphere, cube, cylinder) or to actions associated with the objects (roll, slide, fix), will be discussed later on. For this reason we refer to the linguistic output of the neural network as *linguistic_output_1 [1 0 0]*, *linguistic_output_2 [0 1 0]* and *linguistic_output_3 [0 0 1]*, corresponding to interactions with the sphere, the cube and the cylinder, respectively.

From these values we produce the sequences, one for each object, by setting the sequence element $t + 1$ as target output for the previous element t. In this way, starting from the first pattern, the network has to produce the next pattern. Then, the produced pattern is given as input to the network, which produces the next pattern and so on. This iteration is executed until the end of the sequence is reached.

The complete training set for the present work includes six sequences. Three of these are created in the way just described above, while the other three use the same set of data as before except for the roundness values which is set to 0. The linguistic input is presents in both cases.

During this process, the error produced by the network with respect to the target outputs is accumulated for all three sequences. The synaptic weights are updated according to the global error only after all the sequences have been presented. Therefore, according to the traditional back-propagation notation, the neural network is trained in "batch" mode and not "on-line" (for technical details about the algorithm adopted in this paper and a discussion about computational differences between "batch" and "on-line" training mode in recurrent neural networks see Williams and Zipser, 1995).

To facilitate the training process and to produce a neural network capable of better predicting the sequences, we used a training modality known as *closed-loop training* (Yamashita and Tani, 2008) depicted in **Figure 2A**. In this type of training procedure the input given to the network is the actual output produced by the network itself at the previous time cycle. However, by doing this, the error can accumulate on the input, especially at the beginning of the training process, given that the input is self-generated by a network with random synaptic weights. For this reason, the effect on the learning performance can be heavily affected and can prevent the algorithm to converge to a close to optimal solution. To avoid such a problem, the real input s fed to the network (i.e. the input actually used to calculate the performance), is produced by adding to the self-generated input s+ a small fraction of the recorded input s*, which represents the real input the network should receive. The same is done for the linguistic input m, with the only difference that m* is the linguistic activation fixed by the experimenter:

$$s = 0.1 \cdot s^* + 0.9 \cdot s^+$$

$$m = 0.1 \cdot m^* + 0.9 \cdot m^+$$

The parameters used for training the network used in the following experiments are: Learning rate 0.2; momentum 0.3; initial synaptic weights value between –0.01 and 0.01.

To assure the robustness of the results obtained, 10 replications with different initial random synaptic weights were carried out.

Testing phase

The second part of the experiment is the phase in which the network is tested in *open-loop*. That is, the network is connected to the robot and the input is directly produced by the actual values of its encoders (**Figure 2B**), while the output is used to determine the target angles of the joint for moving the arm. During this phase the robot is placed in front of the desk as before and an object is placed on the desk in front of the robot. By activating the robot, this time the movement of the arm is commanded by the output of the neural network, while the other outputs represent the prediction of the next sensory state.

In this set-up, the only joint that can be directly actuated by the network is the joint on the shoulder, which causes the movement of the arm toward the object. The joints controlling the neck are still commanded by the visual routine that tracks the object in the environment. However, the most interesting part in this experiment is the behavior of the linguistic output. As we will show and discuss in the next session, the interaction between action and language exploited during the training, allows us to better understand what type of sensorimotor contingencies are associated with certain linguistic patterns and how certain categories of action words might be directly grounded in the sensorimotor states of an agent.

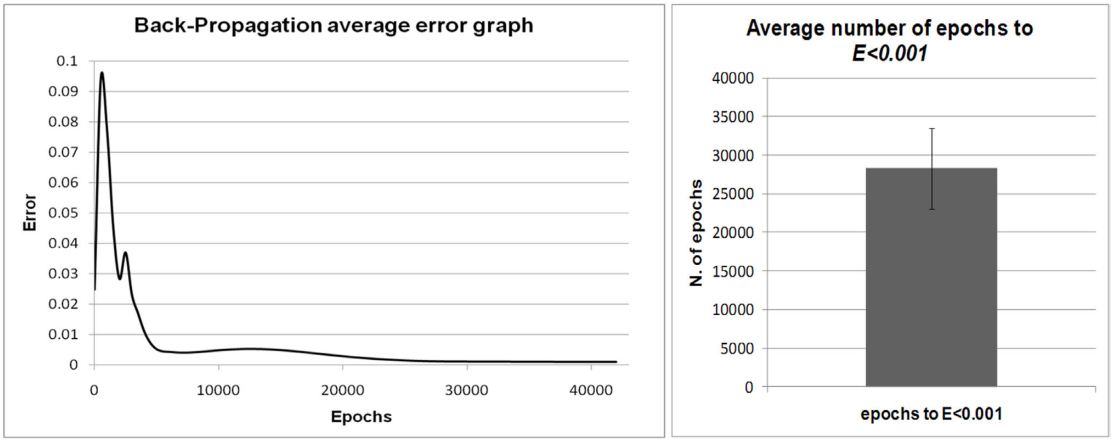


FIGURE 5 | Left: Error graph of the average of 10 replications during the training process. The *x* axis shows the number of epochs and the *y* axis shows the mean error. Right: Average epochs required to reach the error threshold, which was set to 0.001 ($E < 0.001$). Error bar represents Standard deviation.

During the testing phase some parameters of the set-up used for the training have been changed, such as the presence of the linguistic input and the size and position of objects on the desk. In the next section we will describe the tests performed and the results obtained.

RESULTS

After the training of all the 10 neural networks, we obtained 10 controllers that were able to predict the next sensory input state on the basis of the current input state. Previous tests shown that an error *E* smaller that 0.001 produces neural controllers capable of performing the task with a good degree of generalization. To avoid the overtraining of the network, we decided to set the learning threshold to 0.001. Below this threshold the training process is considered completed. Given this threshold, the 10 replications have been carried out by stopping the training as soon as *E* was smaller than the threshold. **Figure 5**, left, shows the average error calculated for the 10 replications during the training process. **Figure 5**, right, shows the average epochs that occurred until an error smaller than 0.001 for the 10 replications was reached. From the integration of the two data sets we can see that after about 28.000 epochs the error is already smaller than 0.001 for the majority of the replications, while for some of them additional epochs are required. The fastest replication reached the error threshold in 22695 epochs and the slowest reached the threshold in 42254 epochs.

The trained neural controllers were tested systematically using the *open-loop* procedure connecting the controller with the simulated robot. A test of the robot under the same conditions as experienced during the training process (i.e. with linguistic input and with/without roundness information) showed that the neural controllers were perfectly trained and were able to move the robot and emit the correct outputs in all training conditions (see condition *Ling* in **Figure 6** for a similar test performed with roundness information. Results without roundness are not shown since the networks were able to reproduce the recorded sequence with a negligible error). After this first test, more comprehensive generalization tests were performed in order to estimate the capability of the controller to cope with different conditions.

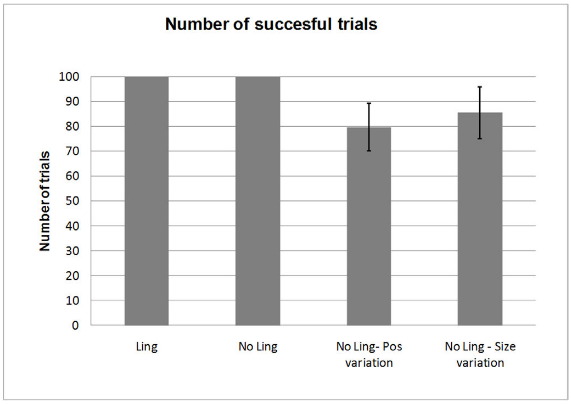


FIGURE 6 | Results of 100 trials from four different testing conditions. Standard deviation is represented as error bar. See text for details.

GENERALIZATION UNDER DIFFERENT CONDITIONS

For this analysis each of the 10 controllers were tested under four different conditions for 100 trials. The number of successful trials was recorded, i.e. the cases in which, at the end of 30 sensorimotor cycles of the neural network controlling the robot, the activation of the output units were the same as the desired output. Given the variation of the initial condition, a deviation of ±0.1 was allowed for every output unit. It should be noted that, given a certain degree of the error, the robot is not able to accomplish the task at all. The four testing conditions are as follows:

Ling. A condition identical to that of the training process, with the linguistic input provided at the beginning of the trial.

No Ling – In this condition the linguistic input is set to zero during the whole duration of the trial. The other parameters are the same of the training process.

No Ling – Pos variation. A condition in which the linguistic input is set to zero during the whole duration of the trial. In addition, at every trial the position of the object randomly varies within a range of ±10 cm.

No Ling – Size variation. A condition in which the linguistic input is set to zero during the whole duration of the trial. In addition, at every trial the global size of the object is randomly scaled within a range of ±20%. For instance, the diameter of the sphere can vary at each trial between a minimum of 9.6 cm and maximum of 14.4 cm.

Information about the roundness was available in all conditions.

Results of the tests are depicted in **Figure 6**. The conditions *Ling* and *No Ling* interestingly are exactly the same for all 10 replications. It should be noted that the neural networks have been trained always with linguistic input. This means that the natural generalization capability of the network is able the reconstruct the input pattern, including the linguistic input, without any loss in terms of performance. In the *No Ling* condition, the robot is placed in front of the object and performs its movement toward it without any linguistic input; still, it is able to produce the correct linguistic pattern after the interaction with the object. This result indicates that the controller can recall and produce the appropriate linguistic output only on the basis of its overall sensory state. *Pos variation* produces slightly worse results and we observe a certain variation among the replications, as indicated by the standard deviation on the graph. It is interesting to note that the worst replication is the one which took more epochs to converge, whilst the best is the one that converged in the fewer number of epochs. Besides the performance decrement, the majority of the replications shows a very high generalization capability, even though in the allowed range of the variation. The same can be observed for the *Size variation* condition, although the results appear slightly better. This can be explained by the fact that the roundness information is, to a certain extent, independent from size variations. Therefore, roundness provides a reliable source of information even in cases of unexpected sensory-motor input in comparison with that experienced during training. This effect has been observed in connection with larger objects.

The tests presented above demonstrate that the neural controller is capable to produce the correct behavior in terms of joint activations and prediction of sensorimotor states, as well as in terms of linguistic activations. Moreover, the correct behavior is performed also without providing a linguistic input. This is also true when the set-up is manipulated to a certain extent.

Nevertheless, this kind of test does not allow us to understand what exactly the information is that is used by the controller to connect an object with its corresponding linguistic label. In order to clarify this issue, additional tests have been performed.

UNDERSTANDING THE MEANING OF WORDS

Further tests and analyzes were carried out in order to better understand the meaning associated to the linguistic labels, which we can imagine as a kind of simplified words, and the relation between the sensorimotor processes triggered by robot-object interactions and these arbitrarily provided words. To analyse the word–meaning mappings that emerge from the current experimental set-up, the dynamics of the activations of the linguistic units and the roundness prediction have been analysed under several conditions in which additional input modifications are explored.

As we stated above, so far we cannot properly link an observed linguistic activation with a particular word. In fact, we still ignore the specific relation between words and meanings created by the controller. Therefore, in the following tests and analyzes we will refer to the linguistic output in very general terms. We will apply a specific word associated to the linguistic output only when the relation between them and their referents will be clarified. The notation for linguistic output identification, already introduced in Section "Training phase", is the following:

linguistic_output_1: correlates with the *robot-sphere* interaction;
linguistic_output_2: correlates with the *robot-cube* interaction;
linguistic_output_3, correlates with the *robot-cylinder* interaction.

Given the differences among the neural controllers in terms of synaptic weights and overall dynamics, the following additional analyzes were carried out using a single controller. The controller of replication 2 was chosen because it demonstrated to be the best one in the previous generalisation tests.

Tests on linguistic outputs

In this test, the robot was placed in front of an object without providing any linguistic input yet with roundness information. During the interaction of the robot with the object, the activations of the linguistic output have been recorded for the usual 30 sensorimotor cycles allowed (15 s). As it is shown in **Figure 7** (left column), linguistic activations vary greatly as the interaction unfolds in time, depending on the object. In the case of the sphere, the roundness provides information that immediately permits *linguistic_output_1* activation, as correctly required by the task. However, after about 3 s, the hand of the robot gets in contact with the object, and for a while the linguistic output changes by activating *linguistic_output_2*, which correlates with robot interacting with the cube. However, as the interaction continues and the sphere rolls away, the robot is then able to produce again the right output pattern after some time. In the graph we can also observe the time dynamics of the roundness prediction, which is affected, like the linguistic output, by the overall sensorimotor state of the robot. The perceived roundness for the sphere is about 0.9, which is correctly predicted by the robot at the beginning. Nevertheless, after the activation of the touch sensor, the prediction switches from "sphere" (~0.87) to "cube" (~0.71), although in input the robot still receive the correct roundness. Finally, the roundness measure returns to the right value after the sphere begins to roll away from the hand.

The dynamics that we observe while the robot is interacting with the cube and the cylinder is very similar. Roundness information, in fact, allows an early recognition of the type of object and the production of the correct linguistic output pattern, i.e., *linguistic_output_2* for the robot-cube interaction and *linguistic_output_3* for the robot-cylinder interaction. In case of the cube, after the contact of the robot with the object, the correct *linguistic_output_2* is triggered and it remains active for the rest of the time. Only a minor activation of the *linguistic_output_3*, related to the cylinder, is observed just after the contact. As for the cylinder, the correct output is emitted by the star and after a brief interference of the *linguistic_output_1*, the correct output, i.e. *linguistic_output_3*, is activated and maintained.

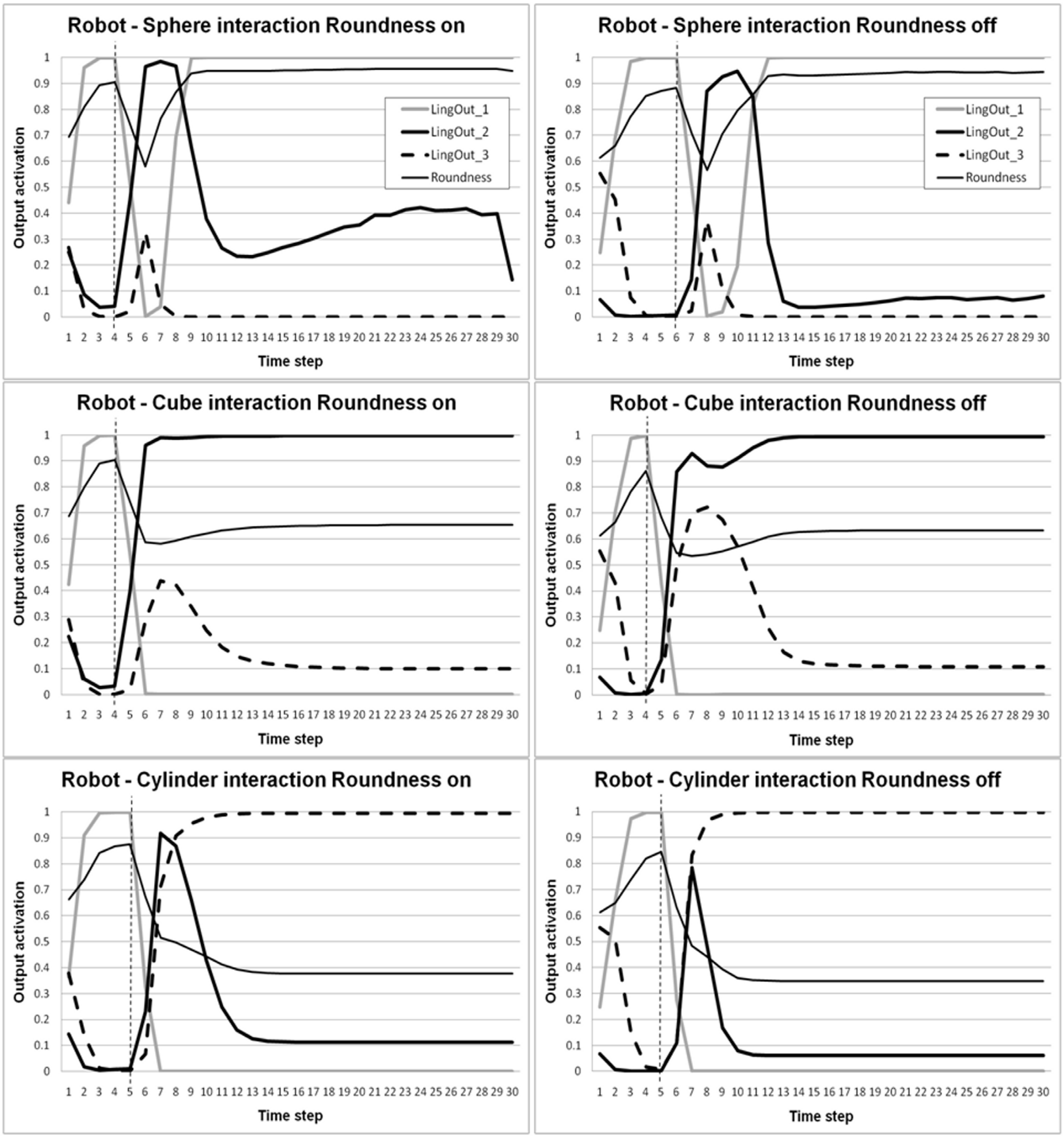


FIGURE 7 | Activations of linguistic output units during the 15-s interaction between the robot and the object on the desk. The graph shows *linguistic_output_1* (interaction with the sphere), *linguistic_output_2* (interaction with the cube), *linguistic_output_3* (interaction with the fixed cylinder), and the roundness prediction. Refer to the legend on the graph. The thin dashed vertical line in the three graphs represents the moment in which the robot's arm touches the object. Left: Condition with roundness information in the input. Right: condition without roundness information in the input.

Figure 7 (right column) shows the same type of analysis as above for a condition in which neither linguistic input nor roundness information are available. This case is much more complex than before because at the beginning of the movement no external information is available. Not surprisingly, the dynamics is also different. When the robot is interacting with the sphere, a presumably stereotypic behavior of the chosen neural controller produces, at the beginning of the movement and without any other information about the object available, a default linguistic output activation, which is the correct one by chance, that is, *linguistic_output_1*. After the robot touches the sphere, similar to what we observed above, the correct output is suppressed and the one corresponding to the cube is activated. During the following interaction, *linguistic_output_2* is active to a smaller extent than in the case with roundness in the input (see corresponding graph on the left column). This effect is probably due to a kind of interaction between the roundness value provided in the input and the linguistic output activation. Roundness values for "sphere" and "cube" are indeed very close. When the robot is interacting with the cube, given the stereotypic behavior of the controller at the beginning, it activates *linguistic_output_1*. However, after the contact the cube slides on the surface and the correct output is produced. For the cylinder, the same interference as before between *linguistic_output_2* and *linguistic_output_3* is observed, but after few seconds the robot produces the correct output.

Generalization tests on linguistic outputs

Results presented so far clearly show that the linguistic input is tightly connected with the sensorimotor dynamics produced by the interaction with the object. The tests demonstrate the ability of the robot to correctly categorize the objects, also in the absence of direct linguist input, and to produce the corresponding linguistic label only on the bases of its sensorimotor state. From this point of view, the observed interaction between the flow of the sensorimotor states and the activations of the linguistic units leads to the hypothesis that the whole sensorimotor state, rather than a single elements such as, e.g., the roundness, is at the core of the controller's ability to categorize the events correctly. In this section, therefore, we performed an additional test to verify this hypothesis and to investigate what the real meaning is on which the linguistic labels are based.

The test consists of three different conditions in which the robot was tested. Again, no linguistic input is provided. The three conditions are the following:

(a) A cylinder very similar to the one used throughout the training process is placed in front of the robot. This time the cylinder is not attached to the table and is free to move. Its starting orientation is parallel to the starting position of the robot arm. Thus, it can roll away when the robot touches it (**Figure 8A** right). The roundness perceived by the robot is the same as for the cylinder.
(b) The same cylinder is placed on the table and free to move. The starting orientation is perpendicular to the robot arm. That is, it is rotated 90° with respect to the previous condition. In this position it can easily slide but not roll (**Figure 8B** right).
(c) A cube is fixed to the table. The perceived roundness is the same of the cube, as during the training, but the cube cannot move if touched (**Figure 8C** right).

Results are shown in **Figure 8**. **Figure 8A** (left) represents the interaction with rolling cylinder, showing that when the cylinder is touched, it starts to roll away. Given the specific sensorimotor dynamics produced by the cylinder, the related pattern dynamics of the linguistic output are very similar to that already seen for the robot-sphere interaction. We may thus conclude that the robot categorizes and labels the rolling cylinder as it categorizes the sphere by activating *linguistic_input_1*. Similarly, the interaction with the sliding cylinder (**Figure 8B** left), given the fact that it slides and produces the same sensorimotor patterns previously seen for the cube, induces the controller to activate *linguistic_input_2*. Finally, **Figure 8C** (left) depicts the linguistic output activations while the robot interacts with the fixed cube. Even though its dimension and perceived roundness are exactly the same as for the cube used during training, the sensorimotor contingency produced by the interaction is identical to that experienced with the cylinder in the training. Not surprisingly, the linguistic activation is the same observed for the cylinder in the previous test: *linguistic_input_3*.

These additional results suggest that the linguistic label are grounded in complex sensorimotor dynamics instead of in the visual features provided by the roundness parameter, despite the fact that roundness information is provided. Specifically, the grounding of the linguistic output can be identified with the dynamics associated to the physical properties of an object.

DISCUSSION AND CONCLUSION

What has been shown so far indicates that the robot is able to extract the sensorimotor contingency of a particular interaction with an object and to reproduce its dynamics by acting on the environment. Moreover, in the absence of linguistic input, the robot is capable of associating a certain temporal sensorimotor dynamics to the learnt linguistic labels. Thus, in the lights of the results provided by the tests, it is now time to ask whether the linguistic label learnt are associated to the objects themselves or whether the label refers to the physical properties of the objects.

The results presented in Section "Tests on linguistic outputs" on the activation of the linguistic output during robot-object interaction, clearly show that the robot is able to correctly categorize and produce the correct label for a given object both in absence of the corresponding linguistic input and roundness information. Moreover, the generalization tests presented in Section "Generalization tests on linguistic outputs" indicate that, irrespective of the roundness information provided and in absence of linguistic input, the linguistic output correlates with the sensorimotor dynamics produced by the specific physical property of the object. Therefore, such results suggest that the linguistic labels are based on an entire sensorimotor dynamics, and not on the visual features provided by the roundness parameter. Specifically, the grounding is exactly the dynamics associated to that physical property.

This interpretation is corroborated by other works that connect active perception with language emergence. For example, in Marocco et al. (2003) (a study based on the previous work on active perception by Nolfi and Marocco, 2002) the evolved robot showed a stereotypic behavior towards the object, which allowed the robot to discover the physical properties of that object and then to categorize it to apply the correct linguistic label. The case we are presenting here is similar to Marocco et al. (2003) in many respects. The iCub robot interacts with the objects in a very stereotypic way and the stratagem by means of which the typology of the object is discovered is based on an active sensorimotor strategy which, given the same exploratory behavior, produces different outcomes. We can therefore speculate that the grounding of the linguistic label is not the actual object, but rather the physical property that allows the object (the patient in this case), when manipulated in a certain way, to produce a specific sensorimotor contingency in the agent.

At this point of the discussion we can definitely affirm that, given the present experimental set-up, the meaning of the labels are not associated to a static representation of the object, but to its dynamical properties. It seems, therefore, that the label that we called *linguistic_output_1* (*[1 0 0]*), is related to the rolling of an object, or in general, to those objects that, when touched, move away from the agent. A corresponding tentative word for this can be "the rolling one", more than "sphere". Similarly, *linguistic_output_2* (*[0 1 0]*) seems to be connected with the affordance of an object

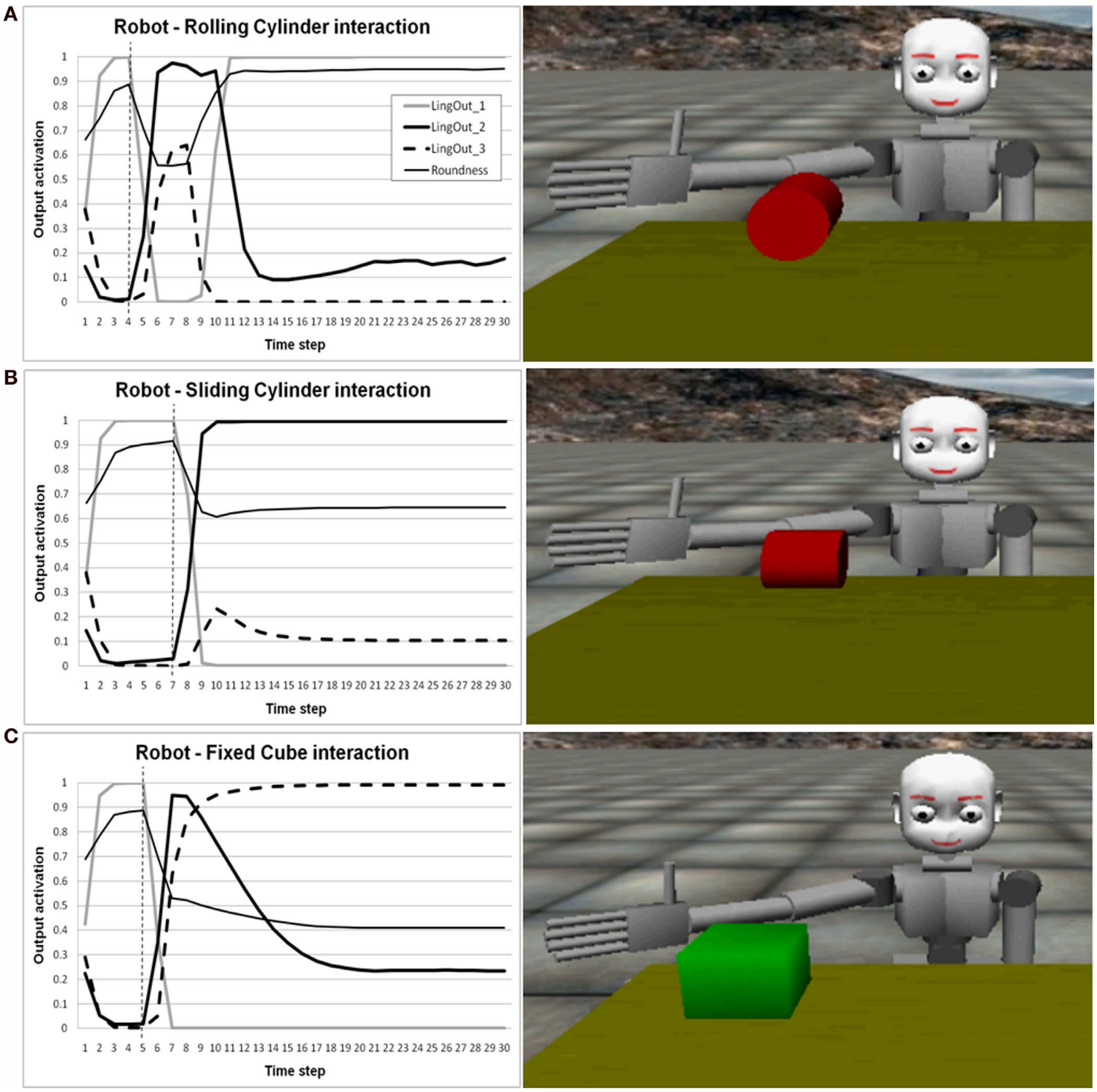


FIGURE 8 | Right column: The three novel conditions used for testing the generalization of the linguistic output units. In the case depicted in **(A)** the cylinder can roll away after being touched by the robot arm. In **(B)** the cylinder tends to slide while in contact with the arm, and in **(C)** the cube is fixed on the table and cannot be moved by the robot. Left column: Activations of linguistic output units during the 15-s interaction between the robot and the object on the desk. The graph shows *linguistic_output_1* (interaction with the rolling cylinder), *linguistic_output_2* (interaction with the sliding cylinder), *linguistic_output_3* (interaction with the fixed cube), and the roundness prediction. Refer to the legend in the graph. The thin dashed vertical line in the three graphs represents the moment in which the robot's arm begins to touch the object.

that can be moved by the agent, for instance, sliding on a surface. Therefore, an appropriate word for this can be "the sliding one" or "the one that slides", rather than "cube". This, indeed, is activated by an object that needs a continuous force applied to it in order to move. *Linguistic_output_3* (*[0 0 1]*), on the other hand, is connected to a fixed object, that is, to an object that does not change its position in space when touched. A word counterpart for this can be "the fixed one". It is interesting to note that "fixed" is not an action. However, it is exactly the property of being fixed (not movable) that, by preventing the robot to accomplish its intended movement, produces the specific sensorimotor contingency that allows the robot controller to identify that particular physical property. All these properties are represented by the control system in terms of the effects produced by the robot itself and dependent on a self-generated movement.

Thus, it appears that in the interpretation of a given dynamic, the robot learns some property related to the force dynamics between objects. This is consistent with Talmy's (1988) cognitive linguistic analysis of the grounding of language in temporal events and, implicitly, of the grounding of action words that describe those events. This concept was explicitly used by Siskind in its software model *LEONARD*, briefly described in the introduction. However, the main difference between our model and Siskind's work is that in Siskind (2001) the force dynamic rules are explicitly embedded in the perceptual system, as well as the events that the system can recognize, e.g. *PICK-UP* or *MOVE*. In a later work (Fern et al., 2002), Siskind added a learning rule to his system that allows *LEONARD* to learn any kind of dynamic events that are shown to the camera. It should be noted, however, that basic force dynamics rules, called *states* by the authors, such as *CONTACTS*, *SUPPORTS* or *ATTACHED* are still predefined by the experimenter. This new model allows *LEONARD* to learn the temporal sequences of *states* observed in a dynamic event. Therefore, any kind of events can be recognized on the basis of predefined force dynamics states.

A similar consideration can be raised with respect to the work by Baillie and Steels (2003). In their model, events are based on a set of predefined detectors in interaction with a kind of top-down reasoning system that allows an agent to create an internal representation of the external world. This internal world, in turn, is the actual grounding of the utterance produced.

In contrast, in our model the robot is able to capture the essence of certain interactions between objects (i.e. its hand and the object of the desk) and to create an embodied representation of those interactions autonomously. Moreover, the embodied knowledge is implemented in the neural control system as specific dynamical patterns of sensorimotor contingencies and does not require an explicit internal representation of the external word.

Yet the results presented here are in line with the work by Cohen et al. (2005) and Cannon and Cohen (2010) on the grounding of action words. In their work, they refer to the concept of energy transfer between agents, which is to some extent connected to the idea of the force dynamics. In our model, we can analyse the way in which the robot categorizes the events in terms of Cannon and Cohen concepts. However, the main difference is that in their work they only refer to visual stimuli, while in our model the grounding of the action words we discovered is deeply rooted in the integration of many sensor stimuli, both visual and proprioceptive.

From a linguistic perspective, the results obtained are also useful for understanding the acquisition of word meanings by young children. That is, the sensorimotor experience of the robot, rather than visual properties of the objects or the linguistic labels used, constitutes the basis for categorization. The word learning by the robot, therefore, depends on the way in which an object behaves when it is manipulated under certain conditions, rather than on its appearance. The results obtained thus support approaches to word meaning that focus on the role of functional affordances of objects in interaction (Clark, 1973; Nelson, 1973; Mandler, 1992). In particular, the meaning of "ball" is not defined on the basis of its perceptual appearance (its roundness in our case) but on its property to roll. "Sphere" and "cylinder", in contrast, are grouped into the same category because of the action-based sensorimotor categories created by the robot during the training. Thus, what we found the robot to do corresponds to what Mandler (1999, p. 305) suggests for infant word learning:

Infants are attracted by and interested in moving objects from birth. Moving objects are the basis of events, which is what infants attend to, and, according to my theory, it is attended events that get analysed into the first conceptual meanings (Mandler, 1992). Understanding events is absolutely central to conceptual life, and it would be surprising indeed if even infants did not have the capacity to generalize across them. It appears, however, that this kind of generalization is more abstract than seeing the commonalities among cats or chairs. It involves generalizing the common roles that different categories of objects play (for example, animals pick-up objects, artifacts get picked up) and this rather abstract understanding forms the basis on which more detailed, concrete understanding of who does what to whom will develop. (Mandler, 1999, p. 305)

Moreover, the research presented is also in line with the body of theoretical and empirical evidence grown in the past years in support of the role of embodiment and sensorimotor factors in language use (e.g. Barsalou, 1999; Glenberg and Robertson, 2000; Feldman and Narayanan, 2004; Gallese and Lakoff, 2005; Pecher and Zwaan, 2005), yet with different perspective. Barsalou (1999), for example, focuses on modality-specific perceptual and simulation processes within the Perceptual Symbol System hypothesis, based on experiences of sensorimotor, proprioceptive and introspective events, and also dynamic mental representations of object interaction. Glenberg and collaborators (e.g. Glenberg and Kaschak, 2002) focus on the action and embodiment component of language by demonstrating the existence of action-sentence compatibility effects that support an embodied theory of meaning that relates the meaning of sentences to human action and motor affordances. The shared aim of these studies is to demonstrate that language processes cannot be fully understood without taking into account an embodied perspective.

Thus, the principal contribution of the results presented with respect to the current literature concerns the computational feasibility of grounding of action words directly in the way in which an agent interacts with the environment and manipulates it. The dynamical properties of external objects, such as being movable, or being fixed, are embodied and directly represented in the way in which the agent experiences the reactions produced from its own, self-generated, active manipulation of the world on its perceptual system. This mechanism has two related, desired side-effects: (a) a word in the input produces the activation of a whole sensorimotor process and, conversely, (b) the experience of a given sensorimotor contingency recalls in the robot controller the associated word. Therefore, the model shows that meanings related to dynamical properties of external objects, such as *roll* or *fix*, can be fully grounded in the embodied experience of the robot. From this perspective the activity performed by the robot itself is the key that allows to uncover the properties of the objects by means of physical interactions.

These findings confirm and extend the large body of work on computational and robotics models that focuses on the sensorimotor bases of language acquisition. In particular, as we have highlighted in the introduction, this is partly due to shifting the attention from actions that are performed by the agent, or the robot, to actions or properties that relate to external objects.

We conclude by acknowledging that this work, as we have already mentioned in the Section "Materials and Methods", has been carried out in simulation. We do not claim here that all the work done can be easily transferred onto the real robot, as we are aware of the difference between simulation and reality. There are, indeed, many reasons that justify this choice. The most important of them concerns the practical difficulties of carrying out a large number of experiments by using a real robotic platform and the fact that the iCub available to us is currently not equipped with touch sensors on the body and compliant motors on the arms, which would allow it to cope with rigid and fixed objects, such as the cylinder stuck on the desk, without breaking the robot. Nevertheless, following Ziemke's (2003) consideration, we believe that robot simulations play an important role in cognitive embodied simulations, although results may be less relevant from an engineering point of view (see also Tikhanoff et al., 2008). Therefore, given the modeling purpose of the present work, the fact that experiments have been carried out in simulation should not diminish the scientific relevance of the results achieved.

ACKNOWLEDGMENTS

This work was supported by the European Commission FP7 Project ITALK (ICT-214668) within the Cognitive Systems and Robotics unit (FP7 ICT Challenge 2) and Apple Research and Technology Support (ARTS).

REFERENCES

Bailey, D., Feldman, J., Narayanan, S., and Lakoff, G. (1997). "Modeling embodied lexical development," in *Proceedings of the 19th Cognitive Science Society Conference,* Mahwah, NJ: Erlbaum.

Barsalou, L. (1999). Perceptual symbol systems. *Behav. Brain. Sci.* 22, 577–609.

Beira, R., Lopes, M., Prac, M., Santos-Victor, J., Bernardino, A., Metta, G., Becchiz, F., and Saltar, R. (2006). Design of the robot-cub (iCub) head. *Proc. IEEE Int. Conf. Robot. Autom.* 94–100.

Bergen, B. (2005). "Mental simulation in literal and figurative language," in *The Literal and Non-Literal in Language and Thought*, eds S. Coulson and B. Lewandowska-Tomaszczyk (Frankfurt: Peter Lang), 255–278.

Bowerman, M., and Choi, S. (2001). "Shaping meanings for language: universal and language-specific in the acquisition of semantic categories," in *Language Acquisition and Conceptual Development,* eds M. Bowerman and S. C. Levinson (Cambridge: Cambridge University Press), 475–511.

Bowerman, M. (2005). "Why can't you "open" a nut or "break" a cooked noodle? Learning covert object categories in action word meanings," in *Building Object Categories in Developmental Time*, eds L. Gershkoff-Stowe and D. H. Rakison (Mahwah, NJ: Erlbaum), 209–243.

Cangelosi, A., and Parisi, D. (2002). *Simulating the Evolution of Language.* London: Springer.

Cangelosi, A., and Riga, T. (2006). An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. *Cogn. Sci.* 30, 673–689.

Cannon, E. N., and Cohen, P. R. (2010). "Talk about motion: the semantic representation of verbs by motion dynamics," in *The Spatial Foundations of Cognition and Language: Thinking Through Space,* eds K. S. Mix, L. B. Smith and M. Gasser (New York: Oxford University Press), 235–258.

Clark, H. H. (1973). "Space, time, semantics and the child," in *Cognitive Development and the Acquisition of Language*, ed. T. E. Moore (New York: Academic Press), 27–64.

Cohen, P. R., Morrison, C. T., and Cannon, E. (2005). "Maps for verbs: the relation between interaction dynamics and verb use," in *Proceedings of the Nineteenth International Conference on Artificial Intelligence (IJCAI 2005).*

Dogar, M., Cakmak, M., Ugur, E., and Sahin, E. (2007). "From primitive behaviors to goal-directed behavior using affordances," in *IEEE/RSJ International Conference on Intelligent Robots and Systems.*

Dominey, P. (2005). Emergence of grammatical constructions: evidence from simulation and grounded agent experiments. *Connect. Sci.* 17, 289–306.

Feldman, J., and Narayanan, S. (2004). Embodied meaning in a neural theory of language. *Brain Lang.* 89, 385–392.

Fern, A. P., Givan, R. L., and Siskind, J. M. (2002). Specific-to-general learning for temporal events with application to learning event definitions from video. *J. Artif. Intell. Res.* 17, 379–449.

Fitzpatrick, P., Metta, G., Natale, L., Rao, S., and Sandini, G. (2003). "Learning about objects through action: initial steps towards artificial cognition," in *IEEE International Conference on Robotics and Automation* (Taipei, Taiwan).

Fritz, G., Paletta, L., Breithaupt, R., Rome, E., and Dorffner, G. (2006). "Learning predictive features in affordance based robotic perception systems," in *IEEE/RSJ International Conference on Intelligent Robots and Systems.* (Beijing, China).

Gallese, V., and Lakoff, G. (2005). The brain's concepts: the role of the sensory-motor system in reason and language. *Cogn. Neuropsychol.* 22, 455–479.

Gelman, S. A. (2009). Learning from others: children's construction of concepts. *Annu. Rev. Psychol.* 60, 115–140.

Gelman, S. A., and Heyman, G. D. (1999). Carrot-eaters and creature-believers: the effects of lexicalization on children's inferences about social categories. *Psychol. Sci.* 10, 489–493.

Glenberg, A. M. (2007). "Language and action: creating sensible combinations of ideas" in *The Oxford handbook of psycholinguistics*, ed. G. Gaskell (Oxford, UK: Oxford University Press), 361–370.

Glenberg, A. M., and Kaschak, M. P. (2002). Grounding language in action. *Psychon. Bull. Rev.* 9, 558–565.

Glenberg, A. M., and Robertson, D. A. (2000). Symbol grounding and meaning: a comparison of high-dimensional and embodied theories of meaning. *J. Mem. Lang.* 43, 379–401.

Halliday, M. A. K. (1985). *An introduction to Functional Grammar.* London: Edward Arnold.

Harnad, S. (1990). The symbol grounding problem. *Physica D* 42, 335–346.

Hatch, E. (1983). Psycholinguistics: a second language perspective. Rowley, MA: Newbury House.

Kaplan, F., Oudeyer, P.-Y., and Bergen, B. (2008). Computational models in the debate over language learnability. *Infant Child Dev.* 17, 55–80.

Karmiloff, K., and Karmiloff-Smith, A. (2001). Pathways to language: from fetus to adolescent. Cambridge, MA: Harvard University Press.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Trans. Evol. Comput.* 5, 102–110.

Kozima, H., Nakagawa, G., and Yano, H. (2002). "Emergence of imitation mediated by objects," in *Second International Workshop on Epigenetic Robotics* (Edinburgh, Scotland).

Lakoff, G., and Johnson, M. (1999). *Philosophy In The Flesh: the Embodied Mind and its Challenge to Western Thought.* New York: Basic Books.

Langacker, R. W. (2008). *Cognitive Grammar*. New York: Oxford University Press.

Lyon, C., Nehaniv, C. L., and Cangelosi, A. (2007). *Emergence of Communication and Language*. London: Springer.

Mandler, J. M. (1992). How to build a baby: II. Conceptual primitives. *Psychol. Rev.* 99, 587–604.

Mandler, J. M. (1999). Seeing is not the same as thinking: commentary on "making sense of infant categorization". *Dev. Rev.* 19, 297–306.

Marocco, D., Cangelosi, A., and Nolfi, S. (2003). The emergence of communication is evolutionary robots. *Philos. Trans. R Soc. Lond. A* 361, 2397–2421.

Metta, G., Fitzpatrick, P., and Natale, L. (2006). YARP: yet another robot platform. *Int. J. Adv. Robot. Sys.* 3, 43–48.

Metta, G., Sandini, G., Vernon, D., Natale, L., and Nori, F. (2008). "The iCub humanoid robot: an open platform for research in embodied cognition," in *Proceedings of IEEE Workshop on Performance Metrics for Intelligent Systems Workshop (PerMIS'08)*, eds R. Madhavan and E. R. Messina (Washington, DC: IEEE).

Montesano, L., Lopes, M., Bernardino, A., and Santos-Victor, J. (2008). Learning object affordances: from sensory motor maps to imitation. *IEEE Trans. Robot.* 24, 15–26.

Nelson, K. (1973). Some evidence for the cognitive primacy of categorization and its functional basis. *Merrill Palmer Q.* 19, 21–39.

Nolfi, S., and Marocco, D. (2002). "Active perception: a sensorimotor account of object categorization," in *From Animals to Animats 7 – The Seventh International Conference on the Simulation of Adaptive Behavior*, eds B. Hallam, D. Floreano, J. Hallam, G. Hayes and J.-A. Meyer (Cambridge: MIT Press), 266–271.

Oudeyer, P.-Y., and Kaplan, F. (2006). Discovering communication. *Connect. Sci.* 18, 189–206.

Oudeyer, P.-Y, Kaplan, F., and Hafner, V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.* 11, 265–286.

Pecher, D., and Zwaan, R. A. (2005). *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking.* Cambridge, UK: Cambridge University Press.

Pulvermüller, F., Haerle, M., and Hummel, F. (2001). Walking or talking? Behavioral and neurophysiological correlates of action verb processing. *Brain Lang.* 78, 134–168.

Roy, D., Hsiao, K., and Mavridis, N. (2003). Conversational robots: building blocks for grounding word meanings. In *Proceedings of the HLT-NAACL03 workshop on learning word meaning from non-linguistic data*.

Rumelhart, D. E., and McClelland, J. L. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. I*. Cambridge, MA: MIT Press.

Sachs, J. (1983). Talking about the there and then: the emergence of displaced reference in parent-child discourse. In *Children's language, Vol. 4*, ed. K. E. Nelson (Hillsdale, NJ: Erlbaum).

Siskind, J. M. (2001). Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *J. Artif. Intell. Res.* 15, 31–90.

Smith, L. B., Jones, S. S., and Landau, B. (1996). Naming in young children: a dumb attentional mechanism? *Cognition* 60, 143–171.

Snow, C. E. (1977). "Mothers' speech research: from input to interaction," in *Talking to Children: Language Input and Acquisition,* eds C. E. Snow and C. Ferguson (Cambridge: Cambridge University Press), 31–49.

Steels, L. (2001). Language games for autonomous robots. *IEEE Intell. Syst.* 16, 16–22.

Steels, L. (2003). Evolving grounded communication for robots. *Trends Cogn. Sci. (Regul. Ed.)* 7, 308–312.

Steels, L., and Baillie, J.-C. (2003). Shared grounding of event descriptions by autonomous robots. *Rob. Auton. Syst.* 43, 163–173.

Steels, L., and Kaplan, F. (2000). AIBO's first words: the social learning of language and meaning. *Evol. Commun.* 4, 3–32.

Steels, L., Kaplan, F., McIntyre, A., and Van Looveren, J. (2002). "Crucial factors in the origins of word-meaning," in *The Transition to Language* ed. A. Wray (Oxford: Oxford University Press), 252–271.

Stoytchev, A. (2005). "Behavior-grounded representation of tool affordances," in *International Conference on Robotics and Automation* (Barcelona, Spain).

Sugita, Y., and Tani, J. (2005). Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. *Adap. Behav.* 13, 211–225.

Talmy, L. (1988). Force dynamics in language and cognition. *Cogn. Sci.* 12, 49–100.

Tikhanoff, V., Cangelosi, A., Fitzpatrick, P., Metta, G., Natale, L., and Nori, F. (2008). "An open-source simulator for cognitive robotics research: the prototype of the iCub humanoid robot simulator," in *Proceedings of IEEE Workshop on Performance Metrics for Intelligent Systems Workshop (PerMIS'08)*, eds R. Madhavan and E. R. Messina (Washington, DC).

Tsagarakis, G., Metta, G., Sandini, D., Vernon, R., Beira, F., Becchi, L., Righetti, J., Santos-Victor, J., Ijspeert, A. J., Carrozza, M. D., and Caldwell, D. G. (2007). iCub: the design and realization of an open humanoid platform for cognitive and neuroscience research. *Adv. Robot.* 21, 1151–1175.

Veneziano, E. (2001). Displacement and informativeness in child-directed talk. *First Lang.* 21, 323–356.

Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proc. IEEE* 78, 1550–1560.

Wierzbicka, A. (1985). *Lexicography and Conceptual Analysis*. Ann Arbor: Karoma.

Williams, R. J., and Zipser, D. (1995). "Gradient-based learning algorithms for recurrent networks and their computational complexity," in *Backpropagation: Theory, Architectures, and Applications*, eds Y. Chauvin and D. E. Rumelhart (Mahwah, NJ: Lawrence Erlbaum Associates), 433–486.

Yamashita, Y., and Tani, J. (2008). Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment. *PLoS Comput. Biol.* 4, e1000220. doi:10.1371/journal.pcbi.1000220

Yu, C. (2005). The emergence of links between lexical acquisition and object categorization: a computational study. *Connect. Sci.* 17, 381–397.

Ziemke, T. (2003). On the role of robot simulations in embodied cognitive science. *AISB J.* 1, 389–399.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 December 2009; paper pending published: 24 February 2010; accepted: 30 April 2010; published online: 31 May 2010.

*Citation: Marocco D, Cangelosi A, Fischer K and Belpaeme T (2010) Grounding action words in the sensorimotor interaction with the world: experiments with a simulated iCub humanoid robot. Front. Neurorobot. **4**:7. doi: 10.3389/fnbot.2010.00007*