The Necessity of Enactive Simulation for Linguistic Comprehension

Acquiring Multi-Modal Semantic Representations

Workshop on Cognitive Architectures for Human-Robot Interaction

Ameer Sarwar
University of Toronto
ameer.sarwar@mail.utoronto.ca

ABSTRACT

This article argues that enactive simulation—the idea that integrated multi-modal representations based on an embodied view of mind-world interaction—is necessary for linguistic comprehension within the modality-specific theoretical framework. Amodal and module-specific theories take language to causally impact categorization, and therefore, theoretical and empirical differences about language comprehension directly impact categorization. As such, empirical psychological evidence is presented, and it is argued that enactive simulation under the broad module-specific framework best accounts for it. A counterargument is evaluated, but it is concluded that it fails to show that enactive simulation is not necessary for linguistic comprehension. The paper concludes by highlighting some directions for research in human-robot natural language interaction.

KEYWORDS

Enactive Simulation; Linguistic Comprehension; Perceptual Symbols; HRI; Integration

The thesis of this article is that enactive simulations under the module-specific framework are necessary for linguistic comprehension. I first present a brief history of seminal approaches to categorization, followed by an account of some unique features of language that any theory must explain. I then present two major opposing frameworks—amodal and module-specific theories. Because they view categorization as a consequence of linguistic comprehension, I review empirical psychological evidence to argue that the latter framework provides the best explanation. Later, I consider a counterargument but show that it manages to refute only a stronger version of my thesis. I conclude by highlighting some implications of my arguments for human-robot interaction with respect to natural language comprehension.

1 A *VERY* BRIEF HISTORY OF CATEGORIZATION

The classical notion of what a 'category' is can be traced back to Plato and Aristotle. This idea entailed that categories are a set of individually necessary and jointly sufficient conditions that delimit its membership. Here, the category is the general class to which its constituents, known as elements, belong. The category of *rabbit*, for instance, includes all and only instances of rabbits. A negative characterization of this example would be that there are no rabbits that do not belong to the category and no non-rabbits that do belong to it. This idea was presupposed up until the 1950's.

Wittgenstein was among the first to raise a major challenge against this understanding. He contended that it is impossible to procure an exhaustive list of individually necessary and jointly sufficient conditions that circumscribe a category [31]. His infamous example is that of a *game*: what conditions are necessary (and, therefore, common) among all games? What conditions are sufficient to make something a game? What are all of the necessary and sufficient conditions that put archery, snooker, and swimming in the category of the *game*? Having rejected this idea, Wittgenstein postulated that 'family resemblance' best captures what we mean by the term 'category.' On this account, there are no strict limits such that all and only members of a category belong to it. Instead, one could here think of a category as positing a relation of similarity or gradation between members of a family [31]. For instance, children may be similar to their parents in that the son may have dad's nose and mom's eyes, whereas the daughter may have dad's hair and mom's ears. So, there are no strict or logical basis for determining category membership; it is rather determined according to a relation of graded similarity between other members in the same category.

Building on Wittgenstein's insights by incorporating empirical findings, Rosch assented that concepts do not adhere to the classical, philosophical notions of categories. They instead exhibit what she called the 'prototypical effects': some members are quintessentially part of the category while others are less so [23]. For example, a sparrow is a more typical member of the category of *bird* than is a penguin. Likewise, a 'typicality gradient' provides, as it were, a hierarchy of membership in a category as a function of any given member's divergence from the ideal member. The closer to the ideal a member is, the more of a member of the category it is (e.g., robin vs. ostrich; apple vs. tomato). Rosch's findings are empirically corroborated and, perhaps, one may say that they are for the most part descriptively accurate as opposed to the seemingly normative undertones that characterize the classical conception. (Other important views include the 'exemplar view' and 'theory-theory view'; see [20] for a detailed review.)

2 LANGUAGE: SOME ESSENTIAL FEATURES

While the nature, direction, and degree of causal influence that language has on thought, and by implication on categorization, is contested, it is uncontroversial that the two are intimately connected.

It is a well-known fact in linguistics that language is *productive*, namely a speaker may utter infinitely many sentence by combining a finite number of simpler constituents. A typical example is writing infinitely many paragraphs using the same twenty-six letters or a finite number of words. The fact about the productivity of language is often explained in terms of its *compositionality*. Here, language is composed of primitive or atomic tokens that are combined in accordance with recursive syntactical rules to engender complex tokens [6]. The atomic tokens may be placed together in a multitude of different ways to generate a plethora of complex tokens (e.g., paragraphs using letters). Therefore, the seemingly endless productivity of language is explained in terms of its compositionality—using simple tokens to make complex ones in accordance with syntactical rules.

Related to the idea of productivity and compositionality is the understanding of our concepts as *generative*, viz. we are capable of understanding a large number of thoughts despite the fact that our cognitive capacities are finite [9]. In the same vein, (linguistic) cognition is *systematic*: the ability to understand the semantics of one sentence comes with the ability of understanding the same semantics despite differences in the surface structure of another sentence. For instance, the sentence "Tom chases Jerry" is understood just as well as the sentence "Jerry is chased by Tom." There is, in other words, semantic symmetry between the two sentences even though they have different surface syntactical forms. It is clear, then, that compositionality of language, which transforms primitive symbols into strings of primitive symbols (thus producing complex symbols) using recursive syntactical rules, can explain the productivity, generativity, and systematicity of (linguistic) cognition. It is unsurprising that these frameworks, which were mostly proposed around and after the 1950's, have proved extremely successful.

3 LANGUAGE AND CATEGORIZATION: TWO OPPOSING THEORETICAL FRAMEWORKS

As previously mentioned, a 'category' is broadly construed as a type or class containing relevantly similar[1] items or elements. The category of *fruit* may have apples, bananas, and oranges as its elements. Because categorization and language comprehension are highly related, let us consider two prominent theories of language comprehension and their implications for categorization.

First are the amodal theories, which interpret the elements of a category independent of their unique, individual details [7, 8]. Mental representation of elements on this account is in an idealized form rather than in the form of an instantiation of any given element. For instance, the color differentials between apples a_1 and a_2 may be "abstracted away" in their mental representation as elements of the category of *apple*. Second, the sensory modalities via which the percept is perceived are seen as secondary—whether one learns about the apple via olfactory, tactile, verbal, or visual modalities is not essential. What matters here is the ultimate, idealized mental representation of the element, not the sensory pathway through which one comes to learn about it. Although it is recognized that sensory information is first received by specific sensory modalities with unique neural pathways, this information is transduced into representations that are amodal, thus enabling people to have inferential capabilities, linguistic productivity, and thoughts [5]. Accordingly, the category of *apple* contains information not about apples a_1 and a_2 or whether one learned about them by seeing them or tasting them; instead, the category information is in the form of an abstracted/ idealized and amodal representation.

The amodality of category representation has two implications for our understanding of language. First, amodal theories explain language's productivity, namely its "ability to construct an unlimited number of complex representations from a finite number of symbols using combinatorial and recursive [syntactical] mechanisms" [3, 6]. This occurs because once we have idealized mental representations of the members of a category, they can be construed as nothing more than symbols or tokens that can be manipulated syntactically. This explains, for instance, the fact that one can use the term 'apple' to describe one's favourite fruit, cheapest item on a grocery list, or the item used in a particular dish—none of this would be possible if the mental representation was not abstract and idealized.

Second, it follows that words are seen as abstract, amodal, and arbitrary. The word chair refers to large and small chairs alike (abstract), it refers to chairs whether it is written down or spoken (amodal), and its phonemic and orthographic characteristics have no physical or functional resemblance to its referent (arbitrary) [27]. Consequently, amodal theories understand language as token manipulations prescribed over syntactic rules; the tokens or symbols here are abstract, amodal, and arbitrary. As the preceding paragraph shows, such manipulations are rendered impossible in cases where one focuses on some given instance of a category membership (i.e., *this* chair), as opposed to focusing on a category's members *sui generis*.[2]

On the other side of the aisle are module-specific theories. These, too, maintain that neural pathways first 'input' sensory information via specific sensory modalities. However, unlike in amodal systems, wherein sensory information is transduced to amodal representations, in module-specific systems this information is stored in neurons that are conjunctive (or adjacent) to the ones involved in receiving sensory stimuli [3, 5]. The conjunctive neurons store information in such a manner that, at a later time, faithful reenactments can be made for use in various mental acts such as language use. In order to allow for the possibility of faithful reenactments, the information is stored in module-specific manner. By implication, the particular details of sensory inputs are not abstracted but stay intact, thereby creating a dynamical relationship between referents (objects in the world) and sensory modalities (particularly, information in conjunctive neurons) to ultimately impact mental representation—therefore changes in

[1] Debates about categorization may be reduced to notions of relevance or similarity—in virtue of what is it the case that items *X* and *Y* both belong to category *Z*. Another important dimension is the agent that makes categorical judgments.

[2] There are crucial questions surrounding the semantics of tokens or symbols. Take a complex symbol *S*, which may be composed of conjunctions between a number of atomic tokens strung together using recursive syntactical rules. The meaning of *S* may depend entirely on the syntax used in composition (i.e., meaning is *in* the programme), it may inhere in primitive tokens, it may 'emerge' due to token-token or token-world relations, or we may have semantics as a result of a relationship between complex tokens and the world. For an influential critique that semantics do not inhere in symbols but are ascribed to them, see [25].

referents cause changes in mental representations [5].[3] Clearly, representations of elements in a category are not abstract and amodal, but they are essentially grounded in embodied experience. What is a representation on this account? The integration of multi-modal sensory input constitutes the representation of some object. In other terms, when perceptual knowledge from various sensory modalities is combined together, the resultant 'image' constitutes a mental representation. (This is often referred to as 'simulation.') For example, the multi-modal integration of a car may have the knowledge of its color from the visual modality, the information about its metallic structure from the tactile module, and so on. The integration of manifold module-specific sensory data lead to the concept of some object. As a result, the mechanism underlying categorization in module-specific framework is multi-modal integration of sense data of some individual element of a category.

A further dimension needs to be understood in order to fully appreciate the conceptualization of categorization within the module-specific framework. Specifically, objects here are construed as having *affordances*—the properties that make them useful or usable relative to some agent. For example, chair has the affordance of 'sit' relative to humans, but a door has this affordance neither for humans nor for zebras. There are interactions between the affordances (of referents) and the mental representations (of embodied agents) that allow the agent to categorize and act meaningfully in the world. Having a body that can interact via sensory and motor modalities with objects in the world enables one to categorize the objects in the manner previously delineated.

Barsalou [3] maintains that successful categorization of an entity makes the categorization of an identical or highly similar entity more rapid. The multi-modal integrated nature of representations of various items enable cognitive agents to create mental simulations that go beyond the finitude of their sensory experience. Accordingly, if the representations successfully capture the affordances of referents, then one can think of interacting with the objects even when they are not physically present [5]. This imaginative act of interaction between physical objects and embodied agents is called enactive simulation—it is 'enactive' because it is an interaction between an embodied agent and physical objects, and it is a 'simulation' because it mentally manipulates multi-modal representations.

4 EMPIRICAL EVIDENCE EVALUATES: WHICH THEORY BEST EXPLAINS THE FACTS?

I will now review some psychological evidence to argue for the necessity of enactive simulation for linguistic comprehension under the module-specific theoretical framework. Reddy, Tsuchiya, and Serre [21] found via functional magnetic resonance imagining (fMRI) that participants in group A imagining an object had neural activations identical with those in group B that saw the object's physical photo. This suggests that there could be module-specific neural firings (of conjunctive neurons) even in the absence of sensory stimuli, a prediction made by the module-specific theories. By comparison, amodal theories neither predict nor explain this finding because abstracted representation (without module-specific informational store in conjunctive neurons) by definition imply that these neurons no longer play a role. In other words, group A members, who were imagining an object, should not have had neural firings, but in fact they did.

Take another example in which participants were given verbal sentences that implied bodily movement. Reaction times for comprehension were faster if the movement implied by the sentences matched the movement one's body can potentially undergo (e.g., "pull the drawer open" implies *toward-agent* motion) than if the movement implied by the sentences was not possible (e.g., "push the drawer open" implies an *away-from-agent* motion) [17, 18]. This suggests two problems: (1) Enactive simulation appears necessary for verbal comprehension, because one has difficulty comprehending sentences implying motion that is bodily impossible. In particular, though neither of the sentences could be comprehended without enactive simulation, comprehension is particularly impaired when physical limitations of the body make enactive simulations impossible. According to amodal theories, (2) there should not be a discrepancy in reaction times, since the embodied interaction with the environment is not at all needed for comprehension. This experimental finding directly counters this thesis.

Furthermore, Kaschak and Glenberg [15] found that affordances most relevant for the comprehension of a given sentence had a faster reaction time than affordances that were not relevant for comprehension. In another experiment, researchers investigated the impact of the orientation of an object implied by a sentence (e.g., sentences "hammer the nail in the wall" and "hammer the nail on the floor" imply horizontal and vertical orientations of the nail, respectively) by later showing the participants a picture of the object and measuring reaction times. If the orientation of the object implied by the sentence and that shown in the picture matched, the reaction times were faster than if there was a mismatch between the implied orientation and one shown in the picture [27]. In a second experiment, Zwaan, Stanfield, and Yaxley [32] investigated people's visualizations of a sentence with respect to a given context. For instance, the sentence "the chef saw the egg in the fridge" almost invariably prompted the visualization of an unbroken, oval-shaped egg even though the word 'egg' itself does not in an *a priori* manner favour the broken-egg or the intact-egg interpretation.

How do the two theories interpret these results? On the one hand, the amodal theory, according to which linguistic comprehension does not require an embodied, dynamical interaction with the environment, cannot explain why the reaction time was faster when the orientation of the nail in the sentence and the picture were identical, as compared to when there was a mismatch. This thesis also cannot explain faster reaction times for comprehension when the affordances are relevant as compared to when they are not relevant, as well as why mental imagery of an object changes in light of changing contextual information (e.g., oval-shaped egg in the fridge vs. broken egg in a skillet). On the other hand, the module-specific thesis can easily account for these results, because

[3] The dynamical relationship described here cannot be accounted for by inherently discrete amodal systems. Because the latter do not change in light of changing environment, amodal accounts have difficulty explaining well-documented phenomena such as the formulation of *ad hoc* categories [2].

it posits an interaction between an embodied agent and an object with affordances relative to that agent, thus creating an interaction between the two that is modulated by sensorimotor capacities to create multi-modal internal representations for the agent. In other words, enactive simulation is needed for comprehending the affordances of objects, understanding their orientations, and visualizing their shapes in environmental contexts.

My assessment of the literature did not, of course, suggest that any and all evidence for embodiment refutes the amodal framework. In another experiment when the participants were presented with targets in an unexpected sensory modality, the reaction times were slower, but if the targets were presented in an expected sensory modality, the reaction times were faster [26]. Though the researchers take this as evidence against the amodal thesis, I take their conclusion to be too strong. Instead, I believe that amodal theories can account for this fact since it is only *after* module-specific sensory input is gained that it is transduced to amodal representations. In the experiment, however, the presentation of targets to some given sensory modality does not confute the amodal thesis, but it only shows that expectation of the appearance of a target at some sensory modality impacts the manner in which the participants interact with it. The temporal nature of the claims made by amodal theorists, namely that it is after gaining perceptual input from various modalities that abstract, amodal representations are made, vindicate them of this experimental result. (See [4] for a detailed review of empirical literature favouring this view.)

5 CRITICISMS AND RESPONSES

Arguing against the notion of module-specific, embodied conceptions of linguistic comprehension, Weiskopf [30] makes distinctions between the strong, medium, and weak enactive simulation hypotheses. The strong hypothesis claims that linguistic comprehension just *is* enactive simulation; here, linguistic understanding is constitutive of enactive simulation. The medium hypothesis holds that linguistic understanding requires but is not identified with enactive simulation; this means that enactive simulation is necessary for language comprehension, but the latter is not constitutive of the former. Finally, the weak hypothesis maintains that linguistic understanding can use but does not require enactive simulation, i.e., enactive simulation is neither necessary nor sufficient for linguistic comprehension; it may just be used as an auxiliary.

Weiskopf tries to show that only the weak hypothesis is plausible. Consider the sentences (1) “The man stood on the street corner” and (2) “The man waited on the street corner.” Weiskopf argues that both sentences employ the same enactive simulation—a man standing on the corner of a street—but maintains that this simulation does not enable us to adjudicate between the different meanings in the two sentences. In (2), there is an implication of intention since one waits *for* something, while in (1) one may stand idly. The meaning cannot be discriminated via enactive simulation so, argues Weiskopf, we should accept amodal conception of language comprehension (or weak simulation).

I agree with Weiskopf that enactive simulation is not sufficient for the adjudication of meaning, and therefore, this example counters the constitutive relationship between enactive simulation and linguistic comprehension suggested by the strong thesis. Indeed, I would further support his claim that causal coupling is not sufficient for establishing a relationship of constitution, as one observes from the causal impact of the circulatory system on the renal system, but this is not taken as the renal system being constitutive of the circulatory system [24]. However, his inference that this example also counters the medium thesis is mistaken. Recall that the medium thesis requires that enactive simulation be necessary for linguistic comprehension without maintaining relationships of sufficiency or constitution. Weiskopf does not show that one could understand *either* of the sentences without enactive simulation (of the man standing on the street corner). If he could demonstrate that comprehension of either of the sentences (let alone the problem of meaning adjudication) can take place without enactive simulation, then I will concede that the medium hypothesis is refuted. In other words, him showing that one could understand the sentence without simultaneously using enactive simulation would thereby counter the claim that enactive simulation is necessary for linguistic comprehension. Formally, if β is a necessary condition or component for *B*, then if there can even in principle be an instantiation of *B* without the presence of β, then it can be deductively concluded that β is not a necessary condition for *B*. In the cases that we are considering, Weiskopf must show that we could understand either of the sentences without enactive simulation; only such a demonstration would succeed in refuting the medium-strength hypothesis, namely that enactive simulation is necessary for, not constitutive of or sufficient for, linguistic comprehension.

A number of further points and clarifications are in order here. Showing that Weiskopf’s failure to refute the medium-strength enactive simulation hypothesis does not logically entail support for this thesis. Rather, the reason Weiskopf’s claims are so important is that if he succeeds in showing that the logical/ theoretical structure of module-specific theses is untenable, then no amount of empirical evidence can be used to support these views. Accordingly, it is essential to show that his attempted refutation fails. Likewise, from the failure of these refutations, it logically follows that the inability to conceptualize either of the cases without enactive simulation serves as a positive reason for favoring them. If one cannot conjure up, as the philosophers like to say, a picture in one’s mind’s eye of either of the two scenarios without using at least some kind of enactive simulation, then its necessity is no longer in doubt. Having said this, I should restate that after logical qualms are put to rest, empirical evidence presented herein may lead one to take more seriously the prospects of module-specific theories.

It may be suggested that, perhaps, there is an intermediate view between the extremes of amodality and modal-specificity. One may have an ‘associationist view’ within which some form of connections or associations between perceptual knowledge and conceptual knowledge remain even though certain aspects of perceptual knowledge are abstracted. I have no disagreements with such a view. Indeed, it bears repeating that the strong enactive simulation thesis above essentially states that conceptual knowledge is nothing other than perceptual knowledge; on a cruder reading, all of the perceptual information is encoded in memory and nothing is abstracted. By contrast, the weak enactive simulation thesis holds that one may use simulations that may aid in thinking, but they are unnecessary, insufficient, and non-constitutive of our thoughts. As such, there would be a strict demarcation between amodal and modality-specific views, namely perceptual knowledge is abstracted to make it amodal. Lastly, the medium-strength view is very much in line with the associationist suggestion. Some aspects of perceptual knowledge must be abstracted (which ones these are and the mechanism underlying them is a question for empirical science); but, not all information can be abstracted, for if it were, then none of the evidence presented above could be explained, thus rendering the theory a mere speculation. Likewise, perceptual knowledge is not sufficient for mental representations, because this implies the lack of all forms of abstraction. As such, cognitive abilities like logical inference would be very difficult to explicate, and as I explained before, the constitution/ identity relation posited by the strong thesis is similarly untenable. So, by this process of *prima facie* elimination, it is most reasonable to think that enactive simulation is necessary for linguistic comprehension.

6 IMPLICATIONS FOR HUMAN-ROBOT INTERACTION

It is unsurprising that classical ideas of categorization are in principle easier to implement in computational systems such as Turing machines than are conceptions that rely on embodiment or grounding. This is because many of the classical understandings of categorization relied on logic, which is tractable in terms of symbol manipulations over recursive syntax. Indeed, that is one of the primary reasons that the Turing test was first conceived of as testing for language ability [28]. Notwithstanding the questions of immanent semantics in the symbols of a computational implementation, an instantiation of a universal Turing machine was often and still may be seen as an instance of *bona fide* intelligence or 'strong AI.' But why think that symbol manipulation leads to cognition? What about other aspects of language processing that rely on gestures, embodiment, perceptual knowledge, and so forth?

Gross, Krenn, and Scheutz [13] have argued for the importance of gestural aspects, such as eye gaze, in human-robot interactions. Indeed, without aid from gestures, robots' 'understanding,' if it could be called this, remains quite limited. Frixione and Lieto propose a 'dual process' theory that tries to mimic the psychological division between system 1 (reflexive, instinctive, quick) and system 2 (deliberative, rational, slow). The idea is that the latter system could be implemented using classical symbolic systems, while the former system may need to be implemented using a hybrid approach, whose details are too technical to explain here. Similarly, it has been argued that while human-robot language communication, sensory motor skills, perception, decision-making, and learning abilities need to be integrated, such integration is highly difficult [16]. Given the focus on modal-specificity accounts of language processing advocated herein, this presents bleak signs for a multi-dimensional, multi-apt robot capable of efficaciously interacting with humans in a natural context.

However, more recent attempts have been made at placing the modules of robots together in such a way that even though different mechanisms underlie workings of each modality, they are all integrated via a common principle [11]. Attempts have also been made to implement this theoretical approach [12]. Other attempts have been made at providing software that are sufficiently general so as to allow forthe possibility of later modifications and additions of further capabilities [1, 19]. However, these attempts have not applied these mechanisms to the case of natural language interaction between humans and robots.

The space available here is, of course, not enough to provide an extensive review of the literature, so I will end with a suggestion in line with that made by Frixione and Lieto. Like their dual processing approach, it does not appear unreasonable to pursue hybrid approaches that combine symbolic processing with motor and perceptual capabilities. In theory, such a machine could have a neural network 'brain' that allows it to detect visual images, moving objects, and other perceptual stimuli. And, after a mechanism is established that 'transcribes' perceptual knowledge into symbolic knowledge, the latter can be, at least in theory, implemented using somewhat more conventional, logical methods. Such an approach, though without a doubt highly quixotic, may bring together some of the theoretical suppositions that seemingly contradict each other. In other words, if neural networks that constitute the representational system of the robot (assuming it to have such a system) garner perceptual knowledge for it, and once this perceptual knowledge is transduced into symbolic knowledge, there is the possibility of linguistic token manipulation that would later be 'communicated' by the robot to the interlocuter in a natural context. Even in this case, though the robot may not have the same type of enactive simulation or mental imagery as humans, in order to meaningfully interact with us it must be receptive to perceptual data, which is made tractable for implementable symbolic manipulations. Hopefully, such fanciful notions are procurable at least in principle, if not in practice as well.

7 CONCLUSION

This paper showed that though both amodal and module-specific theses take language to be directly involved in categorization, and that module-specific theories provide better explanations for the available psychological evidence. In particular, linguistic comprehension requires enactive simulation. Though the counter example refutes the constitution relationship between enactive simulation and linguistic comprehension, it does not show that the former is not necessary for the latter. Moreover, the insights gained through work in psychology and philosophy are applied to human-robot interaction to argue that, perhaps, at least in principle one must consider the prospects of a hybrid approach.

REFERENCES

[1] Tim Baier, Markus Hüser, Daniel Westhoff, and Jianwei Zhang. 2006. A flexible software architecture for multi-modal service robots. In *In Multiconference on Computational Engineering in Systems Applications (CESA*.

[2] Lawrence W. Barsalou. 1983. Ad hoc categories. *Memory & Cognition* 11, 3 (May 1983), 211–227. https://doi.org/10.3758/BF03196968

[3] Lawrence W. Barsalou. 1999. Perceptions of perceptual symbols. *Behavioral and Brain Sciences* 22, 4 (Aug 1999), 637–660. https://doi.org/10.1017/

S0140525X99532147

[4] Lawrence W. Barsalou. 2008. Grounded Cognition. *Annual Review of Psychology* 59, 1 (2008), 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

[5] Lawrence W. Barsalou, W. Kyle Simmons, Aron K. Barbey, and Christine D. Wilson. 2003. Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences* 7, 2 (Feb 2003), 84–91. https://doi.org/10.1016/S1364-6613(02)00029-3

[6] Noam Chomsky. 1959. Review of Verbal behavior. *Language* 35, 1 (1959), 26–58. https://doi.org/10.2307/411334

[7] Noam Chomsky. 1980. Rules and representations. *Behavioral and Brain Sciences* 3, 1 (Mar 1980), 1–15. https://doi.org/10.1017/S0140525X00001515

[8] Jerry A. Fodor. 1975. *The Language of Thought*. Harvard University Press. Google-Books-ID: XZwGLBYLbg4C.

[9] Jerry A. Fodor and Zenon W. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition* 28, 1 (Mar 1988), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

[10] Jen Jack Gieseking, William Mangold, Cindi Katz, Setha Low, and Susan Saegert. 2014. *The People, Place, and Space Reader*. Routledge. Google-Books-ID: b9WWAwAAQBAJ.

[11] B. Goertzel. 2009. Cognitive synergy: A universal principle for feasible general intelligence. In *2009 8th IEEE International Conference on Cognitive Informatics*. 464–468. https://doi.org/10.1109/COGINF.2009.5250694

[12] B. Goertzel. 2009. OpenCogPrime: A cognitive synergy based architecture for artificial general intelligence. In *2009 8th IEEE International Conference on Cognitive Informatics*. 60–68. https://doi.org/10.1109/COGINF.2009.5250807

[13] Stephanie Gross, Brigitte Krenn, and Matthias Scheutz. 2017. The Reliability of Non-verbal Cues for Situated Reference Resolution and Their Interplay with Language: Implications for Human Robot Interaction. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI âĂŹ17)*. ACM, 189–196. https://doi.org/10.1145/3136755.3136795 event-place: Glasgow, UK.

[14] K. M. Heilman, R. Scholes, and R. T. Watson. 1975. Auditory affective agnosia. Disturbed comprehension of affective speech. *Journal of Neurology, Neurosurgery & Psychiatry* 38, 1 (Jan 1975), 69–72. https://doi.org/10.1136/jnnp.38.1.69

[15] Michael P Kaschak and Arthur M Glenberg. 2000. Constructing Meaning: The Role of Affordances and Grammatical Constructions in Sentence Comprehension. *Journal of Memory and Language* 43, 3 (2000), 508–529. https://doi.org/10.1006/jmla.2000.2705

[16] L. S. Lopes and A. Teixeira. 2000. Human-robot interaction through spoken language dialogue. In *Proceedings. 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000) (Cat. No.00CH37113)*, Vol. 1. 528–534 vol.1. https://doi.org/10.1109/IROS.2000.894658

[17] Max M. Louwerse and Patrick Jeuniaux. 2010. The linguistic and embodied nature of conceptual processing. *Cognition* 114, 1 (Jan 2010), 96–104. https://doi.org/10.1016/j.cognition.2009.09.002

[18] Barbara F. M. Marino, Vittorio Gallese, Giovanni Buccino, and Lucia Riggio. 2012. Language sensorimotor specificity modulates the motor system. *Cortex* 48, 7 (Jul 2012), 849–856. https://doi.org/10.1016/j.cortex.2010.12.003

[19] M. Merten and H. Gross. 2008. Highly Adaptable Hardware Architecture for Scientific and Industrial Mobile Robots. In *2008 IEEE Conference on Robotics, Automation and Mechatronics*. 1130–1135. https://doi.org/10.1109/RAMECH.2008.4681459

[20] Gregory Murphy. 2004. *The Big Book of Concepts*. MIT Press. Google-Books-ID: t2jldRsNkgsC.

[21] Leila Reddy, Naotsugu Tsuchiya, and Thomas Serre. 2010. Reading the mindâĂŹs eye: Decoding category information during mental imagery. *NeuroImage* 50, 2 (Apr 2010), 818–825. https://doi.org/10.1016/j.neuroimage.2009.11.084

[22] Holly Robson, James L. Keidel, Matthew A. Lambon Ralph, and Karen Sage. 2012. Revealing and quantifying the impaired phonological analysis underpinning impaired comprehension in WernickeâĂŹs aphasia. *Neuropsychologia* 50, 2 (Jan 2012), 276–288. https://doi.org/10.1016/j.neuropsychologia.2011.11.022

[23] Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General* 104, 3 (1975), 192–233. https://doi.org/10.1037/0096-3445.104.3.192

[24] Robert D. Rupert. 2004. Challenges to the Hypothesis of Extended Cognition. *The Journal of Philosophy* 101, 8 (2004), 389–428.

[25] John R. Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3, 3 (Sep 1980), 417–424. https://doi.org/10.1017/S0140525X00005756

[26] Charles Spence, Michael E. R. Nicholls, and Jon Driver. 2001. The cost of expecting events in the wrong sensory modality. *Perception & Psychophysics* 63, 2 (Feb 2001), 330–336. https://doi.org/10.3758/BF03194473

[27] Robert A. Stanfield and Rolf A. Zwaan. 2001. The Effect of Implied Orientation Derived from Verbal Context on Picture Recognition. *Psychological Science* 12, 2 (Mar 2001), 153–156. https://doi.org/10.1111/1467-9280.00326

[28] Alan M. Turing. 2009. *Computing Machinery and Intelligence*. Springer Netherlands, 23–65. https://doi.org/10.1007/978-1-4020-6710-5_3

[29] Adam F. Wechsler. 1973. The effect of organic brain disease on recall of emotionally charged versus neutral narrative texts. *Neurology* 23, 2 (1973), 130–135. https://doi.org/10.1212/WNL.23.2.130

[30] Daniel A. Weiskopf. 2010. Embodied cognition and linguistic comprehension. *Studies in History and Philosophy of Science Part A* 41, 3 (Sep 2010), 294–304. https://doi.org/10.1016/j.shpsa.2010.07.005

[31] Ludwig Wittgenstein. 2009. *Philosophical Investigations*. John Wiley & Sons. Google-Books-ID: n4hKDwAAQBAJ.

[32] Rolf A. Zwaan, Robert A. Stanfield, and Richard H. Yaxley. 2002. Language Comprehenders Mentally Represent the Shapes of Objects. *Psychological Science* 13, 2 (Mar 2002), 168–171. https://doi.org/10.1111/1467-9280.00430