WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Computational Insights on the Role of Language in Perception and Behaviour

Navender Singh[1], Dr. Rupali Ahuja[2]

[1]MTech Student
[2]HOD CSE Department

ABSTRACT

Words allow us to have meaningful conversations about the properties, interactions, and uses of things and concepts. Robots and virtual environments are now being developed by scientists with human-like language comprehension and response capabilities. To bridge the gap between the abstract realm of language and the concrete realm of real-world references, our research is establishing the basis for a new kind of computational model. It offers an explanation for context-dependent shifts in word meaning, which are difficult for traditional symbolic models to capture. Interesting implications for cognitive modelling arise from grounded systems' capacity to "step into the shoes" of people by directly processing first-personperspective sensory information, since this provides a fresh approach to evaluate various theories of situated communication and learning.

Descriptions of the material world

Since the 1980s, symbolic explanations of language's meaning have been widely used in computational models of language processing [1-5]. Where words are defined in terms of other symbols, as in a dictionary, such models create circular definitions [6,7]. We humans are less inhibited by circular definitions since so many of the words we use have tangible roots in the world.

Unhappy with purely symbolic models of word meaning, researchers have recently tried to build perceptual and robotic systems that ground the meaning of words in terms of their real-world referents. Meanings of words like "round," "push," "heavy," and so on are often established via the use of visual qualities of instances. These techniques provide computational explanations for the development of semantic linkages between words and the experiences that people have.

Although there is a growing amount of literature on the materiality of language [8-10], specific computational concepts for its representations and processes are still difficult to come by [11]. With the use of models of language grounding, the complex crossmodal phenomena that emerge in situated, embodied language usage may now be modelled. These kind of models are particularly useful for comprehending contextual language acquisition [12], since young children's language focuses primarily on things and actions in their immediate physical environment.

The consequences of this find are vast, since it might one day lead to robots that can form and verify their own ideas about the world, as well as have discussions in which they explain and defend their positions using natural language. Automated weather forecasting [13], natural language query for large databases [14], voice control of interactive robots [15–20], and other human-machine communication systems [16, 17] are early examples of such applications.

Both psychologically-inspired models and models that facilitate the development of autonomous systems are discussed in this article. However, these models only account for a subset of how words are learnt and used, despite the fact that many researchers working on them have set language usage as their ultimate goal. There are still unsolved questions about the grammatical and social aspects of language. Some ways for learning new words are also discussed in [21], which provides a more in-depth analysis of this topic.

We begin by revisiting perceptual association models that have been used to determine the meaning of adjectives and spatial terms and to probe how newborns learn new words. Then we go on to models that incorporate both action and perception, which creates richer representations of both verbs and nouns.

Associations between words and mental categories It is possible to build systems that can provide rich descriptions of their observations thanks to the widespread use of language grounding systems that replicate the translation of sensory data into spoken language [13,19,22-26]. All of these models share the ability to take in continuous sensory data, often in the form of feature vectors, and produce linguistically accurate discrete categories and labels. Both generative and discriminative approaches can be taken when modelling classification.

A simple example of how words may have their origins in perceptual categories is provided by color-naming models. Based on studies of human vision, Mojsilovic developed a generative model (described in [26]) that maps colour names to prototypes of colour foci over a feature space. This strategy assumes that words and concepts have a fixed, one-to-one correspondence. Static mappings struggle to capture the nuances of how individuals really use colour terms and other property descriptors in everyday speech. Consider the many shades of red's meaning in common expressions like "red car," "red hair," "red skin," and "red wine." The colour of red wine may be described as purple in one context (when discussing paint colours), whereas the colour of red hair could be described as orange in another. We shift gears to a context-aware model since those with fixed categories, like Mojsilovic's, are unable to capture such regularities.

A Model of Word Context Use

A linear projection of fixed context-independent colour prototypes onto the space of known wine colours is how the model presented by Gardenfors (shown in Figure 1) [29] creates the meanings of red and white wines. This model explains how our eyes and brain work together to determine the meaning of words. It is more common to refer to dark wines as red, not black. "coloured wine" (vino tinto) and "black wine" (vino negro) are two of the many names for wine in Spanish and Catalan, respectively. Whether a person is referred to as "red" (tinto) or "black" (negro) is a matter of custom. However, the scope of what may be termed a standard is constrained by the perceptual colour space. Since dark wines are further from the context-independent prototype of white than are light wines, Gardenfors's idea suggests that a language could never go from using red to white.

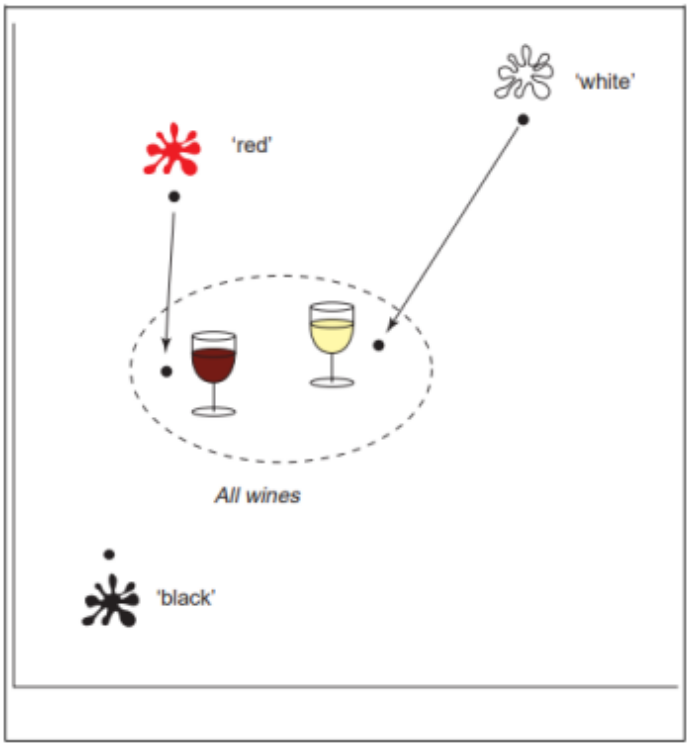


Figure 1. Although red wine is significantly different in color from the context- Independent protype of red ,a geometrical transform s used in Gardenfors model to explain the use of red in the content of wines.

The spatial terms we've seen so far all exhibit a unique form of context dependence. Regier examined ratings rather than binary models of word meaning in his study of the acceptance of geographical terms. Figure 2(a) best illustrates the sentence "the circle is above the block," according to native English speakers, whereas Figure 2(b) and Figure 2(c) are acceptable and inferior alternatives. Above can have several meanings, two of which involve the orientations of lines L1 and L2. L1 connects the masses' centres, whereas L2 connects their nearest points. A and B share the same L2; C and D share the same L1. Therefore, it cannot be attributed only to the orientation of L1 or L2 because people are able to discern between the three configurations. Despite their apparent simplicity, even words like "above" and "near" can convey important yet subtle information about the context. Regier linearly merged these features to develop a model of spatial interactions that was found to be highly reflective of human perceptions [30]. Through the model's analysis of simple movies of objects moving in relation to one another, abstract concepts like "through" and "into" are given concrete visual form. The model predicted [31] that languages would differentiate more precisely between events distinguished by their final states (like inserting a key into a lock) and events distinguished by their beginning structures (like extracting a key from a lock).

A significant shortcoming of Regier's model and other spatial semantics frameworks [32] is their insensitivity to functional contexts [33]. For instance, Regier's concept does not provide a systematic explanation for the nuances in meaning between "clean behind the sofa" and "hide behind the couch" when applied to directions for a robotic vacuum cleaner. As a result of this deficiency, further study into grounded spatial semantics modelling is required.

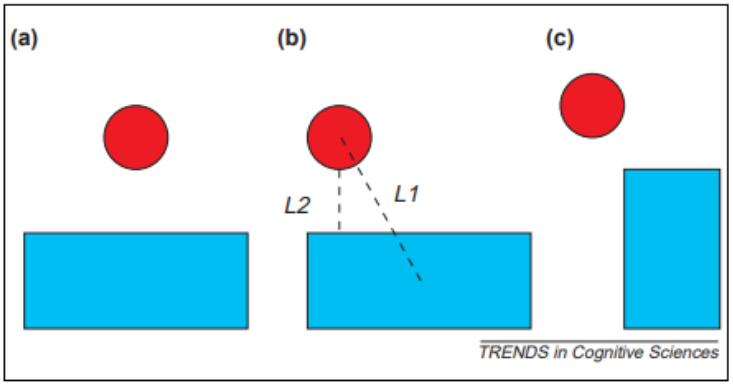


Figure 2. (a) is a good example of '*the circle is above the block*', **(b)** is a less good example, and **(c)** is weaker yet.

The conceptualization-word connecting principle is the basis for all of the presented models. They illuminate the perceptual features that are linguistically salient and the mechanisms that may explain context-dependent word use. Larger systems have been developed that simulate the interaction of visually grounded object descriptors and spatial language to generate phrases and sentences for scene description challenges [19,34,35]. Using perceptually grounded word models, Roy and Mukherjee [8,37,38] recently simulated the visual attention dynamics of situated language comprehension. This method combines text-to-speech translation with contextual observation. As a result of perceptually grounded approaches, sensor-based computational models of infant language learning have evolved, and we now take a look at them.

Models of sensory data processing for infants' early word acquisition from a "first-person perspective"

Sensor-based language systems can be designed to rapidly process recordings generated in natural human settings without the need for manual transcription or coding. These robots can "put themselves in the shoes of humans" and learn new abilities just by monitoring their surroundings. For the first time, researchers have attempted to use audio and video data to learn to segment words and link them with previously acquired visual form and colour categories using a method called cross-channel early lexical learning (CELL) [39]. The model provides a computational rationale for the constraints of language and visual environment on word learning. This approach offers a workaround for a specific form of cross-situational learning [42] due to the fact that it requires data from a broad variety of circumstances to acquire stable audio-visual lexical terms. By seeing films of everyday items alongside untranslated infant-directed speech, CELL was able to acquire a perceptually based lexicon.

CELL simplifies matters by assuming that there is only ever one item in view. To make matters more difficult, any given natural world is likely to contain a large number of things, raising the question of how a language learner is to determine which (if any) of the objects are being referred to by language [43]. The method developed by Yu, Ballard, and Aslin [44] takes into account the speaker's eye gaze direction in addition to the speaker's verbal input and images of other things. Participants in a study were recorded while they narrated their own versions of tales based on illustrations from children's books. Each image showed a number of different things, giving visual depth to the overlapping discussion. A head-mounted eye tracker was used to collect and automatically analyse a speaker's fixation points, or the points at which their eyes remained fixed on a particular portion of the visual picture. Subsets of the visual input were subjected to cross-modal associative learning (CELL-like) depending on the locations of the fixation sites. The model was able to acquire a more concrete visual vocabulary through the use of eye gazing by eliminating superfluous concepts. This model goes far beyond CELL since it takes into account social context, which has been demonstrated to play a crucial role in second-language acquisition [45].

In-depth quantitative studies of many aspects of situational language learning are made possible by these models. To evaluate the impact of the visual context on speech segmentation, we ran a 'blinded' version of the model in CELL. Also, by rerunning Yu's model's association learning algorithms without eye-gaze input, the model may be used to evaluate eye gaze's contribution to vocabulary growth. Regardless of the cognitive plausibility of any model at the level of individual representations and algorithms, sensor-grounded models provide a vital new paradigm for understanding the nature of sensory input from which infants learn.

We'll start with the verbs because they have so many helpful words.

Enhanced realism in the depictions used: fixing the focus on real actions

Verbs denoting concrete physical actions have their origins in conceptualisations that record the passage of time. Using a technology that analyses video sequences of human hands manipulating coloured blocks, Siskind developed a perceptually based model of verb meaning [46]. The contact, support, and attachment interactions in this paradigm are represented visually through the use of hands, blocks, and tables. This set of associations was motivated by Talmy's theory of force dynamics [47]. The semantics of everyday verbs are modelled with temporal schemas that characterise typical sequences of force-dynamic interactions between objects. Some examples of illustrative models for "hand picks up block" include "table supports block," "hand contacts block," "hand attached block," and "hand supports block." The 'Allen relations' [[48]] are used to characterise the temporal interactions between aspects

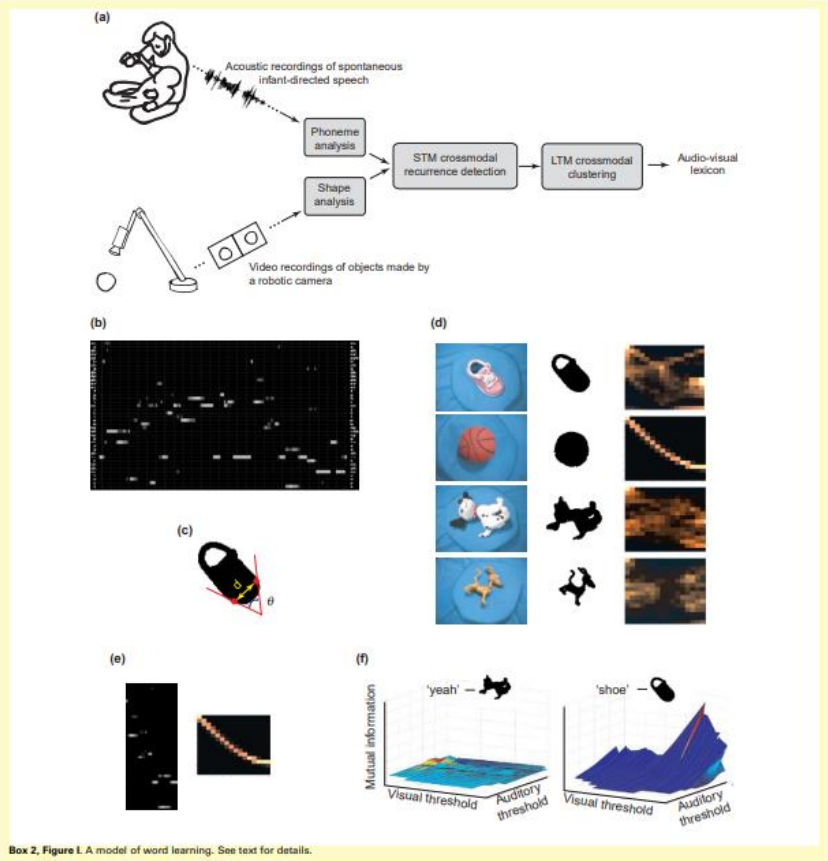
of force dynamics, and they include 13 possible logical relationships between pairs of time periods. The higher-level operations are specified by the lower-level schemas. So, a motion may be thought of as the timing of the schemas that correspond to "pick up" and "put down."

Because it uses logical relations to depict sequences, Siskind's approach has problems identifying the fundamental patterns of motion that give rise to distinctions in meaning between word pairs like "push" and "shove." To address this issue, Bailey et al. [49] devised a system that interprets verb semantics in terms of action control structures, or 'x-schemas,' which specify the order in which a virtual manipulator arm is to carry out a series of actions. X-schemas link action primitives in networks to enable sequential, concurrent, conditional, and recurrent behaviour. Multiple factors can influence an x-scheme's global features, including its force and direction. What distinguishes a verb is its x-schema and its control constraints. In contrast to the structural differences between the x-schemas associated with the verbs pick up and put down, the x-schemas associated with the verbs push and shove are structurally similar but for the force or velocity control parameters employed. To make sense of figurative language used to describe the state of the economy in the news (such as "the economy has hit rock bottom"), Narayanan used x-schema representations [50]. Narayanan takes a fresh approach by trying to characterise rather complex semantics in terms of lower level sensory-motor representations, which was inspired by observations of the pervasiveness of physical metaphor in language [51].

Verbs like "pick up" can be used in the context of either being aware of or taking control of one's behaviour. In contrast to Siskind's approach, which analyses video in order to identify actions that match verbs, Bailey's model is conceptualised as a controller that generates actions for a simulated robot arm. The idea of integrating these two fields is quite promising. Connecting your control and perceptual schemas might be the answer. The alternative is to use a single model to capture both actions and their perceptions. These selections are intriguing because they are consistent with other explanations for how the brain maintains knowledge about actions and objects that have recently been offered [52,53].

There is a tighter connection between perception and action than just in verbs. 'Round' and 'ball' have different meanings, illustrating the importance of action in forming nouns. Perception-based models like CELL, however, should in principle be unable to tell the two apart. Finally, an interactionist model is examined that includes nominal, adjectival, and verbal components.

The Link Between Action and Perception in Noun Roots In order to develop a foundation for language, Roy created structured networks of motor and sensor primitives. This tactic was inspired by the creation of many conversational robots, the most recent of which is Ripley, a robotic manipulator that can translate on-the-spot verbal commands such "hand me the blue one on your right" into the corresponding actions shown in Figure 3 [54]. A dynamic "mental model," or three-dimensional depiction of the robot's immediate physical surroundings (including the tabletop work surface, the robot's own body, and the location of the human communication partner), mediates the robot's vision, manipulation planning, and language. The robot can update its internal model with fresh data it gathers through its sense of hearing, sight, and touch. Using the mental model, Ripley can remember the whereabouts of objects long after they have disappeared from view.



Box 2, Figure I. A model of word learning. See text for details.

Multifunctional sensor demands, such as being aware of a visual zone when the robot looks in the proper spot and the availability of a tactile material when the robot reaches for it, make up the robot's mental picture of an item. Furthermore, the planner expects to be able to relocate objects as they see fit after they have a good grasp on their current locations. Because the location parameters of items in the mental model are updated as a result of manipulation, future visual and tactile expectations are altered in a systematic manner. The robot updates its beliefs as it encounters new information.

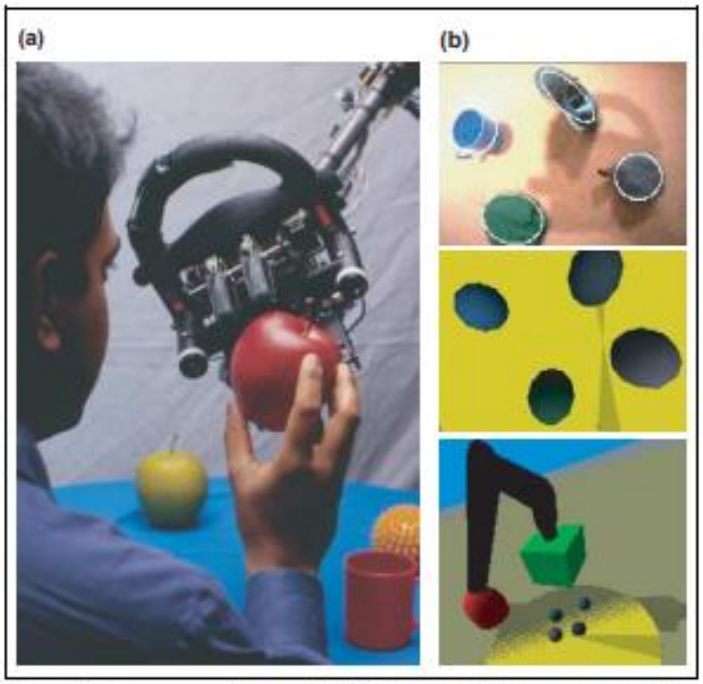


Figure 3. Ripley, a conversational robot. **(a)** Ripley hands its human communication partner an apple in response to the command, 'hand me the one on your right'. **(b)** The top image shows what Ripley sees through its head-mounted video camera when looking down at the table. Thin white lines indicate image regions that the robot's vision system has identified as objects. The second image shows the contents of Ripley's mental model, a rigid body simulator that is dynamically kept in alignment with visual and haptic input. The bottom image shows an alternative visual perspective within the same mental model that the robot is able to generate by moving its 'imagined' perspective by shifting a synthetic camera within the physical simulator. A model of the robot's own body is visible in this view. The ability to shift visual perspective is used by the robot to distinguish, for example, *'my left'* versus *'your left'*. The robot uses a face detection algorithm to track the human's physical position and uses this position to determine the appropriate perspective to simulate to understand 'my left'.

The approach influenced by Ripley's representations and algorithms anchors the meaning of verbs, adjectives, and nouns with respect to physical referents by using a single representational framework [7]. Verbs, like x-schemes, may be traced back to systems that regulate sensory input and motor output. Adjectives describe an object's characteristics based on the sensory expectations associated with a certain activity. Previous theories of perceptual anchoring pale in comparison to this one. For example, the motor programme for maintaining focused attention on an object is linked to the category of red colours. The expectation of exerting one's body weight is what gives the word "heavy" its meaning. This manner, it's possible to connect every facet of perception with the best next step. The user's body shape is used as a data point to record their positions. Objects' positions and properties are stored, but so are the motor affordances that can be used to move or rotate them in the future. In this context, the word "ball" includes not just the sense of "round," which is one of its expected properties along with colour, size, etc., but also any and all actions that could have an effect on the ball. This computational model provides a representation, in line with Piaget's notion of schemas [55] (see also [56,57]), that describes and relates the semantics of words for objects, their attributes, and the actions that may be done on them.

Conclusions

In conclusion, researchers have made significant progress in modelling the interrelationships among language, thought, and action. We've examined methods for naming colours and shapes, defining space, naming verbs and nouns, and discussing how words may have varying meanings based on the circumstances in which they're employed. The models are still in their infancy and only tackle a small portion of the language at this time. Research into how to apply these concepts to areas as diverse as grammatical composition and the social use of language presents significant challenges.

There are several open problems about language grounding that might potentially unite different branches of AI study. Particular subfields of artificial intelligence with well-defined goals, such as computer vision, parsing, information retrieval, machine learning, and planning, have received considerable attention since the 1970s. The need to build machines that can have genuine dialogues with people about their experiences drives artificial intelligence

researchers to blend many fields of study. With the continuing drop in the cost of sensor and robotic technologies and the push towards ubiquitous situated computing [58], new models of language grounding may usher in fresh sorts of situated human-machine communication.

The cognitive theories I've examined all provide potential avenues for addressing the problem of linguistic grounding. Due to the wide variety in these models' and systems' implementations, it is impossible to expect them to give an accurate description of how people think and interact.

References

1 Simon, H. (1980) Physical symbol systems. Cogn. Sci. 4, 135–183

2 Kintsch, W. (1998) Comprehension: A Paradigm for Cognition, Cambridge University Press

3 Miller, G. (1995) Wordnet: A lexical database for english. Commun. ACM 38, 39–41

4 Lenat, D. (1995) Cyc: A large-scale investment in knowledge infrastructure. Commun. ACM 38, 33–38

5 Saint-Dizier, P. and Viegas, E., eds (1995) Computational Lexical Semantics, Cambridge University Press

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

7 Roy, D. Semiotic schemas: A framework for grounding language in action and perception. Artif. Intell. (in press)

8 Tanenhaus, M.K. et al. (1995) Integration of visual and linguistic information during spoken language comprehension. Science 268, 1632–1634

9 Barsalou, L. (1999) Perceptual symbol systems. Behav. Brain Sci. 22, 577–609

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

11 Dennett, D.C. and Viger, C.D. (1999) Sort-of symbols? Behav. Brain Sci. 22, 613

12 Snow, C.E. (1972) Mother's speech to children learning language. Child Dev. 43, 549–565

13 Reiter, E. et al. Choosing words in computer-generated weather forecasts. Artif. Intell. (in press)

14 Barnard, K. et al. (2003) Matching words and pictures. J. Mach. Learn. Res. 3, 1107–1135

15 Reiter, E. and Roy, D., eds Artificial Intelligence: Special Issue on Connecting Language to the World (in press)

16 Yu, C. and Ballard, D. (2004) A multimodal learning interface for grounding spoken language in sensorimotor experience. ACM Trans. Appl. Percept. 1, 57–80

17 Roy, D. (2003) Grounded spoken language acquisition: Experiments in word learning. IEEE Transactions on Multimedia 5, 197–209

18 Skubic, M. et al. (2004) Spatial language for human-robot dialogs. IEEE Trans. Syst. Man Cybern. 34, 154–167

19 Herzog, G. and Wazinski, P. (1994) VIsual TRAnslator: Linking Perceptions and Natural Language Descriptions. Artif. Intell. Rev. 8, 175–187

20 Cohen, P. et al. (2002) Contentful mental states for robot baby. In Proceedings of the 18th National Conference on Artificial Intelligence, Erlbaum

21 Regier, T. (2003) Emergent constraints on word-learning: A computational perspective. Trends Cogn. Sci. 7, 263–268

22 Plunkett, K. et al. (1992) Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. Connection Sci. 4, 293–312

23 Cangelosi, A. et al. (2000) From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories. Connection Science 12, 143–162

24 Steels, L. (2001) Language games for autonomous robots. IEEE Intell. Syst. 16, 16–22

25 Reiter, E. and Sripada, S. (2002) Human variation and lexical choice. Comput. Linguist. 22, 545–553

26 Mojsilovic, A. (2005) A computational model for color naming and describing color composition of images. IEEE Trans. Image Process. 14, 690–699

27 Ng, A. and Jordan, M. (2002) On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Advances in Neural Information Processing Systems (NIPS) (Vol. 14), MIT Press

28 Burges, C. (1998) A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2, 121–167

29 Ga¨rdenfors, P. (2000) Conceptual Spaces: The Geometry of Thought,

MIT Press

30 Regier, T. (1996) The Human Semantic Potential, MIT Press

31 Regier, T. and Zheng, M.M. (2003) An attentional constraint on spatial meaning. In Proceedings of the 25th Annual Meeting of the Cognitive Science Society (Alterman, R. and Kirsh, D., eds), Erlbaum

32 Matsakis, P. and Wendling, L. (1999) A new way to represent the relative position between areal objects. IEEE Trans. Pattern Anal. Machine Intell. 21, 634–643

33 Coventry, K. and Garrod, S. (2004) Saying, Seeing and Acting, Psychology Press

34 Gorniak, P. and Roy, D. (2004) Grounded semantic composition for visual scenes. J. Artif. Intell. Res. 21, 429–470

35 Roy, D. (2002) Learning visually-grounded words and syntax for a scene description task. Comput. Speech Lang. 16, 2002

36 Roy, D. and Mukherjee, N. (2005) Towards situated speech understanding: Visual context priming of language models. Comput. Speech Lang. 19, 227–248

37 Spivey, M.J. et al. (2001) Linguistically mediated visual search. Psychol. Sci. 12, 282–286

38 Chambers, C.G. et al. (2004) Actions and affordances in syntactic ambiguity resolution. J. Exp. Psychol. Learn. Mem. Cogn. 30, 687–696

39 Roy, D. and Pentland, A. (2002) Learning words from sights and sounds: A computational model. Cogn. Sci. 26, 113–146

40 Robinson, T. (1994) An application of recurrent nets to phone probability estimation. IEEE Trans. Neural Netw. 5, 298–305

41 Cover, T.M. and Thomas, J.A. (1991) Elements of Information Theory, Wiley-Interscience

42 Pinker, S. (1989) Learnability and Cognition, MIT Press

43 Quine, W.V.O. (1960) Word and Object, MIT Press

44 Yu, C. et al. The role of embodied intention in early lexical acquisition. Cogn. Sci. (in press)

45 Bloom, P. (2000) How Children Learn the Meanings of Words, MIT Press

46 Siskind, J. (2001) Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. J. Artif. Intell. Res. 15, 31–90

47 Talmy, L. (1988) Force dynamics in language and cognition. Cogn. Sci. 12, 49–100

48 Allen, J. (1983) Maintaining knowledge about temporal intervals. Commun. ACM 26, 832–843

49 Bailey, D. et al. (1997) Embodied lexical development. In Proceedings of the 19th Annual Meeting of the Cognitive Science Society, Erlbaum

50 Narayanan, S. (1999) Moving right along: A computational model of metaphoric reasoning about events. In Proceedings of the National Conference on Artificial Intelligence AAAI-99, AAAI

51 Lakoff, G. and Johnson, M. (1980) Metaphors We Live By, University of Chicago Press

52 Gallese, V. and Lakoff, G. The brain's concepts: The role of the sensorymotor system in conceptual knowledge. Cogn. Neuropsychol. (in press)

53 Mahon, B. and Caramazza, A. The orchestration of the sensorymotor systems: Clues from neuropsychology. Cogn. Neuropsychol. (in press)

54 Roy, D. et al. (2004) Mental imagery for a conversational robot. IEEE Trans Syst Man Cybern B Cybern 34, 1374–1383

55 Piaget, J. (1954) The Construction of Reality in the Child, Ballentine

56 Arbib, M.A. (2003) Schema theory. In The Handbook of Brain Theory and Neural Networks (2nd edn) (Arbib, M.A., ed.), pp. 993–998, MIT Press

57 Bates, E. (1979) The Emergemce of Symbols, Academic Press

58 Weiser, M. (1999) The computer for the 21st century. ACM SIGMOBILE Mobile Computing and Communications Review 3, 3–11

59 Grosz, B. and Sidner, C. (1986) Attention, intentions, and the structure of discourse. Comput. Linguist. 12, 175–204

60 Cohen, P.R. and Perrault, C.R. (1979) Elements of a plan-based theory of speech acts, Cognitive Science

61 Allen, J. and Perrault, R. (1980) Analyzing intention in utterances. Artif. Intell. 15, 143–178

62 Zwaan, R.A. and Radvansky, G.A. (1998) Situation models in language comprehension and memory. Psychol. Bull. 123, 162–185

63 Johnson-Laird, P.N. (1983) Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness, Cambridge University Press