Review TRENDS in Cognitive Sciences Vol.9 No.8 August 2005

Grounding words in perception and action: computational insights

Deb Roy

The Media Laboratory, Massachusetts Institute of Technology, 20 Ames Street, Cambridge, MA 02142, USA

We use words to communicate about things and kinds of things, their properties, relations and actions. Researchers are now creating robotic and simulated systems that ground language in machine perception and action, mirroring human abilities. A new kind of computational model is emerging from this work that bridges the symbolic realm of language with the physical realm of real-world referents. It explains aspects of context-dependent shifts of word meaning that cannot easily be explained by purely symbolic models. An exciting implication for cognitive modeling is the use of grounded systems to 'step into the shoes' of humans by directly processing first-person-perspective sensory data, providing a new methodology for testing various hypotheses of situated communication and learning.

Words about the physical world

Over the past few decades computational models of language processing have focused on symbolic explanation of linguistic meaning [1–5]. Such models define word meanings in terms of other symbols, producing circular definitions much like those found in a dictionary [6,7]. Humans are less hindered by circular definitions because we ground many words in physical experience in the world.

Researchers dissatisfied with purely symbolic models of word meaning have recently sought to build perceptual and robotic systems that ground the meaning of words in terms of their real-world referents. Thus the meaning of *round* is grounded in visual features of exemplars, *push* in motor control structures, *heavy* in haptic features, and so on. These systems provide computational explanations of how words acquire meaning through their connections with perception and action.

Although the embodied nature of language has received significant recent attention [8–10], computational hypotheses formulated in terms of specific representations and processes remain elusive [11]. Models of language grounding open a new avenue for modeling complex crossmodal phenomena arising in situated, embodied language use. Such models are of particular interest for understanding situated language acquisition because early language tends to be primarily about objects and activities in the child's immediate physical environment [12].

Corresponding author: Roy, D. (dkroy@media.mit.edu).
Available online 11 July 2005

A long-term implication of this work is the possibility of machines that are able to autonomously acquire and verify beliefs about the world, and to communicate in natural language about their beliefs. Early applications along these lines are already emerging, including automated generation of weather forecasts [13], large-scale image database retrieval by natural language query [14], verbal control of interactive robots, and other human–machine communication systems [15–20].

This article reviews a range of work, from psychologically motivated models evaluated mainly for their ability to explain human behavior, to models which support the development of autonomous systems. Although the ultimate goal of many researchers developing such models is to understand situated language use, at present the models address only limited aspects of word meaning, learning and use. Eventually, grammatical and social aspects of language must also be addressed, but currently stand as open questions. Furthermore, although some approaches to word learning are discussed, a more detailed review can be found in [21].

We begin by reviewing models of perceptual association for grounding the meaning of adjectives and spatial terms, and for studying strategies of infant word learning. We then shift to models that integrate action with perception providing richer representations underlying verbs and nouns.

Associations between words and perceptual categories

Many language grounding systems model the translation of sensory input into natural language descriptions, leading to systems that can talk about what they observe [13,19,22–26]. A common element in these models is that they sort continuous sensory input represented as feature vectors into discrete categories that are associated with labels according to linguistic convention. Categorization may be modeled through either generative or discriminative methods (Box 1).

Models of color naming provide a simple example of word grounding through association with perceptual categories. Motivated by considerations of human visual perception, Mojsilovic developed a generative model that associates color terms with prototypes of color foci defined over a feature space [26]. This model assumes that the mapping from words to perceptual categories is fixed. In reality, however, people use color terms and other property descriptors in flexible ways that are not easily captured by static mappings. For example, consider the shift in meaning

 1364-6613/$ - see front matter doi:10.1016/j.tics.2005.06.013

Box 1. Generative and discriminative models of categorization

Consider a simple model in which visual regions are represented by two-dimensional feature vectors consisting of measures of height and width (see Figure I). A discrete shape category (dashed line circle) can be defined by selecting a prototype vector combined with a threshold value (a). Two prototypes can 'compete' (b), leading to a category boundary along points of equal distance from both prototypes (if non-Euclidean distance measures are used, non-linear boundaries may emerge). Categories may also be modeled by explicitly representing categorical boundaries. In (c), a linear model, $f(height) = A * width + B$, encodes the same categorical distinction as the prototypes in (b). Methods (a) and (b) are examples of generative models whereas (c) is a discriminative model of categorization (for a discussion on generative and discriminative methods in a probabilistic setting, see [27]). An important advantage of generative models, as their name suggests, is that they provide a natural basis for generating examples of categories by using prototypes as targets for behavioral processes. This may be particularly important in representations related to motor control that will be used not only to recognize but also generate output. Some of the most powerful machine learning methods such as support vector machines [28], however, operate on discriminative models.

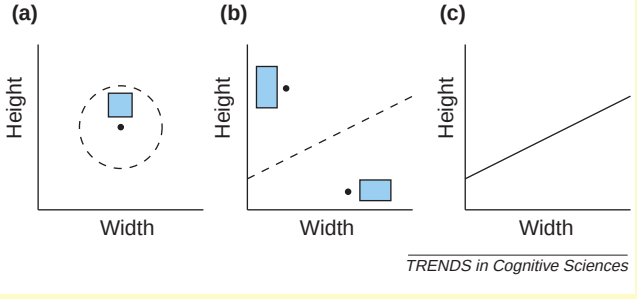


Figure I. Different kinds of models. **(a)** and **(b)** are generative models whereas **(c)** is a discriminative model of categorization. See text for details.

of *red* in the contexts '*red car*', '*red hair*', '*red skin*', or '*red wine*'. The color of red wine might be called *purple* in another context (e.g. discussing colors of paint), the color of red hair *orange*, and so on. Fixed category models such as Mojsilovic's are unable to account for such patterns, so we now turn to a model that addresses context sensitivity.

Modeling context-dependent word use

Gardenfors proposes a model (illustrated in Figure 1) in which the meanings of *red* and *white* in the context of wines are produced by starting with fixed context-independent color prototypes which are linearly projected into the space of known wine colors [29]. This model explains how linguistic convention and visual perception combine to determine word meanings. The choice of *red* versus, say, *black* to describe dark colored wines is in part owing to convention. Spanish speakers will say '*vino tinto*' (literally, 'colored wine') and in Catalan speakers would call the same wine '*vino negro*' ('black wine'). The choice of *red* versus *black* (*tinto* or *negro*) is thus a matter of arbitrary linguistic convention. Nonetheless the perceptual color space constrains possible conventions. According to Gardenfors's model, it would be impossible for a language to reverse the use of *red* and *white* because the distance from the context-independent prototype of white is farther from dark wines than from light ones.

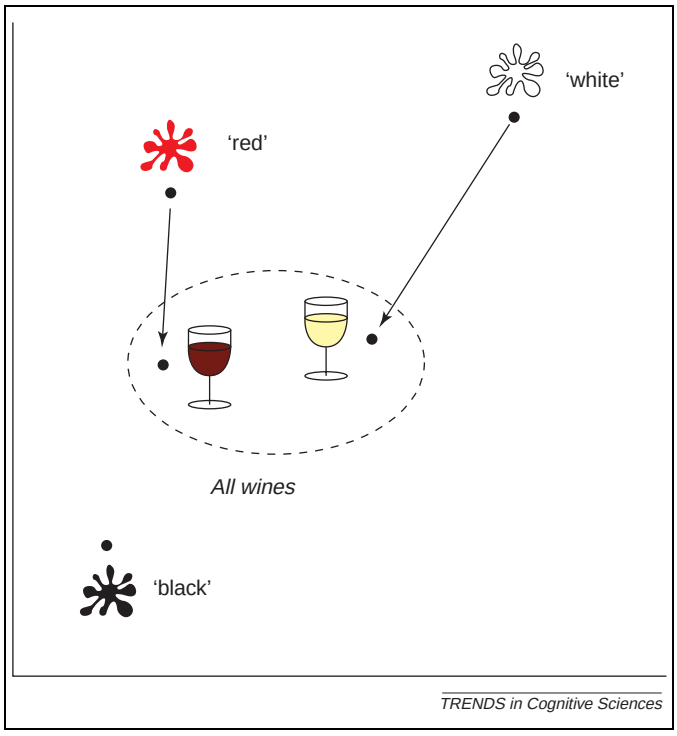


Figure 1. Although red wine is significantly different in color from the context-independent prototype of 'red', a geometric transform is used in Gardenfors' model to explain the use of 'red' in the context of wines.

A different kind of context dependence appears in the use of spatial terms such as *above*. Rather than model word meaning as having all-or-none applicability, Regier studied graded acceptability judgments of spatial terms. English speakers agree that in Figure 2, configuration (a) is a good example of 'the circle is above the block', (b) is acceptable but not as good (a), and (c) is weaker yet. Two possible features underlying the meaning of 'above' are the orientations of the lines L1 and L2. L1 connects the centers of mass of the regions whereas L2 connects the closest points between regions. L2 is identical in (a) and (b); L1 is identical in (b) and (c). Thus, the orientation of neither L1 nor L2 alone can explain the fact that humans differentiate each of the three configurations. This example demonstrates that apparently simple words such as *above* and *near* frequently encode non-obvious features of the environment. Regier developed a model of spatial relations based on a linear combination of both

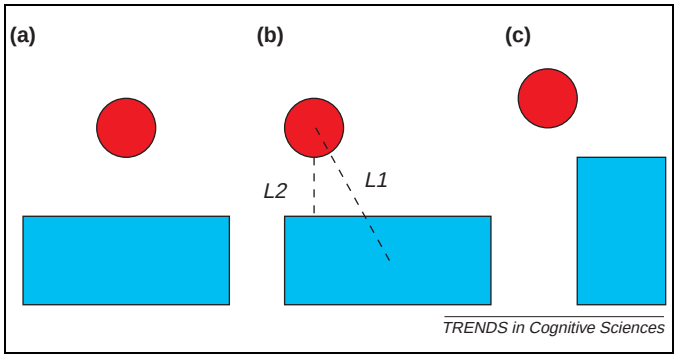


Figure 2. (a) is a good example of '*the circle is above the block*', **(b)** is a less good example, and **(c)** is weaker yet.

features which was found to closely match human judgments [30]. Furthermore, the model processes simple movies of objects moving relative to one another to visually ground words such as *through* and *into*. The model correctly predicted that owing to preferential attention to endpoints of spatial events, languages will make finer semantic distinctions when referring to events distinguished by their endpoints (e.g. putting a key into a lock) compared with events distinguished by their initial structure (e.g. removing a key from a lock) [31].

An important omission in Regier's model and other models of spatial semantics [32] is that they are insensitive to functional contexts [33]. For example, if we command a robotic vacuum cleaner to '*clean behind the couch*' versus '*hide behind the couch*', Regier's model is unable to systematically model the shift in meaning of 'behind'. This limitation suggests important future work on modeling grounded spatial semantics.

The models reviewed thus far are based on the idea of associating linguistic labels with perceptual categories. They provide insights into linguistically salient perceptual features and suggest possible mechanisms underlying context-dependent word use. Larger systems have been developed that model the composition of visually grounded object descriptors and spatial language to generate phrases and sentences in scene description tasks [19,34,35]. Recently, Roy and Mukherjee used perceptually grounded word models in a scene description understanding system that integrates speech interpretation with visual context [36], modeling visual attention dynamics of situated language comprehension [8,37,38]. Perceptually grounded approaches have also recently led to sensor-grounded computational models of infant language learning, which we now review.

Models of infant word learning that process 'first-person-perspective' sensory data

The implementation of sensor-grounded language systems opens the door to a new kind of cognitive model that is able to directly process recordings from natural human environments without the need for manual transcription or coding. These systems are able to 'step into the shoes' of humans and learn from natural sensory data. The first effort of this kind is the cross-channel early lexical learning (CELL) model which learns to segment and associate spoken words with acquired visual shape and color categories based on speech and video input [39] (Box 2). The model provides a computational account of how visual context and speech constrain the process of word learning. The model solves a form of cross-situational learning [42], as evidence from numerous situations must be combined to learn stable audio-visual lexical items. In evaluations, CELL successfully acquired a vocabulary of perceptually grounded words by listening to untranscribed infant-directed speech paired with video images of everyday objects.

A simplifying assumption in CELL is that visual input consists of only one object at a time. In reality, infants face a much more difficult learning problem because any natural environment is typically cluttered with numerous objects, raising the question of how a language learner is to decide which (if any) of the objects are being referred to by language [43]. Yu, Ballard and Aslin developed a system that processes spoken input paired with visual images of multiple objects combined with the speaker's eye gaze direction [44]. In an experiment, speakers were recorded as they narrated stories in their own words based on illustrations in a book for young children. The illustrations contained multiple objects so that for any co-occurring speech there were multiple visible referents. A head-worn eye tracker recorded detailed eye-movements of the speakers which were automatically analyzed to detect fixation points at which the speaker's eyes remained focused on a particular part of the visual scene. The location of fixation points was used to select specific regions from the visual input, which were then subjected to crossmodal associative learning similar to CELL. The use of eye gaze reduced ambiguity of possible referents and enabled the model to successfully acquire a visually grounded lexicon. This model is a significant extension beyond CELL in that it makes use of social information that is known to be crucial in language acquisition [45].

These models enable fine-grained quantitative study of various aspects of situated language learning. CELL was used to quantify the impact of visual context for segmenting speech by re-running a 'blinded' version of the model. Similarly, the impact of eye gaze on word learning may be measured using Yu's model by re-running its association learning algorithms without the benefit of eye-gaze input. Regardless of the cognitive plausibility of each model at the level of specific representations and algorithms, sensor-grounded models provide an important new methodology for understanding the nature of sensory input from which infants learn.

Let us now shift our attention to a particularly important class of words: verbs.

Richer representational structures: grounding verbs in physical action

Verbs that refer to physical actions are naturally grounded in representations that encode the temporal flow of events. Siskind developed a perceptually grounded model of verb meaning as part of a system that analyzes video sequences of human hands manipulating colored blocks [46]. The model uses visually derived features that express the contact, support and attachment relationships between hands, blocks and tabletops. This choice of relationships is motivated by Talmy's theory of force dynamics [47]. The semantics of basic verbs are modeled using temporal schemas that define expected sequences of force dynamic interactions between objects. For example, the meaning of '*hand picks up block*' is modeled by the sequence: *table-supports-block*, *hand-contacts-block*, *hand-attached-block*, *hand-supports-block*. Temporal relations between force dynamics features are specified using 'Allen relations', which encode 13 possible logical relations between pairs of time intervals [48]. For intervals A and B, Allen relations include: *A ends after B starts*, *A ends exactly as B starts*, *A and B start together but A ends first*, and so on. Time durations are not specified by the schemas, enabling the model to classify observations across varying timescales. Higher level actions are defined

Box 2. A model of learning words from sights and sounds

The cross-channel early lexical learning model or CELL is a computational model of sensor-grounded word learning [39] that learns to segment speech at word boundaries, form visual categories, and acquire semantically appropriate associations between spoken words and visual categories. Speech recordings were made of six mothers as they played with their pre-verbal infants using common toys. These recordings were paired with video of the same objects recorded by a robot, providing multisensory input for the model (Figure Ia, next page).

A recurrent neural network (RNN) [40] extracts phonemic features from input speech; **(b)** shows the representation of, '*Oh, you can make it bounce too!*' The brightness of each row corresponds to the probability of 40 English phonemes as they evolve over time from left to right. Columns with multiple bright bands indicate phonemic confusions.

Visual features used for shape analysis are illustrated in **(c)**. The silhouettes of the objects are found using background color segmentation. The distance, d, and angle, θ, formed by each pair of points along the boundary of silhouettes, are measured. Distances are normalized by the largest distance between any two points on the object's boundary. Thus, the two-dimensional vector remains constant as the object is rotated in-plane and rescaled.

All pairs of boundary points are analyzed and aggregated in a two-dimensional histogram. Several objects from the infant experiment are shown in **(d)** along with their silhouettes and shape histograms (right column). Values of d and θ are binned along the vertical and horizontal axes of the histograms. The three-dimensional structure of an object is captured by sets of histograms derived from multiple views.

On the assumption that caregivers tend to repeat salient words that refer to the environment, speech and images are analyzed together to find recurrent segments of speech in similar visual contexts (a). A first-in, first-out short-term memory (STM) stores the last five spoken utterances paired with co-occurring shapes. Because there are no acoustic equivalents of spaces between printed words in natural speech, CELL systematically compares *all* pairs of segments of speech across all pairs of utterances in STM. When a recurring segment of speech is found, the shape histograms that co-occurred with the segments are compared. If the visual contexts are also similar, a crossmodal recurrence is detected. The recurrent speech segment and shape may be thought of as the system's guess of a possible word and its meaning, a 'lexical candidate'. Over time, lexical candidates found in STM accumulate in long-term memory (LTM).

An example of a lexical candidate is the speech segment *bounce* paired with the shape of a ball (**e**). This candidate would be produced if the speech segment *bounce* occurred within multiple utterances in STM in the context of similarly shaped round objects.

Lexical candidates are unreliable for two reasons. First, as sensory processes are prone to noise, many detected recurrences will lead to incorrect lexical candidates. Second, recurrence analysis will often generate semantically inappropriate candidates. For example, if a caregiver repeats '*yeah*' while playing with a toy dog, the semantically incorrect hypothesis of '*yeah*' paired with dogs will be obtained. To address these problems, only clusters of lexical candidates with non-coincidental crossmodal structure are retained. Visual and auditory thresholds are combined with each candidate prototype in LTM to generate crossmodal categories. Mutual information (MI) [41] is used to measure the association strength between the resulting speech and visual category. The MI for a range of possible auditory and visual thresholds is computed, yielding a MI surface. In **(f)**, the MI surface for the '*yeah*'-to-dog pairing is low, but the '*shoe*'-to-shoe pairing yields a high peak value. Lexical candidates that lead to high MI values such as *shoe* in (f) constitute the final output of the model.

In evaluations, CELL learned a small vocabulary of shape names such as *ball*, *shoe* and *truck* from six different mothers' input. It also learned several semantically appropriate examples of onomatopoeic sounds such as '*ruf-ruf*' for dogs and '*vroooom*' for cars. In comparative tests in which the systems was 'blinded', a lack of visual input led to over a 50% drop in word discovery accuracy, demonstrating the value of crossmodal structure for word learning.

Although this initial experiment focused solely on learning shape names, the recurrence and crossmodal clustering algorithms extend without modification to learn names for multiple perceptual domains. CELL has been integrated into an interactive word learning robot that learns to use color and shape names to interpret and generate two-word descriptive phrases of objects [17].

in terms of these lower level schemas. Thus *move* is defined as the ordered sequence of the schemas corresponding to '*pick up*' followed by '*put down*'.

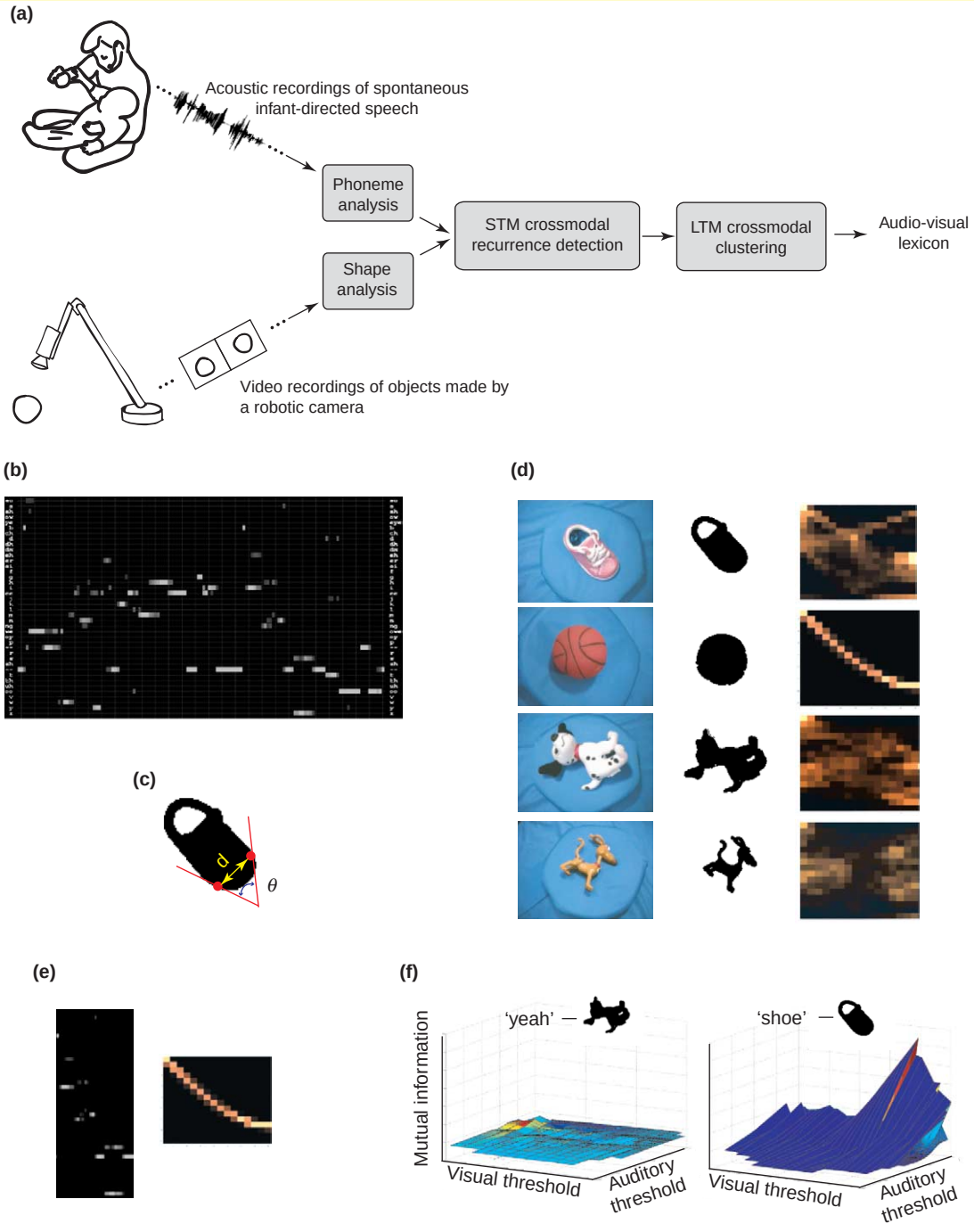
The use of logical relations to encode sequences makes it difficult to use Siskind's approach to distinguish manners of motion that underlie the difference in meaning of word pairs such as *push* versus *shove*. Bailey *et al.* addressed this issue by developing a system that learns verb semantics in terms of action control structures, called 'x-schemas', which control sequences of movements of a simulated manipulator arm [49]. X-schemas organize action primitives into networks that allow for sequential, concurrent, conditional and repetitive action. A set of control parameters specify high level attributes of x-schemas such as force and direction. A verb is defined by its associated x-schema and control parameters. The verbs *pick up* and *put down* are distinguished by the structure of their associated x-schemas, whereas *push* and *shove* are distinguished by different force or velocity control parameters applied to the structurally identical x-schema. Narayanan used x-schema representations as a basis for interpreting physical metaphors in news stories (e.g. '*the economy has reached rock bottom*') [50]. Narayanan's approach is unique in its attempt to model relatively abstract semantics in terms of lower level sensory-motor representations, inspired by observations of the prevalence of physical metaphor in language [51].

Verbs such as 'pick up' refer to both the perception and control of action. Siskind's model watches video and recognizes actions corresponding to verbs whereas Bailey's model is conceived as a controller that generates actions for a simulated robot arm. An interesting future direction will be to bring these two lines of research together. One possibility is to link perceptual and control schemas using some sort of bridging structure. An alternative is to develop a single action–perception representation that unifies functions of both models. Interestingly, these options correspond to distinct current hypotheses regarding the neural representation of actions and objects [52,53].

The intertwined nature of perception and action is not limited to the domain of verbs. To see the relevance of action in grounding nouns, consider the difference between the meaning of 'round' versus 'ball'. Perceptually grounded models such as CELL are unable in principle to distinguish their meanings. The final model we review integrates action and perception in an interactionist representation of verbs, adjectives and nouns.

Integration of action and perception in grounding nouns

Roy developed a framework for grounding words in terms of structured networks of motor and sensor primitives. This approach arose from building a series of conversational robots, the most recent of which is Ripley, a robotic manipulator that is able to translate spoken commands



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

such as '*hand me the blue one on your right*' into situated action (Figure 3) [54]. The robot maintains a dynamic 'mental model', a three-dimensional model of its immediate physical environment (including its table top work surface, the robot's own body, and the location of the human communication partner) that it uses to mediate perception, manipulation planning and language. The contents of the robot's mental model may be updated based on linguistic, visual, or haptic input. The mental model endows Ripley with object permanence, remembering the position of objects when they are out of its sensory field. From the robot's point of view, the meaning of an object in its mental model is a multimodal sensory expectation: if the robot looks at the appropriate location, its visual system expects to find a visual region; if the robot reaches to the same location, it expects to touch and grasp the object. Furthermore, the planner expects control over object locations once they have been grasped. Thus manipulation updates location parameters of objects in the mental model, leading to systematic shifts in future visual and haptic expectations. Violations of expectations cause the robot to update its mental model.

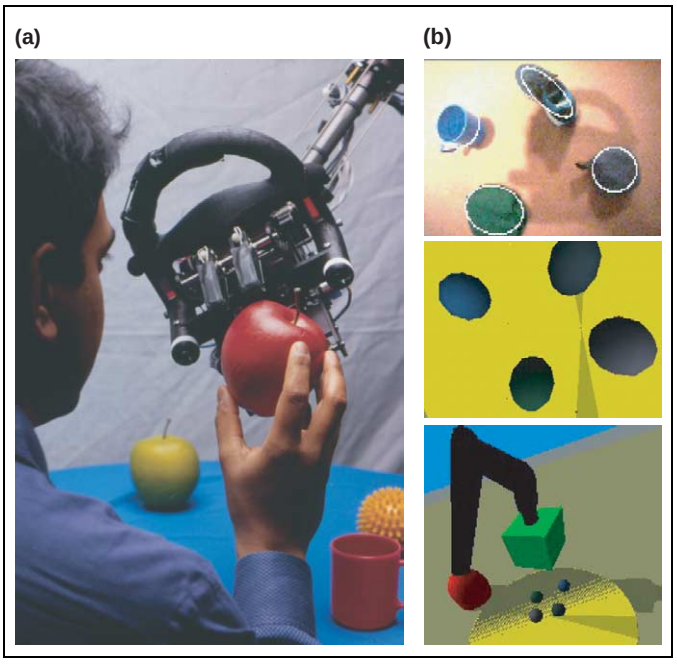


Figure 3. Ripley, a conversational robot. **(a)** Ripley hands its human communication partner an apple in response to the command, 'hand me the one on your right'. **(b)** The top image shows what Ripley sees through its head-mounted video camera when looking down at the table. Thin white lines indicate 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'.

Ripley's representations and algorithms led to an approach that grounds the meaning of verbs, adjectives and nouns referring to physical referents using a unified representational framework [7]. Verbs are grounded in sensory-motor control programs similar to x-schemas. Adjectives describing object properties are grounded in sensory expectations relative to specific actions. This is a significant extension of earlier models of perceptual grounding. For example, the meaning of *red* is not simply a color category, but rather a color category linked to the motor program for directing active gaze towards an object. *Heavy* is grounded in haptic expectations associated with lifting actions. In this way, all perceptual properties are related to appropriate actions. Locations are encoded in terms of body-relative coordinates. Objects are represented as bundles of properties tied to a particular location along with encodings of motor affordances for affecting the future location of the bundle. In effect, the meaning of *ball* according to this model subsumes both the meaning of *round* (which is one of its expected properties along with color, size, etc.), and all of the actions that may affect the ball. This computational model is consistent with Piaget's notion of schemas [55] (see also [56,57], and provides a representation that distinguishes and relates the semantics of words for objects, their properties, and actions that can be taken on them.

Conclusions

To summarize, researchers have made significant recent progress in modeling the interactions between word use, perception, and action. We have reviewed models of color and shape naming, spatial language, verbs and nouns, all of which are able to relate words to real-world referents, often providing novel explanations of context-dependent word use. The models are at a very early stage and address only a small fragment of language. Many difficult research challenges lie ahead to bring these ideas together with other aspects of language such as grammatical composition and the functional use of language in social contexts (see Box 3 for some examples).

The numerous open challenges of language grounding provide an opportunity to re-unite sub-fields of artificial intelligence (AI). Since the 1970s, many AI researchers have shifted their focus to sub-fields of AI with well defined goals such as computer vision, parsing, information retrieval, machine learning and planning. Language grounding provides the impetus for AI researchers to integrate these sub-fields to address – and exploit the capability of building machines that can converse about what they observe and do in human-like ways. With the continuing drop in cost of sensor and robotic technology, and the trend towards ubiquitous situated computing [58], models of language grounding may pave the way to exciting new forms of situated human–machine communication.

From a cognitive modeling perspective, each model I have reviewed suggests possible strategies to address aspects of language grounding. We cannot expect that such models and systems will directly explain how people think and communicate: both design and implementation differ dramatically. Nonetheless it seems clear that these computational models, together with behavioral

Box 3. Questions for future research

- How might computational models of planning and discourse [59–61] be combined with models of sensory-motor processes to create models of physically situated discourse?
- How can we extend models of language grounding to study the role of situation models in conversation [62,63] – in particular the role of perception of shared physical environments in helping maintain alignment [64,65] between communication partners' individual models?
- How can words that refer to physical affordances be modeled? For example, the sensory-motor grounding of the verb *open* depends on both the physical structure of a situation and an understanding of goals to be achieved.
- How might sensory-motor grounded models of language be augmented with metaphor or analogy-making processes [66] to explain the semantics of various physical metaphors that are ubiquitous in language [51]?
- Dropping prices and miniaturization of digital sensing and recording technology mean that extremely dense recordings of longitudinal language development will soon become feasible – perhaps tens of thousands of hours from single subjects. With the demand for increased longitudinal language development data [67] soon met, however, the resulting rise in cost of manual annotation of speech and especially video will become unaffordable. How can computational models of language grounding be scaled to analyze massive multisensory observation recordings and empirically test hypotheses of language acquisition?

and brain imaging studies, can provide tangible steps towards such explanations.

Acknowledgements

I thank Michael Fleischman, Peter Gorniak, Kai-yuh Hsiao, Stefanie Tellex, Michael Arbib, Art Glenberg, Rupal Patel, Geoff Arnold, Talia D'Abramo, Felipe De Brigard, Robert Briscoe, and the anonymous reviewers for valuable comments on earlier drafts of this paper.

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 Gärdenfors, 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 sensory-motor system in conceptual knowledge. *Cogn. Neuropsychol.* (in press)
53 Mahon, B. and Caramazza, A. The orchestration of the sensory-motor 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
64 Clark, H. (1996) *Using Language*, Cambridge University Press
65 Pickering, M. and Garrod, S. (2004) Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* 27, 169–190
66 French, R. (2002) The computational modeling of analogy-making. *Trends Cogn. Sci.* 6, 200–205
67 Tomasello, M. and Stahl, D. (2004) Sampling children's spontaneous speech: how much is enough? *J. Child Lang.* 31, 101–121