Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Artificial Intelligence 167 (2005) 1–12
Artificial Intelligence
www.elsevier.com/locate/artint

Connecting language to the world

Deb Roy [a], Ehud Reiter [b,*]

[a] *Media Laboratory, MIT, USA*
[b] *Department of Computing Science, University of Aberdeen, UK*

Received 21 June 2005; accepted 22 June 2005

Available online 21 July 2005

1. Language in the world

How does language relate to the non-linguistic world? If an agent is able to communicate linguistically and is also able to directly perceive and/or act on the world, how do perception, action, and language interact with and influence each other? Such questions are surely amongst the most important in Cognitive Science and Artificial Intelligence (AI). Language, after all, is a central aspect of the human mind—indeed it may be what distinguishes us from other species.

There is sometimes a tendency in the academic world to study language in isolation, as a formal system with rules for well-constructed sentences; or to focus on how language relates to formal notations such as symbolic logic. But language did not evolve as an isolated system or as a way of communicating symbolic logic; it presumably evolved as a mechanism for exchanging information about the world, ultimately providing the medium for cultural transmission across generations. Motivated by these observations, the goal of this special issue is to bring together research in AI that focuses on relating language to the physical world. Language is of course also used to communicate about non-physical referents, but the ubiquity of physical metaphor in language [21] suggests that grounding in the physical world provides the foundations of semantics.

Systems that connect language to the world may be called *situated* to emphasize the links to non-linguistic situational context. These systems also address the symbol grounding problem [17] and may thus be called *grounded*. The topic of this special issue is

* Corresponding author.
E-mail addresses: dkroy@media.mit.edu (D. Roy), ereiter@csd.abdn.ac.uk (E. Reiter).

0004-3702/$ – see front matter
doi:10.1016/j.artint.2005.06.002

situated, grounded systems. This includes systems which translate sensory signals into language and language into physical actions, systems which learn how to use language in this manner, and systems that use non-linguistic data when making linguistic decisions. Non-linguistic interaction is generally anchored in sensors or effectors which are connected to the physical world (or to simulations of the physical world). Although much current work on sensor grounding emphasizes visual perception, other modalities ranging from thermal to haptic are also being explored.

The numerous open challenges of language grounding provide an opportunity to bring together many sub-fields of AI. While early AI researchers often investigated many aspects of machine intelligence together, in recent years there has been a tendency for researchers to focus on specific sub-fields of AI with well defined goals, such as computer vision, parsing, information retrieval, machine learning, and planning. Language grounding provides an impetus for AI researchers to integrate these sub-fields, so that they can attempt to build machines that can converse about what they observe and do in human-like ways. Early applications along these lines are already emerging, including:

- Automatic generation of textual reports grounded in real-time numerical data such as weather forecasts, financial reports, and sports summaries (for example, [34] and Reiter et al.'s paper in this special issue);
- Multimedia information retrieval and management (for example, [2] and Barnard and Johnson's paper in this special issue);
- Natural language interfaces to robots (many everyday objects and environments such as cars and houses may be treated as robots in the sense that they have sensors, actuators, bodies, and control systems) (for example, [19,43] and Roy's paper in this special issue);
- Natural language interfaces to virtual reality systems and games (for example, [16] and Kelleher et al.'s paper in this special issue);
- Situated NLP for mobile devices (e.g., location-dependent web search queries);
- Intelligence analysis that combines language with contextual cues to interpret otherwise ambiguous or noisy communication signals.

Broadly speaking, the long term implications of this work include 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. Although at a relatively early stage, we see the work in this issue as taking steps in this direction.

The growing interest in situated language processing systems closely parallels the rise in study of embodied cognition and cognitive linguistics [3,6,9,14,20,46,47]. Behavioral and neural studies are increasingly uncovering the rich interplay between language, action, and perception. These findings bring into question strongly modular theories of mind that posit stringent information encapsulation between modules. Insights emerging from the construction of situated/grounded language processing system may lead to computational models that are relevant to understanding human cognitive processes at a functional level [36], and to help us understand how language evolved and is learned [30,44].

2. Key technical challenges

A system that connects language to the world must bridge the symbolic realm of words with the non-symbolic realm of sensory-motor interaction. This requirement raises difficult and often subtle issues that do not arise in purely symbolic approaches to natural language processing, nor in purely non-symbolic approaches to perception and action. We highlight below some of the challenges in building such systems. For each of these broad areas, the specific challenge of cross-modal processing (language, action, perception) is key to building situated/grounded language processing systems.

We illustrate many of the these challenges with the problem of mapping visual input data to linguistic color terms, which is one of the simplest and perhaps best understood language-and-world tasks. A robot which is using visual data purely for internal decision making (such as deciding which widgets coming off a production line need to be inspected) might simply feed camera data into a neural network that is trained to distinguish 'good' from 'bad' widgets, without attempting to explicitly model color, or indeed even separate color from other visual data. On the other hand, a machine translation system which is translating English to French might simply treat color as a set of semantic primitives, without attempting to model what these mean; all such a system needs to know is that RED is a color primitive which is lexicalised as *red* in English and *rouge* in French. Such simple approaches can work well in systems which only operate in one modality, but building systems which link visual color data to linguistic color terms requires us to solve many additional problems.

First of all, we need *cross-modal representations* (and associated reasoning algorithms) that support both the linguistic and sensory-motor sub-systems. In practice, construction of AI systems have led to a large variety of representations and reasoning algorithms that are targeted towards specific niches; that is, representations which work well for language processing, vision, expert systems, or some other AI niche. Unfortunately, in most cases representations that work well for one niche do not work well for others. For example, a neural network that identifies faulty widgets from visual data may have internal nodes which in some way encode color information, but these nodes are unlikely to be useful for choosing linguistic color terms. Another example is that many modern vision systems use "bag of feature" models that lead to robust object detection, but ignore the spatial structure (shape) of objects. Although such approaches lead to good performance on strictly visual tasks, they provide no obvious basis for grounding important aspects of natural language semantics. For example, modifiers in language can be used to specify part-whole modularity of objects (*cup without a handle*). A visual representation that does not preserve spatial structure and part-whole relations will not be able to link to these kinds of natural language phrases. Similarly, approaches to motor control and action representation that do not preserve appropriate temporal structure will be unable to link to adverbial modifiers in natural language. In general, the challenge is to design representations that work robustly in sensory-motor contexts, yet provide the appropriate structural "hooks" for language.

Once we have established a representation that encodes the necessary non-linguistic information, we need to *associate words with perceptual and action categories*. Drawing from established methods in pattern classification (e.g. [28]), words can be treated as labels for sensory-grounded categories. As is well known in the field of pattern clas-

sification, feature selection is crucial to success. In particular, we should choose sensory features which are similar to those encoded by natural language. For example, a popular choice for visual color features is a cognitively motivated three-dimensional color encoding [22,23]. Standard generative or discriminative classification techniques may then be used to model categories within this three-dimensional space. In more complex domains, the choice of perceptual features is often not as obvious. For example, the meaning of spatial terms such as *above* have proven to depend on subtle interactions between the shapes of objects involved and led to extensive research in appropriate choices of features [31]. Feature selection is only part of the challenge, however.

Linguistic word choice depends on *context* as well as the actual sensor data. For example, the meaning of a modifier may depend on the category of objects it is modifying; thus the visual association of *red* shifts widely as it is used in differing contexts such as *red wine*, *red hair*, or *red car*. One approach to this problem might be to separately model possible colors of each object class which are geometrically combined with context-independent color prototypes [13]. Color words can also convey non-linguistic information; for example, *green banana* suggests a banana that is not ripe as well as a banana which is visually green. Such context effects are ubiquitous in natural language. To take just one non-color example, consider adding voice commands to a mobile robotic vacuum cleaner. The difference in meaning of "behind" in "clean behind the couch" and "hide behind the couch" depends on complex interactions between the physical environment and the functional meaning of "cleaning" and "hiding". These and numerous other kinds of context effects stand as open challenges for future research.

Another issue is deciding *how specific and detailed* the linguistic description should be; for example, is it better to use a broad color term such as *red* or a more specific one such as *crimson*? Most objects are not uniformly one color—does this need to be mentioned (for example, *red car* vs. *red car with silver trim*). This often depends on domain knowledge (we would not usually say *red car with black wheels*, as wheels by default are black). In most cases linguistic descriptions are summaries of sensor data (since we cannot communicate megabytes of sensor data in a few words), so we must decide what to include in the summary. For example, an agent seeing a cup on a table will have a large amount of information about the cup's color, size, orientation, precise position, and so forth, which is not communicated in the linguistic summary *there is a cup on the table*. Specificity and detail decisions may depend on the task (context again!), and algorithms have been proposed for these decisions in specific constrained tasks such as reference generation [10]. However, we do not know of methods to make such decisions in general.

To ground *verb* meanings, systems must represent temporal structure of actions. Beyond simply labeling sequences of movement, verbs often encode causal structure (who did what to whom). Thus, ideally, representations of action would on one hand link to perception and control of action in the physical environment, and on the other provide structural hooks for the argument structure of verbs. The intertwined nature of verbs and actions leads to larger scale challenges in designing *planning* algorithms for situated language processors. A scene description system, for example, needs to plan word choice such that possible listener ambiguities arising from the current physical context can be anticipated and avoided. More challenging yet, is the problem of planning with a mix of communicative and motor actions. For example, consider a cooperative robot that helps its human partner in physical

tasks (e.g., lifting large objects) and that uses language to coordinate joint action. The robot must plan its words and its motor actions in a coordinated fashion. Methods from robot planning and discourse planning must be integrated to achieve such behaviors.

The above discussion focuses on representations and algorithms; but how do we actually get the data to create specific language-to-world rules? For example, how do we actually decide which color values correspond to *red hair*? In general researchers have to date assumed that most of this data is *learned from examples and feedback*. Hence we must decide how this learning should be done for each of the challenges discussed above. For learning perceptual associations of words, established methods of parameter estimation and feature selection may be used. Learning how to plan across modalities or integrate ambiguous sources of knowledge might be cast as a reinforcement problem. In general, many learning problems will involve not only parameter estimation but also structure acquisition. The complexity of situated language systems, such as those described in this volume, suggest that any "blank slate" learning approach is likely not to scale due to the enormous search space size. Thus, we anticipate structured learning approaches, i.e., learning methods in which manually designed biases constrain learnability, will play an critical role.

If language-to-world rules are learnt rather than explicitly communicated, it is likely that the rules learnt by different agents will be different to some degree. Indeed, it is clear that different people associate different meanings with words [33]. For example different people associate different color values with the word *red*, even in identical contexts. Humans who are talking to each other align their language to each other [7,26], and computer language-to-world systems may wish to likewise *align with their human conversational partner*.

The fact that different agents use different language-to-world rules suggests that it is possible that the overall set of rules used by a community of agents may change over time, especially if old agents are regularly replaced by new agents, who again must learn and align language rules. Many researchers are interested in using simulations of such agents to study *language evolution*, and gain insights as to how human language evolved.

Last but not least, an important methodological issue is how language-and-world systems should be *evaluated*. For example, if we have built a system that generates color words from visual data, how can we determine if this system does a good job or not? The papers in this special issue use a very diverse range of evaluation techniques, including performance on a held-out test set, psychological experiments with human subjects, soundness and completeness measures, user questionnaires, and simulations. This diversity may reflect the fact that researchers in this area come from many different subfields of AI, which have their own expectations and conventions about evaluation. While in some ways this diversity is exciting, it can make it more difficult for readers to understand and compare evaluations. Indeed, we note as editors that the criticisms of evaluation were the most common complaints made by referees about the content of papers submitted to this special issue. Hence, agreeing on appropriate evaluation techniques is an important challenge for the language-and-world community as a whole.

The challenges we have laid out are broad and are meant to provide an overall guide to the issues at stake. We now highlight selected previous work to provide some historical context.

3. Examples of related prior work

A comprehensive survey of work in situated/grounded language processing is beyond the scope of this introduction, so instead we highlight a few threads of research that are representative. Readers interested in more thorough surveys may refer to collections of related work [4,8,27] and reviews of research on word grounding, learning and evolution of language [30,36,44].

Winograd's seminal SHRDLU system demonstrated the importance of integrating world models with language planning and understanding [48]. The system could engage in natural language dialog to control the actions of a simulated robot arm in a blocks world, ask clarifying questions, and generate explanations of its actions. Although SHRDLU did not deal with problems of sensory-motor categorization, cross-modal ambiguity, or learning, it was nonetheless a milestone in the history of AI and a clear example of situated language processing. In the same period, the first robots that connected machine perception to symbolic descriptions were being developed by Nilsson [25]. Although natural language was not the focus of this work, many of the issues related to sensory-motor categorization and planning mentioned above were central to this early work.

More recently, Siskind has explored the links between language and perception through the construction of a series of visually-grounded language systems [40–42]. Building on insights from cognitive linguistics [45], he has developed a temporal representation that encodes the causal relations between objects inferred from visual observation. He has demonstrated implementations that translate video input into structured representations. Due to the emphasis on causal relationships, his approach provides a natural basis for linking argument structure of verbs to objects that fulfill semantic roles (e.g. agent, patient) in the physical world. Learning from positive examples has been demonstrated using this representation [12].

The Visual Translator (VITRA) is one of the most ambitious end-to-end visually grounded scene description systems built to date [18]. VITRA was able to generate natural language descriptions of traffic scenes and segments of soccer matches. Visually-grounded models of spatial relations and actions operated on video input, which were then translated into verbal descriptions used a set of domain-dependent generation rules. The generation system included a listener ambiguity model that was used to eliminate potential listener confusions by generating descriptions that were unlikely to match distractor referents in visual scenes.

The "L0 Project" was created with the goal of developing computational models of situated language acquisition motivated by the question, "How could we learn to describe what we see" [11]. This effort led to a series of projects that addressed different aspects of physically situated language acquisition and use [1,5,24,29]. Bailey and Narayanan developed a structured representation of action underlying verbs that was used to control a simulated robot arm [1] and as a basis for understanding physically grounding metaphors [24]. Regier explored geometric visual features that underlie spatial relations and that seem to be at play across languages. This led to his later work with colleagues on linguistically motivated vector-based representations of spatial relations [31], and insights into the role of attention in spatial relations [32].

Roy and his colleagues have developed a series of systems that relate words, descriptive phrases, and commands to physical environments [15,35,37–39]. The cross-channel early lexical learning (CELL) model was used to learn words by processing acoustic recordings of infant-directed speech paired with video images of objects [39]. Later work focused on visually-guided grammar acquisition for scene description [35], modeling spatial language in scene descriptions [15], and visual context sensitive speech understanding [38]. Roy, Hsiao, and Mavridis developed an interactive manipulator robot named Ripley that is able to translate spoken commands such as *hand me the blue thing on your right* into physical actions. The robot maintains a mental model of its table top environment, providing a cross-modal representation for binding verbal commands, visual perception, and motor control. Roy's paper in this volume synthesizes many of the theoretical insights that emerged from this body of work.

4. Papers in this volume

4.1. Barnard and Johnson: Word sense disambiguation with pictures

Barnard and Johnson show that visual information can help in the classic Natural Language Processing (NLP) problem of word sense disambiguation (WSD). Many words of course have multiple senses (meanings), and WSD systems attempt to determine which sense of a word is meant. For example, whether *bank* refers to a financial institution or to the edge of a river. Existing algorithms for this task use purely linguistic information. Barnard and Johnson show that when the text is accompanied by an image which has visual correlates with previous uses of the word (that is, the image might be a city street scene or a natural scene with water), visual analysis of this image can increase the accuracy of the WSD system. They use images from a standard corpus, not images hand-crafted to assist in the WSD task. Their algorithm is based on a technique for predicting likely words for images, which is inspired by statistical machine translation techniques; in other words, they apply ideas developed for translating French to English to the task of 'translating' images to English.

From the perspective of Section 2 challenges, a primary contribution of this paper (and of previous work by Barnard and collaborators [2]) is in the area of cross-modal representations. The authors show how to extend a technique developed for NLP (statistical machine translation) so that it also works with visual data; and that it is possible to develop integrated algorithms and representations, for an important real-world task, which work well with both linguistic and non-linguistic data. From an applications perspective, the authors show that is possible to use visual information to assist in an NLP task (WSD).

4.2. Dominey and Boucher: Learning to talk about events from narrated video in a construction grammar framework

Dominey and Boucher describe a system that learns how language is used to describe events in a simple microworld, by observing how humans talk about events in this world. They focus on learning sentence structures, and in particular propose that this be done

using simple template-like rules for sentences, instead of complex compositional grammars. They experimentally show that their system does a reasonably good job of learning language even when the observational data comes from naive human subjects who know nothing about the system.

From the perspective of Section 2 challenges, this paper is a contribution to learning, and also perhaps to evaluation (since previous systems in this area have tended to use observational data provided by the developers themselves, who knew how the system worked). The paper also shows how psychological insights about how children learn language can be incorporated into a computer language-learning system.

4.3. Kelleher et al.: Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context

Kelleher et al. show that visual information can assist in the NLP task of generating and interpreting referring expressions (noun phrases that identify objects) such as *the blue house* or *the tall tree*. In principle the choice of referring expression is strongly affected by context and salience, including both discourse context (what has been mentioned in previous utterances) and perceptual context (what speaker and hearer see or otherwise perceive). Kelleher et al show how these two kinds of context (and two types of salience) can be integrated and jointly used for reference interpretation and generation in a virtual reality system.

From the perspective of Section 2 challenges, perhaps the major contributions of this paper are in the areas of cross-modal representations and on the use of non-linguistic data in adjudicating the specificity of linguistic descriptions. The authors show how linguistic and visual data about context and salience can be integrated and used for NLP tasks; and how such a cross-modal integration leads to a better understanding of the general phenomena of salience. From an applications perspective, they show that language-and-world research can play an important role in the new area of virtual reality.

4.4. Needham et al.: Protocols from perceptual observations

Needham et al. show how an autonomous agent can learn how to play a simple game which includes both visual and linguistic aspects. Audio and video recordings are made of two humans playing a game, and these are analyzed to learn the rules of the game; these rules are then given to a computer agent. Learning is done in two stages. First, the audio and video systems learn classifiers which identify distinct visual objects and distinct audio words. These classifiers are used to convert recorded games into a symbolic representation; this is essentially a temporally-ordered sequence of states, combined with a symbolic description of the state of the game at each state. Inductive logic programming is then used to learn the rules (protocols) of the game from this information.

From the perspective of Section 2 challenges, this paper is clearly a contribution to learning. In particular, it shows how symbolic and non-symbolic learning can be combined, so that an agent can learn both linguistic and visual aspects of a real-world activity. From a more applied perspective, the authors shows that it is possible for agents to learn how to participate in real-world multi-modal interactions with humans.

4.5. Reiter et al.: Choosing words in computer-generated weather forecasts

Reiter et al. focus on the problem of choosing words to communicate numerical weather prediction data; this indeed is the only paper in the special issue which does not attempt to connect language to vision. They present an extensive empirical analysis of how humans (weather forecasters) perform this task, focusing on the fact that there are substantial differences between the forecasters in which words they prefer to use, and indeed in the meanings they associated with words; for example some forecasters used *late morning* to mean 9AM, while others used this phrase to mean noon. Reiter et al. then describe their SumTime-Mousam weather-forecast generation system, which in fact is operationally used by a forecasting company to generate several kinds of forecasts. At a lexical level, SumTime-Mousam is programmed to avoid words whose meaning varied substantially across forecasters, and words only used by a small minority of forecasters. An evaluation of wind descriptions (part of the weather forecast) showed that human forecast readers preferred SumTime-Mousam texts over texts written by human forecasters; qualitative comments from the users suggest that this is partially because SumTime-Mousam texts contain fewer idiosyncratic or ambiguous words.

From the perspective of the challenges presented in Section 2, this paper's most important contribution is perhaps in the area of alignment. The authors show that there is considerable difference in the language used by different people (that is, in "idiolects"), that a computer text-generation system can be programmed to avoid many idiolect-specific misunderstandings, and that this seems to enhance the quality of the generated texts. They also show that it is possible to build a complete data-to-language system which is good enough to be used operationally, and which produces texts that are as good as (perhaps even better than) human-written texts, at least by some metrics.

4.6. Roy: Semiotic schemas: A framework for grounding language in action and perception

Roy presents a theoretical framework for grounding the meaning of verbs, adjectives, and nouns referring to physical referents using a unified representational scheme that "provides a computational path from sensing and motor action to words and speech acts". He defines grounding as a cycle that relies on both "bottom-up" sensor-grounded perception and "top-down" agent-driven action on the physical environment. Rather than start with an ontological distinction between objects and events, Roy takes a constructivist approach by suggesting a common set of representational primitives that are used to construct complex events, objects, and object properties. As a result, the conceptual grounding of verbs, nouns, and adjectives are expressed as networks of sensory-motor primitives called *semiotic schemas*. The internal structure of schemas provides a basis for relating and combining concepts underlying words—thus the framework provides a subsymbolic level of explanation of conceptual structures that ground symbolic (linguistic) activity. The framework arose from—and provides a guide for future work in—the construction of robotic and virtual systems that connect situated language to machine action and perception.

In terms of Section 2 challenges, Roy's framework is a contribution to cross-modal representation and processing that is shaped by the relationship between natural language and embodiment.

4.7. Vogt: The emergence of compositional structures in perceptually grounded language games

Vogt is somewhat different from the other papers in this special issue because he is interested in understanding how language evolved—not in building computer systems that interact with human users or analyze human-authored documents. He explores language evolution by creating a simulated world where agents interact linguistically in a shared environment, in particular by playing "language games" where an "adult" agent which already has a linguistic model interacts with a "child" agent which is learning the model. Vogt is interested in what happens to the language system over the course of many generations, and in particular if grammatical structures (such as compositionality) evolve and remain stable over the course of time in the language system, and how the evolution of compositional structures is related to (the modeling of) semantic development.

From the perspective of Section 2 challenges, this paper is a contribution to language evolution, and also to learning. In particular, Vogt shows how complex compositional rules can evolve in an agent population, as well as basic sense-data-to-word associations.

5. Conclusions

Understanding how language relates to the world is one of the grand challenges of cognitive science, and building automated systems that connect the symbolic world of language to the non-symbolic world of sensory input and effector control is one of the great challenges of AI. As the papers in this special issue show, researchers are beginning to develop techniques to address the problems described in Section 2, and also beginning to build systems that link language and the world in sophisticated ways, in quite a variety of application contexts. These systems often operate in limited domains and/or assume input data that is relatively noise-free, but nonetheless they demonstrate that even our current limited understanding of the scientific issues involved enables us to build systems that do a good job at real tasks such as generating weather forecasts and word sense disambiguation.

Research in this area is especially exciting because it requires integrating various subfields of AI, including vision, robotics, pattern analysis, knowledge representation, learning, and natural language processing. Current AI research often feels like a collection of subfields which rarely communicate with each other. While such specialization has in many ways helped the subfields progress, we believe that the subfields could benefit from interacting more, and also that this would benefit the AI and cognitive science research agenda as a whole. The papers in this volume show that tangible progress in the theory and application of situated, grounded language processing systems is well underway. We hope this special issue encourages more people to get involved in this growing research area.

References

[1] D. Bailey, J. Feldman, S. Narayanan, G. Lakoff, Embodied lexical development, in: Proceedings of the Nineteenth Annual Meeting of the Cognitive Science Society, Erlbaum, Mahwah, NJ, 1997.

[2] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, M. Jordan, Matching words and pictures, J. Machine Learning Res. 3 (2003) 1107–1135.

[3] L. Barsalou, Perceptual symbol systems, Behavioural and Brain Sciences 22 (1999) 577–609.

[4] R. Barzilay, E. Reiter, J.M. Siskind, in: HLT-NAACL03 Workshop on Learning Word Meaning from Non-Linguistic Data, 2003.

[5] B. Bergen, N. Chang, S. Narayanan, Simulated action in an embodied construction grammar, in: Proceedings of the 26th Annual Meeting of the Cognitive Science Society, 2004.

[6] A. Clark, Being There: Putting Brain, Body and World Together again, MIT Press, Cambridge, MA, 1997.

[7] H. Clark, Using Language, Cambridge University Press, Cambridge, 1996.

[8] S. Coradeschi, A. Saffiotti, Special issue of robotics and autonomous systems on perceptual anchoring, 43 (2–3) (2003).

[9] K. Coventry, S. Garrod, Saying, Seeing and Acting, Psychology Press, 2004.

[10] R. Dale, E. Reiter, Computational interpretations of the gricean maxims in the generation of referring expressions, Cognitive Sci. 19 (2) (1995) 233–263.

[11] J. Feldman, G. Lakoff, D. Bailey, S. Narayanan, T. Regier, A. Stolcke, Lzero: The first five years, Artificial Intelligence Rev. 10 (1996) 103–129.

[12] A.P. Fern, R.L. Givan, J.M. Siskind, Specific-to-general learning for temporal events with application to learning event definitions from video, J. Artificial Intelligence Res. 17 (2002) 379–449.

[13] P. Gärdenfors, Conceptual Spaces: The Geometry of Thought, MIT Press, Cambridge, MA, 2000.

[14] A. Glenberg, M. Kaschak, Grounding language in action, Phychonomic Bull. Rev. 9 (3) (2002) 558–565.

[15] P. Gorniak, D. Roy, Grounded semantic composition for visual scenes, J. Artificial Intelligence Res. 21 (2004) 429–470.

[16] P. Gorniak, D. Roy, Speaking with your sidekick: Understanding situated speech in computer role playing games, in: Proceedings of Artificial Intelligence and Interactive Digital Entertainment, 2005.

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

[18] G. Herzog, P. Wazinski, VIsual TRAnslator: Linking perceptions and natural language descriptions, Artificial Intelligence Rev. 8 (1994) 175–187.

[19] J. Juster, D. Roy, Elvis: Situated speech and gesture understanding for a robotic chandilier, in: Proceedings of the International Conference on Multimodal Interfaces, 2004.

[20] G. Lakoff, Women, Fire, and Dangerous Things, University of Chicago Press, Chicago, IL, 1987.

[21] G. Lakoff, M. Johnson, Metaphors We Live By, University of Chicago Press, Chicago, IL, 1980.

[22] J.M. Lammens, A computational model of color perception and color naming, PhD thesis, State University of New York, 1994.

[23] A. Mojsilovic, A computational model for color naming and describing color composition of images, IEEE Trans. Image Process. 14 (5) (2005) 690–699.

[24] S. Narayanan, Moving right along: A computational model of metaphoric reasoning about events, in: Proceedings of the National Conference on Artificial Intelligence AAAI-99, Orlando, FL, 1999.

[25] N. Nilsson, Shakey the robot, Technical Report 323, A.I. Center, SRI International, 1984.

[26] M. Pickering, S. Garrod, Toward a mechanistic psychology of dialogue, Behav. Brain Sci. 274 (2004) 169–226.

[27] P. McKevitt (Ed.), Integration of Natural Language and Vision Processing, Kluwer Academic, Dordrecht, 1995.

[28] P.E. Hart, R.O. Duda, D.G. Stork, Pattern Classification, Wiley, New York, 2001.

[29] T. Regier, The Human Semantic Potential, MIT Press, Cambridge, MA, 1996.

[30] T. Regier, Emergent constraints on word-learning: A computational perspective, Trends Cognitive Sci. 7 (6) (2003) 263–268.

[31] T. Regier, L. Carlson, Grounding spatial language in perception: An empirical and computational investigation, J. Experimental Psychol. 130 (2) (2001) 273–298.

[32] T. Regier, M. Zheng, An attentional constraint on spatial meaning, in: R. Alterman, D. Kirsh (Eds.), Proceedings of the 25th Annual Meeting of the Cognitive Science Society, 2003.

[33] E. Reiter, S. Sripada, Human variation and lexical choice, Computational Linguistics 28 (2002) 545–553.
[34] J. Robin, K. McKeown, Empirically designing and evaluating a new revision-based model for summary generation, Artificial Intelligence 85 (1996) 135–179.
[35] D. Roy, Learning visually-grounded words and syntax for a scene description task, Comput. Speech Language 16 (3) (2002).
[36] D. Roy, Grounding words in perception and action: Computational insights, Trends in Cognitive Sciences, in press.
[37] D. Roy, K. Hsiao, N. Mavridis, Mental imagery for a conversational robot, IEEE Trans. Systems Man, Cybernet. B 34 (3) (2004) 1374–1383.
[38] D. Roy, N. Mukherjee, Towards situated speech understanding: Visual context priming of language models, Comput. Speech Language 19 (2) (2005) 227–248.
[39] D. Roy, A. Pentland, Learning words from sights and sounds: A computational model, Cognitive Sci. 26 (1) (2002) 113–146.
[40] J. Siskind, Naive physics, event perception, lexical semantics, and language acquisition, PhD thesis, Massachusetts Institute of Technology, 1992.
[41] J. Siskind, Grounding language in perception, Artificial Intelligence Rev. 8 (1995) 371–391.
[42] J. Siskind, Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic, J. Artificial Intelligence Res. 15 (2001) 31–90.
[43] M. Skubic, D. Perzanowski, S. Blisard, A. Schultz, W. Adams, M. Bugajska, D. Brock, Spatial language for human-robot dialogs, IEEE Trans. Systems Man, Cybernet. C 34 (2) (2004) 154–167.
[44] L. Steels, Evolving grounded communication for robots, Trends Cognitive Sci. 7 (7) (2003) 308–312.
[45] L. Talmy, Force dynamics in language and cognition, Cognitive Sci. 12 (1988) 49–100.
[46] L. Talmy, Toward a Cognitive Semantics, MIT Press, Cambridge, MA, 2000.
[47] M.K. Tanenhaus, M. Spivey-Knowlton, K. Eberhard, J. Sedivy, Integration of visual and linguistic information during spoken language comprehension, Sci. 268 (1995) 1632–1634.
[48] T. Winograd, A Process Model of Language Understanding, Freeman, New York, 1973, pp. 152–186.