Semantics in the Wild

Robert J. Glushko (glushko@ischool.berkeley.edu)
School of Information, University of California,
Berkeley CA 94720

Paul P. Maglio (pmaglio@almaden.ibm.com)
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120-6099

Teenie Matlock (tmatlock@ucmerced.edu)
Cognitive Science Program, University of California, Merced
P.O. Box 2039, Merced, CA 95344

Lawrence W. Barsalou (barsalou@emory.edu)
Psychology Department, Emory University
532 Kilgo Circle, Atlanta GA 30322

Keywords: Language; categorization; semantics; tagging; interoperability; Web 2.0.

Introduction

Traditionally, cognitive science has focused on the mental representation of abstract and concrete concepts through laboratory experiments (e.g., Smith, Shoben, & Rips, 1974; Rosch & Mervis, 1975). Subsequent more applied research in semantics – especially in corporate settings at Bell Labs, IBM, Xerox PARC, and elsewhere – involved field studies on how people naturally describe objects and computing processes (e.g., Furnas, Landauer, Gomez, & Dumais, 1987). Because different people often use different categories and words to refer to the same things, and they use the same ones to refer to different things, library science sought to train professional indexers and cataloguers to follow precise rules with controlled vocabularies (e.g., Taylor, 2004). But library science was late to recognize the potential impact of computing technology on the vocabulary problem. The most significant innovations in information organization and retrieval emerged from the cognitive and computer sciences, including the use of embedded thesauri and ontologies in information systems and techniques for latent semantic indexing (see Dumais, 2003).

Yet in today's world of ubiquitous computing and ubiquitous information resources, we interact daily with a bewildering variety of information types, and we constantly make choices about whether and how to organize them. It is now impossible to rely on professionals to describe and catalog information resources, which proliferate exponentially as web pages, office documents, and multimedia objects that often include photos and videos from digital cameras and cell phones. And though sophisticated indexing by web search engines, such as Google, can compensate for the lack of explicit description of information resources, much of the information we encounter and use is in fact never indexed by such systems. So rather than relying on professionals or automatic computational methods, many people have begun to impose their own semantic structure on the information and processes they encounter by *tagging* information with their own keywords and categories and then sharing the tags broadly or even publicly (see Hammond, Hannay, Lund, & Scott, 2005). Distributed or social categorization systems include **del.icio.us** for bookmarking and tagging web pages, **flickr** for storing and sharing photos, and **youtube** for videos.[1] These new, rich information environments containing semantically tagged content would seem to provide a perfect opportunity for research on semantics in the wild.[2] And so we think it is time to reconsider the nature of research on semantics.

Symposium Structure

The symposium will include a series of four interrelated talks on semantics in the wild. The first two will focus on practical issues in technology and business, and the second two will focus on scientific issues in linguistics and psychology. The goal is to begin a new conversation in semantics research that is grounded in the problems presented by and the opportunities afforded by modern computational and business environments.

We now describe each of the participants and their potential presentation topics in turn.

Robert J. Glushko is an adjunct professor in the School of Information at the University of California, Berkeley. After receiving a Ph.D. in Cognitive Psychology from UC San Diego in 1979, he spent over twenty years in corporate R&D, in consulting, and as a Silicon Valley entrepreneur before returning to the university.

Glushko's interests lie in methods and tools for the design, development, and deployment of information-

[1] See http://del.icio.us/, http://www.flickr.com/, and http://www.youtube.com/.

[2] With apologies to Hutchins (1996).

intensive applications and services. He has been developing "document engineering" analysis and modeling methods for achieving semantic robustness and interoperability (Glushko & McGrath, 2005), using consumer web applications and business-to-business systems as natural laboratories for understanding how people and organizations establish semantic and lexical equivalence when they interact with information resources and each other. His presentation will focus on the design and usage of "tags" and "tag sets." The former includes considerations like their number, level of abstraction, and naming conventions for tags. The latter includes frequency of tag use and how semantic and lexical differences are resolved, or not resolved.

Paul P. Maglio is senior manager of Service Systems Research at the IBM Almaden Research Center. He received a bachelor's degree in computer science and engineering from MIT and a Ph.D. in cognitive science from UC San Diego. Maglio has worked on programmable web intermediaries, attentive user interfaces, multimodal human-computer interaction, and human aspects of autonomic computing. He holds 13 patents and has published more than 70 scientific papers in computer science and cognitive science.

Maglio's presentation will focus on how people communicate in business and work settings. Drawing on field studies of work practices in technical and non-technical jobs, his presentation will address how language use changes over time to create specific and idiosyncratic business vocabularies and business categories. Technologies for semantic tagging of web-based and other content will be discussed in the context of business practices and of the need to support vocabulary and category development.

Teenie Matlock is an assistant professor of cognitive science and founding faculty at UC Merced, the newest University of California campus. She did graduate work in linguistics at UC San Diego and completed a Ph.D. in cognitive psychology at UC Santa Cruz. After doing postdoctoral research at Stanford University, she established the Cognitive Science Program at UC Merced in 2004. Matlock studies language as interaction, especially its interface with spatial thought, visual processing, and physical action. Some of her work investigates conceptual knowledge of abstract or novel domains, including the web.

Matlock's presentation will focus on user experience with interactive web applications. Drawing on cognitive linguistics, especially conceptual semantics (e.g., Lakoff & Johnson, 1987) and psycholinguistics, especially everyday language use (e.g., Clark, 1996), Matlock will discuss the way people talk about and think about social networks on the web. Key questions include: How do people naturally anchor their understanding of these large, abstract domains? How do they express that knowledge to others? What are the implications for application, including developing optimal tagging methods?

Lawrence W. Barsalou is Samuel Candler Dobbs Professor of Psychology at Emory University. He received a bachelor's degree in Psychology from UC San Diego in 1977, and a Ph.D. in Psychology from Stanford University in 1981. Since then Barsalou has held faculty positions at Emory University, the Georgia Institute of Technology, and the University of Chicago, returning to Emory in 1997. Barsalou's research addresses the nature of human knowledge, and its roles in perception, memory, language, and thought. The current theme of his research is that the human conceptual system is grounded in the brain's modality-specific systems. Other themes include the situated character of knowledge, the dynamic online construction of conceptual representations, the development of ad hoc categories to support goal achievement, the structure of knowledge, and category learning.

Barsalou's presentation will review basic research that bears on recent attempts in industry to study semantics in the wild. Relevant research includes the roles of goals and background knowledge in making information relevant, the importance of situations in organizing information, the dynamic construction of categories that cross-classify the same referents, and the infinite construals of referents that are possible. Barsalou's presentation will also point to gaps in basic science where future research could be performed to support semantics in the wild.

References

Clark, H.H. (1996). *Using Language*. Cambridge University Press.

Dumais, S. T. (2003). Data-driven approaches to information access. *Cognitive Science*, *27*, 491-524.

Furnas, G. W., Landauer, T.K., Gomez, L. M., & Dumais S. T. (1987). The vocabulary problem in human-system communication, *Communications of the ACM, 30*, 964-971, 1987.

Glushko, R. & McGrath, T. (2005). *Document Engineering*. Cambridge, MA: MIT Press.

Hammond, T., Hannay, T., Lund, B, & Scott. J. (2005). Social bookmarking tools. *D-Lib Magazine*, 11(4).

Hutchins, E. (1996). *Cognition in the Wild.* Cambridge, MA: MIT Press.

Lakoff, G., & Johnson, M. (1987). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. HarperCollins Publishers.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573-605.

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review, 81*, 214 241.

Taylor, A. (2004). *The Organization of Information*. Libraries Unlimited.