Natural Language and Spatial Reasoning

Stefanie Tellex

October 13, 2009

Thesis Proposal for the degree of
Doctor of Philosophy in Media Arts and Science
at the Massachusetts Institute of Technology

Advisor:

Deb Roy
Associate Professor of Media Arts and Sciences
AT&T Career Development Professor
Chair, Academic Program
Director, Cognitive Machines
MIT Media Lab

Reader:

Boris Katz
Principal Research Scientist
MIT Computer Science and Artificial Intelligence Laboratory

Reader:

Yuri Ivanov
Senior Research Scientist
Mitsubishi Electric Research Laboratories

Executive Summary

What semantic structures can enable a system to understand and use spatial language in realistic situations? This thesis will develop a model of spatial semantics that can enable a system to understand and use spatial language in real-world domains: finding video clips that match a spatial language query, and enabling a robot to understand natural language directions. Understanding spatial language expressions is a challenging problem because linguistic expressions, themselves complex and ambiguous, must be connected to real-world objects and events. I will build systems that bridges this gap by modeling the meaning of spatial language expressions hierarchically, first capturing the semantics of spatial prepositions, and then composing these meanings into higher level structures. Spatial prepositions will be instantiated as classifiers which can recognize examples of that preposition. The classifiers will be trained using a library of features computed from a geometric model of the situation. A key contribution of this work will be creating and characterizing this library and showing that it is able to correctly model a variety of spatial relations in realistic domains. Next, these models will be composed to create higher level models of spatial relations. In robotic direction understanding, this will be done by creating a Markov Random Field model of a path through a space, and finding the path that maximizes the likelihood of the directions. For surveillance video search, data can be collected from annotators describing a person's movement over many seconds or minutes of video, and a similar model can be used to find and score video clips that match the query. The result will be a computational model of spatial semantics that connects symbols to real-world paths and movements of people. Landau and Jackendoff (1993), Talmy (2005) and others have characterized semantic structures underlying spatial language without an implemented computational model. Regier (1992), Carlson and Covey (2005) and others have created computational models of spatial semantics focusing on schematic visual contexts, rather than realistic situations such as video and path descriptions. This work will move beyond previous work by creating implemented computational models of spatial language expressions, and then evaluating those models on two realistic, real-world domains. Through this work I will create new models of spatial semantics that are faithful to human judgements and robust to the noise and ambiguity inherent in real-world situations.

1 Introduction

Building a computer that can understand natural language has been a dream of artificial intelligence since the Turing test was first proposed in 1950. Language is uniquely human: through language humans unify and integrate many disparate abilities. But many of these abilities seem to have nothing to do with language, such as vision, movement, and social reasoning. One fruitful strategy when faced with a large problem is to divide it into smaller subproblems. This approach has been hugely successful, leading to amazing progress in individual problem areas of artificial intelligence, from web search to autonomous robotic navigation. However there has been less work towards integrating results from different subfields into a consistent and coherent framework. Models from computational linguistics often model only words, and not the non-linguistic components of semantics. In this thesis I will divide the problem of language understanding not horizontally, but vertically, focusing on a narrow subset of language, but *grounding* that language in data collected from a real world. This strategy has two benefits. First, it decreases the scope of the language understanding problem, making it more tractable. Second, by choosing a semantically deep core domain, it offers an opportunity to explore the connection between linguistic and non-linguistic concepts. A model that spans domains may be more powerful and generalizable to new situations and contexts, a key challenge in creating intelligent systems.

When pursuing this strategy, the choice of sub domain is key. In this thesis I will study spatial language. Reasoning about movement and space is a fundamental competence of humans and many animals. Humans use spatial language to tell stories and give directions, abstracting away the details of a complex event into a few words such as "across the kitchen." A system that understands spatial language could be directly useful to people by finding video that matches spatial language descriptions, or giving natural language directions. In order to reason in this domain, the system must somehow map information about the world into the language of interest. This thesis will implement a model for one way that this mapping could be carried out, and apply the model to two realistic domains.

Applying the model to a realistic domain is important to show that it works. Working with more than one domain means the model must be more generalizable, making it more likely the ideas will apply in a wide array of contexts. Moreover, a system that works in a realistic domain is useful in its own right, making it more likely that the ideas behind it will have impact. The choice of domains depends on several factors. First, the applications should be compelling, with use cases that have impact. Second, the scope of the problem should be neither too small nor too large. Finally, the two domains should be similar enough to reuse the same modeling machinery, but different enough to require interesting generalizations.

In this thesis I will focus on natural language video retrieval and direction giving and understanding for robots. Video retrieval is a compelling application: in the United States alone, there are an estimated 30 million surveillance cam-

eras installed, which record four billion hours of video per week.(Vlahos, 2008) Analyzing and understanding the content of video data remains a challenging problem. A spatial language interface to video data can help people naturally and flexibly find what they are looking for in video collections. For example, a system installed in the home of an elder could help a health worker diagnose problems the person is having as they age, and design effective interventions to enable them to continue living at home. Studying language used to give directions could enable a robot to understand natural language directions. People talk to robots even if they do not have microphones installed (Kinzer, 2009), and it makes sense to build systems that understand what they say. A robot that understands natural language is easy for anyone to use without special training.

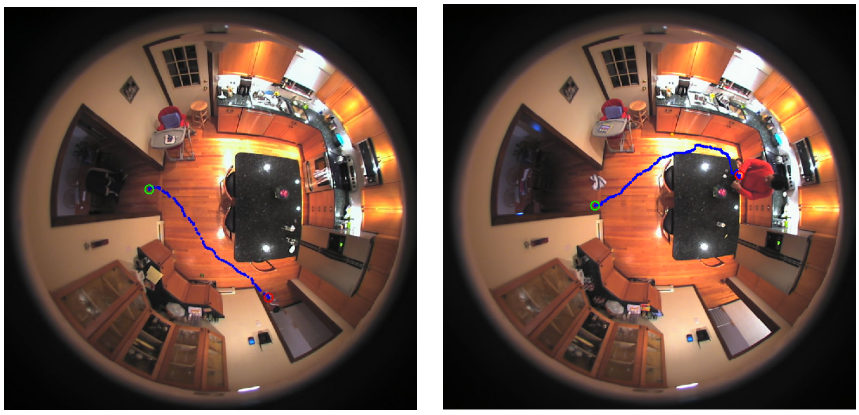
Both these domains require a system that can connect language to concrete non-linguistic contexts. A corpus can be collected that maps to specific non-linguistic events. In the case of video retrieval, I have collected a corpus of natural language descriptions of video clips. The video corpus consists of data recorded from a fish-eye camera installed in the ceiling of a home. Sample frames from this corpus, retrieved by the system for the query "across the kitchen," are shown in Figure 1. To associate video clips with a natural language description, annotators were asked to write a short phrase describing the motion of a person in the clip. Using this corpus of descriptions paired with video clips, I trained models of the meanings of some spatial prepositions, and explored their semantics by analyzing which features are most important for good classification performance. Figure 2 shows the most frequent descriptions that appeared in this corpus. These corpora provide a way to train and evaluate models of spatial semantics in a restricted, but still realistic context.

For robotic language understanding, I am working with a corpus collected by Nick Roy's lab at CSAIL. Subjects were given a tour of a room and asked to write down directions between two locations inside a building. The corpus consists of paragraph length descriptions containing sentences. One set of instructions from the corpus is shown in Figure 4. They have also collected images and sensor data from a robot that drove through the environment, allowing it to create a map of the environment populated with many landmarks. Figure 3 shows the most frequent spatial relations from this corpus. This data offers an opportunity to map the language to the nonlinguistic spatial context.

2 System

Although these two domains seem different, the machinery to parse language, extract spatial relations, and map it to geometric contexts can be shared between them. For video retrieval, a system can model the clip as a schematic diagram representing a person's path together with reference objects. For direction understanding, the robot could create a map, together with the probable locations of landmarks. In both contexts, the system must map linguistic structures to structures in these domains.

This work will decompose the problem of understanding spatial language into



(a) (b)

Figure 1: Frames from two clips returned for the query "across the kitchen."

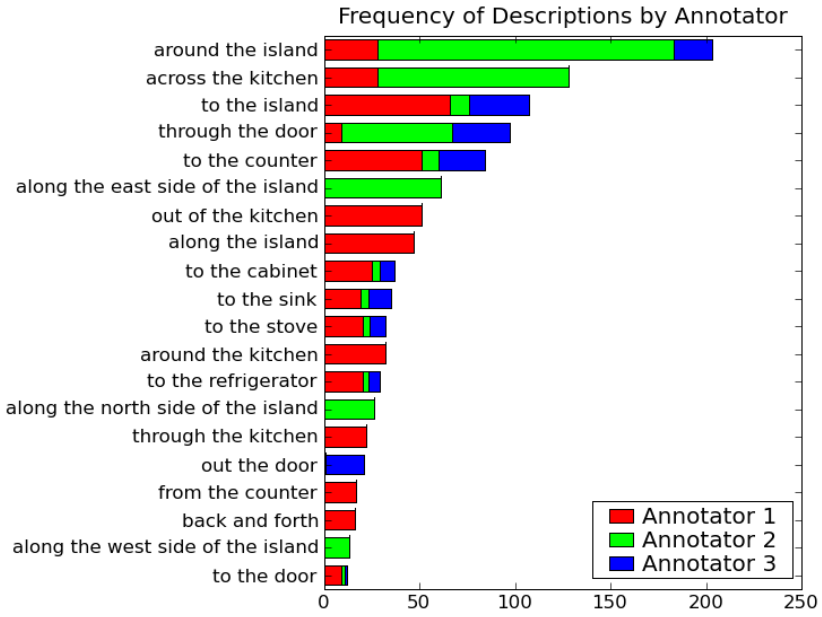


Figure 2: Histogram of descriptions of video clips. Annotators watched a several second long clip and completed the sentence "The person is going ..."

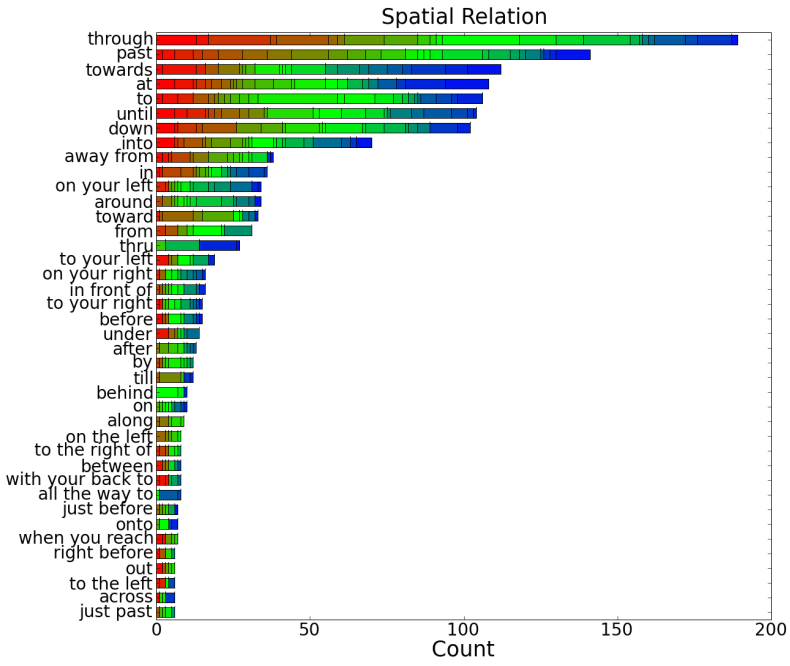


Figure 3: Histogram of the most frequent spatial relations in the route instructions corpus. Each color represents a separate annotator.

With your back to the windows, walk straight through the door near the elevators. Continue to walk straight, going through one door until you come to an intersection just past a white board. Turn left, turn right, and enter the second door on your right (sign says "Administrative Assistant").

Figure 4: A set of directions through an office space from the route instruction corpus.

two parts: understanding spatial expressions, and composing those expressions together into compound structures. Spatial expressions such as "across the kitchen" will be modeled by visual classifiers that take spatial paths as input. These classifiers are trained using labeled path examples. Continuous geometric paths are converted into a set of features motivated by theories of human spatial language (Talmy, 2005; Jackendoff, 1983; Landau and Jackendoff, 1993).

Spatial expressions will be composed together by abstracting spatial language expressions as a list of spatial description clauses. Each spatial description clause consists of four fields: figure, verb, spatial relation, and landmark. Figure 5 shows the spatial description clauses for a sentence from the route instruction corpus. Spatial description clauses can be automatically extracted from the linguistic structure of the text. A set of route directions, or a description of a person's movement, can be modeled as a sequence of spatial description clauses. Each element in the sequence maps to some part of the movement. A Markov Random Field (MRF) model will be created to model probability of a path given a sequence of SDCs and probable locations of landmarks in the environment. The path that maximizes this probability can be used by a robot seeking to follow natural language directions or return video clips that match a natural language description. Classifiers that model the semantics of spatial prepositions are a key component, specifying how each entry in the mapping connects to the geometric context.

After applying this model to the route instruction corpus, I will apply it to information retrieval. Annotators could describe a person's movement over a minute of video. A system could extract entity relations from these paragraphs, and an MRF could score video clips based on how well they match the description. This method could be used for video retrieval by finding clips that match more complicated natural language descriptions.

3 Related Work

There is a long history of systems that understand natural language in small domains, going back to Winograd (1970). This work builds on previous work by bringing the system in contact with realistic data: it is created and evaluated using data from a corpus of naive annotators who are not otherwise involved with the development of the system, and because the language is about real situations, describing routes in office buildings and people's movement. The system must tackle unsanitized language not tuned and filtered by author of the system. A constant theme in the work will be the struggle to balance open-ended natural language understanding with the limitations arising from the sensing and understanding capabilities of the system.

The linguistic structure extracted from spatial language expressions and many of the features in the model are based on the theories of Jackendoff (1983), Landau and Jackendoff (1993) and Talmy (2005). This work aspires to be a computational instantiation of some of the ideas in their theories. For example, Talmy (2005) says that for a figure to be across a particular ground,

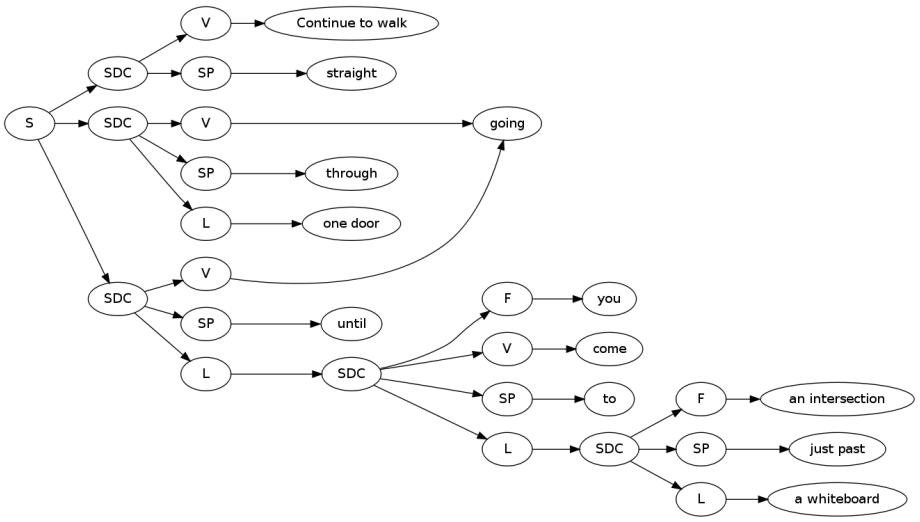


Figure 5: Entity structure for the sentence "Continue to walk straight, going through one door until you come to an intersection just past a white board."

among other things, the axes of the two objects must be "roughly perpendicular." The implementation of "across" in this work extends his definition by giving an algorithm for computing the axes a figure imposes on a ground, and a set of features which quantify "roughly perpendicular," using a machine learning algorithm to fine-tune the distinctions by training on labeled data.

Others have implemented and tested models of spatial semantics. Regier (1992) built a system that assigns labels such as "through" to a movie showing a figure moving relative to a ground object. Kelleher and Costello (2009) and Regier and Carlson (2001) built models for the meanings of static spatial prepositions such as "in front of" and "above." Building on their paradigm of testing the semantics of spatial prepositions against human judgements, this work focuses on realistic situations, requiring the model to be robust to noise, and enabling an analysis of how the semantics of spatial prepositions change in different real-world domains.

Katz et al. (2004) built a natural language interface to a video corpus which can answer questions about video, such as "Show me all cars leaving the garage." Objects are automatically detected and tracked, and the tracks are converted into an intermediate symbolic structure based on Jackendoff (1983) that corresponds to events detected in the video. This work focuses on handling complex spatial prepositions such as "across" while they focus on understanding a range of questions involving geometrically simpler prepositions.

Researchers have developed video retrieval interfaces using non-linguistic input modalities which are complementary to linguistic interfaces. Ivanov and Wren (2006) describe a user interface to a surveillance system that visualizes information from a network of motion sensors. Users can graphically specify

patterns of activation in the sensor network in order to find events such as people entering through a particular door. Yoshitaka et al. (1996) describe a query-by-example video retrieval system that allows users to draw an example object trajectory, including position, size, and velocity, and finds video clips that match that trajectory. The natural language query interface that will be developed in this work would complement these interfaces in several ways. First, queries expressed as text strings are easily repeatable; in contrast, it is difficult to draw (or tell someone else to draw) the exact same path twice in a pen-based system. Second, language can succinctly express paths such as "towards the sink," which would need to be drawn as many radial lines to be expressed graphically. The combination of a pen-based interface and a natural language interface is more powerful than either interface on its own.

3.1 Understanding Natural Language Directions

Many authors have proposed formalisms similar in spirit to spatial description clauses for reasoning about the semantics of natural language directions. Many of these representations are more expressive than SDCs, but correspondingly more difficult to automatically extract from text, to the point where many authors sidestep this problem by using human annotations. SDCs capture many of the semantics of natural language directions, while still being simple enough to extract and reason about automatically.

For example, Levit and Roy (2007) describes a probabilistic model for finding likely paths described by dialogs from the MapTask corpus. Semantic units called navigational informational units (NIUs) are annotated in the text, and the system finds paths given a seequence of NIUs. This formulation is the most similar to SDCs of the frameworks reviewed here. For a phrase like "move two inches towards the house," an NIU contains a path descriptor ("move... towards"), a reference object ("the house"), and a quantitative description ("two inches"). Spatial description clauses break down the instruction in a similar way, separating the path descriptor into a verb and spatial relation, and not explicitly modeling the quantitative description, since it appears so infrequently in our corpus. The possible path descriptors of their formalism correspond to spatial relations in our framework. The SDC formalism explicitly represents the structure common to any spatial referring expression, whether it refers to a position, an orientation, a move, or a compound reference to a landmark such as "the doors near the elevators."

Bugmann et al. (2004) identified a set of 15 primitive procedures associated with clauses in a corpus of spoken natural language directions. This work follows their methodology of corpus-based robotics, working with natural language directions given by a human for another human. An individual spatial description clause in our framework could correspond to one of their primatives actions. Spatial description clauses explicitly represents the structure common to all of their primatives, enabling a factorization of the system into a spatial-relation processing module, and a landmark processing module, both of which can be used in other applications.

Macmahon (2006) built a system that follows natural language directions created by a human for another human through a simulated environment. His system represents each clause in a set of directions as one of four simple actions: move, turn, verify, and declare-goal. A parser extracts these simple actions from text, and forms compound actions, consisting of a simple action plus pre-conditions, while-conditions, and post-conditions. A compound action in his formalism is roughly equivalent to an SDC. This framework is more expressive than SDCs: a compound action can have more than one pre-, post-, and while-conditions. For example for "Follow the atrium all the way to the right," "follow the atrium" can be seen as a while-condition, while "all the way to the right" describes a post-condition for the path segment. However, clauses involving more than one pre-, post-, or while-conditions are relatively rare in the corpus of anatural language directions. When they occur, they are modeled as separate spatial description clauses.

Klippel et al. (2005) created a formalism for representing route knowledge called wayfinding choremes. At each decision point in the route, a possible direction to take is discritized into one of seven equidistant directions. (The directions can be lexicalized as sharp right, right, half right, straight, half left, left, sharp left. Back is a special case.) A sequence of wayfinding choremes can be chunked together to create higher-order direction elements. In this model turning actions are seen as primary. Like Klippel et al. (2005), the SDC model discritizes orientation. However, rather than treating turning as primitive, in this model landmarks are the key feature used to connect between natural language directions and the external world. Each SDCs describes a transition between two viewpoints, almost always with respect to a landmark: only 21% of the SDCs in the corpus appear without an associated landmark. Landmarks are a key part of natural language directions, so the formalism represents them explicitly.

Dzifcak et al. (2009) created a language understanding system that simultaneously builds semantic structures representing both the goal of a sentence such as "Go to the breakroom," as well as the action needed to achieve that goal. A combinatory categorial grammar (CCG) parser extracts both structures from the input text. The CCG formalism enables the robot to understand complex commands going beyond following directions, such as "Go to the breakroom and report the location of the blue box." This work takes a different strategy: rather than trying to extract the entire linguistic structure from natural language input, and understanding it completely, the system extracts a simplified, domain specific representation. Because the representation is simple and domain specific, the extraction is robust to ungrammatical sentences such as "go to hallway," and can follow directions from untrained users with high accuracy.

4 Contributions

The key scientific contribution of this thesis is a model of spatial semantics that enables a system to understand and use spatial language in real-world

domains. Spatial prepositions in English will be defined in terms of a set of features extracted from the two-dimensional geometry of a scene. I will apply this lexicon of spatial relations to two real-world problems: natural language video retrieval and natural language direction understanding. This effort will show the effectiveness of the features and provide an opportunity to analyze their performance in order to study which ones perform best. The thesis will advance the state of the art in natural language understanding and grounding by connecting spatial language to real-world domains.

5 Schedule

- September 14, 2009 - Submit a paper about understanding natural language directions to HRI (Human-Robot Interaction).
- September 2009 - Proposal defense.
- October 7, 2009 - Collect corpus of paragraph descriptions of video.
- October 12, 2009 - Demo of generating natural language descriptions for video clips. (For sponsor week.)
- November 14, 2009 - Implement SDC model on that corpus.
- November 30, 2009 - Evaluate model.
- January 31, 2010 - Submit a paper about video retrieval with longer descriptions to SIGIR.
- February 1, 2010 - Thesis outline to Deb.
- February 7, 2010 - Complete introduction and contributions.
- February 14, 2010 - Complete related work chapter.
- January 22, 2010 - Submit paper analyzing spatial prepositions in direction understanding and video retrieval, comparing their meanings, to ACL.
- February 28, 2010 - Complete chapter about all the features and their importance.
- March 7, 2010 - Complete chapter about the different models and inference.
- March 14, 2010 - Complete chapter about evaluation.
- March 21, 2010 - Complete chapter about corpus and data collection.
- March 31, 2010 - Draft to committee.
- May, 2010 - Thesis defense

References

Bugmann, G., E. Klein, S. Lauria, and T. Kyriacou (2004). Corpus-Based robotics: A route instruction example. *Proceedings of Intelligent Autonomous Systems*, 96—103.

Carlson, L. A. and E. S. Covey (2005). How far is near? Inferring distance from spatial descriptions. *Language and Cognitive Processes 20*, 617–631.

Dzifcak, J., M. Scheutz, C. Baral, and P. Schermerhorn (2009). What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *IEEE International Conference on Robotics and Automation (ICRA-2009)*, pp. 4163–4168.

Ivanov, Y. A. and C. R. Wren (2006). Toward spatial queries for spatial surveillance tasks. In *Pervasive: Workshop Pervasive Technology Applied Real-World Experiences with RFID and Sensor Networks (PTA)*.

Jackendoff, R. S. (1983). *Semantics and Cognition*, pp. 161–187. MIT Press.

Katz, B., J. Lin, C. Stauffer, and E. Grimson (2004). Answering questions about moving objects in surveillance videos. In M. Maybury (Ed.), *New Directions in Question Answering*, pp. 113–124. Springer.

Kelleher, J. D. and F. J. Costello (2009, June). Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics 35*(2), 271–306.

Kinzer, K. (2009). Tweenbots: Robot/people art. http://www.tweenbots.com.

Klippel, A., H. Tappe, L. Kulik, and P. U. Lee (2005). Wayfinding choremes–a language for modeling conceptual route knowledge. *Journal of Visual Languages & Computing 16*(4), 311–329.

Landau, B. and R. Jackendoff (1993). "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences 16*, 217–265.

Levit, M. and D. Roy (2007). Interpretation of spatial language in a map navigation task. *Systems, Man, and Cybernetics, Part B, IEEE Transactions on 37*(3), 667–679.

Macmahon, M. (2006). Walk the talk: Connecting language, knowledge, and action in route instructions. *In Proc. of the Nat. Conf. on Artificial Intelligence (AAAI*, 1475—1482.

Regier, T. and L. A. Carlson (2001, June). Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology. General 130*(2), 273–98. PMID: 11409104.

Regier, T. P. (1992). *The Acquisition of Lexical Semantics for Spatial Terms: A Connectionist Model of Perceptual Categorization.* Ph. D. thesis, University of California at Berkeley.

Talmy, L. (2005). The fundamental system of spatial schemas in language. In B. Hamp (Ed.), *From Perception to Meaning: Image Schemas in Cognitive Linguistics.* Mouton de Gruyter.

Vlahos, J. (2008). Welcome to the panopticon. *Popular Mechanics 185*(1), 64.

Winograd, T. (1970). *Procedures as a representation for data in a computer program for understanding natural language.* Thesis, Massachusetts Institute of Technology. Thesis (Ph. D.)–Massachusetts Institute of Technology, Dept. of Mathematics, 1970.

Yoshitaka, A., Y. Hosoda, M. Yoshimitsu, M. Hirakawa, and T. Ichikawa (1996). Violone: Video retrieval by motion example. *Journal of Visual Languages and Computing 7*, 423–443.

6 Acknowledgements

The work on natural language direction understanding was done in collaboration with Thomas Kollar.

7 Biography

Stefanie Tellex has been interested in artificial intelligence since she first started programming her graphing calculator as a high-school student. After spending the dot-com years working part-time for Cambridge-area startups and an undergraduate degree in computer science at MIT, she worked with Boris Katz at the MIT Artificial Intelligence Laboratory studying passage retrieval algorithms for factoid question answering. This work won the Best Student Paper award at SIGIR 2003, and formed the basis of her Masters of Engineering thesis. After leaving MIT, she worked for one year at LifeHarbor, Inc, a startup building software for managed accounts, and then started graduate school at the Media Lab, working with Professor Deb Roy. During her time at the Lab, she has developed semantic models of spatial reasoning, applying them to a speech-contorlled vehicle, a real-time strategy game and a humanoid robot. For her Ph.D thesis she is creating a spatial language video retrieval system and a robotic direction understanding system. In her spare time she enjoys programming, sailing, reading books, watching fish, and playing with cats.