CoSLI 2011
Computational Models of Spatial Language Interpretation and Generation
– Preface –

Joana Hois[1], Robert J. Ross[2], John D. Kelleher[2], and John A. Bateman[3]

[1] University of Bremen, Germany
joana@informatik.uni-bremen.de
[2] Artificial Intelligence Group, Dublin Institute of Technology, Ireland
robert.ross@dit.ie,johnd.kelleher@dit.ie
[3] English Department and Research Center on Spatial Cognition (SFB/TR8), University of Bremen, Germany
bateman@uni-bremen.de

Competence in spatial language modeling is a cardinal issue in disciplines including Cognitive Psychology, Computational Linguistics, and Computer Science. Within Cognitive Psychology, the relation of spatial language to models of spatial representation and reasoning is considered essential to the development of more complete models of psycholinguistic and cognitive linguistic theories. Meanwhile, within Computer Science and Computational Linguistics and Engineering, the development of a wide class of so-called situated systems such as robotics, virtual characters, and Geographic Information Systems is heavily dependent on the existence of adequate models of spatial language use.

Achieving competence in spatial language requires that appropriate meanings be assigned to spatial terms used in language, such as location, motion, orientation, perspective, projective, topological, distance, or path descriptive markers. The computational modeling of such spatial language meanings in turn supports the interpretation of an intended spatial meaning as well as the generation of adequate linguistic expressions in certain situations and contexts. While early computational models for spatial language interpretation and generation primarily focused on a geometric understanding of spatial terms, it is now widely recognized that spatial term meaning depends on functional and pragmatic features in many ways. Competent models of spatial language interpretation and generation must thus draw on complex models of situated meaning by developing heterogeneous approaches with qualitative and quantitative models and by combining geometric, functional, pragmatic, and cognitive features in multi-modal contexts and applications.

Drawing together theories and results in spatial language modeling is a critical research topic for a range of research disciplines. These includes not only Psychology where computational theories can be used to bind experimental results and models, but also disciplines from the wider community, including: Artificial Intelligence, Computational Linguistics, Human-Robot Interaction, Ontology Engineering, the Semantic Web, and Geographic Information Systems.

The main objective of the CoSLI-2 workshop is to foster computational formalisms and approaches for interpreting or generating spatial language that take into account cognitive, functional, or embodiment criteria in modeling. In particular, this years workshop theme is "*Function in Spatial Language: From evidence to execution*", and we welcome in particular any contributions which aim to address the issues of modeling function or pragmatic features in spatial language interpretation or generation. More generally, the workshop also welcomes contributions that address symbolic and embodied spatial language interpretation and generation. This topic remains an ongoing issue in both natural language processing and cognitive science, and novel work is encouraged. Such work includes both formal and empirical models of spatial language templates and linguistic calculi, corpus-based and statistical methods, combinations of symbolic and sub-symbolic representations, and aspects of sensory-motor and multi-modal information. Contributions to spatial language interpretation and generation that integrate results from empirical and psychological frameworks for spatial language and that can improve and support situated natural language systems are also particularly welcomed.

Workshop Organization

Co-Chairs

Joana Hois	University of Bremen, Germany
Robert J. Ross	Artificial Intelligence Group, Dublin Institute of Technology, Ireland
John D. Kelleher	Artificial Intelligence Group, Dublin Institute of Technology, Ireland
John A. Bateman	English Department and Research Center on Spatial Cognition (SFB/TR8), University of Bremen, Germany

Program Committee

Marios Avraamides	University of Cyprus, Cyprus
Kenny Coventry	Northumbria University, UK
Alexander Klippel	Penn State, USA
Alexander Koller	University of Potsdam, Germany
Gérard Ligozat	University of Paris-Sud, France
Matt Mac Mahon	Google, USA
Amitabha Mukerjee	Indian Institute of Technology Kanpur, India
Philippe Muller	Université Paul Sabatier, France
Robert Porzel	University of Bremen, Germany
David Schlangen	University of Potsdam, Germany
Emile van der Zee	University of Lincoln, UK
Joost Zwarts	Universiteit Utrecht, Netherlands

Invited Speaker

Kenny Coventry	Northumbria University, UK

Acknowledgements

We acknowledge generous financial support from the DFG-funded Research Center on Spatial Cognition (SFB/TR 8) situated at the Universities of Bremen & Freiburg, Germany, and from the Artificial Intelligence Group situated at the Dublin Institute of Technology, Ireland. We would like to thank the PC members for their timely reviewing work, our invited speaker, Kenny Coventry, for delivering the keynote presentation at the workshop, and especially Barbara Landau and Kenny Coventry for contributing to our panel session during the workshop.

We would also like to thank the organizers of the 33rd annual meeting of the Cognitive Science Society for hosting the CoSLI-2 workshop, in particular, Duncan Brumby, Kevin Gluck, Andy Stull, and Nicole Dillon for their support.

July 2011

J. Hois, R. Ross, J. Kelleher, J. Bateman
CoSLI 2011 Program Chair

Workshop Schedule

9:00 - 9:05	Opening
9:05 - 10:00	Invited Talk: Kenny Coventry "Spatial Language and the Dynamics of Meaning"
10:00 - 10:30	Kalyan Moy Gupta, Abraham R. Schneider, Matthew Klenk, Kellen Gillespie, and Justin Karneeb "Representing and Reasoning with Functional Knowledge for Spatial Language Understanding"
10:30 - 11:00	Coffee Break (poster setup)
11:00 - 11:30	Yunhui Wu and Stephan Winter "Interpreting Destination Descriptions in a Cognitive Way"
11:30 - 12:00	Masoud Rouhizadeh, Daniel Bauer, Bob Coyne, Owen Rambow, and Richard Sproat "Collecting Spatial Information for Locations in a Text-to-Scene Conversion System"
12:00 - 13:00	Poster Session
13:00 - 14:00	Break
14:00 - 14:30	Alice Ruggeri, Cristina Battaglino, Gabriele Tiotto, Carlo Geraci, Daniele Radicioni, Alessandro Mazzei, Rossana Damiano, and Leonardo Lesmo "Where should I put my hands? Planning hand location in sign languages"
14:30 - 15:00	James Pustejovsky, Marc Verhagen, Anthony Stefanidis, and Caixia Wang "Geolocating Orientational Descriptions of Landmark Configurations"
15:00 - 15:30	Coffee Break
15:30 - 16:00	Alexander Klippel, Sen Xu, Jinlong Yang, and Rui Li "Spatial Event Language across Geographic Domains"
16:00 - 17:00	Panel Discussion & Closing

Table of Contents

Representing And Reasoning with Functional Knowledge for Spatial Language Understanding

Kalyan Moy Gupta[1], Abraham R. Schneider[1], Matthew Klenk[2], Kellen Gillespie[1], and Justin Karneeb[1]

[1]Knexus Research Corporation; Springfield, VA 22153
[2] Naval Research Laboratory, Washington, DC 20745
firstname.lastname@knexusresearch.com |

Abstract. One of the central problems in spatial language understanding is the polysemy and the vagueness of spatial terms, which cannot be resolved by lexical knowledge alone. We address this issue by developing a representation framework for functional interactions between objects and agents. We use this framework with a constraint solver to resolve and recover the meanings of spatial descriptions for object placement tasks. We describe our approach in a virtual environment with an example of object placement task.

1 Introduction

Virtual scene (re)construction or object placement is a vital task in many practical applications such as background layout in 3-D animated movies, accident or crime scene simulation, and navigation maps for video game development. Using natural language (NL) commands can be a natural and efficient alternative to an otherwise effort intensive manual placement of objects in a virtual world (e.g., Coyne and Sproat, 2001; and Dupuy, 2001). However, machine understanding of natural language commands is notoriously difficult due to polysemy and vagueness of spatial terms. Only considering lexical semantic knowledge of spatial terms is clearly insufficient for this task; world knowledge and pragmatics must be considered for understanding language in a form that can be acted upon by autonomous agents.

Over the past couple of decades much research in spatial term semantics has focused on developing computational models that map utterances to semantics (e.g., Regier and Carlson, 2001; Coventry et al., 1994). Although, such research recognizes the need for pragmatic and functional knowledge about objects, the development of computational models for representing and using such knowledge has received little attention. To address this gap, we present a framework for representing world knowledge that can be effectively translated into spatial constraints to resolve vague and underspecified natural language commands. We present an algorithm that utilizes such knowledge for interpreting natural language commands and to perform valid least cost object placements.

We organize the remainder of this paper as follows. We explain the nature of linguistic underspecification in object placement tasks in the next section. We follow this with a description of our representation framework and an algorithm that performs linguistically commanded single object placement task. Next, we illustrate our approach with an example. Finally, we discuss the strengths and limitations of our approach and conclude the paper.

2 Vagueness in NL driven Object Placement

Consider the task of generating a static scene described by text utterances in a 3D virtual environment. For example, generating a scene with "a chair *in front of* the table" and subsequently placing a "printer *on* the table." The desired rendering of the scene is shown in Figure 1.

Figure 1. Example scene imagination based on linguistic description

The central issue in such a task is interpreting *vague* spatial prepositions such as `on` and `in-front-of` into *valid* object placements. The utterance "printer on the table" can only be judged as vague when attempting to place the printer in the World. For instance, the possible placements on the table are to the left, right, front, and behind the monitor. However, the placements in front and back of the monitor are functionally invalid for a human user. The utterance also does not specify the suitable orientation of the printer. Without such a specification, the printer could be oriented in numerous ways in relation to the monitor and the chair, only some of which would be valid. For example, the orientation shown in Figure 1 is a valid one. However, the orientations of the printer such as upside down or facing the wall would be invalid.

Clearly, functional knowledge of interaction between objects must be considered for generating valid placements. The question is what should be the content of such functional and world knowledge and how can it be utilized to recover the unspecified elements and generate a complete and valid specification for object placement. We answer this question in the next section.

3 Representation and Reasoning for Linguistically commanded Object Placement

Problem Task: Given a world, W, containing a set of objects, O, located in various places in the world and an underspecified linguistic command requesting to place a target object, o^t in W, find a location with the least *interaction cost* to place o^t. We return to the notion of interaction cost later in this section.

Functional Knowledge Representation. We introduce an autonomous agent, α, as the central element of a functional representation of objects, O, and their parts in the World. Given our goal of building agents that interact with humans, our representation encodes spatial constraints accordingly. We assume that α is human-like and interacts with objects using a set of *primitive actions* or perceived affordances (Gibson, 1977, Norman, 2002). We introduce a set of the following primitive actions:

1. *Reach*: the agent reaches for objects to manipulate and interact with them. Given our assumption that α is human-like, we subcategorize the *reach* interaction as follows:
    1.1. *Reach.Arm*: the agent reaches for objects with arms fully extended.
    1.2. *Reach.Forearm*: the agent reaches for objects with only the forearm extended.
    1.3. *Reach.Foot*: the agent reaches for object with its foot.
    1.4. *Reach.Assisted*: the agent reaches for objects with tools.
2. *See*: the agent obtains visual information from objects to perform reach actions. For an agent to see objects it must be oriented toward the objects. In certain situations the agent must be able to read the information present on the object. We represent this with the read action, a tighter constraint than see:
    2.1. *Read*: an agent reads the information present on the object such as signs or writing. Clearly, this can be subcategorized to read fine print, read normal print, read large print, read poster print etc.

We further assume that the agent performs these activities while it is located at certain places in W called *activity stations*, S. In addition, we assume that an agent has the following human-like *poses*; sitting, standing, and lying down.

We categorize the functional relation between objects into the following three types:

1. *Support*: this is a functional relation typically implied by the preposition "on" in English. For example, a table `supports` a printer and a printer is `supportedBy` a table.
2. *Contain*: this is a functional relation typically implied by the preposition "in". For example, a box `contains` the printer and a printer is `containedBy` a box.
3. *Group*: relates multiple objects into a spatial group. For example, a computer keyboard and display monitor may be related to each other by a spatial group relation.

Table 1. Example representation of functional interaction constraints for a Printer

Object/parts	Object	Agent		
	Interaction	Interaction	Pose	Activity Station
Printer/Parent		• Reachable.Arm • Visible	• Stand • Sit	Perimeter
Control panel	supportedBy(parent)	• Readable		
Connection panel	supportedBy(parent)	• Reachable.Arm		
Paper tray	containedBy(parent) contains(paper)	• Reachable.Arm		

An object and its various parts may solicit different functional interaction constraints for agents. A representation of functional interaction constraints for a printer is shown in the Table 1. We assume a canonical geo-orientation for the printer, that is, it is upright. The table specifies that the printer control panel must be readable to the agent, for example.

We introduce the notion of a *possible interaction space, PIS*, for an agent at an activity station (e.g., see Kurup & Cassimatis, 2010). As a simplification in this paper, we assume that the possible interaction space is a two dimensional region. Figure 2 shows the *PIS* with reachability, visibility, readability spaces.

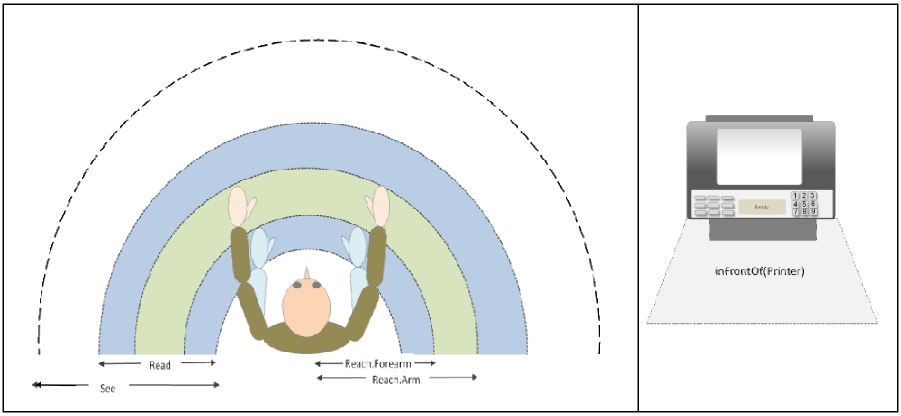


Figure 2. Possible interaction spaces for agent α and possible linguistically constrained space for infrontOf(Printer)

We introduce the notion of a *possible linguistically constrained space, PLCS,* as well. For example, Figure 2 shows the region selected by the function `inFrontOf(Printer)`. The possible space resulting from multiple constraints is the intersection of individual possible spaces. We will use this approach in the linguistically commanded single object placement algorithm presented next.

Linguistically commanded single object placement algorithm.
The algorithm begins by generating the set of potential activity stations to identify the smallest subset of that satisfies the spatial constraints in the World. The functional interaction knowledge of the objects in the World is transformed into spatial constraints. Next, it uses possibility spaces to identify the candidate placements and selects the one with least cost. We detail these steps below:

Inputs

1. O, set of objects in W.
2. o^t, the target object to be placed (e.g., Printer).
3. *lcs,* linguistically expressed placement constraint (e.g., on the table).
4. *KB*, the functional interaction knowledge base containing the agent and object interaction knowledge covering all objects in W (O and o^t).
5. α^{psp} the possibility space parameter for α for which the minimal cost placement is to be performed.

Output

1. P, a set of placements with minimum cost of functional interaction for agent α.

Processing steps

1. *Find the smallest set of activity stations,* S^{min}, that satisfes the functional interaction constraints for all objects $o^i \in O$; the constraints are retrieved from the *KB* for a given category of object. The candidate activity stations for an object o^i are located around its perimeter.
2. *Set the candidate placements,* $CP = \phi$, *placement cost,* pc=0
3. *Set candidate stations,* $S^c = S^{min}$
4. *For each activity station* $s^j \in S^c$

a. *Compute Possible Placement Space,* PPS^s for the target object o^t as the intersection of PIS, possible interactions space at the activity station and the LCS, linguistically constrained space:

$$PPS^s = PIS \cap PLCS$$

b. *Generate candidate placements (cp) and compute their cost, c*: The candidate placements are possible placements PPS^s if it is not empty. Only those placements that satisfy all the interaction constraints of o^t without violating any of the existing constraints satisfied by s^j are retained. The candidate placements are a combination of locations and orientations. For simplicity, we only consider 4 orientations of o^t relative to the orientation of the agent at the activity station. The cost of a placement is 0 when one of the existing stations is used for the placement.

5. *Select the minimum cost placements* P.
   IF $CP \neq \phi$ THEN
      Return minimum cost placements $P \subset CP$
   ELSE,
      Generate new activity stations (S^{new}) in the neighborhood of stations in S^c.
      set S^c=S^{new}, and
      IF *pc* is 0 set *pc*=1, i.e., cost of placement increases with the number of activity stations.
      go to Step 4

End.

Example

Consider a world W that includes a `table` placed against a `wall` with a `monitor` on it. In addition, it includes a `chair` located in front of the monitor (see Figure 3). The placement agent receives a linguistic command to place a `printer`, o^t, in this world; "Place the printer on the table".

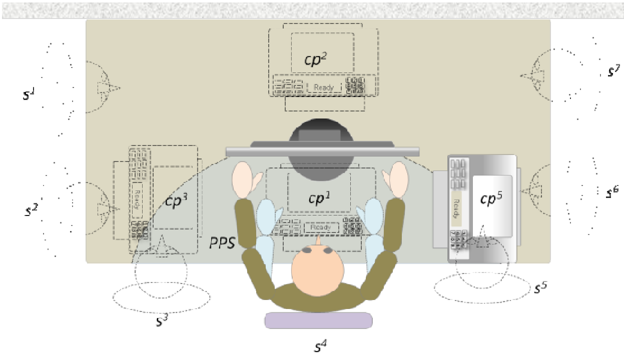


Figure 3. Place the printer on the table

We assume that this linguistic command (i.e., *lsc*) is interpreted into a semantic form and its $PLCS$ is computed, which is the entire surface of the table. We begin with step 1 to find S^{min}. The algorithm generates the potential activity stations in the world, for example, stations s^1 through s^7. It is easy to see that stations s^1, s^2, s^6 and s^7 do not satisfy the reachability and readability constraints for the monitor. Similarly, stations s^3 and s^5 fail to satisfy the readability constraint of the monitor. Notice that the alternative orientations of these stations would also fail on reachability constraints of various objects. Activity station s^4 is the only one that satisfies the reachability and readability constraints for the monitor and the reachability constraint of the chair (i.e., S^{min} ={ s^4}=S^c). We perform Step 4 and obtain *PPS* for s^4 shown in grey obtained by the intersection of *PIS* and *PLCS* (on table). Since *PPS* is not empty, we create candidate placements cp^1 through cp^4. Although cp^1 satisfies the printer's reachability and readability constraints, it violates the monitor's readability constraints for station s^4. Similarly, cp^2 fails to satisfy the readability constraint for the printer. Note that reorienting cp^3 to face the agent will create a valid placement. The candidate placement cp^3 satisfies all the constraints and is a valid. Placement cp^4 is not in the *PPS* space and is shown here for illustration only. Our example illustrates how the algorithm using functional knowledge about object and agent interactions produces two valid placements for a printer given a highly underspecified placement directive.

4 Discussion

Recent research on spatial language understanding has pointed out the need for functional representations for understanding spatial utterances. For example, Coventry and Garrord (2004) present a functional geometric framework which includes geometric and dynamic kinematic routines, and object knowledge. Our approach also considers the dynamic interactions and object knowledge. However, we explicitly consider the role of an agent along with a very small set of interaction primitive affordances specialized for the object placement task. Further, we present an inferencing algorithm that utilizes the world knowledge to perform valid object placement. Lockwood (2009) also emphasizes the need for functional knowledge but focuses on structure mapping as a means learning functional knowledge for a scene labeling task. However, she did not include an interpretation method to recover meanings of underspecified utterances. In contrast, we manually encode the affordances to recover underspecified spatial semantics in object placement tasks. We intend to develop methods of acquiring the interaction knowledge in our future work.

Although, we demonstrated the use of functional knowledge for generating valid object placement, we did not consider the pragmatic and contextual elements such as plans, goals, and the situation of the agent requesting object placements. For instance, the directive "put the printer on the table" would carry different functional constraints with it if the requester were a mover in an office building or a warehouse instead of a worker in an office building. We plan to extend our models to include constraint selection based on the requesting agent's goals and intentions.

5 Conclusion

Interpretation of spatial descriptions and commands, such as those for an object placement, poses significant challenges due to polysemy and underspecification of spatial term semantics. To address this issue, we developed a functional interaction knowledge representation framework with a very small number of agent action primitives and object to object interaction primitives. We described a cost based constraint satisfaction algorithm for utilizing world knowledge for object placement. In our future work, we will implement and evaluate the performance our algorithm with varying number of objects in the scene and consider aspects of visual attention to resolve residual ambiguities and diectics (e.g., see Kelleher, 2003). Additionally, we will extend our approach to include the role of goals and intentions of the requesting agent in selecting the appropriate spatial constraints for object placement.

Acknowledgements
This research was supported by the Office of Naval Research. Matthew Klenk was supported by NRC Postdoctoral Research Fellowship. We thank the two anonymous reviewers for their comments and suggestions.

References

Coventry, K.R. & Garrod, S.C. (2004) *Saying, Seeing and Acting: The Psychological Semantics of Spatial Prepositions.* Hove: Psychology Press.

Coventry, K.R., Carmichael, R. & Garrod, S.C. (1994). Spatial prepositions, object-specific function and task requirements, *Journal of Semantics*, 11, 289-309.

Coyne, B., & Sproat, R., (2001) WordsEye: An automatic text-to-scene conversion system, SIGGRAP'01, *Proceedings of the 28th annual conference on computer graphics and interactive techniques*, pp. 487-496, New York, NY: ACM.

Dupuy, S. (2001). Generating a 3-D simulation of a car accident from a written description in natural language: The CARSIM system. *Proceedings of the Workshop on Temporal and Spatial Information Processing*, pp.1-8.

Gibson, K.J. (1977). The theory of affordances. In R. Shaw and J Bransford (Eds.), Perceiving, acting, and knowing: Toward and ecological psychology (pp. 67-82). Hillsdale, NJ: Erlbaum.

Kelleher, J.D. (2003). A perceptually based computational framework for the interpretation of spatial language, *Ph.D Thesis*, Dublin, Ireland: School of Computing, Dublin City University.

Kurup, U., & Cassimatis, N.L., (2010). Quantitative spatial reasoning for general intelligence, *Proceedings of the Third Conference on Artificial General Intelligence Conference*, pp. 1-6, Lugano, Switzerland: AGF.

Lockwood, K. (2009). Using analogy to model spatial language use and multimodal knowledge capture, *PhD Thesis*, Department of Computer Science, Evanston, IL: Northwestern University.

Norman, D. (2002). *The design of everyday things*, New York, NY: Basic Books.

Regier, T., & Carlson, L. (2001). Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General, 130*, 273-298.

Interpreting Destination Descriptions in a Cognitive Way

Yunhui Wu* and Stephan Winter

Department of Infrastructure Engineering, University of Melbourne
Parkville, VIC 3010, Australia
y.wu21@pgrad.unimelb.edu.au,winter@unimelb.edu.au

Abstract. This paper proposes a cognitively motivated approach to interpreting destination descriptions without computing spatial relations. In contrast to other computational approaches, this approach is based on a few assumptions drawn from human communication behavior. Although this cognitively motivated approach is relatively simple, the performance of the approach is almost as good as other computational approaches.

Keywords: Destination description, spatial reasoning, spatial relation

1 Introduction

People provide destination descriptions when they specify where to go to. Destination descriptions are referring expressions [1] of the form "x related to y", where x is the destination, and y is a reference feature. A destination description is a reflection of the speaker's conceptual map of the environment in their mind. In geographic environments, people perceive salient features (landmarks), anchoring their mental representations of the environment [2]. They update their knowledge by linking new experiences of other features to the existing ones. Therefore it is natural for people to describe the location of features by addressing their spatial relation to other, more salient features. Using landmarks in destination descriptions is also a way to set up the common ground between parties in the communication: the speaker expects that the listener knows the landmarks due to their salience in the urban environment, and then, through the spatial relation with the landmarks, figure out where the destination is. This paper focuses on the spatial reasoning of using spatial relations in human destination descriptions, and proposes an approach to interpret these descriptions automatically to smarten the user interaction of navigation services.

Although *humans* have the capability of understanding the spatial relations in destination descriptions, making sense of spatial relations is not an easy task for *computational systems*. The major challenge is interpreting the qualitative

* This research was supported under Australian Research Council's Discovery Projects funding scheme (project number 0878119).

relations frequently used in destination descriptions. Characterizing or interpreting *topological* relations, such as *in*, requires access to the spatial extents of the involved spatial features. *Orientation* directions, such as *in front of*, require size and shape information. *Distance* relations, such as *near*, require taking contextual factors into account such as the size of the features, the purpose of the located feature, the distance from the observer, the functional relationship and interaction between the two features, and the asymmetry from the order in which locations are retrieved in memory [3–5]. However, so far there is no comprehensive computational model able to handle all these factors for qualitative spatial relations. Given these difficulties, this paper studies the spatial reasoning behind human communication behavior, and suggests a cognitively motivated approach to interpreting destination descriptions. This approach does not require any of the additional information, but instead is built on point locations (as given in standard gazetteers), and a salience model of geographic features. We will in particular demonstrate that the cognitively motivated approach can identify destinations without computing any other spatial relations than neighborhoods based on salience.

2 Related Work

Common ground is the basis of joint actions by speakers and listeners [6]. People do things based on individual beliefs or assumptions about what is common ground between each other. Clark identifies two kinds of common ground: communal common ground and personal common ground. Communal common ground is based on factors, such as communication parties' nationality, residence, education, occupation, and religion. Personal common ground is based on joint personal experiences. This paper assumes that the speaker refers to communal common ground, as in talking to strangers, such that spatial databases can be used to enable the interpreting process of destination descriptions.

Research has been made on formalization and computational modeling of spatial relations. Models for characterizing topological relations exist (e.g., [7]). A cognitive and computational model for *nearness* has been developed before [8]. Schlieder et al. propose to encode neighborhood relations in gazetteers for retrieving qualitative information [9]. Freksa develops an approach for representing qualitative spatial reasoning using orientation information [10], which is later developed into reasoning toolboxes [11]. But yet a comprehensive computational model for qualitative relations is not found.

3 Cognitive Motivated Approach to Interpreting Destination Descriptions

If destinations are hard to recognize, ambiguous or lacking in the common ground, people usually refer to the most salient landmark nearby according to their knowledge, which is chosen from potentially large numbers of spatial features available. From the speaker's perspective, the more salient the landmark is in the environment, the more likely it is to be known to the listener. However people perceive the urban environment variously. Petrol stations are more meaningful, thus salient, to car drivers than to walkers. Therefore it is more likely that car drivers refer to petrol stations in destination descriptions than walkers. It appears that the choice of a landmark relies more on its salience than on the spatial relation between the destination and the addressed landmark. Furthermore, spatial relations are mental connections, or characterizations of the configuration of spatial features at particular locations [12]. Therefore when different speakers refer to the same landmark, they may use different terms to depict the spatial relation or even different relations. Here the first assumption is:

– It is always the most salient landmark chosen among others, no matter what the type of the spatial relation between the landmark and the destination is.

This assumption establishes a basis for interpreting spatial relations without computing them explicitly. By saying "the most salient landmark among others", there must be implied a spatial restriction from which landmarks are selected. This restriction can be derived from the principle of relevance [13], which we apply here by a second assumption. We expect that the landmark has to include the destination within their neighborhood – a concept that needs to be further formalized. If the destination is not in the neighborhood of the landmark, the relationship is too weak to use in the destination description, since the relationship to another landmark is stronger. So the second assumption is:

– The landmark is chosen only if the destination is within the landmark's neighborhood.

By referring to a chosen landmark, the speaker wants to ensure that the listener can figure out the destination effectively and unambiguously. If there are two pizza shops in the neighborhood of the landmark, the speaker has some choices to disambiguate. They can specify the name of the target one, such as "the *Pizza Hut* next to the petrol station" (i.e., not *Domino's*), or employ another landmark to avoid such confusion, like "the pizza shop *opposite the church*" (which is also next to the petrol station, but does not apply for *Domino's*). Or they can name a disambiguating spatial relation, like "the pizza shop *left of* the petrol station" (instead of the one right of the petrol station). Except for the third case, it can be inferred that in destination descriptions the destination is unique within the neighborhood of the chosen landmark. The third case can be discovered either from inflection (where the relation would be stressed), or from discovering the ambiguities in the interpretation. The third case requires special treatment, but for the other cases we can make our third assumption:

– The landmark is chosen because it is sufficient enough to disambiguate the destination.

These assumptions require a computational model of neighborhood. Moulin et al. advise that the influence area of a spatial feature defines the portion of neighborhood in which every other features are spatially related to the located feature in a qualitative way [14]. The influence of a spatial feature can be used to define its neighborhood in the environment. Salience of landmarks represents their influence: the more salient a landmark, the larger its influence area. Winter et al. suggest a method of generating a hierarchical partition establishing the neighborhoods of landmarks at different levels of salience (or context) [15]: landmarks are grouped by their salience at different levels in a hierarchy, and then Voronoi cells representing the neighborhood of landmarks are created at each level (Figure 1).

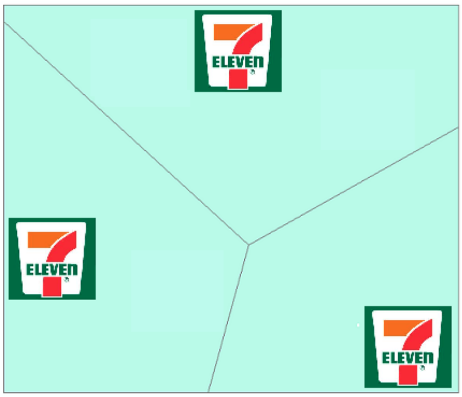


Fig. 1. Illustration of Voronoi diagram between landmarks of similar salience.

The interpretation of a destination description starts with a list of identified potential destination candidates. In the example above, it would start with a list of all known pizza shops. If there is no destination (no pizza shop) found in the database, the interpretation fails, similar to lacking common ground in human-to-human communication. If there is only one pizza shop found in the database, the interpretation completes successfully, and the relation to the landmark can only be used in an affirmative way. But if there are multiple destination candidates found, then the assumptions above will allow the disambiguation of destination candidates. This disambiguation process will use only the locations of landmarks, the salience of landmarks, and the location of destination candidates. The interpretation process computes a second list, namely a list of landmark candidates, e.g., all petrol stations. If this list is empty, no common ground could be established. If exactly one landmark candidate is found, the nearest destination is considered as a solution. If multiple landmarks are found, their neighborhoods are computed [15], and the one that has a unique destination candidate in its neighborhood identifies the destination.

This algorithm discovers automatically the third case – where the spatial relation is used for disambiguation – when no landmark has a unique destination candidate in their neighborhood. In this case the algorithm has to fall back to computing the spatial relations (which is possible, but not addressed in this paper).

This interpretation process offers an approach that avoids in many cases computing of spatial relations. The next section explains by example how this approach works.

4 Example and Discussion

Angela wants to meet her friend for lunch, and says "let's meet at the pizza shop next to the 7-Eleven". The pizza shop is the destination (x), and the 7-Eleven is chosen as the landmark (y). This section demonstrates how the cognitively motivated approach interprets the spatial relation in this destination description, and finds "the pizza shop". At first it is supposed that the spatial restriction of this communication is known from context (the area shown in Figure 2). In this area three 7-Eleven and three pizza shops are found (Figure 2, left). The 7-Eleven are of similar salience, therefore no hierarchy is created. The neighborhoods of three 7-Eleven are defined by Voronoi cells. In Figure 2, left, the 7-Eleven at the

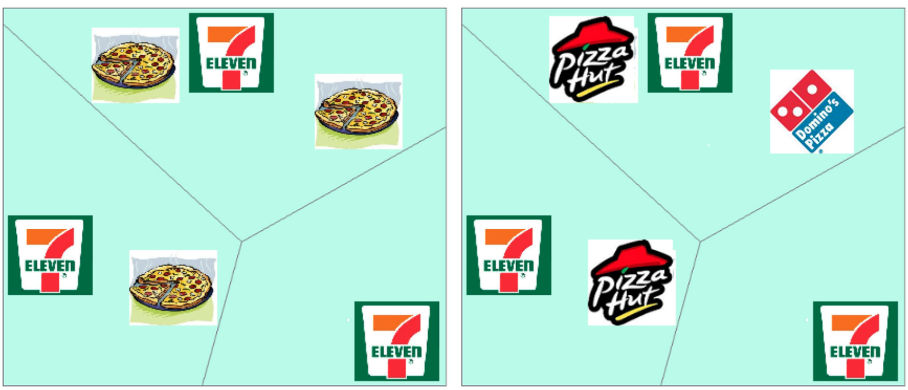


Fig. 2. Left: Three 7-Elevens and three pizza shops found in neighborhoods of the landmarks; Right: Two pizza shops are *Pizza Hut*, and one is *Domino's*.

right-bottom corner has no pizza shop in its neighborhood, thus does not define any destination; the 7-Eleven on the top has two pizza shops in its neighborhood, thus destination candidates are found ambiguous; and the 7-Eleven on the left has a unique pizza shop in its neighborhood. Assuming the rules of relevance theory, this pizza shop would be the target destination. Therefore, the hypothesis is proven. This is the general process of the cognitively motivated approach, and the computation complexity is $O(n)$. In comparison, other computational approaches need to compute the spatial relation between each 7-Eleven and pizza shop, and then check whether the relationship is a "next to" relation. If this process identifies any nearest pizza shop to a 7-Eleven, it could omit less relevant pizza shops. Thus the computation complexity is $O(n^2)$.

In some cases, the general process cannot obtain a unique result. For example, Angela specifies the name of the pizza shop by saying "let's meet at the Pizza Hut next to the 7-Eleven". Figure 2, right, shows a unique Pizza Hut in two 7-Eleven's neighborhood separately. In this case, the two 7-Eleven (on the top and on the left) are results through this cognitively motivated approach, and further refinement to resolve the remaining ambiguity is needed.

Computational approaches require separate computation algorithms for various types of spatial relations, i.e., topology, orientation, distance relations. As natural language is flexible, categorizing spatial relation in destination descriptions may introduce error. Mismatching between the identified types from descriptions and the preset types in algorithms will also cause failure. This cognitively motivated approach only checks the uniqueness of destination candidates within neighborhood of each landmark candidates, therefore avoids these risks. However the precision of this approach depends on the appropriate definition of landmark neighborhood. Imprecision may also be produced when spatial relations are used to disambiguate destinations.

5 Experimental Evaluation

To evaluate the performance of the cognitively motivated approach, a gazetteer was built, including 36,134 instances. The gazetteer data is based on the point of interest data from Whereis[1] and VicNames data[2] over the entire area of Victoria, Australia. Each instance consists of three essential attributes: the place name, the category of place, and a geographic location provided by the data sources [16].

The only other attribute needed is a salience value. Salience is derived here by a method suggested by Duckham et al. [17] utilizing the categories of gazetteer instances. Each category in a gazetteer is assessed by an expert on nine criteria (physical size, proximity to road, visibility, difference from surroundings, ubiquity, nighttime vs. daytime salience, permanence, length of description and spatial extents) in two ways: the average salience of individual instances in a category (suitability) and their standard deviation (typicality). The final salience of each category is then normalized in the range [0,1]: 1 represents the highest suitability, and 0 represents the lowest suitability.

From salience, the influence areas of all instances are computed at all levels of a salience hierarchy, according to Winter et al. [15]. This concludes the preprocessing of generating a suited gazetteer.

After preparing the gazetteer data, we collected 57 destination descriptions given by participants in an interview experiment. Examples of these collected destination descriptions are "Yarra Bend Park near Alphington", and "Lorne on the Western Coast Road, between Geelong and Apollo Bay, about half way

[1] www.whereis.com.au

[2] http://services.land.vic.gov.au/vicnames/

between each". In total, 80 individual places were mentioned in the collection of destination descriptions, including street names, suburb names, names of stations, restaurants, shopping centers, hospitals, clubs, universities, and parks. Besides individual places, there are 15 paraphrased places found in the collection, such as "the library". In the collected data, there are 38 descriptions (67%) including spatial relations and reference place names.

For this experiment the cognitively motivated approach was implemented to interpret the given destination descriptions. Participants were asked to judge the interpretation results. For comparison, we also developed an approach computing topological, orientation and distance relations. According to their judgement, 27 destination descriptions were interpreted correctly by the cognitively motivated approach, and 28 by the approach computing relations explicitly (the cognitively motivated approach was not allowed to fall back on the explicit computation of relations). The results show that the performance of the cognitively motivated approach is almost as good as the approach with explicit computation of relations.

6 Conclusions

Destination descriptions can use various qualitative spatial relations, thus computing spatial relations can be computationally expensive. This may be one of the reasons why no commercial navigation system has implemented methods for interpreting spatial relations (Google Maps, for example, ignores any given relationship and imposes a 'near' relationship on any destination description, of which the semantics remains opaque, of course). Compared to other computational approaches, this cognitively motivated approach is relatively simple, because no computation of individual spatial relation needed.

This paper proposes a cognitively motivated approach to interpreting destination descriptions without computing spatial relations. This approach is based on disambiguating combinations of the multiple destination and landmark candidates found in gazetteers. Given the context, further discussion is needed to retrieve relevant destination candidates and salient landmark candidates from gazetteers. Furthermore, the cognitive adequacy of the construction of the hierarchy of neighborhoods also needs further study.

References

1. Reiter, E., Dale, R.: Building Natural Language Generation Systems. Cambridge University Press, Cambridge, UK (2000)
2. Couclelis, H., Golledge, R.G., Gale, N., Tobler, W.: Exploring the anchorpoint hypothesis of spatial cognition. Journal of Environmental Psychology **7** (1987) 99–122
3. Morrow, D.G., Clark, H.H.: Interpreting words in spatial descriptions. Language and Cognitive Processes **3**(4) (1988) 275–291
4. Ferenz, K.S.: The role of nongeometric information in spatial language. dissertation, Dartmouth College, Hanover, NH (February 2000)

5. Mcnamara, T.P., Diwadkar, V.A.: Symmetry and asymmetry of human spatial memory. Cognitive Psychology **34**(2) (11 1997) 160–190
6. Clark, H.H.: Using Language. Cambridge University Press, Cambridge, UK (1996)
7. Egenhofer, M., Franzosa, R.: Point-set topological spatial relations. International Journal of Geographical Information Systems **5**(2) (1991) 161–174
8. Duckham, M., Worboys, M.: Computational structure in three-valued nearness relations. In Montello, D., ed.: Spatial Information Theory. Volume 2205 of Lecture Notes in Computer Science. Springer, Berlin (2001) 76–91
9. Schlieder, C., Vögele, T., Visser, U.: Qualitative spatial representation for information retrieval by gazetteers. In Montello, D.R., ed.: Spatial Information Theory. Volume 2205 of Lecture Notes in Computer Science. Springer, Berlin (2001) 336–351
10. Freksa, C.: Using orientation information for qualitative spatial reasoning. In Frank, A., Campari, I., Formentini, U., eds.: Theories and Methods of Spatio-Temporal Reasoning in Geographic Space. Volume 639 of Lecture Notes in Computer Science. Springer, Berlin (1992) 162–178
11. Wallgrün, J.O., Frommberger, L., Wolter, D., Dylla, F., Freksa, C.: A toolbox for qualitative spatial representation and reasoning. In Barkowsky, T., Knauff, M., Ligozat, G., Montello, D.R., eds.: Spatial Cognition V: Reasoning, Action, Interaction. Volume 4387 of Lecture Notes in Artificial Intelligence. Springer, Berlin (2007) 39–58
12. Van der Zee, E., Nikanne, U., eds.: Cognitive interfaces: constraints on linking cognitive information. Oxford University Press, New York, USA (2000)
13. Tomko, M., Winter, S.: Pragmatic construction of destination descriptions for urban environments. Spatial Cognition and Computation **9**(1) (2009) 1–29
14. Moulin, B., Kettani, D., Gauthier, B., Chaker, W.: Using object influence areas to quantitatively deal with neighborhood and perception in route descriptions. In Hamilton, H., ed.: Advances in Artificial Intelligence. Volume 1822 of Lecture Notes in Computer Science. Springer, Berlin (2000) 69–81
15. Winter, S., Tomko, M., Elias, B., Sester, M.: Landmark hierarchies in context. Environment and Planning B **35**(3) (2008) 381–398
16. Hill, L.: Core elements of digital gazetteers: Placenames, categories, and footprints. In: Research and Advanced Technology for Digital Libraries. Volume 1923 of Lecture Notes in Computer Science. Springer, Berlin (2000) 280–290
17. Duckham, M., Winter, S., Robinson, M.: Including landmarks in routing instructions. Journal for Location-Based Services **4**(1) (2010) 28–52

Collecting Spatial Information for Locations in a Text-to-Scene Conversion System

Masoud Rouhizadeh[1], Daniel Bauer[2], Bob Coyne[2], Owen Rambow[2], and Richard Sproat[1]

[1] Oregon Health & Science University, Portland OR, USA,
[2] Columbia University, New York NY, USA

Abstract. We investigate using Amazon Mechanical Turk (AMT) for building a low-level description corpus and populating VigNet, a comprehensive semantic resource that we will use in a text-to-scene generation system. To depict a picture of a location, VigNet should contain the knowledge about the typical objects in that location and the arrangements of those objects. Such information is mostly common-sense knowledge that is taken for granted by human beings and is not stated in existing lexical resources and in text corpora. In this paper we focus on collecting objects of locations using AMT. Our results show that it is a promising approach.

Keywords: Text-to-Scene Systems, Amazon Mechanical Turk, Lexical Resources, VigNet, Location Information, Description Corpora

1 Introduction

Our aim is to populate **VigNet**, a comprehensive semantic resource that we will use in a text-to-scene generation system. This system follows in the footsteps of Coyne and Sproat's WordsEye [2], but while WordsEye did only support a very limited number of actions in a static manner and mostly accepted low-level language as input (*John is in front of the kitchen table. A cup is on the table. A plate is next to the cup. Toast is on the plate*) the new system will support higher-level language (*John had toast for breakfast*).

VigNet is based on FrameNet[1] and contains lexical, semantic and spatial/graphical information needed to translate text into plausible 3D scenes. In VigNet frames are decomposed into subframes and eventually into primitive spatial relations between frame participants (frame elements), describing one way a frame can be depicted graphically. We call a frame that is decomposable into such primitives a **vignette**. Even though the technical details are not crucial to understand this paper we refer the interested reader to [4].

This paper deals with the collection of spatial information to populate VigNet. Even though VigNet contains vignettes for actions and other events, complex objects and situations, this paper focuses only on the induction of **location vignettes**. Knowledge about locations is of great importance to create detailed scenes because locations define the context in which an action takes place. For instance when someone takes a shower he usually does so in the bathroom, interacting with the 'affordances' provided by this room (i.e. shower cabin, curtain, shower head, shower tap etc.) in a specific way. Note that location vignettes can, but do not have to be evoked by lexical items. We can say *John took a shower in the bathroom*, but this seems redundant because bathrooms are the preferred location for *taking a shower*. VigNet records knowledge of this type that can be accessed in the text-to-scene generation process.

In this paper we propose a methodology for collecting semantic information for locations vignettes using Amazon Mechanical Turk (AMT). The next section first discusses location vignettes in more detail. We then review related work in section 3. We describe how we use AMT to build an image description corpus and collect semantic information for locations in section 4 and compare different methods in an evaluation. Section 5 concludes.

2 Location Vignettes

As mentioned before, location vignettes are important because they provide the context in which actions can take place. Locations involve the spatial composition of several individual objects. For example, in *'John sat in the living room'*, we might expect the living room to contain objects such as a sofa, a coffee table, and a fireplace. In addition, these objects would be spatially arranged in some recognizable manner, perhaps with the fireplace embedded in a wall and the coffee table in front of the sofa in the middle of the room. In order to represent such locations graphically we are adding knowledge about the typical arrangements of objects for a wide variety of locations into VigNet.

Any given location term can potentially be realized in a variety of ways and hence can have multiple associated vignettes. For example, we can have multiple location vignettes for a *living room*, each with a somewhat different set of objects and arrangement of those objects. This is analogous to how an individual object, such as a *couch*, can be represented in any number of styles and realizations. Each location vignette consists of a list of constituent objects (its frame elements) and graphical relations between those objects (by means of frame decomposition). For example, one type of living room (of many possible ones) might contain a couch, a coffee table, and a fireplace in a certain arrangement.

LIVING-ROOM_42(left_wall, far_wall, couch, coffee_table, fireplace)
TOUCHING(figure:couch, ground:left_wall) FACING(figure:couch, ground:right_wall) FRONT-OF(figure:coffee_table, ground: sofa) EMBEDDED(figure:fire-place, ground:far_wall)

The set of graphical primitives used by location vignettes control surface properties (color, texture, opacity, shininess) and spatial relations (position, orientation, size). This set of primitive relations is sufficient to describe the basic spatial layout of most locations (and scenes taking place in them). Generally we do not record information about how the parts of a location can be used in an action, but rather consider this knowledge to be part of the action.

3 Related work

Existing lexical and common-sense knowledge resources do not contain the spatial and semantic information required to construct location vignettes. In a few cases, WordNet [5] glosses specify location-related information, but the number of such entries with this kind of information is very small, and they cannot be used in a systematic way. For example, the WordNet gloss for *living room* (*a room in a private house or establishment where people can sit and talk and relax*) defines it in terms of its function, not its constituent objects and spatial layout. Similarly, the WordNet gloss for *sofa* (*an upholstered seat for more than one person*) provides no location information. FrameNet [1] is focused on verb semantics and thematic roles and provides little to no information on the spatial arrangement of objects.

More relevant to our project is OpenMind [8] where online crowd-sourcing is used to collect a large set of common-sense assertions. These assertions are normalized into a couple dozen relations, including the typical locations for objects. The list of resulting objects found for each location, however, is noisy and contains many peripheral and spurious relations. In addition, even the valid relations are often vague and represent different underlying relations. For example, a *book* is declared to be located *at a desk* (the directly supporting object) as well as *at a bookstore* (the overall location). In addition, like most existing approaches, it suffers from having objects and relations being *generalized* across all locations of a given type and hence is unable to represent the dependencies that would occur in any given *specific* location. As a result, there's no clear way to reliably determine the main objects and disambiguated spatial relations needed for location vignettes.

LabelMe [7] is a large collection of images with annotated 2D polygonal regions for most elements and objects in a picture. It benefits from the coherence of grounding the objects in specific locations. It suffers, though, from the lack of differentiation between main objects and peripheral ones. Furthermore, it contains no 3D spatial relations between objects.

One of the well-known approaches for building lexical resources is automatic extracting lexical relations from large text corpora. For a comprehensive review of these works see [6]. However, a few works focus specifically on extracting semantic information for locations, including [10] and [11], which use the vector-space model and a nearest-neighbor classifier to extract locations of objects. Also directly relevant to this paper is work by Sproat [9] which attempts to extract associations between actions and locations from text corpora. This approach provides some potentially useful information, but the extracted data is noisy and requires hand editing. In addition, it extracts locations for actions rather than the objects and spatial relations associated with those locations.

Furthermore, much of the information that we are looking for is commonsense knowledge that is taken for granted by human beings and is not explicitly stated in corpora. Although structured corpora like Wikipedia do mention associated objects, they are often incomplete. For example in the Wikipedia entry for *kitchen* there is no mention of a *counter* or other surface on which to prepare food but the picture that goes with the definition paragraph (labeled "A modern Western kitchen") clearly has one.

In this paper we investigate using Amazon Mechanical Turk (AMT) for building a low-level description corpus for locations and for directly collecting objects of locations vignettes. We will compare the accuracy of collected data to several gold standard vignettes generated by an expert. We show that we can tune our information collection method to scale for large number of locations.

4 Using AMT to build location vignettes

In this section we discuss how we use Amazon Mechanical Turk (AMT) to build a *location description corpus* and for collecting the *typical objects of location vignettes*. AMT is an online marketplace to co-ordinate the use of human intelligence to perform small tasks such as image annotation that are difficult for computers but easy for humans. The input to our AMT experiments are pictures of different rooms. By collecting objects and relations grounded to specific rooms we capture coherent sets of dependencies between objects in context and not just generalized frequencies that may not work together. In each task we collected answers for each room by five workers who were located in the US and had previous approval rating of 99%. Restricting the location of the workers increases the chance that they are native speakers of English, or at least have good command of the language. We carefully selected input pictures from the results of image searches using the Google and Bing search engines. We selected photos that show 'typical' instances of the room type, e.g. room instances which include typical large objects found in such rooms. Photos should show the entire room. We then defined the following task:

Task 1: Building low-level location description corpus: In this task, we asked AMT workers to provide simple and clear descriptions of 85 pictured room. We explicitly asked AMT workers that their descriptions had to be in the form of naming the main elements or objects in the room and their positions in relation to each other, using verbs such as *is* or *are* (i.e. *linking verb*). Each description had to be very precise and 5 to 10-sentence long. Our collected description corpus contains around 11,000 words.

In order to extract location information from the low-level location description corpus, the text is first processed using the NLP module of WordsEye. We extracted the objects and other elements of locations which are mainly in the form of RELATION–GROUND–FIGURE and extract the objects and elements which are represented as FIGURE or GROUND. We then further processed the extracted locations as is explained in sub-section 4.1.

Task 2: Listing functionally important objects of locations: According to this criterion, the important objects for a room are those that are required in order for the room to be recognized or to function in this way. One can imagine a *kitchen* without a *picture frame* but it is rarely possible to think of a *kitchen* without a *refrigerator*. Other functional objects include a *stove*, an *oven*, and a *sink*. We asked workers to provide a list of functional objects using an AMT hit such as the one shown in figure 1. We showed each AMT worker an example room with a list of objects and their counts. We gave the following instructions:

" Based on the following picture of a **kitchen** list the objects that you really need in a **kitchen** and the counts of the objects.

1. In each picture, first tell us how many room doors and room windows do you see.
2. Again, don't list the objects that you don't really need in a **kitchen** (such as magazine, vase, etc). Just name the objects that are absolutely required for this **kitchen**. "

Task 3: Listing visually important objects of locations: For this task we asked workers to list large objects (furniture, appliances, rugs, etc) and those that are fixed in location (part of walls, ceilings, etc). The goal was to know which objects help define the basic structural makeup of this particular room instance. We used the AMT input form shown in figure 1 again, provided a single example room with example objects and and gave the following instruction:

" What are the main objects/elements in the following **kitchen**? How many of each?

1. In selecting the objects give priority to:
   - Large objects (furniture, appliances, rugs, etc).
   - Objects that are fixed in location (part of walls, ceilings, etc).

   The goal is to know which objects help define the basic makeup and structure of this particular kitchen.
2. In each picture, first tell us how many room doors and room windows do you see. "

4.1 Post-processing of the extracted object names from AMT

We post-processed the extracted objects from the location description corpus and the objects that were listed in tasks 2 and 3 in the following steps:

1. Manual checking of spelling and converting plural nouns to singular.
2. Removing conjunctions like"*and*", "*or*", and "/". For example, we converted "*desk and chair*" to "*desk*" and "*chair*".
3. Converting the objects belonging to the same WordNet synset into the most frequent word of the synset. For example we converted *tub*, *bath*, and *bathtub* into *bathtub* with frequency of three.
4. Finding the intersection from the inputs of five workers and selecting the objects that listed three times or more

Fig. 1. AMT input form to collect functionally important objects (task 2) or visually important (task 3) objects in locations. Workers are asked to enter the name of each object type and the object count.

5. Finding major substrings in common: some input words only differ by a space or a hyphen character such as *night stand*, *night-stand*, and *nightstand*. We convert such variants to the simplest form i.e. *nightstand*.
6. Looking for head nouns in common: if the head of the compound noun input such as *projector screen* can be found in another single-word input i.e. *screen*, we assume that both refer to the same object i.e. *screen*.
7. Recalculating the intersections and selecting the objects with frequency of three or more.

4.2 Evaluation

For evaluating the results we manually built a set of gold standard vignettes (GSVs) for 5 rooms which include A) a list of objects in each room, and B) the arrangements of those objects. Selected objects for GSVs are the ones that help define the basic makeup and structure of the particular room. We are comparing the extracted object from AMT tasks against the list of objects in the GSVs.

Table 1 shows the comparison of the AMT tasks against GSVs. The "Extracted Objs" row shows the number of objects we extracted from each AMT tasks for 5 rooms. The "Correct Objs" row shows the number of extracted objects from AMT that are present in our GSVs of each room and the precision score derived based on that. The "Expected Objs" row shows the number of all the objects in GSVs that we expected the workers to list, and the recall score based on that.

AMT Task	Free Description		Functional		Visual	
Extracted Objs	39		32		32	
Correct Objs	26	Pre: 67%	28	Pre: 87%	29	Pre: 91%
Expected Objs	33	Rec: 79%	33	Rec: 85%	33	Rec: 88%

Table 1. The accuracy of each AMT tasks for the objects of 5 rooms compared to GSVs. (See the above paragraph for the definition of rows and columns.)

5 Conclusion and future work

In this paper we explored different approaches to populate VigNet, a resource containing spatially grounded lexical semantics, with locational information (location vignettes) using Amazon Mechanical Turk. In one approach we used AMT to collect a low-level description corpus for locations. We then used the WordsEye NLP module to extract the objects from each description. For comparison we asked AMT workers to directly list objects of locations shown in photographs, either based on visual or on functional criteria. We then post-processed the extracted objects from each experiment and compared them against gold standard location vignettes.

We have shown that we can extract reasonably accurate objects from processing the description corpus as well as spatial relations and arrangements of objects. The results achieved using the functional and visual object listing tasks approximate the gold standard even better, with the visual elicitation criterion outperforming the functional one.

In current work, due to the good results on the small training set we are using the visual object listing paradigm to induce descriptions of 85 rooms. We are planing to collect vignettes for a variety of other indoor and outdoor locations.

Location vignettes also contain the spatial arrangement of objects. In addition to the extracted relations from the description corpus, we also designed a series of AMT tasks for determining the arrangements of objects in different locations using the objects that we collected in the present work. For each room we ask AMT workers to determine the arrangements of the previously collected objects in that particular room. For each object in the room, workers have to determine its spatial relation with A) *one wall* of the room and B) *one other object* in the room. We did not include the results in this paper since we are still exploring methods to evaluate the *spatial arrangements* task. The gold standard location vignettes include arrangements of objects, but it is difficult to directly compare the gold standard to the AMT workers' inputs as there are different possibilities to describe the same spatial layout.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0904361.

References

1. Baker, C., Fillmore, C., Lowe, J.: The Berkeley FrameNet Project. COLING-ACL (1998)
2. Coyne, B., Sproat, R.: Wordseye: An automatic text-to-scene conversion system. In Proceedings of the 28th annual conference on Computer graphics and interactive techniques, Los Angeles, CA, USA, pp. 487- 496, (2001)
3. Coyne, B., Rambow, O., Hirschberg, J., Sproat, R.: Frame semantics in text-to-scene generation. In R. Setchi, I. Jordanov, R. Howlett, and L. Jain (Eds.), Knowledge-Based and Intelligent Information and Engineering Systems, Volume 6279 of Lecture Notes in Computer Science, pp. 375-384. Springer Berlin / Heidelberg (2010)
4. Coyne, B., Bauer, D., Rambow, O.:VigNet: Grounding Language in Graphics using Frame Semantics. In ACL Workshop on Relational Models of Semantics, (2011)
5. Fellbaum, C.: WordNet: An Electronic Lexical Database. Bradford Books (1998)
6. Girju, R., Beamer, B, Rozovskaya, A., Fister, A., Bhat. S.:A knowledge-rich approach to identifying semantic relations between nominals. Information Processing and Management, vol. 46, no. 5, pp. 589-610, (2010)
7. Russell, B. C. , Torralba, A., Murphy, K. P., and Freeman, W. T.: LabelMe: a database and web-based tool for image annotation. International Journal of Computer Vision, vol. 77, no. 13, pp. 157173, May (2008).
8. Havasi, C., Speer, R., Alonso, J.: ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge. Proceedings of Recent Advances in Natural Languge Processing (2007)
9. Sproat, R.: Inferring the environment in a text-to-scene conversion system. First International Conference on Knowledge Capture, Victoria, BC (2001)
10. Turney,P.,Littman,M.:Corpus-based Learning of Analogies and Semantic Relations. Machine Learning Journal 60 (1-3), pp. 251-278. (2005)
11. Turney, P.: Expressing implicit semantic relations without supervision. In: Proceedings of COLING-ACL, Australia (2006)

Where should I put my hands? Planning hand location in sign languages

Alice Ruggeri[1], Cristina Battaglino[1], Gabriele Tiotto[2], Carlo Geraci[1], Daniele Radicioni[1], Alessandro Mazzei[1], Rossana Damiano[1], Leonardo Lesmo[1]

[1] Dipartimento di Informatica, Univeristà degli Studi di Torino, Italy
[2] Dipartimento di Automatica e Informatica, Politecnico di Torino, Italy
lesmo@di.unito.it

Abstract. This paper describes a corpus-based study addressing the spatial positioning of hands in Italian Sign Language (LIS). The corpus includes LIS sentences extracted from TV news concerning weather forecasting. The results shows that spatial relations depend on contextual and geographical features. These results have been exploited in the implementation of a planning module that generates a sequence of commands that drive the generator of a virtual character. The planner is part of an Italian to LIS translation system whose goal is the automatic translation of weather forecast programs.

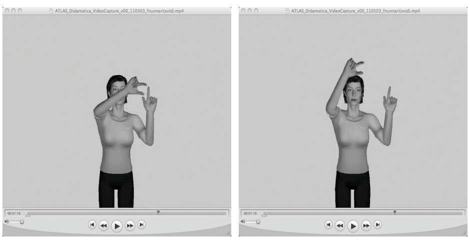
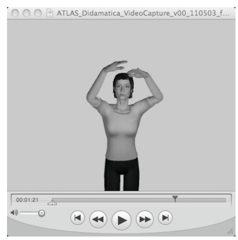
Keywords: Sign languages, automatic translation, signs place of articulation.

1 Introduction

This paper describes the module of an Italian to LIS (Italian Sign Language) translator that takes care of hand positioning in the animation of the virtual character. The overall organization of the translator, the ATLAS project (Automatic Translation into the Language of Signs), includes various software modules taking care of the linguistic translation and of the virtual character animation. Most of these modules, in particular the ones devoted to language interpretation, have been used in various projects and aim at a wide coverage of linguistic structures. However, the structure of the planner, which which covers hands position, is based on the results of a corpus-based study on spatial relations in LIS, and takes care of the general principles of hand positioning in sign languages.

Sign languages (SL), namely the languages used by Deaf people for everyday exchanges, are visuo-spatial languages in the sense that they use the visual apparatus to perceive the linguistic input and use space as a crucial component to transmit linguistic meaning. Signers (like speakers) may adopt two different perspectives to convey spatial information: either they use a route perspective where the viewpoint is within the scene, or they use a survey perspective where the viewpoint is outside the environment [1]. In the route perspective the signing space reflects the individual's perspective of the environment in a 3-D like representation, while in the survey viewpoint it takes a fixed bird's eye view over a horizontal plane. Weather forecasting provides an interesting case study of the use of space in SLs. Weather conditions always refer to geographic areas and a common way to conceptualize them, in the SL use, is by adopting survey perspective. The domain of weather forecasting offers a slightly different survey perspective from the bird's eye view. Specifically, the perspective adopted in this situation is mapped onto imaginary geographical maps displayed along the vertical axis, reflecting the standard way of illustrating forecasting on TV news. For instance, the sign ALPS[1] in LIS (i.e. the mountain chain in the north of Italy) is iconically realized in the upper part of the signing space (i.e. in the north of the imaginary map), as illustrated in (1a). Accordingly, meteorological events happening in that part of the country would be articulated in the same spatial location. For instance, this could be the case of the sign CLOUD-GATHERING, as in (1b). If the cloud gathering had happened in the south of Italy, say in Sicily, the place of articulation of that sign would have changed accordingly.

In the next section we describe the architecture of the translator; section 3 presents the corpus analysis, section 4 sketches the plan-based implementation, while section 5 concludes the paper.

(1) a. ALPS b. CLOUD-GATHERING

2 System architecture

The translator is based on a traditional rule-based approach, where the input sentences are interpreted in terms of an ontology-based logical representation, which acts as input to a linguistic generator that produces the corresponding LIS "sentence" in a language we called AWLIS (Atlas Written Italian Sign Language). A LIS sentence is a sequence of glosses, annotated with some syntactic pieces of information; the sequence is sent to a planner that takes a decision about the position where a sign must be articulated and its output is analyzed by a character animator that visualizes the movements of the virtual character on different media (see Fig.1).

The syntactic analysis is carried out by the Turin University Parser (TUP), which returns syntactic structures in dependency format. Dependency grammars represent syntactic relations by means of labeled binary relations between pairs of words (a Head and a Dependent). The parser includes a chunking step, where chunks (usually nominal groups) are collected and a verbal analysis step, where the chunks are connected to verbs on the basis information about verbal subcategorization (e.g. transitive vs. intransitive). Semantic interpretation is based on an ontology describing

[1] Throughout the paper, we use all-capital words to refer to the glosses of the LIS signs.

the concepts of the weather forecasting domain. The interpreter analyses the syntactic tree and, by accessing the ontology, builds a logical formula that takes advantage of thematic relations in order for assessing the semantic role of the verbal dependents. This formula is used as input for a CCG (Combinatory Categorial Grammar) linguistic generator [2]. This module uses rules describing the AWLIS grammar that have been designed by exploiting the straightforward syntax-semantics interface that is one of the main features of CCG grammars. These rules account for various linguistic phenomena, as the morphological realization of plural, and coordination.

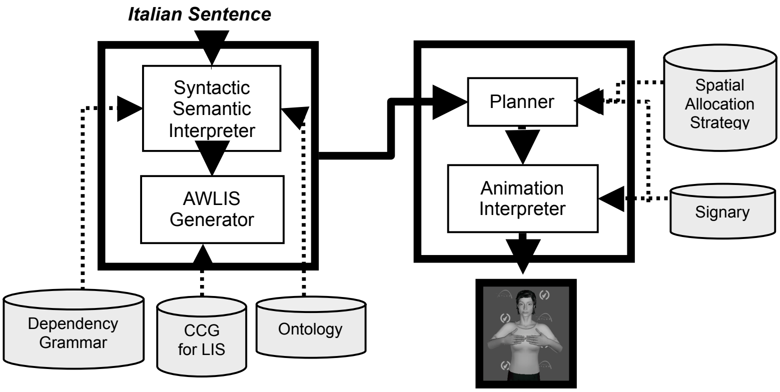


Figure 1: Architecture of the Atlas system

The resulting AWLIS sentence is fed to a hierarchical planner that produces a representation of the signed LIS where various visual features (as facial expression, hand movements, body position, etc.) are explicitly encoded (see section 4). Finally, the planner output is used by a character animation module that generates the actual character movements on the screen.

3 Corpus analysis

Within the ATLAS project, a corpus has been collected, consisting in the translation from Italian into LIS of 40 weather forecasting news recorded from the national broadcasting network. The corpus has been annotated with ALEA, a web application developed for the annotation of video content [3]. The following steps compose the annotation process:

- Sentence segmentation and tokenization (sequential ID association and linking of each token with a lemma entry in the LIS signary)
- Phonological tagging (manual and non-manual components)
- Morphological tagging (part of speech and lexical field)
- Syntactic tagging (syntactic dependencies)
- Discourse tagging (topic and focus identification)

In particular, the annotators indicated Sign Spatial Location (SSL). If the sign is in its standard location (as encoded in the signary) the SSL is left empty. In the case of a location diverging from the citation form, two options are possible: either the sign is relocated in the neighborhood of another (preceding or following) sign (attractor), or the sign is relocated absolutely, i.e. without an overt attractor. The former case is illustrated by the example in (1b) above, where the sign for CLOUD-GATHERING is relocated in the same spatial position where the sign for ‘Alps’ is produced. In presence of an attractor, the annotator encoded the sequential ID number of the attractor, and the relative position of the relocated sign with respect to the attractor (top, bottom, left, etc.). Absolute relocation is illustrated by the examples in (2a-b). In (2a), the sign for LIGURIA, a region in the northwest of the country, is produced in its citation form, i.e. the one you find in standard LIS dictionaries. Its place of articulation is in the middle of the signing space in front of the signer. However, the variant in (2b) is relocated on the left side of the signer, positioning the region on the northwest part of an imaginary map of Italy. In this case, there is no sign working as an attractor. In the case of absolute relocation, the notation simply indicates the position toward which the sign location has shifted (up, down, left, etc.).

(2) a. LIGURIA (citation form) b. LIGURIA (absolute relocation)

Two studies on sign spatial location based on 10 out of 40 weather forecasting news have been conducted. In the first study we investigated the potential causes triggering relocation, while in the second study we addressed the issue of what triggers assimilation-style relocation or absolute relocation.

For the first study, we excluded from the 1027 annotated items those signs that never undergo the relocation process. Among the 513 remaining items (“relocatable”) the most frequent option is to maintain the standard place of articulation (61% of the cases), while in the 39% of the cases signs are produced in a relocated position. Several multivariate analyses, considering five potential predictors for the distribution of relocated vs. non-relocated signs: part of speech, lexical field, absence/presence of non-manual components, one/two-handed sign, and absence/presence of second-hand activity (i.e. whether the second hand was anticipating or perseverating the handshape of another signs in the utterance). The results reported in table 1 show the two predictors emerged as significant ($p<0.05$) in the VARBRUL analysis [4]: lexical field and number of hands.

Table 1. Input value: 0.381, total chi-square = 4.6363 (chi-square/cell = 0.5795), Log likelihood = –326.326.

Predictor	Level	Application value: relocated yes	
		Factor Weight[2]	%
Lexical field	Geographic	.666	52%
	Meteorological	.569	46%
	Functional	.477	35%
	Other	.307	23%
Hand-number	Two-handed	.534	40%
	One-handed	.422	37%

Lexical field had the strongest effect on the distribution of relocated and non-relocated signs (range = 0.359). Geographic and meteorological terms favors relocation (FW= .666 and FW= .569, respectively), while functional signs and the rest of the signs disfavor relocation (FW= .477 and FW= .307, respectively). A weaker but still significant effect is found for number of hands (range = 0.112): Two-handed signs slightly favor relocation (FW = .534); while one-handed signs disfavor it (FW= .422).

For the second study we concentrated only on those items showing relocation (i.e. on the 200 relocated signs). The results show that most relocated signs have an attractor preceding it in the clause, surfacing as cases of anticipatory assimilation (73%), while relocated signs with a following attractor are very rare (4%). A good portion of tokens shows absolute relocation (27%). For the purposes of the statistical analysis we only considered cases of perseverative assimilation and cases of absolute relocation and we found a significant effect of lexical field, as illustrated in table 2.

Table 2. Input value: 0.752. Total chi-square = 0.000 (chi-square/cell = 0.000). Log likelihood = –104.908.

Predictor	Level	Application value: anticipatory assimilation	
		Factor Weight	%
Lexical field	Functional	.804	93%
	Other	.531	77%
	Meteorological	.518	77%
	Geographic	.292	56%

The main effect of lexical field is due to the extreme behavior of functional and geographic signs. The former strongly favor anticipatory assimilation (FW= .804),

[2] Factor weights (also known as factor probabilities) are a numerical measure of the strength of each level of a predictor. Values that are close to 1 favor the application value (i.e. the level of the dependent variable used as ‘baseline’), values close to 0 disfavor it, while values around 0.5 are neutral. In this case, values close to 1 favor relocation, while values close to 0 disfavor it. Factor weights can be easily converted into logits [7].

while the latter strongly disfavor anticipatory assimilation (FW= .292), therefore favoring absolute relocation. Other signs favor assimilation over absolute relocation (FW= .531), as well as meteorological terms, although only weakly (FW= .518).

The fact that geographic and meteorological signs favor relocation can be interpreted as the result of a mapping of spatial relations onto an imaginary geographical map stretched on the vertical plane, forcing displacement of signs somewhere on the map. In a sense, the domain of weather forecasting makes the iconic component of these signs extremely relevant forcing a schematic isomorphism between the aspects of the linguistic signal and aspects of the spatial scene [5]. However, geographic terms and meteorological signs crucially differ on the type of relocation in which they are involved. The former favor absolute relocation as the result of a direct link with the imaginary map, while the latter depend upon an already established spatial relation (e.g. the presence of a geographic sign, as in 'it's raining on the Alps'). Interestingly, absolute relocation is extremely unlikely in the case of functional signs. This is quite expected since the iconic component is virtually absent from these signs.[3]

4 The planner

The planner module has the specific role of organizing the sign flow in the signing space. Its input is the sequence of the lemmas and information on the semantic roles as produced by the generator module. The output returns the same sequence enriched with information on where hands has to be positioned in the signing space.

We use the SHOP2 planning system [6], which relies on the formalism of Hierarchical Task Networks (HTN). Given the sequence of signs, the planning component accounts for the use of the signing resources, namely hands, facial expression, torso, head, and the organization of the signing space in order to accomplish the animation of the given sign sequence. A library of linguistic plans describes how signs can be adapted to the context of a specific sentence (encoded in the AEWLIS input representation), given the constraints provided by the communicative situation and the interpreter's configuration (signing resources, availability of the resources, etc.) [8].

Figure 2 represents the top-level portion of the hierarchical task networks (HTN). The high-level task (*LIS-sign*) decomposes into the task of assigning the signs to the interpreter's hands (not shown), finding the location for each sign (*Localize* in the figure, achieved through the *Find-position* sub-task), then performing the sign (*Make-sign*). Also, the planner determines the initial and final location of the signs that have a parameterized trajectory, such as movement verbs (*Sign-relation* task).

---

3 Pointing signs, whose deictic nature might force absolute relocation, would represent a relevant exception within the class of functional signs. However, the number of pointing signs was extremely low in our corpus possibly due to the formal register used by the interpreters, therefore we couldn't test this prediction.

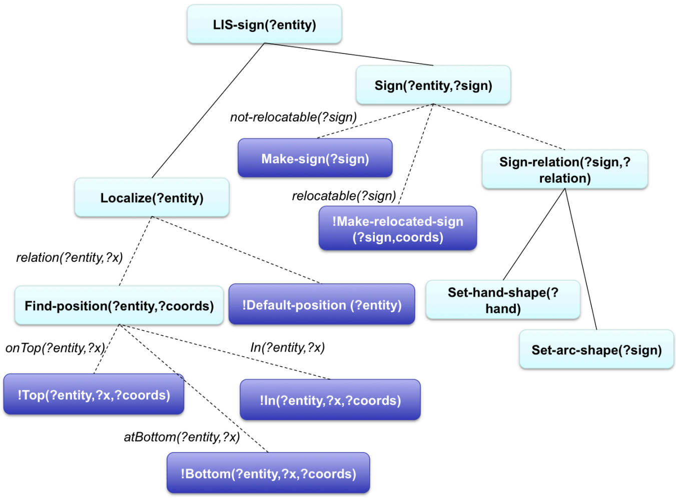


Figure 2: the top level of HTN encoding the signing strategy of the planning component. Dark boxes represent primitive actions.

For each sign the planner produces an ordered pair of spatial coordinates: the first one corresponds to the location of the sign in its citation form, the second one is generated by the planner itself and corresponds to the position in the signing space where the sign will be actually articulated. These latter coordinates are the result of the planning process, which currently take into account only spatial relations lexically expressed by specific lemmas (e.g. NORTH 'in the north', LEFT 'on the left', BOTTOM 'at the bottom of…', etc.), as provided by the generator module. This, however, fails to provide the correct spatial location coordinates for cases where signs are relocated without an overt indication of the spatial relation. These are precisely the situations of absolute relocation and assimilation of the place of articulation described in section 3. In order to fix this problem, we implement the planner by adding two further conditions on the generation of the final coordinates. The first condition introduces additional contextually salient information like the presence of a geographic map that in turns generates an imaginary map in the signing space. This will be used for the purposes of absolute relocation. The second condition capitalizes on the semantic information provided by the generator module. Specifically, we use the semantic role of location (in addition to lexically specified phonological constraints) to identify potential attractors that drive the assimilation of place of articulation. We illustrate here this latter case by considering the sign CLOUD-GATHERING, already introduced above. The first spatial coordinates produced for the sign CLOUD-GATHERING are those of its citation form. If no attractor is present in the utterance, the planner will use these coordinates as the final coordinates for the sign. If an attractor (e.g. ALPS) is present, the coordinates of the attractor are copied as the final coordinates for the sign CLOUD-GATHERING.

Conclusions

In this paper we presented the architecture of a generation module of the Italian into LIS automatic translator ATLAS. One of the crucial aspect of the grammar of SLs is the used of space in order to convey linguistic meaning. In the sign stream, signs can be articulated in positions different from their citation vocabulary-like position. We identified two modes of relocation: relative and absolute. A corpus study on weather forecasting revealed that both articulatory and grammatical factors are responsible for relocation. In particular, the mental representation of spatial relations is projected onto a virtual 'geographical' map along the vertical axis that is used as a guide for relocation. We implemented these findings as a set of rules that drive the behavior of a planner. The planning module gets as input a "symbolic" representation of the sentence in a written form (AWLIS) and takes the decision about where to put the hands, according to the principles identified in the corpus study.

References

1. Emmorey, K., Tversky, B., and Taylor, H.A. 2001. Using space to describe space: Perspective in speech, sign and gesture. *Spatial Cognition and Computation 2*, 157--180.
2. White, M. 2006 Efficient realization of coordinate structures in combinatory categorial grammar. Researh on Language and Computation, 39--75.
3. Barberis, D., Garazzino, N., Piccolo, E., Prinetto, P., and Tiotto, 2010. G. A Web Based Platform for Sign Language Corpus Creation. In proceedings of the International Conference in Computers Helping People with Special Needs, ICCHP 2010, Vol.2, 193--199.
4. Rand, D., and D. Sankoff. 1990. *Goldvarb 2.1: A Variable Rule Application for the Macintosh*. Montreal: Centre de Recherches Mathématiques, University of Montreal. Version 2.
5. Emmorey, K. 2007. The psycholinguistics of signed and spoken languages. In G. Gaskell (ed.), *The Oxford Handbook of Psycholinguistics*. Oxford University Press, 703--721.
6. D. Nau, T. Au, O. Ilghami, U. Kuter, J. Murdock, D.Wu, and F. Yaman, "SHOP2: An HTN planning system," Journal of Artificial Intelligence Research, vol. 20, no. 1, 379--404, 2003.
7. Morrison, G. 2005. Dat is what the PM said: a quantitative analysis of Prime Minister Crétien's pronunciation of English voiced dental fricatives, *Cahiers linguistiques d'Ottawa 33*, 1--21.
8. V. Lombardo, F. Nunnari, and R. Damiano. 2010. A Virtual Interpreter for the Italian Sign, in *Intelligent Virtual Agents: 10th International Conference*, IVA 2010, Philadelphia, PA, USA. Proceedings, Springer.

Geolocating Orientational Descriptions of Landmark Configurations

James Pustejovsky[1], Marc Verhagen[1], Anthony Stefanidis[2], Caixia Wang[2]

[1] Department of Computer Science
Brandeis University, Boston, MA
{jamesp, marc}@cs.brandeis.edu
[2] Center for Geospatial Intelligence
George Mason University, Fairfax, VA
{astefani, cwangg}@gmu.edu

Abstract. In this paper we outline how to translate verbal subjective descriptions of spatial relations into metrically meaningful positional information, and extend this capability to spatiotemporal monitoring. Document collections, transcriptions, cables, and narratives routinely make reference to objects moving through space over time. Integrating such information derived from textual sources into a geosensor data system can enhance the overall spatiotemporal representation in changing and evolving situations, such as when tracking objects through space with limited image data. We focus on landmark identification, since it proves to be a more tractable problem than open-domain image recognition.

Keywords: Spatial language, geolocating, spatial configurations, landmarks.

1 Introduction

The relation between language and space has long been an area of active research. Human languages impose particular linguistic constructions of space, of spatially-anchored events, and of spatial configurations that relate in complex ways to the spatial situations in which they are used. Establishing tighter formal specifications of this relationship has proved a considerable challenge and has so far eluded general solutions. One reason for this is that the complexity of spatial language has often been ignored. In much earlier and ongoing work, language is assumed to offer a relatively simple inventory of terms for which spatial interpretations can be directly stated. Examples of this can be found not only in accounts that focus on formalizations of particular tasks, such as path and scene descriptions, navigation and way-finding, but also in foundational work on the formal ontology of space, on qualitative spatial calculi, and on cognitive approaches.

Visual information in human experience is frequently accompanied by a linguistic description of the image or scene. Consider, for example, the image in Figure 1. If the goal is to identify the region of the image where one should look for the lost keys, one first must identify the correct tree. If this image is automatically segmented using a stock library of images for trees and entrances (Millet et al., 2005; Hollink et al., 2004), several candidate regions for “tree” and “entrance” will be identified. Each candidate region may then be ranked with respect to how likely it is to correspond to a tree or an entrance, producing two ranked lists of candidate regions, T = (T1; T2; : : :)

and E = (E1;E2; : : :), where Ti are the candidate regions for "tree", Ti ranks higher than Ti+1, and Ej are the candidate regions for "entrance". The associated verbal description invokes the "left of" relation, thereby restricting the search for the appropriate pair of candidate regions by imposing the corresponding spatial constraint: LEFT_OF(Ti;Ej). The (Ti;Ej) pairs that do not satisfy the specified spatial relation are given lower ranking, thus increasing the likelihood of identifying correctly the relevant region in the image.

Fig. 1: Speaker: *See the tree to the left of the entrance? I dropped my keys under that tree.*

Over the past decade, image annotation has been the focus of attention within several research areas, in particular, in the context of content-based image retrieval (CBIR). Some research, including the work done within the TRIPOD project at Sheffield, examines the different ways that geo-referenced images can be described (Edwardes et al., 2007), though different approaches, such as the ESP Game can also be used to address this problem. Much of the work on text-based image retrieval has relied on extracting information about the image from image captions, as well as the surrounding text and related metadata, such as filenames and anchor text extracted from the referring web pages, as for example, in Yahoo!'s Image Search. Another kind of image annotation data has become available with the rise of "citizen geography". User-annotated geo-referenced digital photo collections allowing for image content labeling and annotation are being generated in distributed environments, such as Flickr and GoogleEarth. Images are indexed with user-supplied labels that typically form a particular language subset (Grefenstette, 2008). Under such schemes, however, detailed image content annotation is not provided. A notable exception is the "Flickr notes" feature that allows users to annotate regions within images. This and other adaptations of the Fotonotes image annotation standard and the associated software provide an opportunity for detailed annotation of images with both captions and extended free text associated with each annotated image region.

2 Geolocating Descriptions of Landmark Configurations

While such efforts as those discussed above are useful metadata encodings over images, there remain significant problems with unconstrained object recognition. Hence, in this paper, we will focus on linguistic descriptions of *landmark configurations*. Landmarks are visually identifiable objects with fixed spatial locations, which carry semantic meaning for large groups of individuals. They are typically large man-made or physical structures (e.g. buildings, communication antennas, hills) and play an important role in navigation and wayfinding decisions (see e.g. Werner et al, 1998; Steck and Mallot, 2000). For example, routes can be expressed as sequences of landmarks (Duckham et al., 2010) and paths connecting them. The saliency of different landmarks can be expressed in terms of their perceptive, cognitive, and contextual value (Caduff and Timpf, 2008). In this section we address their role for geolocating an observer describing their relative orientational properties in his/her view of a scene.

Let us consider the scene depicted in Fig. 2, taken from Google StreetView, of the intersection of Huntington and Mass. Avenues in Boston. In it we can identify reference landmarks, namely three buildings: Horticultural Hall (HC), Prudential Center (PC) and the Christian Science Monitor building (CSM). It also comprises various other objects, for example a white van, a black truck, and a red car. Our interest is in geolocating the observer of this scene by using orientational descriptions of the relative appearance of the landmarks contained in it.

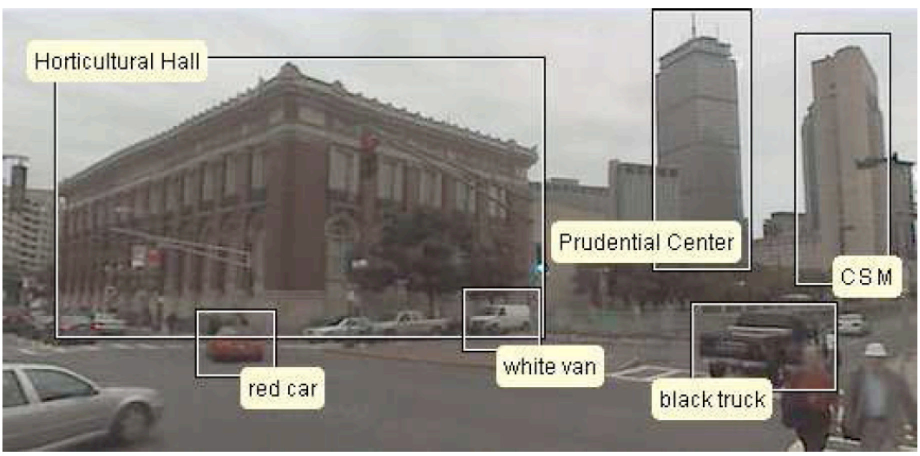


Fig. 2 Landmarks and objects identified in a ground-view.

Assuming that a narrator is familiar with these three landmarks, he/she could describe the scene as follows:

I see the SW and SE sides of Horticultural Hall, and
to the right of it I see the SW and SE sides of the Prudential Center, and
to the right of it I see the Christian Science Monitor Building.

In this situation the narrator has described the scene through three types of statements:

- *explicit reference to specific landmarks*, positioning the scene in their vicinity,
- *explicit description of orientational properties* expressing the relative positions of these landmarks in an observer-centric system[1], and
- *implicit visibility declarations*, whereby she indicates that she can observe specific façades of landmark buildings.

The orientational properties are modeled using ISO-Space (Pustejovsky et al., 2011). ISO-Space distinguishes two major types of elements: entities and relations. Entities include *location*, *spatial entity*, *motion*, *event* (or *spatial state*), and *path*. The two main relations between these entities are the *distance relation* and the *qualitative spatial relation*, which can be either a topological or a relative spatial relation.

Relations such as "to the right of" are annotated as a relative spatial relation between two elements, the figure and the ground, and the viewer perspective is accounted for by two further attributes on the link tag: *rframe*, with values absolute, relative and intrinsic, and *viewer*, which contains a variable indexed to the viewer (Levinson, 2003, Freksa 1992, Ligozat, 1998). Using the three kinds of information above (landmarks, relative positions and visibility declarations), we can identify the three landmarks in a GIS (Fig. 3), and proceed to estimate the location of the observer through a series of view analysis and visibility polygon overlays as we describe below.

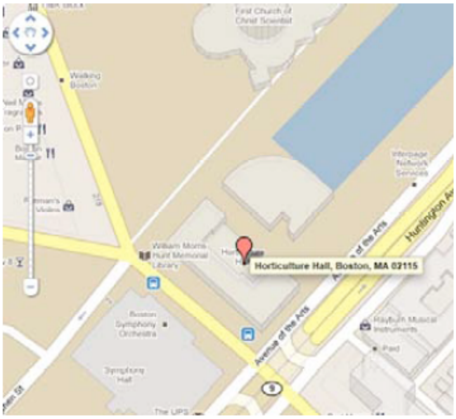


Fig. 3 Map location of Horticultural Hall.

For every visibility statement we can identify a *visibility zone* through *viewshed analysis*, using the local GIS information (Kim et al., 2004). The 2D visibility zone of a specific façade (or any other object in space) is the locus of all points from which at least a part of this façade is visible. For example, in Fig. 4, the visible zone of façade

[1] An alternative would be to use the intrinsic orientation of the landmark, in which case "to the right" would be interpreted relative to the landmark and not relative to the observer. Clearly, both options would need to be explored down-stream.

F_1 is shown as the gray-shaded area. From any point outside this area it would be impossible to see façade F_1.

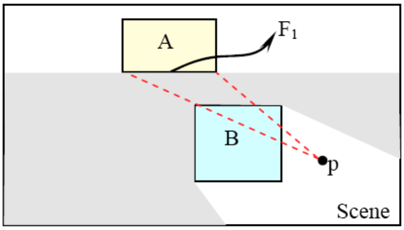


Fig. 4. The visibility zone (gray shaded area) for façade F_1 of Building A.

Each additional visibility statement introduces additional visibility zone information, and the location of the narrator can be eventually determined through the intersection of the corresponding visibility zones through polygon clipping techniques, such as Weiler-Atherton (1977). Fig. 5 shows the implementation of this process for the scene of Fig. 1, through a progressive assessment of visibility conditions for HC, CSM, and PC.

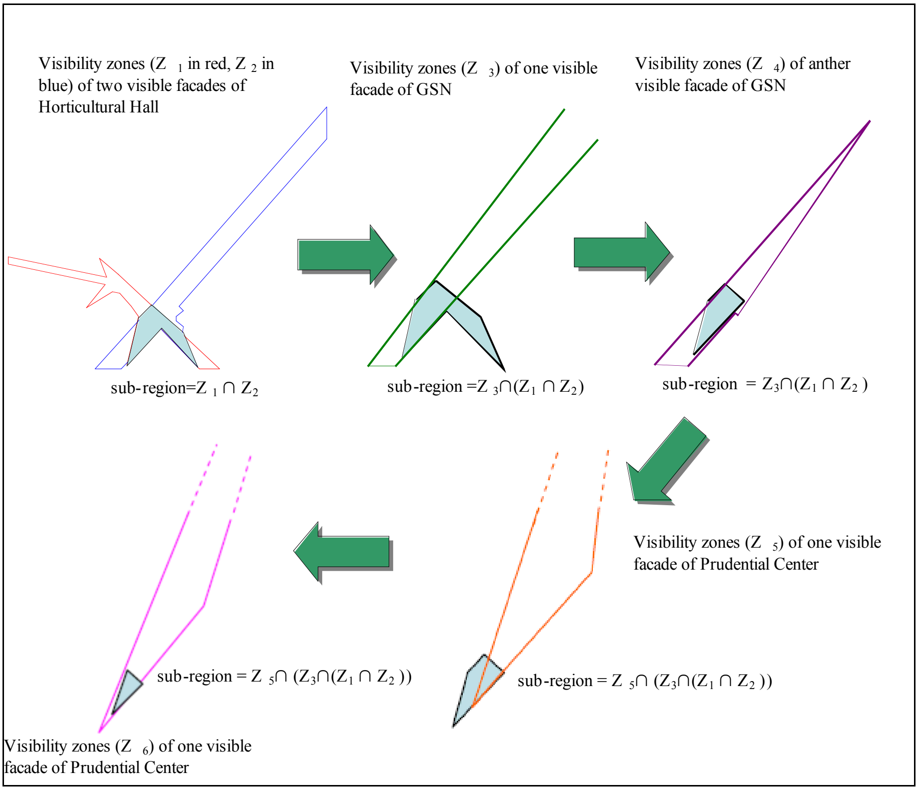


Fig. 5. The progressive visibility intersection process

The narrator position estimated through the process visualized in Fig. 5 is shown on the local map in Fig. 6, marked as a red triangle. The triangle corresponds to all positions from which the narrator would have a view of our scene that would be comparable to the one depicted in Fig. 2 in terms of the orientational relationships of the three depicted landmarks.

Fig. 6. The estimated location of the narrator, indicated as a red triangle, and the views used to estimate it.

3 Conclusion

In this paper, we discuss the integration of multi-source data analysis for spatial knowledge extraction from images. In particular, we focused on the specific contribution of verbal subjective descriptions of spatial relations involving orientation, and how these can be translated into metrically interpretable positional statements within a GIS environment. We concentrated on the more tractable subproblem of landmark identification. Orientational information in language was modeled with ISO-Space annotation, providing both qualitative spatial relations and anchored GPS values, once geolocating is performed.

This work is ongoing research aimed to allow for the integration of information available from different sources, better addressing the evolving needs of the geoinformatics community. Our preliminary results suggest that scene content information provided by verbal description can be mapped faithfully to metrically grounded information. As this is preliminary work, there are clearly many details to be worked out. For example, we have not yet precisely defined how orientational relations are used to identify a landmark in a GIS, especially with anonymous landmarks, a problem exacerbated when an ambiguity between intrinsic and observer-based relative orientation cannot be easily resolved.

One of the ultimate goals of this research is the development of algorithms that take an image and accompanying verbal utterances and maps these to a partition of a 2D grid. This application would be tuned to deal with more natural utterances than the somewhat stilted verbal descriptions given with Fig. 2 above.

4 Acknowledgements

This research was funded under the NURI grant HM1582-08-1-0018 by the National Geospatial Agency.

References

Caduff D., Timpf S. On the Assessment of Landmark Salience for Human Navigation. Cognitive Processing, 9(4), 249-267 (2008).

Duckham M., Winter S., Robinson M. Including Landmarks in Routing Instructions. J. Location Based Services, 4(1), 28-52 (2010).

Edwardes, R. Purves, S. Birche, and C. Matyas. Deliverable 1.4: Concept ontology experimental report. Technical report, TRIPOD Project (2007).

Freksa, Christian. Using orientation information for qualitative spatial reasoning. In A. Frank, I. Campari, and U. Formentini, eds, Theories and methods of spatiotemporal reasoning in geographic space, pages 162–178. Springer, Berlin, (1992).

Grefenstette, G. Comparing the Language Used in Flickr, general Web Pages, Yahoo Images and Wikipedia. In OntoImage 2008, LREC, pages 6–11, (2008).

Hollink, L., G. Nguyen, G. Schreiber, J. Wielemaker, B. Wielinga, and M. Worring. 2004. Adding spatial semantics to image annotations. In Proceedings of 4th International Workshop on Knowledge Markup and Semantic Annotation, 3rd International Semantic Web Conference.

Kim Y.-H., Rana S., Wise S. Exploring Multiple Viewshed Analysis using Terrain Features and Optimisation Techniques. Computers & Geosciences, 30(9-10), pp. 1019-1032 (2004).

Levinson, S. C. Space in Language and Cognition, Cambridge University Press, (2003).

Ligozat, G. Reasoning about cardinal directions. *Journal of Visual Languages and Computing*, 9:23-44. (1998).

Millet, C., I. Bloch, P. Hede, and PA Moellic. Using relative spatial relationships to improve individual region recognition. In Proc. 2nd Eur. Workshop Integration Knowledge, Semantics and Digital Media Technology, pages 119–126. (2005).

Pustejovsky, J., J. Moszkowicz, and M. Verhagen. ISO-Space: The Annotation of Spatial Information in Language, in Proceedings of ISA-6: ACL-ISO, Oxford, England, (2011).

Steck S., Mallot, H.: The Role of Global and Local Landamarks in Virtual Environment Navigation. Presence, 9(1), 69-83 (2000).
Weiler K., P. Atherton. Hidden Surface Removal using Polygon Area Sorting. ACM SIGGRAPH Computer Graphics, 11(2), 214-222 (1977).
Werner S., Krieg-Brueckner B., Mallot H., Schweizer K., Freska C.: Spatial Cognition: The Role of Landmark, Route, and Survey Knowledge in Human and Robot Navigation. In: Jarke M., Pasedach K., Phl K. (eds.) Informatik '97, pp. 41-50, Springer Verlag (1997).

Spatial Event Language across Geographic Domains

Alexander Klippel, Sen Xu, Jinlong Yang, Rui Li

Department of Geography, GeoVISTA Center,
The Pennsylvania State University, PA, USA
{klippel,sen.xu,jinlong,rui.li}@psu.edu

Abstract. We present first results of an analysis of a corpus of linguistic descriptions that were collected in controlled experiments. This corpus and its analysis add to the body of knowledge on formal models for spatial language, language interpretation and generation. The experiments are grounded in qualitative formalisms (RCC and Intersection Models, IM) that have a long standing tradition as means to bridge formal and linguistic descriptions of space and spatial relations. Our experiments address dynamically changing spatial relations (movement patterns/geographic events). By keeping the formal spatial characterizations identical across experiments but changing the semantics (that is, we used movement patterns across seven different geographic domains such as a hurricane in relation to a peninsula, plus two geometric figure domains) we contribute to disentangling spatial and domain specific aspects of spatial (event) language. We briefly discuss here two aspects: First, we hand examine the corpus by selecting participants that show the same conceptual behavior as identified through RCC/IM; second, we analyze the domain specific sub-corpora to address similarities and dissimilarities between individual domains.

Keywords: Event language, topology, corpus analysis.

1 Introduction

Formal models of spatial language play an import role in several disciplines addressing questions of (natural) language processing, natural language generation, the automatic description of spatial scenes, or the design of unifying frameworks for multimodal information systems and processing [1–4]. While we are in the age of spatio-temporal representation and reasoning, the four-dimensional treatment of spatial language (and information in general) is still a hotly debated topic. With respect to language, research shows that naming of events is more challenging than naming of object [5] and it is therefore not surprising that the insights gained from describing static spatial relations linguistically need to be carefully evaluated and extended to the dynamic domain. This contribution is addressing this issue by combining approaches to model events employing qualitative spatial formalisms with linguistic analysis.

2 Approach

We have developed an experimental paradigm that allows us to evaluate the influences of domain semantics on the conceptualization of movement patterns as well as how movement patterns are linguistically described. Here we focus on the linguistic descriptions. Our framework is based on a topologically defined conceptual neighborhood graph [6–8]. Figure 1 provides an overview of the different semantic domains that we have subjected to behavioral validation. In a nutshell: We distinguish movement patterns on the basis of formal path characteristics as identified by the conceptual neighborhood graph. The shortest path (in each scenario) is a single topological relation, DC (disconnected), the longest path (in each scenario) is defined as follows: DC-EC-PO-TPP-NTPP-TPP-PO-EC-DC. To give an example, a boat that never touches or crosses an area of shallow water will always be disconnected (DC) from it. In contrast, a boat that makes it completely across an area of shallow water will exhibit the long path characteristics with the start and end relation being identical (DC). Our participants have to perform a grouping task as a way to elicit conceptual knowledge. After performing this task, participants are presented with the groups that they created again and are asked to provide linguistic descriptions: a short label and a longer description detailing the grouping rational.

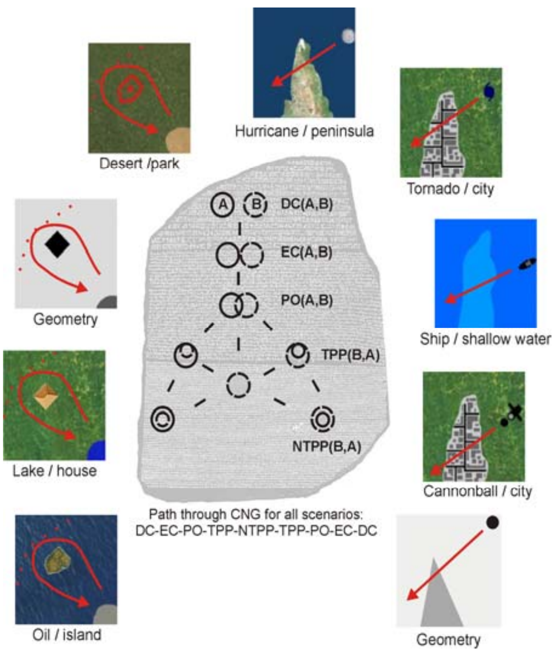


Fig. 1. Nine scenarios from our experiments. Left: four scaling movement patterns: An extending desert in relation to a recreational park, two geometric figures showing a static diamond and an extending/shrinking circle, a lake extending in relation to a house, and an oil slick extending in relation to an island. Right: five translation scenarios: A hurricane in relation to a peninsula, a tornado in relation to a city, a ship in relation to a body of shallow water, a cannonball in relation to a city, and two geometric figures. They are arranged around the Rosetta Stone because all movement patterns in all our experiments are characterized by *topologically equivalent paths* through the conceptual neighborhood graph (which is overlaid on top of the Rosetta Stone).

3 Some results

As rich as our data set is, the flexibility of natural language has made it a challenging task to analyze it. We are presenting two approaches. First we had a look into linguistic descriptions for specific paths. Here we show results from four domains, two from our translation movement patterns (geometry and hurricane) and two from scaling movement patterns (geometry and lake). This path (DC) could be described as *a hurricane not making landfall* or *a lake not flooding the house*. Our goal was to analyze the variety of linguistic descriptions that participants use to this relatively simple scenario. Table 1 provides some representative examples. The important distinction that we made for both scenarios is whether the spatial information (about the movement patterns) in these two scenarios is linguistically encoded using *spatial language*, or, whether this information is encoded using *domain specific language*. The two corresponding geometry scenarios serve as a reference as they obviously do not easily allow for using domain semantics.

With respect to the spatial language we find very diverse ways of conveying spatial information. We do believe that this diversity is fostered by the fact that our research is addressing geographic events / spatio-temporal information (rather than static spatial relations). Especially in the hurricane example we find the following strategies: relative reference frames focusing on the end relations of the geometric characteristics of figure and ground; qualitative distance-based descriptions; negation of what the path does not do; absolute reference; (experiment) context specific descriptions; explicit topological descriptions; intrinsic reference induced by the movement. Interestingly, the explicit spatial descriptions in the lake scenario seem to be less varied, indicating a potential difference between scaling and translation movement patterns that are indistinguishable from a topological perspective.

In both scenarios we also find descriptions that are encoding the spatial event in terms of domain specific language (to different degrees). While, for example, a statement such as “no hit”, “weak hit”, “no landfall” are still rather explicit, a statement such as “weak hurricane” relies heavily on background knowledge of a scenario and is open to interpretation. In case of the lake, the descriptions are much less varied, again, and in most cases refer to a flood not happening.

Table 1. Linguistic descriptions for the shortest possible path (DC).

Hurricane	**Lake**	**Geometry translation**	**Geometry scaline**
Right side stopping circles	Below house	Outside right	Under the box
	Away from the house	Outside right	Up and didn't get too far
Hurricane stops short of land	Not touched	Any part outside the triangle	Team grow
	Not covered		Bellow box
Path doesn't cross	Water reaches short of house	Balls outside triangles	Circle grows beneath box.
Completely off east coast		Before	
		Fully outside	Half way to the diamond
Right side		Off to the right	
Outside right		Right	Straight 1/4
Before land	Not flooded	Stopped on right of triangle	Fall short
	No flood		Expand before
Don't make it	Dry house	To the right	Below
Hurricanes that never made it to shore	Short flood	Far outside on the right	Before diamond stop
	No flooded house	Ball on triangle	Expanding short stop.
No hit or weak hit	No house flood	Too short	Expand halfway
Calm right before the storm	No flooding	Outside	No contact
	Lower risk	Outside right	Stop short
Weak hurricane	Tiny lakes		Grow stop between
Weak hurricanes			Out not close to square
No landfall			Far away
No landing			
Pre-landfall hurricanes			

These findings led us to explore differences between the sub-corpora (the nine different scenarios). First, some domain-corpus properties can be extracted using AntCont [9]. The token occurrences are visualized using Wordle (http://www.wordle.net/), seeFigure 2.

Fig. 2. Frequently appearing words in 9 corpora

From the tag clouds in Figure 2, we can see that top frequent words are mostly related to domain specific semantics. For example, *city*, *desert*, and *island* are referring to objects illustrated in each scenario. It is not surprising that participants make use of the domain semantics for reference to objects in the scenario, as it is a direct and succinct way to describe an object and distinguish it from surroundings. However, for the topological change depicted in different icons, participants would have to use more complicated descriptions such as verb phrases and prepositional phrases. This is the reason that spatial language terms such as *on*, *middle*, *at*, *outside*, *left*, *right*, *through*, and *ended* also appear prominent in the tag cloud. Our analytical question is: given scenarios where only domain semantic is different, how different will the descriptions be?

In the next step, we used the Stanford POS tagger [10]. We investigate the most frequently appearing nouns, verbs, adjectives, and prepositions:

- The most frequently appearing nouns are domain specific ones (see Figure 1). Domain specific nouns with top frequency in one corpus are often never found in other corpora, such as *tornado*, *oil*, and *desert*. Nouns that can be found across domains are common referral terms, such as *side*, *icons*, *middle*, and *bottom*.
- Frequently appearing verbs seem to be not as domain specific as nouns. Common verbs are various forms of *be*, *end*, *touch*, *have*, and *go*. However, there are a few verbs that appear frequently in some corpora but not in others. *Hit* and *miss* frequently appear in the Cannon, Hurricane, and Tornado corpus. *Cover*, *expand* and *grow* frequently appear in the Desert, Oil, and Lake corpus. It is not surprising because *hit* and *miss* can be naturally used for describing “translation” while *cover*, *expand* and *grow* naturally relate to “scaling”, which is the major difference in the above two corpora sets. There are also cases where verbs are specific to a domain. *Landed* used as a verb frequently appears

in Cannon and Hurricane. *Flooded* used as a verb appears exclusively in Lake. *Sailed* exclusively appears in Ship. This shows that domain semantic also influences verb usage, but not as explicit as nouns. More examples are *recede*, *retreat*, *leave*, *surrounded*, *shrink*, and *disappear*.

- Adjectives seem to even less domain specific. Common adjectives across all corpora are *middle*, *same*, *right*, and *lower*. The few cases where adjective are domain specific are the use of colors. *Blue*, *grey*, and *red* appears as to provide additional referral information respectively in Ship, Geometry, and Desert corpus. Exclusively in the Ship corpus, *shallow*, *light* and *dark* are frequently used to refer to the boundaries or the center of the water body. Adjectives about size were also used. *Large* appear more often in Oil.
- Prepositions are the least domain specific lexical category. Few prepositions are domain specific. *Across* frequently appears in translation scenarios but not in scaling ones.

In sum, POS-tagging offers possibilities to examine linguistic usages by lexical categories. Examining the nine corpora, frequently appearing nouns are highly domain specific; a few verbs and adjectives are domain specific and a general difference in translation vs. scaling can be found; prepositions are least domain specific, only the word “across” is found to be differentiable between translation scenarios vs. scaling scenarios.

The last analysis step here involves topic modeling [11,12]. It is a method for discovering “topics” shared among documents within a corpus. It can be viewed as cluster analysis for documents. Applying topic modeling to all documents (one for each participant, 20 documents per scenario) in the nine corpora (180 documents in total), we can evaluate whether documents might be clustered based on their domain. Mallet (Machine Learning for LanguagE Toolkit) [13] is used to realize topic modeling. Setting the “number of topics” to be nine, we can see if the nine topic models correlate with the nine domains (scenarios). Because topic models are data-driven and don’t imply any predefined knowledge, we want to compare the topic modeling result with domain semantics and see if they are comparable to each other. Each topic is defined by the keywords appearing most frequently and most distinctively.

Table 1. Keywords for the nine topic models (TopicID).

Topic ID	Keywords
0	bottom diamond circle stops top stop back grows expand box expands halfway corner middle diamond grey expanded moves touches
1	middle ended lower blue light ships upper boats side boat left top corner chose screen section hand cross horizontal
2	area half touch past stopping retreat point square part retreats short tan position contact full expanding small space pass
3	left side triangle inside ball end center circles middle ends line landed start high dot touching fell location images
4	water house flood back entire recedes reaches flooded touching halfway spread lake lakes front receded show past starting receeding
5	group desert reserve stopped fully put based touched nature partially chose groups animations criteria reached red shape sand choose
6	island oil covers completely covered cover spill stop reach ocean tip covering islands barely large pattern reaches spills animation
7	city edge icons tornado cannon balls region tornadoes border gray grouped tornados east boundary enter missed southwest town block
8	shallow land hit peninsula hurricanes hurricane moving made mid close west coast ship part central move hits low

Table 1 shows the keywords that identify each topic model. Unsurprisingly, domain specific nouns are distributed across topic models. These topic models can be used to evaluate the probability of one document (descriptions created by one participant) being associated with a specific topic model (ideally catching the domain). Assigning the most probable topic ID to a document allows for using topic models for document classification. To evaluate the correlation between topics and domain semantics further, we use the already built topic models to classify each document. The results and evaluations are shown in Table 2.

Table 2. Matching nine topic models to the nine domains in a confusion matrix.

Domain \ Topic ID	0	1	2	3	4	5	6	7	8
Geometry_translation	**16**	0	0	1	1	0	1	1	0
Desert	1	**14**	0	0	1	0	2	0	2
Lake	0	0	**19**	0	0	0	1	0	0
Cannon	3	0	0	3	3	1	0	**10**	0
Hurricane	0	0	0	1	**16**	0	1	1	1
Geometry_scaling	0	4	0	1	0	**14**	1	0	0
Oil	5	0	1	3	0	0	**11**	0	0
Tornado	0	0	0	0	3	0	2	**15**	0
Ship	0	0	0	0	0	0	1	0	**19**

Out of 20 documents from each domain, we evaluated the proportion of documents being classified into the same topic model, which ideally should correspond to the domain (this correspondence worked except for tornado and cannon, where most of both are assigned to Topic ID 7). The **bold** numbers in Table 2 shows the topic model (see also Table 1) that most documents from a domain are assigned to. As shown in Table 2, it is reasonable to match each topic ID to one domain semantic and the matching proportion (sum of diagonal cells divided by total) is 70.56%. Cross-examining the domain semantic with keywords from corresponding topic models (see Table 1) we find that a large proportion of documents are classified correctly.

However, the above matching of topic models and domain semantics may be skewed by the high volume of domain specific nouns. Hence, as a comparison, we removed all the domain specific nouns from all corora and rebuilt the topic models.

Table 3. Keywords for nine topic models (excludes domain specific nouns).

TopicID	Keywords
0	area hit grows box half gray touch tan past enter missed grow leaving boxes consumes retreat green direction paths
1	half covers cover covered fully entire recedes tip touch reach ocean covering recede oil starting sand island retraction affected
2	land icons touching chose center grouped landed hand border put close section location barely didn mass shore passed landing
3	left side middle top bottom end corner ends start high starts drop adjacent moved inbetween flush receds hang till
4	shallow stopped blue light moving made screen based cross horizontal vertical route low make angle map body path sailed
5	edge inside ended city line mid dot west east fell ball boundary central criteria south images portion impacts southwest
6	completely point flood stopping square touched icons retreat receded show house groups large pattern space receeding grass part recession
7	back stops stop halfway reaches past short grey retreats expanded front full moves touches position hits expanding reached disappears
8	group lower upper expand region part expands spread animations straight red shape shrink partially grew diamond slightly icon movement

As shown in Table 3, because all domain specific nouns are excluded, the keywords are not as clearly correlated to domain semantics. Nouns that are not domain specific, verbs, adjectives, prepositions and all other words are still kept and were used for building another topic model. In the following we analyze if these words can create document clusters that correlate to domain semantics, too.

Table 4. Matching nine topic models (excluding domain specific nouns) to the nine domains in a confusion matrix.

Domain \ Topic ID	0	1	2	3	4	5	6	7	8
Geometry_scaling	5	5	0	1	1	2	6	0	0
Hurricane	1	1	4	3	4	2	1	4	0
Tornado	1	0	10	1	1	2	4	0	1
Lake	3	1	0	7	0	2	3	0	4
Ship	8	3	0	0	9	0	0	0	0
Desert	0	0	0	1	0	5	3	1	10
Geome-try_translation	1	2	1	1	1	4	2	7	1
Cannon	2	2	6	1	1	1	0	6	1
Oil	3	0	0	5	0	3	1	0	8

From Table 4, we can see that document from each domain are being classified as belonging to various topic models. It is a stretch to relate topic IDs from this topic model to the nine domains and the highest possible matching proportion is only 29.44%. This result shows that excluding domain specific nouns lets the correspondance between topic models and domain semantics disappear.

4 Conclusions

Two observations are important: In the first part of this paper we showed an analysis by hand that allows for relating a qualitative formal description of a movement pattern to a linguistic description. The linguistic descriptions are varied and participants used manifold strategies to characterize formally identical movement patterns. However, we seem to be able to clearly reveal domain specific differences, especially if we look into whether or not domain semantics is present. In the second part of this paper we tried to use this insight and compared the documents from each domain (one document with all linguistic descriptions per participant, 20 documents in each domain). We found that figure and ground (moving entity and reference entity) are the dominating linguistic features used and that these nouns allow for classifying documents largely correctly. However, once we remove these obvious, domain specific features, classification and identification of documents becomes very inaccurate despite the differences we found in the first part.

There could be a number of reasons for this. Instead of comparing all documents of a particular domain, which contains linguistic descriptions of several, topologically distinguishable paths, we may need a finer granularity for the analysis. For example, we could extract all DC descriptions from all domains and focus only on these. Likewise, we could extract all descriptions for movement patterns that could be labeled *across* in the translation scenarios and *expand-and-retreat* in the scaling scenarios. We could perform this analysis for all topologically equivalent movement patterns that we used to design our experiments.

It also could be that the topic modeling approach we used needs refinement. Topic models make use of terms and co-occurrences with documents to discover topics. It is an effective method for knowledge discovery from large corpora without predefined knowledge. However, we are specifically looking for spatial language usage in this study. In order to reduce the influence of domain specific nouns, we use a crude method which is removing the domain specific nouns. Integrating predefined knowledge (in our case, specific target language and contexts) into topic models would allow an analysis to focus on certain term usages, which would enhance the capability of topic modeling.

To sum up, we presented a first exploratory analysis of a corpus that is the result of the conceptualization of movement patterns in different semantic domains. The unique aspect of our experiments is that grounding the design in qualitative spatial representation and reasoning frameworks allows for keeping the spatial information identical across domains only changing the semantic (domain specific) context. We are hopeful that this corpus can contribute to a better understanding of the relation between formal/computation models and spatial language across different domains.

References

1. Bateman, J.A.: Language and space. a two-level semantic approach based on principles of ontological engineering. International Journal of Speech Technology **13**, 29–48 (2010).
2. Galton, A.: Spatial and temporal knowledge representation. Earth Science Informatics **2**, 169–187 (2009).
3. Kordjamshidi, P., Otterlo, M. von, Moens, M.-F.: From language towards formal spatial calculi. In: Ross, R.J., Hois, J., Kelleher, J. (eds.) Computational Models of Spatial Language Interpretation (CoSLI) Workshop at Spatial Cognition 2010, Mt. Hood, Oregon, pp. 17–24. CEUR Workshop Proceedings (2010).
4. Ross, R.J., Hois, J., Kelleher, J. (eds.): Computational Models of Spatial Language Interpretation (CoSLI) Workshop at Spatial Cognition 2010, Mt. Hood, Oregon. CEUR Workshop Proceedings (2010).
5. Gentner, D., Boroditsky, L.: Individuation, relativity, and early word learning. In: Bowerman, M., Levinson, S.C. (eds.) Language acquisition and conceptual development, pp. 215–256. Cambridge Univ. Press, Cambridge (2001).
6. Egenhofer, M.J., Al-Taha, K.K.: Reasoning about gradual changes of topological relationships. In: Frank, A.U., Campari, I., Formentini, U. (eds.) Theories and methods of spatio-temporal reasoning in geographic space, pp. 196–219. Springer, Berlin (1992).
7. Freksa, C.: Temporal reasoning based on semi-intervals. Artificial Intelligence **54**, 199–227 (1992).
8. Randell, D.A., Cui, Z., Cohn, A.G.: A spatial logic based on regions and connections. In: Proceedings 3rd International Conference on Knowledge Representation and Reasoning, pp. 165–176. Morgan Kaufmann, San Francisco (1992).
9. Anthony, L.: AntConc (version 3.2.2). Waseda University, Tokyo, Japan (2011). available from http://www.antlab.sci.waseda.ac.jp/.
10. Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of HLT-NAACL 2003, pp. 252–259 (2003).
11. Blei, D.M., Lafferty, J.D.: Topic models. In: Srivastava, A.N., Sahami, M. (eds.) Text mining. Classification, clustering, and applications, pp. 71–93. CRC Press/Taylor & Francis, Boca Raton, Fla (2009).
12. Steyvers, M., Griffiths, T.L.: Probabilistic topic models. In: Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (eds.) Handbook of Latent Semantic Analysis. Lawrence Erlbaum, Mahwah, NJ (2007).
13. McCallum, A.K.: MALLET. A machine learning for language toolkit (2002). http://mallet.cs.umass.edu/.

Supporting inferences in space – A wayfinding task in a multilevel building

Carsten P. J. Gondorf [1], Cui Jian[2]

[1] Transregional Collaborative Research Center SFB/TR 8 Spatial Cognition, University of Bremen, Enrique-Schmidt-Str. 5, D-28359 Bremen, Germany
[1]gondorf@uni-bremen.de, [2]ken@informatik.uni-bremen.de

Abstract. The explorative study presented in this paper investigates indoor wayfinding strategies and inferences of users in an unknown building. Participants were asked to find two consecutive targets in a multilevel building with the help of written route instructions. Routes were generated by a dialogue system and tested against adjusted instructions containing additional architectural information. The empirical findings suggest that selectively adding structural information can help (1) to build up a participant's cognitive map and support inferences about returning paths, (2) a guidance system optimize its wayfinding process to avoid redundancy with respect to the human-friendly principle, and (3) to improve to a great extent the effectiveness and efficiency of the system itself, and generate more adaptive and intuitive route instructions.

Keywords: indoor wayfinding strategies, spatial inferences, cognitive map

1 Introduction

The purpose of the study was to test whether selectively adding information about the architectural structure of a building to a set of instructions provides additional inferences about the returning path, leading the user to choose another, perhaps shorter way back.

As for the instructions we used a dialogue system that automatically generates indoor route instructions when asked about locations, using text-based natural language input and output (cf. [2]). We are interested in how the behavioral findings could help (1) to improve route instruction generation according to cognitive principles underlying human route descriptions [6], and (2) to improve, to some extent, the effectiveness and efficiency of the system itself and generate more adaptive and intuitive route instructions.

Recent approaches to indoor wayfinding have encountered difficulties in their investigation in conference centers [3], libraries [1], and others. Hölscher et al. [4, 5] for instance, suggest the following taxonomy in cases of incomplete spatial information for complex multilevel buildings (a) central point strategy – more likely to be used by first-time-visitors/unexperienced users – hanging on to well-known parts of the building (e.g. main entry hall, main stairs etc.); (b) direction strategy; (c) floor strategy.

According to Kuipers [7] places can be connected by associated movement responses; concatenating such place recognition-triggered responses then constructs a linear route. If two or more crossing routes are merged together, a network occurs, i.e. an internal mental representation of the environment. This representation is dynamic enough to describe intuitive spatial relationships and relative positions of places.

2 Infokiosk

The written instructions given in-advance were derived from a dialogue system – called *Infokiosk* – which was developed and implemented in our research group I5-DiaSpace.

The route instructions generator of the current system runs on a combined computational model that consists of three sequential processing steps:

- GUARD (Generation of Unambiguous, Adapted Route Directions, cf. [2]) generates the “context-specific” low level route directions out of raw route paths that lead to a given destination.
- GOHLI (Generation of High Level Instructions, cf. [2]) segments the low-level route directions coming from GUARD and generates the high-level route instructions on the basis of major direction changes.
- GOSLRI (Generation of Structuring Landmarks Related Instructions), which is the primary focus in this paper, takes high-level route instructions from GOHLI as input, and generates High-level structuring landmarks-based route instructions.

3 Experiment

32 participants (20 women/12 men) with little or no prior familiarity were asked to undertake wayfinding tasks in a multi-level building (GW2 – University of Bremen). They were all German native speakers and university students with an average age of 22.91 (age range 20-35, SD=4.518). For participation they received either course credit or were paid €6 Euro each. With the given detailed written paper-based route instructions, they had to find two consecutive targets - as well as their way back to the starting point.

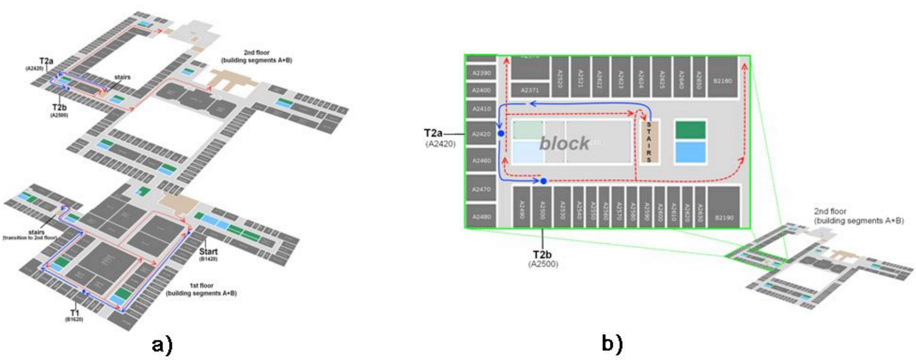


Fig. 1. Map views of the test environment, GW2, University of Bremen. (a) Shows building segments A/B, 1st and 2nd floor. Continuous lines indicate the path from the start point to first target (T1) and to target (2a) and (2b) on 2nd floor; dashed lines present a continuum of chosen returning paths. (b) Block scenario on 2nd floor.

Instructions were always given with respect to the point of departure. The movement commands in the instructions were all aligned to an egocentric frame of reference, e.g. "Turn around and go straight until the next hallway on the left-hand side." Adjusted versions were preceded by a short description, containing the floor and section of the building, in or on which the target is determined. Furthermore and most importantly for the scenario, on the 2nd floor we provided the explicit naming of an additional structuring path+landmark: *go around the block*. The *block* is a salient, structuring element of this particular environment (22x7meter), already perceivable while going up the stairs (see Fig. 1b).

The first target to reach was situated on the 1st floor and stayed the same for each subject – room B1620. The only difference was that each participant was given only either a system-generated (IS) or an adjusted instruction (IA). The second goal was on the 2nd floor and was divided into target (2a), room A2420, and target (2b), room A2500 (see Fig. 1b) – selection was done by taking into account the block structure and visibility of the stairs. That is, participants with target 2b had to walk further around the block (see Fig. 1b). The transition took place at a staircase in building segment A. The task started and ended on the 1st floor at room B1420. For this, as well as for the whole wayfinding task the users were instructed to think aloud and verbalize their thoughts and considerations. After reaching target (2a/b) the participants were instructed to hand back the instructions and walk to the starting point (they were not given any restrictions or assignments). After the task was completed, each participant was asked to fill two questionnaires – one for individual differences in spatial orientation, and the other asked about the performed task.

4 Results

All participants reached the two described targets (1, 2a/b) and got back to the starting point. By coding the chosen paths back by total numbers it was feasible to apply a set of nonparametric tests to the collected data. Analysis was carried out particularly for the type of instruction, i.e. target to reach on 2nd floor (T2a/b), and floor.

A total number of 23 participants chose the same returning path on the 2nd floor. Just 6 out of 32 participants continued walking around the block and used it as shortcut on the 2nd floor – 5 of them got the adjusted instruction (Target 2b) vs. one with a system generated instruction (T2b). A chi-square test was used for analysis based on floors and by comparable conditions, i.e. IS1 → T2a vs. IA1 → T2a on the one hand, and IS2 → T2b vs. IA2 → T2b on the other hand. For 1st floor for conditions IS(1) and IA(1) from target (2a) to the starting point (χ^2 =1.3, *df*=4, *p*>0.85; Cramér's *V*=0.28), and for IS(2) and IA(2) from target (2b) ((χ^2 =0.34, *df*=2, *p*>0.84; Cramér's *V*=0.14). For 2nd floor from target (2a) (χ^2 =1.1, *df*=2, *p*>0.58; Cramér's *V*=0.25), from target (2b) ((χ^2 =4.26, *df*=1, *p*<0.05; Cramér's *V*=0.51). This result supports the above described distribution of returning paths by total numbers for respective floors. Participants who handed out instruction IA2 were more likely to take the shortcut on their way back on 2nd floor in order to reach the stairs, i.e. to go back to the starting point, compared with their counterparts with IS(2).

This holds also for the evaluation of the walking distance of IS(2) compared with IA(2). A one-way ANOVA revealed a significant effect for walking distance on 2nd floor: F (1, 14)=5.091, p<0.05; η^2=0.27. This is mainly due to the fact that 7 out of 8 participants in condition IS(2) chose the same and thus longer returning path (~29.4m) – compared to 3 in condition IA(2). The shortcut (~21m) further around the block and thus in the direction of the stairs was selected by four participants in both conditions.

Participants' walking paths used for their return on the 1st floor back to the starting point showed a wide continuum of total returning paths across all conditions (see Fig.1a). This suggests a combination of central-point strategy and recognition-triggered response. The first review of the elicited data confirmed the behavioral findings for 1st floor: the central-point strategy was applied by 17 and recognition-triggered response by 13 people.

Furthermore, the analysis of question (9a) of the general questionnaire revealed, that participants who started with the system generated instruction (IS2) would change their walking preferences if confronted with a bird's-eye view of the scenario on the 2nd floor. Six out of eight would choose the shortcut (χ^2 =8.0, *df*=2, *p*<0.05; Cramér's *V*=1.0).

5 Discussion

The task was to reach two consecutive targets in a multilevel building with the help of a given route instruction. Both types of instruction served their purpose – all participants successfully reached their respective target on the 2nd floor and starting point. The qualitative analyses support the assumption that, due to the additional structural cue *block* + path, the participants built up (or faster updated) a more detailed internal mental representation of the environmental setting and thus were more likely to choose another, shorter returning path.

Therefore, the presented block structure is believed to be a good landmark, regarding the choice of a cognitively efficient return path. This in turn could be utilized for the generation of route descriptions by the presented guidance system. By accordingly annotating maps and implementing these kind of block structures as salient structuring landmarks it is not only possible to save user/body turns and make the generated adaptive instructions briefer and thus easier to recall, but also help novice building users with this cue to build up (i.e. faster update) their internal mental representation.

One key feature of an adaptive wayfinding guidance system is that, route instructions should be generated according to the cognitive principles underlying human route descriptions [6]. However, good general cognitive principles that can provide useful wayfinding information specifically suitable to human users within certain situations are in fact very difficult to discover. The empirical finding involving structuring landmarks such as blocks in this paper can help a guidance system optimize its wayfinding process to avoid redundancy with respect to the human-friendly principle, and therefore, improve to a great extent the effectiveness and efficiency of the system itself and generate more adaptive and intuitive route instructions. Specific landmarks such as blocks will be instantiated as salient structuring landmarks and used by the GOSLRI component to generate high-level structuring landmark-based route instructions.

Further research will address the degree of complexity and accuracy regarding additional architectural information implemented in indoor wayfinding instructions, e.g. regarding block size etc.

Acknowledgments

Thanks go to the German Science Fund (DFG) for funding and the research team of I5-DiaSpace of SFB/TR8 Spatial Cognition at the University of Bremen.

References

1. Carlson, L.A., Hölscher, C., Shipley, T.F., Conroy Dalton,R.: Getting Lost in Buildings Current Directions in Psychological Science 2010, pp. 19–284 (2010)
2. Cuayáhuitl, H., Dethlefs, N., Frommberger, L., Richter, K-F., Bateman, J. : Generating Adaptive Route Instructions Using Hierarchical Reinforcement Learning. In: Hölscher, C., et al. (eds.) Spatial Cognition V - Reasoning, Action, Interaction, pp. 319–334. Springer, Berlin (2010)
3. Hölscher, C., Brösamle, M., Vrachliotis, G.: Challenges in Multi-level Wayfinding: A Case-study with Space Syntax technique. Environment and Planning B: Planning & Design (in press)
4. Hölscher, C., Büchner, S., Meilinger, T., Strube, G.: Wayfinding Strategies and Map Use in a Multi-Building Ensemble. In T. Barkowsky et al. (eds.): Spatial Cognition V - Reasoning, Action, Interaction., pp. 365–380. Springer, Berlin (2007)
5. Hölscher, C., Meilinger, T., Vrachliotis, G., Brösamle, M., Knauff, M.: Finding the Way Inside: Linking Architectural Design Analysis and Cognitive Processes. In: Freksa, C., et al. (eds.) Spatial Cognition IV - Reasoning, Action, Interaction, pp. 1–23. Springer, Berlin (2005)
6. Klippel, A., Hansen, S., Richter, K.-F., & Winter, S. Urban Granularities - A Data Structure for Cognitively Ergonomic Route Directions. GeoInformatica, 13 (2), pp. 223-247 (2009)
7. Kuipers, B.: Modeling spatial knowledge. Cognitive Science, 2: pp. 129–153 (1978)

Semantic Ambiguity of Spatial Relational Nouns in Japanese

Sumiyo Nishiguchi

Department of Management
Tokyo University of Science
500 Shimo-kiyoku, Kuki-city, Saitama 346-8512, Japan
nishiguchi@rs.tus.ac.jp

Abstract. This paper discusses spatial terms in Japanese. Common nouns such as *ue* "on/over/above" and *naka* "inside" are used in Japanese to represent spatial and temporal locations, as *front* in *in front of*, or *center* in *in the center of* in English. I consider Japanese common nouns that represent spatial locations to be relational nouns that are two-place predicates, one of whose argument slots is filled by the entity represented by the other NP in the NP_1-*no* NP_2 construction. Since the corpus data [1] suggest that spatial nouns are often semantically ambiguous among physical, metaphorical, and temporal locations, the unified semantic entry in the Generative Lexicon (GL) [2] proves to be useful for handling the semantic ambiguity.

1 Spatial Relational Nouns

Languages such as Chickasaw in North America use relational nouns to express locations [3], rather than prepositions such as *in*, *on*, *under*, or *between* as in English. In (1a), *pakna'* "top" is a relational noun, that follows its possessor *chokka'* "house." Japanese is another language that expresses locations using relational nouns such as *naka* "inside," *ue* "on/above," and *shita* "under" as in (1b).

(1) a. chokka' pakna'
house top
"the top of the house (the house's roof)"

[3, 4]

b. mune-no mae-de tenohira-o awase (4179) [1]
chest-GEN front-LOC palms-ACC hold
"Put your palms together in front of your chest"

Mae "front" is a relational noun that does not stand alone semantically; therefore, it always means "the front of something," for example, *musuko* which means "son" always stands for "someone's son" (e.g., "Bill's son"). *Naka* "inside," *ue* "on/above," and *shita* "under" are two-place holders, and nouns such as *aida* "between" that require another argument are three-place predicates.

[1] The numbers in parentheses indicate the sentence IDs of the output of the data in the Yahoo! Chiebukuro section of [1] using ChaKi.NET 1.2β.

(2) a. $[\![ue\text{“}on/top\text{”}]\!] = \lambda x \lambda y[\mathrm{on}(y)(x)]$
b. [$_{VP}$ [[$_{NP}$ kohi-no ue]-ni] [miruku-o] [$_{V}$ ireru]]
coffee-GEN on-DAT milk-ACC put
“put milk on (the surface of) coffee”
c. $[\![kohi-no\_ue\text{“}on\_coffee\text{”}]\!] = \lambda x[\mathrm{on}(\varepsilon y.\mathrm{coffee}(y))(x)]$

2 Ambiguity among Physical, Metaphorical and Temporal Locations

Table 1 indicates that Japanese relational nouns are ambiguous among three types of readings—physical location, metaphorical location, and temporal sequence. For example, the most frequent relational word *ho* “toward” is generally used for comparisons and indicates preference as in (3a). Such meaning is a metaphorical extension of literal physical directions as in (3b). On the other hand, *mae* “front/before” is ambiguous between physical and temporal locations; for example, *shuppatsu-no mae* “before departure” (4000) and *mune-no mae* “in front of the chest” (4179).

Spatial Noun	Translation	Instances	Share	Physical Direction(Share)	Metaphor(Share)	Time(Share)
ho	toward	54	0.338	6(0.111)	48(0.889)	
naka	in	34	0.213	21(0.618)	13(0.382)	
aida	between/among	10	0.063	6(0.273)	1(0.1)	3(0.3)
ue	on	9	0.05	5	1	2
mae	in front of/before	6	0.037	5		1
shita	under	6	0.038	6(1)		
ue-no	top	6	0.038		6(1)	
ato	after	4	0.025			4(1)
chikaku	near	4	0.025	4(1)		
TOTAL		160	1	75	74	11

Fig. 1. Distribution of Spatial Nouns among 3083 Occurrences of “NP1-no NP2” in the *Yahoo! Chiebukuro* portion of [1]

(3) a. Chunichi-yori Hanshin-no ho-ga tsuyoi (2219)
Chunichi Dragons-than Hanshin Tigers-GEN direction-NOM strong
“The Chunichi Dragons are stronger than the Hanshin Tigers”
b. (neko-ga) watashi-no ho-e ki-masu. (5177)
cat-NOM me-GEN direction-GOAL come-HON
“Cats (usually) come toward me.”

3 Modeling Lexical Ambiguity of Spatial Language

3.1 Formal Semantics

This section formalizes the spatial terms in Japanese. Most of them are two-place holders except *aida* “between” which is a three-place predicate.

(4) a. $[\![mae]\!] = \lambda$x,y[in-front-of(x)(y)]
b. $[\![mae]\!] = \lambda$t,t'[before(t)(t')]

(5) a. $[\![mune - no_mae``in_front_of_the_chest"]\!] = \lambda$y.in-front-of($\varepsilon$x.chest(x))(y)
b. $[\![shuppatsu - no\_mae``before\_departure"]\!]$
$= \lambda$e'.$\exists$e[before(time(e))(time(e'))&departure(e)]

(6) a. $[\![ho]\!] = \lambda$x,y[toward(x)(y)]
b. $[\![ho]\!] = \lambda$x,y[to(x)(y)]

(7) a. $[\![(physical)naka]\!] = \lambda$x,y.in(x)(y)
b. $[\![(metaphorical)naka]\!] = \lambda$x,y.among(x)(y)

(8) a. $[\![nabe - no_naka``inside_the_pot"]\!] = \lambda$y.in($\varepsilon$x.pot(x))(y)
b. $[\![reshipi - no_naka``amongrecipes'']\!] = \lambda$y.among($\varepsilon$x.recipe(x))(y)

(9) a. $[\![aida]\!] = \lambda$x,y,z[between(x)(y)(z)]
b. $[\![aida]\!] = \lambda$x,y,z[among(x)(y)(z)]
c. $[\![aida]\!] = \lambda$t,t'[t'= during(t)]

(10) a. Ha-to ha-no aida atari-ga chairoku naru-no-desu-ka. (2906)
tooth-and tooth-GEN between vicinity-NOM brown become-GEN-HON-Q
"Have the gaps between your teeth turned brown?"
b. Geinojin-no aida-de hayat-teiru daietto-shokuhin (427)
entertainer-GEN among-LOC popular-PROG diet-food
"The diet food popular among TV entertainers"
c. Koko sukagetsu-no aida (3201)
this a few months-GEN period
"during these few months"

3.2 Lexical Ambiguity in the Generative Lexicon

Contrary to the previous section which listed two-way or three-way ambiguous lexical entries, the GL [2] has the means to provide unified lexical entries for a single spatial term, due to its elaborate lexical semantic information. In particular, the Lexical Conceptual Paradigm (LCP) [4, 2] is a powerful tool for resolving semantic ambiguity.

The formal quale in GL contains ontological information. In (11), *coffee* is a drink according to its formal quale, and its higher ontological category is a physical entity, which implies that *ue* "on" is interpreted physically. The unification profess is described in the following manner:

(11)

$$\begin{bmatrix} \textbf{COFFEE} & \\ \text{ARGSTR} = & \begin{bmatrix} \text{ARG1} = \boxed{x}\textsc{drink} \\ \text{D-ARG1} = \boxed{y}\textsc{human} \\ \text{D-E1} = \boxed{e1}\textsc{process} \end{bmatrix} \\ \text{QUALIA} = & \begin{bmatrix} \text{FORMAL} = \textsc{liquid}(\boxed{x}) \\ \text{TELIC} = \textsc{drink\_act}(\boxed{e1}, \boxed{y}, \boxed{x}) \end{bmatrix} \end{bmatrix} \begin{bmatrix} \textbf{UE ``ON''} & \\ \text{ARGSTR} = & \begin{bmatrix} \text{ARG1} = \boxed{x}\textsc{physical\_object} \\ \text{ARG2} = \boxed{y}\textsc{physical\_object} \\ \text{D-E1} = \boxed{e1}\textsc{state} \end{bmatrix} \\ \text{QUALIA} = & \begin{bmatrix} \text{FORMAL} = \textsc{on}(\boxed{e1x}, \boxed{x}, \boxed{y}) \end{bmatrix} \end{bmatrix} \begin{bmatrix} \textbf{KOHI-NO UE ``ON COFFEE''} & \\ \text{ARGSTR} = & \begin{bmatrix} \text{ARG1} = \boxed{x}\textsc{physical\_object} \\ \text{ARG2} = \boxed{y}\textsc{coffee} \\ \text{D-E1} = \boxed{e1}\textsc{state} \end{bmatrix} \\ \text{QUALIA} = & \begin{bmatrix} \text{FORMAL} = \textsc{on}(\boxed{e1}, \boxed{x}, \boxed{y}) \end{bmatrix} \end{bmatrix}$$

Mae “in front/before” is lexically ambiguous between physical and temporal locations. Lexical ambiguity calls for a meta-entry, that is, the LCP, which is a Cartesian product of the different concepts represented by a single lexical item [2, 5] as in (12). For example, *book* is a Cartesian product of a physical entity and the information contained within it, thus, (13a,b) are both grammatically correct.

(12) mae.lcp = {location.time, location, time }

(13) a. The book is on the table.
b. That book was right. An earthquake did happen as it had predicted.

(14)

$$\begin{bmatrix} \textbf{MAE} & \\ \textbf{``IN FRONT/BEFORE''} & \\ \text{ARGSTR} = & \begin{bmatrix} \text{ARG1} = \boxed{x}\,\text{PHYSICAL\_OBJECT} \\ \text{ARG2} = \boxed{y}\,\text{PHYSICAL\_OBJECT} \\ \text{E1} = \boxed{e1}\,\text{PROCESS} \\ \text{E2} = \boxed{e2}\,\text{STATE} \\ \text{D-E1} = \boxed{e3}\,\text{STATE} \end{bmatrix} \\ \text{QUALIA} = & \begin{bmatrix} \text{LOCATION.TIME\_LCP} \\ \text{FORMAL} = \text{R} \\ \begin{pmatrix} \text{LOCATION}\left(\boxed{e3}\text{:IN FRONT OF}\left(\boxed{e3}, \boxed{x}, \boxed{y}\right)\right), \\ \text{TIME}\left(\boxed{e2}\text{: BEFORE}\left(\text{TIME}\left(\boxed{e2}\right), \text{TIME}\left(\boxed{e1}\right)\right)\right) \end{pmatrix} \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} \textbf{MUNE-NO MAE} & \\ \textbf{``IN FRONT OF CHEST''} & \\ \text{ARGSTR} = & \begin{bmatrix} \text{ARG1} = \boxed{x}\,\text{PHYSICAL\_OBJECT} \\ \text{ARG2} = \boxed{y}\,\text{BODY\_PART} \\ \text{D-E1} = \boxed{e1}\,\text{STATE} \end{bmatrix} \\ \text{QUALIA} = & \begin{bmatrix} \text{LOCATION.TIME\_LCP} \\ \text{FORMAL} = \\ \text{LOCATION}\begin{pmatrix} \boxed{e1}: \\ \text{IN FRONT OF}\left(\boxed{e1}, \boxed{x}, \boxed{y}\right) \end{pmatrix} \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} \textbf{SHUPPATSU-NO MAE} & \\ \textbf{``BEFORE DEPARTURE''} & \\ \text{ARGSTR} = & \begin{bmatrix} \text{E1} = \boxed{e1}\,\text{DEPARTURE} \\ \text{E2} = \boxed{e2}\,\text{STATE} \end{bmatrix} \\ \text{QUALIA} = & \begin{bmatrix} \text{LOCATION.TIME\_LCP} \\ \text{FORMAL} = \text{TIME}\left(\boxed{e2}\text{:BEFORE}\left(\boxed{e2}, \boxed{e1}\right)\right) \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} \textbf{AIDA} & \\ \textbf{``BETWEEN/AMONG/DURING''} & \\ \text{ARGSTR} = & \begin{bmatrix} \text{ARG1} = \boxed{x}\,\text{LOCATION\_HUMAN\_TIME} \\ \text{ARG2} = \boxed{y}\,\text{LOCATION\_HUMAN} \\ \text{ARG3} = \boxed{z}\,\text{LOCATION\_HUMAN} \\ \text{E1} = \boxed{e1}\,\text{STATE} \end{bmatrix} \\ \text{QUALIA} = & \begin{bmatrix} \text{LOCATION.MENTAL\_LOCATION.TIME\_LCP} \\ \text{FORMAL} = \text{R} \\ \begin{pmatrix} \text{LOCATION}\left(\boxed{x}, \boxed{y}, \boxed{z}\right), \\ \text{MENTAL\_LOCATION}\left(\boxed{x}, \boxed{y}, \boxed{z}\right), \\ \text{TIME}\left(\boxed{e1}\text{:DURING}\left(\boxed{e1}, \boxed{x}\right)\right) \end{pmatrix} \end{bmatrix} \end{bmatrix}$$

Argument structure also needs to have metaentries since *mae* “front/before” and *aida* “between/among/during” combine with different types of semantic arguments.

4 Conclusion

In this paper, spatial language in the form of “NP_1-GEN NP_2” constructions in Japanese was taken from [1] and classified into literal, temporal, and metaphorical meanings. Spatial terms are semantically ambiguous relational nouns. Lexical meta-entries in the GL effectively handle the semantic ambiguity of the most common spatial nouns.

References

[1] BCCWJ: Balanced Corpus of Contemporary Written Japanese, BCCWJ2009 edition. The National Institute of Japanese Language (2009)
[2] Pustejovsky, J.: The Generative Lexicon. MIT Press, Cambridge (1995)
[3] Lillehaugen, B.D., Munro, P.: Prepositions and relational nouns in a typology of component part locatives (2006)
[4] Pustejovsky, J., Anick, P.: On the semantic interpretation of nominals. In: Proceedings of COLING-1988, Budapest (1988)
[5] Carpenter, B.: The Logic of Typed Feature Structures. Cambridge University Press, Cambridge (1992)

Analyzing directionality: From paths to locations

Sander Lestrade

Research group Sprachkontakt und Sprachvergleich &
SFB/TR 8 Spatial Cognition, University of Bremen, Germany

Abstract. This paper proposes an alternative to the currently prevalent analysis of directionality in terms of paths. It is argued that directionality should be understood as the temporal specification of locative modification in its stead. The proposal is compatible with both geometric and functional representations of space, is corroborated with typological findings, and meets the requirements for the careful development of a spatial ontology.

Keywords: directionality, locations, paths, event structure, motion

1 Against paths as the primitives of directionality

Spatial expressions are predicates that map a thing or event onto a location.[1] This location is specified by the configuration function in terms of a (geometrically or functionally defined) region with respect to a ground. In *The ball is under the table.*, *the ball* is said to be in the location specified by *under the table*. The predication of locations often is temporarily restricted (cf. [9]). Generally, an object is mapped to some region for a restricted interval of time only as its position may change at a later stage. In the present proposal, this change of configuration is the realm of the directionality function.

In the currently prevalent analysis of directionality, *viz.* the one by Jackendoff ([7]), directionality is a function that refers to ordered stretches of space, so-called *paths*. For example, the FROM path *off* refers to a path that has an ON configuration as its starting point. The first and most important objection to the analysis of Jackendoff concerns the methodology in the collection of the data that is said to constitute the directionality domain [7, 168–169]. There is no independent evidence that the used examples actually are examples of the same phenomenon, i.e. of directionality. Indeed, some of Jackendoff's directionality expressions probably express something completely different. For example, TOWARD is said to belong to the type of paths called *directions*, which, unlike a *bounded* path such as TO, does not include (the region with respect to) the reference object but would do so if the path were extended by some unspecified distance. In a non-trivial sense, we probably only want to allow for extensions in approximately the same direction (otherwise, any direction could be turned

[1] Thanks to John Bateman, Carl Schultz, Thora Tenbrink, and two anonymous reviewers for valuable comments and discussion.

into a TO path). Now, imagine an enclosure around point A with an opening at its south side and point B to its north. Because of the enclosure, one can only go from *A to B* going southwards, through the opening. To go from *A toward B*, however, one should go north. Crucially, the TOWARD path in this situation cannot be extended in the same direction to become a bounded *to B* path. Thus, instead of directionality, *toward* rather seems to expresses orientation. When modifying a motion event with this expression, the moving object of course ends up closer to the ground. And by continuing along this direction, one will generally end up at this ground too. But this need not be, as this example shows.

As a second objection, the path reference that is assumed by Jackendoff may follow from world knowledge instead of being part of the lexical semantics of directionality (cf. the procedure for the development of a spatial ontology in [1] and the *principle of conceptual abstraction* in [16, 595]). We know that it takes a path from A to B to go from A to B as we cannot but traverse all points in between when going there. Crucially, however, such paths are not necessarily what is referred to by directionality expressions. In fact, directionality expressions are probably better analyzed as predicates (cf. [16], [17]).

Finally, if directional PPs referred to paths, it should be possible to combine an expression of duration with the continuation along such a path. But this is not possible with Goal directionality as illustrated by the ungrammaticality of *He is walking into the building (*for hours).* (cf. also [18]). Note that this is not due to the semantics of the verb *to walk*: If explicit reference *is* made to a path, by substituting *through* for *into* in the example, it is possible to use an expression of duration. Thus, the contrast between these sentences shows that (Goal) directionality is probably *not* about paths.

2 Directionality in terms of locations

A more careful procedure than the one used by [7] can be followed in the collection of the data. It has been observed that more grammatical means of expression tend to make less idiosyncratic meaning distinctions ([14, 178], [6, 178], [1, 1035]). Also, it has been found that spatial cases primarily express directionality ([8], [13], [2]). Finally, cross-linguistic agreement is said to suggest relative uniformity in the way people conceptualize a domain [4]. So not only can we indeed expect directionality distinctions to be expressed by spatial case systems, also, we can expect whatever directionality distinctions that are made by spatial case to be of a more fundamental, conceptual type, especially when they show up in language after language. In a cross-linguistic study of spatial case inventories that is thus motivated, Lestrade [10] identifies three basic distinctions of directionality: Place, Goal, and Source.

This kernel of directionality could be described in terms of paths, but, arguably, it is preferable to use locations only as we need these anyway for the configuration function. Then, Goal and Source directionality denote a change of location, and Place denotes an absence of such a change. To define Goal and

Source, we need some ordered dimension: Goal directionality denotes a change into some location, Source does the opposite.

An ordered dimension can be provided for free by the extended event structure of the verb. Pustejovsky ([12]) argues that Davidsonian event arguments may have internal structure. For our present purposes, only the structure in which there is a strict partial order between the two subevents is relevant:

(1) a. $[_{e_3}\ e_1 <_\alpha e_2\ ] \ =_{def} \ <_\alpha (\{e_1, e_2\}, e_3)$
b. $\forall e_1, e_2, e_3 [<_\alpha (\{e_1, e_2\}, e_3) \leftrightarrow e_1 \preceq e_3 \wedge e_2 \preceq e_3 \wedge e_1 < e_2 \wedge$
$\forall e[e \preceq e_3 \rightarrow e = e_1 \vee e = e_2\ ]]$

In this definition, event e_3 is a complex event structure that consists of two subevents, e_1 and e_2, where e_1 and e_2 are temporally ordered such that each is a logical part of e_3, the first subevent precedes the second, and there is no other event that is part of e_3 [12, 69]. For example, the verb *build* is analyzed into a development process and a resulting state.

Pustejovsky [12, 74] explicitly allows for adverbial phrases to take scope over both the entire event and over individual subevents. Thus, we have three logical possibilities for spatial modification of motion verbs, which nicely corresponds to the empirically established kernel: the spatial modification of the entire event is called *Place* directionality (note the different use of this term here from the one by Jackendoff in the above); the modification of the first subevent is called *Source*, and the modification of the second subevent is called *Goal*. For example, depending on the type of directionality that is imposed by the spatial modifier and assuming the structure in (1), a walking event e_3 of subject x modified by location y can be decomposed as follows: $[walk(e_3, x) \wedge loc(e_3, x, y)]$ for Place, $[walk(e_1, x) \wedge loc(e_2, x, y)]$ for Goal, and $[loc(e_1, x, y) \wedge walk(e_2, x)]$ for Source.

In principle, the explicit spatial modification of one subevent by location y does not exclude the additional implicit modification of the second subevent by this same location. Following a suggestion of Hendriks et al. [5, chapter 8], we can ensure a change of location in a system of pragmatic contrasts (cf. also [15]): The speaker would have modified the whole event if the location had scope over the whole event, so if she only modifies the first subevent, we know that the locative function does not apply to the second one by pragmatic implicature.

By only using existing structures that have been established independently from present purposes, the proposal meets the criterion of Bateman et al. [1] to exclude the contribution of world knowledge in the development of a spatial ontology. Also, the account straightforwardly accounts for syncretism patterns in directionality systems. It has been observed that such syncretisms occur between Place and Source, between Place and Goal, or between all three distinctions, but not between Source and Goal to the exclusion of Place (cf. [2], [10], [11]). This naturally follows from the present proposal: If a language has a special marker for the spatial modification of the first subevent, the second subevent and the entire event will be treated uniformly as its complement (and the other way around), but taking together the two subevents would render the entire event (cf. (1-b)) and therefore could not be distinguished from it. Finally, the temporal specification of a spatial modification does not impose any specific ontological category to this modification and is thus compatible with both geometrical and functional representations of space [3].

In conclusion, it was argued that directionality is best analyzed as the locative modification of an extended event structure. This accounts for the empirically established kernel of directionality, correctly predicts attested syncretism patterns, and does not stipulate any additional machinery.

References

1. Bateman, J.A., Hois, J., Ross, R., and Tenbrink, T.: A linguistic ontology of space for natural language processing. Artificial Intelligence **174**, 14 (2010) 1027–1071
2. Creissels, D.: Spatial cases. In Malchukov, A., and Spencer, A., The Oxford Handbook of Case, Oxford: Oxford University Press (2009) 609–625
3. Garrod, S.C., and Sanford, A.J.: Discourse models as interfaces between language and the spatial world. Journal of Semantics **6** (1988) 147–160
4. Gentner, D., and Bowerman, M.: Why some spatial semantic categories are harder to learn than others.In J. Guo et al., Crosslinguistic approaches to the psychology of language, New York: Psychology Press (2009) 465–480;
5. Hendriks, P., de Hoop, H., Krämer, I., de Swart, H., and Zwarts, J.: Conflicts in Interpretation. London: Equinox publishing (2010)
6. de Hoop, H., and Zwarts, J.: Case in formal semantics. In: Malchukov, A., and Spencer, A. (eds), The Oxford Handbook of Case, Oxford: Oxford University Press (2009) 170–184
7. Jackendoff, R.: Semantics and Cognition (ed. 8). Cambridge: MIT Press (1983)
8. Kilby, D.: Universal and particular properties of the Ewenki case system. Papers in Linguistics **16**, 3/4, (1983) 45–74
9. Kracht, M. On the semantics of locatives. Linguistics and Philosophy **25** 2 (2002) 175–232
10. Lestrade, S.: The kernel of directionality: A spatial case study. Manuscript, University of Bremen (2011)
11. Pantcheva, M.: The syntactic structure of locations, goals, and sources. Linguistics **48**, 5, (2010) 1043–1082
12. Pustejovsky, J.: The generative lexicon. Cambridge, MA and London: The MIT Press. (1995)
13. Stolz, Th.: Lokalkasussysteme. Aspekte einer strukturellen Dynamik. Wilhelmsfeld: Gottfried Egert Verlag (1992)
14. Talmy, L.: Toward a Cognitive Semantics. Cambridge and London: MIT Press (2000)
15. Wilkins, D.P. and Hill, D.: When "go" means "come": Questioning the basicness of basic motion verbs. Cognitive Linguistics **6** (1995) 209–259
16. Wunderlich, D.: How do prepositional phrases fit into compositional syntax and semantics? Linguistics **29** (1991) 591–621
17. Zwarts, J.: Vectors as relative positions: A compositional semantics of modified PPs. Journal of Semantics **14** (1997) 57–86
18. Zwarts, J.: Aspects of a typology of direction. In Rothstein, S. (ed.), Theoretical and crosslinguistic approaches to the semantics of aspects, Amsterdam: John Benjamins (2008) 79–106

Applying spatial knowledge from a scene description task to question answering

Simon Dobnik

Department of Philosophy, Linguistics and Theory of Science
Box 200, 405 30 Göteborg, Sweden
simon@dobnik.net
http://flov.gu.se

This presentation extends our previous work in which we build and test a mobile robot which learns grounded semantic representations of spatial concepts from human descriptions and its own perception through sensors of a room containing real objects. The learning is performed offline as a machine learning classification task. In [1] we show that the learning of spatial concepts is successful when the classifiers are evaluated. In [2] we argue that classifier evaluation is not enough to show that the robot acquired human-like spatial knowledge which generalises to new spatial configurations. We therefore integrated the classifiers in our own NL generation system (*pDescriber*) which produces grounded descriptions of spatial scenes such as "The table is to the left of the chair" and allows humans to observe the acquired spatial knowledge. The discourse setting in which these descriptions are made is identical to the one in which they were sampled before they were learned. In this contribution we examine whether we can use data-sets and classifiers from the scene description task to answer questions that (A) locate objects: "Where is the table?" - "The table is to the left of the chair"; (B) confirming object description: "Is the table to the left of the chair?" - "No, the table is near the chair."; (C) find objects: "What is to the left of the chair?" - "The pillars, the tyres and the wall are to the left of the chair"; and (D) reference objects: "What is the chair to the left of?" - "The chair is to the left of the table, the desk and the wall". We see the task as an experimentally constrained form of dialogue which contains only two dialogue acts (information request and answer) which are always performed by the same illocutionary partner: a human directs questions to the robot. Since the dialogue is situated both spatially and in discourse we do expect to find effects of semantic coordination of human observers when interpreting the robot's responses.

Generating question answers (*pDialogue*) requires more steps than generating descriptions and hence more factors may influence the evaluation of spatial knowledge. User utterances must be interpreted as questions and their content must be matched against dialogue rules which specify how to answer them. Most dialogue rules require an application of ML classifiers that take linguistic descriptions and predict perceptual properties rather than reverse (*pDescriber*). The classification tells us what state of perception corresponds to a description. The dialogue rules must then issue commands that bring the robot to this state or find a configuration of objects that holds in the state. The resulting knowledge is used to generate natural language sentences.

The system was individually evaluated by 13 non-expert volunteers in a room environment different from that used in data collection for ML. Each evaluator considered the robot's answers to 55 questions which were scripted and were automatically asked by the evaluation software at four distinct room locations (L1 to L4). This ensured that various spatial and linguistic conditions were covered. The evaluators' task was to indicate whether each robot's answer is an intuitive or natural description on a scale from 1 (bad) to 5 (best). Each run took between 45 to 60 minutes to complete. We estimated evaluator agreement by calculating Pearson's correlation coefficient between the scores of each evaluator per particular question-answer pair and the mean of such scores over all other evaluators. The overall agreement of 0.583 (the mean of correlation coefficients from all 13 folds) shows that there is a considerable consensus between the evaluators on the performance of the system.

To estimate the accuracy of the system the evaluator scores were normalised to values between 0 and 1 (1=0, 2=0.25, 3=0.5, 4=0.75, 5=1) and summed. The accuracy per question type is as follows: A - 43.5%, B - 54.2%, C - 54.7%, D 56.9% and mean - 52.3%. The steps involved in answering questions A are identical to generating a description in *pDescriber* (59.3%) [2] – one or two objects are selected at random and the relation between them is classified – but the estimated performance of *pDialogue* on questions A is lower by 15.8%. The result suggests that a new discourse setting affects the interpretation of spatial descriptions. When the system generates a description on its own, a human hearer understands it as a general statement about the scene that both are observing. However, when an agent in conversation asks a question, they expect an informative and relevant answer which helps them to interpret the scene. Choosing a salient reference object is particularly important. Objects can be salient in their visual properties (visual-salience) or through being previously discussed and located in dialogue (discourse salience). The modelling of both kinds of salience is an object of our future investigations.

We also tested two other properties affecting the semantics of spatial descriptions in a situated discourse. The difference in evaluation scores for questions-answer pairs that involved (a) objects that were in the robot's visual field (L1 and L2) and those that were not (L3) is statistically significant (t-test: $a > b$; $\alpha = 2P = 0.000$). The interlocutors expect the robot to change its orientation towards objects referred to in questions and answers. Secondly, the difference in evaluation scores (a) where the spatial description in questions was unambiguous in terms of the reference frame (C at L1: "What is to the left of you?" – intrinsic) and (b) where a question could be answered using an alternative reference frame ("What is to the left of the chair?) is not statistically significant (t-test: $a = b$; $\alpha = 2P = 0.61 > 0.05$.). This shows that human observers align to the reference frame chosen by the robot (relative to itself) and do not insist on changing it.

References

1. Dobnik, S.: Learning spatial referential words with mobile robots. In: Proceedings of the 9th Annual CLUK Research colloquim. The Open University, Milton Keynes, United Kingdom (2006)
2. Dobnik, S., Pulman, S.: Human evaluation of robot-generated spatial descriptions. In: Proceedings of the Workshop on Computational Models of Spatial Language Interpretation (CoSLI). Portland, Oregon, USA (2010)

Timelines: Conceptual Integration, Emotion, and Poetic Effects

Cristóbal Pagán Cánovas[1], Seana Coulson[2], Max Jensen[3],

[1] Classics Department, University of Murcia, Spain
cpcanovas@um.es
[2] Cognitive Science Department, University of California San Diego, California, USA
coulson@cogsci.ucsd.edu
[3] Cognitive Linguistics & World Literature, Case Western Reserve University, Cleveland, Ohio, USA
mfj4@case.edu

Abstract. One of the most broadly investigated topics in the literature on conceptual mappings is the importance of spatial construals for thinking and talking about time. In two forthcoming articles [1] [7] we explore how people understand timelines – both as graphical objects, in discourse about timelines taken from newspapers and the web, and in poetic examples. A comparison with metaphors incorporating circular patterns shows that temporal and affective meanings can change dramatically when they arise from different spatial structures.

Keywords: timeline metaphors, time-space mappings, generic templates of conceptual integration, material anchors for conceptual blends, image schemas, emotion.

1 The Timeline

When instantiated graphically, the timeline serves as a material anchor [4] in a conceptual integration network [2] representing partial cognitive models of time, lines, objects, and a hybrid model known as a *blend*. When understood with respect to this network, the analogue properties of the line give it novel computational properties that facilitate inferences about the events it represents.

The history of the modern timeline reflects a distributed cognitive process involving multiple individuals over a large span of time. It illustrates the cultural development of conceptual integration networks. Conventional mapping schemas are best viewed not as determining the interpretation of timelines, but rather as providing soft constraints that help guide meaning construction.

2 Anchoring the Time-Space Scene: Computational Properties of Timelines

A detailed analysis shows that the cognitive linguistic research on time-space metaphors does not merely describe a set of mappings from space to time; it rather describes a particular spatial scene with temporal meaning, which recurs across many metaphoric expressions. This scene is framed by a simple, familiar spatial event: an object or objects travel towards a reference point or observer.

The temporal version of the motion scene has restrictive and often contradictory properties. For instance, in this schematic narrative all observers are on the same spot, all objects are aligned to travel along the same path, objects cannot overtake one another, arrive at the same time or from different directions, change trajectory, etc. None of these occur in our normal experience of motion through space. These properties rather originate from cognitive constraints set by our existing knowledge of time. Instead of being the result of direct space-time projections, they emerge from successive integrations of a variety of conceptual materials, including event structure, motion, and a cultural mechanism to measure duration [3]. The properties of the line comply with these constraints, and provide an adequate topology for the blend, although they clash with many other basic aspects of our experience of traversing paths. Thus the timeline has properties distinct from those of the cognitive models in each of its inputs.

Although it instantiates some of the mappings in the TIME IS SPACE metaphor [6] [5], the timeline itself is an integrated construct with computational affordances that differ from those available in the input domains. For example, studying a timeline might enhance one's memory for the sequence of salient events, or allow us to more easily recognize the most productive periods via the density of points. Researchers in the field of information visualization recommend timelines because their visual properties facilitate inferences about temporal events (such as temporal and causal contingency) that are either difficult or impossible to make using other representational formats [8].

Much of the emergent structure of the timeline and its novel computational properties result from the compression of temporal relationships to spatial ones, as well as from the congregation in the blend of structures from multiple inputs. Rhetorical goals are also crucial, as shown by everyday metaphoric expressions providing further emergent properties: timelines can be cut or compressed into analogous but shorter ones, years can be taken away from them, they can be accelerated, etc.

3 Linear and Circular Patterns in Poetic Time Metaphors

In addition to the analysis of the computational properties of timelines and the metaphoric language related to them [1], Pagán Cánovas and Jensen [7] compare time metaphors by Borges, Kavafis, Heraclitus, Manrique, Lorca, Quevedo, Paz, and Shakespeare, and by recent prose writers Ian McDonald and Karen Russell. These metaphors exhibit a linear pattern (such as a river) or a circular pattern (such as a winding labyrinth).

Analysis of this corpus suggests that static lines and circles can acquire narrative properties, be instantiated according to relevant cultural frames and rhetorical goals (e.g. a line can be a snake, a circle a magnifying glass), blend with the self, with emotional scenarios, with motion along a path, etc., while still retaining their temporal values. Although straight lines or circles, or time itself for that matter, are not by themselves loaded with emotion or intentionality, the image-schematic properties of these blends can be opportunistically exploited on-line for further integrations with contextual and background knowledge, in order to produce emergent affective meanings.

In our poetic examples we see that the linear pattern is more suitable to function as a material anchor, which helps ground conceptualization on a perceptual structure. Unlike the timeline, the circular pattern is not so appropriate to provide spatial landmarks on which to ground temporal relations. Past, future, periods of human life, duration differences, or remaining time available are not so easily "seen" at a glance in the circle.

References

1. Coulson, S., Pagán Cánovas, C. Understanding Timelines: Conceptual Metaphor and Conceptual Integration. Cognitive Semiotics, Special issue on Conceptual Metaphor Theory (forthcoming)
2. Fauconnier, G., Turner, M. The way we think. Conceptual blending and the mind's hidden complexities. Basic Books, New York (2002)
3. Fauconnier, G., Turner, M. Rethinking Metaphor. In: Gibbs, R. (ed.): Cambridge Handbook of Metaphor and Thought. Cambridge University Press: New York (2008)
4. Hutchins, E. Material anchors for conceptual blends. Journal of Pragmatics 37 (2005) 1555–1577
5. Lakoff, G. The contemporary theory of metaphor. In: Ortony, A. (ed.): Metaphor and thought. Cambridge University Press: Cambridge (1993)
6. Lakoff, G., Johnson, M. (1980). Metaphors we live by. Chicago and London: University of Chicago Press.
7. Pagán Cánovas, C. Jensen, M. Kavafis' Candles, Manrique's Rivers, Life Rounded with a Sleep: Linear and Circular Patterns for Affective Time (in review)
8. Phan, D., Gerth, J., Lee, M., Paepcke, A., Winograd, T. Visual Analysis of Network Flow Data with Timelines and Event Plots. VizSEC 2007 85-99 (2008)

Author Index