Situating Spatial Templates for Human-Robot Interaction

John Kelleher and **Robert Ross** and **Brian Mac Namee** and **Colm Sloan**
Artificial Intelligence Group, Dublin Institute of Technology, Ireland.
johnd.kelleher,robert.ross,brian.macnamee@dit.ie, colmsloan@gmail.com

People often refer to an object by describing its spatial location relative to another object. Due to their ubiquity in situated discourse, the ability to use such **locative expressions** is fundamental to human-robot dialogue systems. Computational models of spatial term semantics are a key component of this ability. These models bridge the grounding gap between spatial language and sensor data. Within the Artificial Intelligence and Robotics communities, spatial template based accounts, such as the Attention Vector Sum model (Regier and Carlson 2001), have found considerable application in mediating situated human-machine communication (Gorniak and Roy 2004; Brenner et al. 2007; Kelleher and Costello 2009).

Through empirical validation and computational application these template-based models have proven their usefulness. We argue, however, that important contextual features are being ignored; resulting in over-generalization and failure to account for actual usage in situated context. Such over-simplifications are a natural consequence of the experimental design taken in acquiring template-based models. These experimental designs used simplified scenes and reduced 2-dimensional survey-based object configurations. While this is understandable given the original aims of these studies, we nevertheless believe that this is not sufficient justification for the direct application of idealized spatial templates to situated communication.

This critique of template based models is similar in spirit to critiques already put forward by a number of researchers: Coventry and Garrod (2004) have stressed the need to account for functional effects; Kelleher and Costello (2009) highlighted the need to account for the effects introduced by distractors. Here, we argue that the models must also be extended to incorporate perspective effects.

Situating Spatial Templates

Following previous psycholinguistic data (Logan and Sadler 1996; Carlson-Radvansky and Logan 1997; Kelleher and Costello 2005), template-based models of directional spatial term semantics essentially model spatial term acceptability across a region as a function of the distance from the landmark and the angular deviation of a point from a direction vector defining the canonical direction of the spatial term in a given frame of reference. In these models, as in the data, acceptability drops as angular deviation and distance increase. In situations where multiple frames of reference are applicable, multiple functions may be combined. While such models have been shown to hold for canonical and 90° off canonical landmark orientations, the mechanisms by which such templates may be distorted by situated use - as would be the case for human-robot interaction - has yet to be systematically investigated. The next section describes a study conducted to determine how acceptability ratings altered for oblique landmark orientations.

Participants, Stimuli & Procedure

Participants for the study were recruited online and compensated. 42 participants were native English speakers and their data was retained for analysis. Participants were asked to rate their agreement with a series of paired linguistic and visual stimuli. Linguistic stimuli situated a trajector with respect to a landmark. Each linguistic stimulus was of the form 'The A box is REL of the B box', where A and B were substituted by color words (explained below) and REL was one of three directional spatial terms, i.e., 'in front', 'to the right', and 'to the left'. Visual stimuli were 2.5 dimensional images of a scene consisting of a rectangular landmark and cylindrical trajector. The landmark object was 8 units wide by 6 units deep by 2 units high, while cylinders were one unit in diameter and one unit high. The landmark was situated obliquely to the participant's viewing angle. While the landmark position and viewing angle were fixed, trajector position could be moved to one of ten locations. The landmark, trajector positions, and viewer angle are depicted to scale in Figure 1[1]. Note that in order to reduce repetition effects, each scene configuration was produced in accordance with two different coloring schemes (trajector:yellow landmark:red and trajector:red landmark:blue).

After being given written instructions describing the procedure, but not priming for any discourse or spatial phenomena, participants were presented with a randomly ordered set of visual and linguistic stimuli pairings. For the 'in front of' linguistic stimulus, the visual stimulus could be drawn from

[1]Examples of the rendered visual scene stimuli can be viewed at http://www.speaking-systems.com/stimuli/

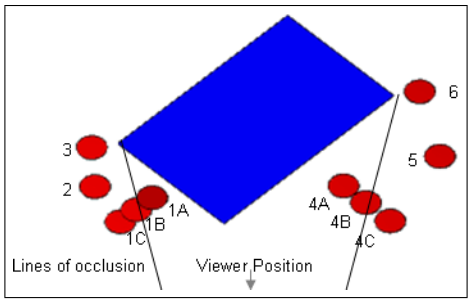

Figure 1: The configurations used in the visual stimuli.

Table 1: Mean acceptance ratings for stimuli with standard deviations: 0 = Strong disagreement. 6 = Strong agreement.

	Front		Right		Left	
Position	Mean	sd	Mean	sd	Mean	sd
1a	2.96	1.81	-	-	4.04	1.50
1b	2.93	1.74	-	-	3.87	1.65
1c	3.22	1.80	-	-	4.43	1.53
2	2.41	1.82	-	-	4.83	1.40
3	1.52	1.38	-	-	4.54	1.62
4a	4.43	1.46	3.26	1.91	-	-
4b	3.91	1.67	3.39	1.84	-	-
4c	4.28	1.68	4.26	1.73	-	-
5	2.63	1.78	4.57	1.49	-	-
6	1.48	1.41	4.85	1.28	-	-

any possible trajector position. For the 'to the right of' and 'to the left of' linguistic stimulus, the visual stimulus was limited to location sets 4, 5, 6 and 1, 2, 3 respectively. For each stimulus pairing, participants were required to indicate their level of agreement through a 7 point Likert scale. Each participant rated each applicable visual stimulus for each linguistic stimulus, and color schemes for each pairing were selected randomly from the available pairings.

Results & Analysis

Table 1 presents the mean acceptance ratings and standard deviations for each stimulus pairing. As a control test of participant understanding, we expected, and found, high ratings for trajector positions 2 and 3 for stimuli including 'to the left', and high ratings positions 5 and 6 for stimuli including 'to the right'. Also expected and found, was that participants rated the acceptability of 'in front of' higher for positions oriented with the long face of the landmark higher than the short face (compare set 1 and 4). We attribute this to the effect of the landmark's intrinsic frame of reference.

In comparison with (Logan and Sadler 1996; Carlson-Radvansky and Logan 1997; Kelleher and Costello 2005), however, there are some anomalies in our data. Most striking is the fact that the acceptability ratings do not systematically drop with angular deviation from any plausible direction vectors, nor with distance from the landmark. For example, 4A,4B,4C have the same angular deviation from both the intrinsic and the viewer-centered direction vectors. However, their acceptance ratings vary and this variation does not correlate with distance from the landmark. The same holds for 1A,1B,1C. We posit that this is caused by the participants' oblique perspective on the landmark causing a distortion in the spatial templates for the different frames of reference.

Conclusions & Future Work

In Human-Robot dialogue systems spatial template models bridge the grounding gap between spatial language and sensor data. To date, however, the effect of interlocutor perspective on spatial templates has not been systematically examined. To this end, we conducted an experiment where the landmark object was presented at an oblique angle to the participants. We interpret our results to indicate that interlocutor perspective on a landmark may distort directional spatial templates anchored on the object. While this is not in itself a surprising result, it does highlight an issue with current template-based accounts of spatial term semantics; namely, that these models focus on the geometric relationships between the trajector and the landmark and largely omit interlocutor perspective as a feature. As such they are incomplete. In future work we aim to take the computational models used by robotic systems and move them towards these more complete accounts.

References

Brenner, M.; Hawes, N.; Kelleher, J.; and Wyatt, J. 2007. Mediating between qualitative and quantitative representations for task-orientated human-robot interaction. In *Proc. of the* 20^{th} *Int. Joint Conference on Artificial Intelligence*.

Carlson-Radvansky, L., and Logan, G. 1997. The influence of reference frame selection on spatial template construction. *Journal of Memory and Language* 37:411–437.

Coventry, K. R., and Garrod, S. C. 2004. *Saying, seeing and acting. The psychological semantics of spatial prepositions*. Essays in Cognitive Psychology series. Psychology Press.

Gorniak, P., and Roy, D. 2004. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research* 21:429–470.

Kelleher, J., and Costello, F. 2005. Cognitive representations of projective prepositions. In *Proc. of the 2nd Workshop on The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications.*

Kelleher, J. D., and Costello, F. J. 2009. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics* 35(2):119–149.

Logan, G. D., and Sadler, D. D. 1996. A computational analysis of the apprehension of spatial relations. In Bloom, P.; Peterson, M. A.; Nadel, L.; and Garrett, M. F., eds., *Language and Space*. Cambride, MA: MIT Press. 493–530.

Regier, T., and Carlson, L. A. 2001. Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General* 130(2):273–298.