

# A Scene Corpus for Training and Testing Spatial Communication Systems

Michael Barclay and Antony Galton<sup>1</sup>

**Abstract.** It is argued that a ‘scene corpus’ would be a useful tool for the training and testing of systems for grounded spatial communication in the same way that text corpora have been used for training and assessing other language processing systems. Such a scene corpus would need to allow a full range of spatial relationships to be expressed over a range of scale spaces. The scenes should be sufficiently complex to allow sequential spatial descriptions to be constructed. The integration of listener models and reference frame variation will be required. The interface design will need to allow for different implementations and corpus extensions. Initial steps in the design of a scene corpus are described.

## 1 INTRODUCTION

To see that spatial communication is among the most fundamental and important forms of communication in which humans engage, consider only the sentence, “There is a lion under those trees!”. Before the emotional, political, financial, academic or other realms were invented as subjects for discourse the physical realm existed and required description and discussion. Spatial metaphors influence and structure more areas of human communication than any other [13].

Spatial communication is often multi-modal communication: the words ‘those trees’, in a typical environment, may only be meaningful if distinguished from other groups of trees by a gesture. Direction is indicated by pointing, limits on spatial location or extent may be indicated by two-handed indications (or arm sweeps) and turns or direction changes by representational hand movements.

Even the simplest of spatial phrases, containing two nouns, linked by a preposition conceals much complexity depending on the form of the objects involved [8, 10, 21]. If some functional relationship between the objects is also involved the difficulties of machine generation of spatial language become even more apparent [7, 5, 15, 9, 4]. A cup may be *usefully* described as ‘on the table’ even if it is actually ‘on’ a saucer which is ‘on’ a mat which is ‘on’ a tablecloth which is ‘on’ the table. The cup might not *usefully* be described as ‘on’ the saucer, since the saucer is as mobile as the cup and does not help a listener find the cup. In any case the cup might be ‘in’ the saucer (not on it) and the tablecloth ‘over’ the table. To form even a simple spatially locative phrase acceptably requires a broad knowledge of physical laws (in the naive sense), along with concepts of support, containment, mobility and persistence. Knowledge of how objects interact (hardness and softness, deformation, permeability etc) and how they are conventionally used also plays an important part.

This complexity has meant that the best systems to date for analysing and generating spatial language have concentrated on elements of the problem rather than on its entirety [20, 19, 11, 22, 16].

These systems have typically used their own sets of scenes, although in most cases they could have been adapted to use a standard corpus.

Currently comparison between systems and their methods and algorithms would be difficult because of this concentration on different elements of spatial language generation. Over the next few years, however, the development of more complete spatial communication systems, including those with multi-modal output, is anticipated. An agreed scene corpus which is designed to address all the elements of spatial communication will be essential.

## 2 CONTENT REQUIREMENTS FOR A SCENE CORPUS

### 2.1 Scope of the corpus

To encompass all aspects of spatial language in practice means that the scene corpus must contain examples of all expressible spatial relationships between as wide a variety of objects as possible, to enable training and testing of comprehensive spatial language systems. Given the task of describing the location of object(s) in a scene (the ‘located’ object(s)), typically with respect to one or more ‘reference’ or ‘ground’ object(s), these aspects of spatial language can be summarised as follows:

1. Selection of appropriate reference object(s)
2. Adoption of appropriate reference frame
3. Use of correct spatial prepositions
4. Incorporation of gesture, emphasis or other non-verbal communication
5. Integration of listener models
6. Strategies for construction of multi-phrase descriptions

How these are to be incorporated in the corpus is discussed in the following subsections. Note the problem being addressed is not that of referring expression generation [6] in which disambiguation is the aim and for which the ‘tuna’ corpus [24] was designed. The located object is assumed to be unambiguously identifiable but its location must be described (relative to a reference object).

### 2.2 Reference object selection

The general problem of reference object selection does not seem well addressed in the literature, although there is a specific body of work on landmark selection [2, 23, 1, 18]. Generalising and extrapolating from this work, the factors influencing reference object selection, so far identified, are as follows;

1. Reference object locatability, comprising

---

<sup>1</sup> University of Exeter, UK, email: mjb231@ex.ac.uk

- (a) Visibility
  - (b) Unambiguousness (Uniqueness)
  - (c) Persistence (if listener not present)
  - (d) User acquaintance with reference
2. Search space optimisation, comprising
- (a) Reference object location
  - (b) Reference geometry
  - (c) Scale of located and reference objects
  - (d) Listener location

The corpus must contain scenes that allow these influences to be discriminated and their weights compared. Consider for instance a scene where a car (whose location is to be described) was parked beside a row of identical houses but across the road from a bus stop. The best located, most visible and persistent landmark would be the nearest house but this being ambiguous (there are many houses) the bus stop might be the best reference (along with the preposition ‘opposite’ possibly). Cases where the house might or might not be the best reference when disambiguated by a gesture accompanying the verbal description would also be needed.

### 2.3 Reference frame selection

The choice of reference frame is crucial to the acceptability [3] and effectiveness [17] of spatial communication. Reference frames can be briefly described as;

1. Speaker-centred (deictic). ‘Left’, ‘right’, ‘in front of’, ‘behind’, etc., are relative to the speaker.
2. Absolute (extrinsic). Typically this frame will relate to a ‘previously agreed’ outer reference object, such as the earth in the case of North/South etc.
3. Object-centred (intrinsic). This reference frame may be chosen when a reference object has a distinct orientation defined either by its function (e.g., a car) or by convention (e.g., some buildings).
4. Listener-centered. This is usually equivalent to either object-centered, if the listener is the reference (as in ‘it’s in front of you’), or speaker-centered, as in a typical route description where the speaker is talking a listener through a route as though they were together.

The scene corpus will need to include objects that have functionally or conventionally defined orientations as well as objects that do not. Currently it is thought that these orientations will need to be explicitly noted as well as indicated by features on the objects in question. Scenes will also have to have a defined external reference orientation (e.g., a North pointer) and defined positions for the listener and speaker.

### 2.4 Preposition assignment

The number of English spatial prepositions is small (some 70 are listed in [10]) compared to the number of expressible spatial relationships. Although some duplication is apparent (e.g., ‘above’ and ‘over’ can be interchanged in some examples), there is even more ‘overloading’ of meaning on some prepositions (even excluding metaphorical usage) as shown by the discussion of ‘over’ in [14] or the discussion of ‘in’ in [4].

No attempt has yet been made to devise even a partially grounded, trainable system, capable of acceptable use of all of the prepositions listed in [10] and it is an open question how large a scene corpus would need to be to enable this training. Minimally the corpus should include scenes in which some objects have spatial relationships that can unambiguously be mapped onto each of the prepositions in [10]. A wide range of representations of the common geometric prepositions will inevitably be included.

### 2.5 Non-verbal communication

The corpus can and should be designed to train and assess the use of gesture to distinguish and disambiguate objects in conjunction with verbal communication (as well as simply to indicate objects, locations and ranges). More work will be required to decide how far a scene corpus can be taken in this respect. For example, if the use of intonation or emphasis to indicate the degree of belief in the location of an object is required, ‘partial’ information from a source outside the scene corpus as currently envisaged will be needed.

### 2.6 Listener location and listener models

Currently three elements of a listener model are assumed to be included in the scene corpus;

1. Whether the listener is present at the scene (important to test discernment of the relevance of persistence in a reference object)
2. The listener’s location in the scene if he is present (to test reference frame selection and preposition assignment)
3. Listener acquaintance with specific reference objects (to test the use of less visible references if their location is already known)

Aspects of a listener model such as preference and cognitive capacity (as discussed in [12]) are outside of the scope of the scene corpus.

The speaker model is currently limited to location which is coincident with the ‘openGL camera’ location for the scene.

### 2.7 Complex phrases and multi-phrase descriptions

At least three classes of complex description forms can be identified which are potentially important for a spatial communication system to be able to handle:

1. Complex locative statement. A locative phrase with more than one reference such as “The vase is in the living room, on the table under the window”
2. Path and route descriptions. These are possibly the most important for multi-modal systems. Descriptions such as “the man came from between the shops, ran along the road and disappeared down the alley by the church” are seldom unaccompanied by gestures.
3. Sequential scene descriptions. These are linked descriptive phrases such as “Behind the shops is a church, to the left of the church is the town hall. In front of the town hall is a fountain”

Strategies for sequential scene description are discussed in [12]. It would be difficult to capture all the necessary considerations and *design* a corpus to comprehensively test these behaviours. The complexity of the scenes in the corpus, however, should enable systems to generate acceptable phrases and gesture sequences of the sort outlined in the list above.

### 3 USAGE ISSUES

#### 3.1 Scene representation and partial information

The scenes in the corpus as currently envisaged will contain objects that may not be visible to an observer from a given point of view. It is not at present intended to enable the 'removal' of these objects from the scene as presented to the spatial communication generation system. Thus the information in a scene should be thought of as analogous to a 'speaker cognitive model', possibly composed from many visual images, rather than as analogous to a 'view from a point'.

How this 'cognitive model' might have been arrived at in practice or how a system would deal with partial or missing information are not addressed by the corpus at present.

#### 3.2 Bias and information sources

It was suggested above that the corpus should be designed so that it can deliberately be made to contain all of the expressible spatial relationships, along with the other requirements in section 2. This raises the potential problem that the distribution of spatial relationships expressed in the corpus will not be representative of the real world; this could lead to bias when training a spatial communication system. It is uncertain whether any series of real world images, selected by individuals, could in practice be less skewed, however, or indeed whether a child would learn spatial language in a 'representative' environment.

Any effects of this and possible remedies will need to be the subject of further work.

### 4 DESIGN AND IMPLEMENTATION ISSUES

#### 4.1 Dimension and representation

Many of the systems for spatial language generation have used 2-dimensional images, often represented as bit-maps. It is clear from the requirements however that a 3-dimensional representation will be required and this will be beyond the practical limits of bit-mapping.

A vertex-list representation that is OpenGL compatible has been adopted at this point. Although it is not entirely optimal, in that it is difficult to avoid line-segment duplication during analysis, it combines ease of visualisation with reasonable simplicity of analysis.

Note this does not preclude the use of 2-dimensional scenes. The representation of maps, in particular, might be a useful addition to the corpus.

Animation of scenes is also required to allow proper mapping on to motion prepositions such as 'through' and 'towards'.

The objects in the scenes as currently envisaged are defined as solid regions of arbitrary complexity immersed in a medium (assumed to be 'air'). They may be convex or concave and may entirely enclose regions of the medium. Care must be taken as currently spaces or parts of the medium cannot be named. If a ball is to go "through a window" a window must be provided, not simply a gap in a wall.

Typical objects in a scene are constructed from primitives which can be labelled. So although geometrically a table may be treated as a single object a phrase such as "the ball rolled between the table legs" could be correctly constructed from the information provided.

Typical scenes from the initial corpus, with example description strings, are shown in figures 1 and 2.

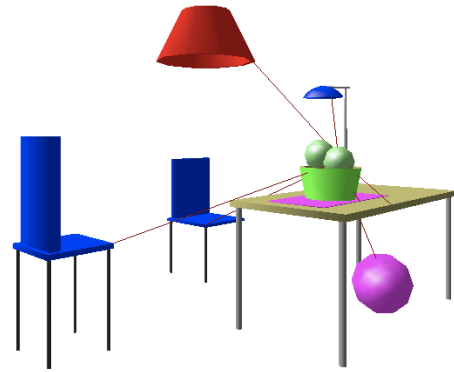


Figure 1. A table-top scene "the apples are on the table"

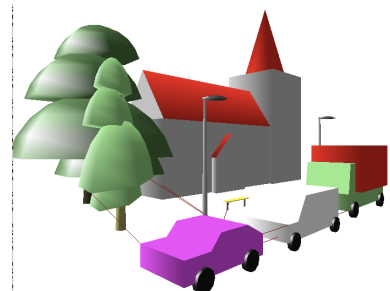


Figure 2. A street-scale scene "the car is in front of the church"

#### 4.2 Corpus interfaces

The key interfaces between the corpus and associated systems are shown in figure 3. A detailed description of the information presented at these interfaces is not possible here but in summary the XML file defining a scene contains the following:

1. The object list
2. Animation vectors
3. Description strings
4. OpenGL drawing information

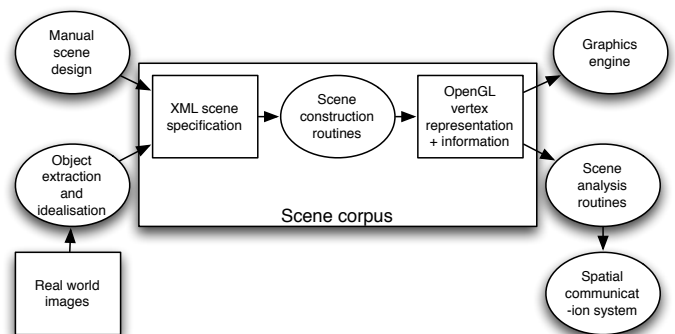


Figure 3. Scene corpus interfaces

The description strings contain acceptable spatial communications relating to the scene. These are in English at present but there is no reason why other languages and non-verbal communications could not be incorporated as well. The XML files could be generated from 'real' images by image processing systems, employing object recognition and vector extraction, to provide extensions to the corpus as initially designed.

The constructed scene passed to the analysis routine contains the OpenGL representation of the object list (as a sequence of animated frames if required) and the description strings.

The designed section of the corpus will be more or less manually constructed.

### 4.3 Scale space coverage

To ensure coverage of a full range of spatial relationships and mappings on to prepositions in particular, a range of environments needs to be represented in the corpus. The word 'beyond', for instance, is used more frequently in large scale environments than within rooms.

As currently envisaged the corpus will include scenes at 'table-top', 'room' 'building' and 'street' scales. Some scenes will need cross-scale representations for reference object selection testing in particular. Extensions to what might be termed 'landscape' scale might be required.

### 4.4 Complexity and computational load

Construction of the vertex representation of the scenes is relatively trivial, analysing the geometry and topology of the scene is more time-consuming. In the current implementation the complexity of analysis for a scene containing  $n$  objects each with  $m$  facets is:  $O(m^2n^2)$ . The most time-consuming aspects of the analysis are creation of a qualitative spatial relation matrix and calculation of closest approach vectors for all objects. Analysing a scene of 8 spheres each composed of 180 facets takes about 0.5 seconds on a 'standard' PC. A 'maximum complexity' scene of 50 objects of 180 facets each would take about 20 seconds and thus a corpus of 1000 of these scenes would require 5.5 hours (without animation). It is thought that this could be improved by a factor of 4 with more adept pruning and a further factor of 4 by attention to the algorithms.

The time taken for the spatial communication system to be trained or to produce the required descriptions is clearly system specific and not included here.

### 4.5 Corpus size

Considering the potentially diverse nature of systems to be tested it is not entirely glib to say the corpus should be as large as possible. With 70 prepositions, 4 reference frames and 8 complex factors influencing reference choice it is probable that a corpus of less than 1000 scenes would be inadequate even though many relationships can be expressed in a single scene.

## 5 NEXT STEPS

A corpus for the current research task will be constructed along the lines described but in parallel with this, input from other interested researchers, with a view to constructing a generally useful corpus, would be welcomed.

## REFERENCES

- [1] G. E. Burnett, D. Smith, and A. J. May, 'Supporting the navigation task: characteristics of good landmarks', in *Proceedings of the Annual Conference of the Ergonomics Society*. Taylor and Francis, (2001).
- [2] G. E. Burnett, 'Turn right at the King's Head'. *Drivers' requirements for route guidance information.*, Ph.D. dissertation, Loughborough University, 1998.
- [3] L. A. Carlson-Radvansky and Logan G. D., 'The influence of reference frame selection on spatial template construction.', *Journal of Memory and Language*, **37**, 411–437, (1997).
- [4] K.R. Coventry, 'Spatial prepositions, functional relations, and lexical specification', in *Representation and Processing of Spatial Expressions*, eds., Patrick Olivier and Klaus-Peter Gapp, 1–35, Laurence Earlbaum Associates, (1998).
- [5] K. R. Coventry, M. Prat-Sala, and L Richards, 'The interplay between geometry and function in the comprehension of over, under, above, and below.', *Journal of Memory and Language*, **44**(3), 376–398, (2001).
- [6] R. Dale and E. Reiter, 'Computational interpretations of the gricean maxims in the generation of referring expressions.', *Cognitive Science*, **19**, 233–263, (1995).
- [7] M. I. Feist and D. Gentner, 'Factors involved in the use of in and on', in *Proceedings of the Twenty-fifth Annual Meeting of the Cognitive Science Society.*, (2003).
- [8] K.P. Gapp, 'An empirically validated model for computing spatial relations', *Künstliche Intelligenz*, 245–256, (1995).
- [9] S. Garrod, G. Ferrier, and S. Campbell, 'In and on: investigating the functional geometry of spatial prepositions', *Cognition*, **72**, 167–189, (1999).
- [10] A Herskovits, 'Schematization', in *Representation and Processing of Spatial Expressions*, eds., Patrick Olivier and Klaus-Peter Gapp, 149–162, Laurence Earlbaum Associates, (1998).
- [11] Tanja Jording and Ipke Wachsmuth., 'An anthropomorphic agent for the use of spatial language.', in *Spatial Language. Cognitive and Computational Perspectives*, eds., K. R. Coventry and P. Olivier, 69–85, Dordrecht: Kluwer Academic Publishers, (2002).
- [12] R. Klabunde and R. Porzel, 'Tailoring spatial descriptions to the addressee: a constraint-based approach.', *Linguistics*, **36**(3), 551–577, (1998).
- [13] G. Lakoff and Johnson M., *Metaphors we live by*, University of Chicago Press, 1980.
- [14] G. Lakoff, *Women, Fire and Dangerous Things*, University of Chicago Press, 1987.
- [15] K. Lockwood, K. Forbus, D. Halstead, and J. Usher, 'Automatic categorization of spatial prepositions', *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, (2006).
- [16] K. Lockwood, K. Forbus, and J. Usher, 'Spacecase: A model of spatial preposition use', in *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, (2005).
- [17] S. D. Mainwaring, B. Tversky, Mohoto Ohgishy, and D. J. Schiano, 'Descriptions of simple spatial scenes in english and japanese', *Spatial Cognition and Computation*, **3**(1), 3–43, (2003).
- [18] C. Nothegger, S. Winter, and M Raubal, 'Computation of the salience of features.', *Spatial Cognition and Computation*, **4**, 113–136, (2004).
- [19] R. Porzel, M. Jansche, and R Klabunde, 'The generation of spatial descriptions from a cognitive point of view', in *Spatial Language. Cognitive and Computational Perspectives*, eds., K. R. Coventry and P. Olivier, 185–207, Dordrecht: Kluwer Academic Publishers, (2002).
- [20] Terry Regier, *The human semantic potential: Spatial language and constrained connectionism.*, MIT Press, 1996.
- [21] T Regier and L. Carlson, 'Grounding spatial language in perception: An empirical and computational investigation.', *Journal of Experimental Psychology: General*, **130**(2), 273–298, (2001).
- [22] D. K. Roy, 'Learning visually-grounded words and syntax for a scene description task', *Computer Speech and Language*, **16**(3), (2002).
- [23] M. Sorrows and S. Hirtle, 'The nature of landmarks for real and electronic spaces', in *Spatial Information Theory: Cognitive and Computational Foundations of GIS*, eds., C. Freska and D. Mark, Springer-Verlag, (1999).
- [24] K. van Deemter, I. van der Sluis, and A. Gatt, 'Building a semantically transparent corpus for the generation of referring expressions.', (2006).