Strategies in the Use of Referring Expressions to Describe Things Urban

William Mackaness[1], Phil Bartie[2] and Philipp Petrenz[1]

[1]The University of Edinburgh, School of Geosciences
[2]University of Stirling, Biological and Environmental Sciences

Summary

In the context of wayfinding technologies, there is increasing interest in dialogue based systems that use description of landmarks as a way of guiding people through cities. In the absence of maps and photographs, the challenge for automated systems is the production of descriptions of things in the field of view that are unambiguous and easily interpreted. We are therefore interested in the mechanisms used by humans to create and interpret descriptions of things in the urban vista. Here we report on a web based experiment in which we explored the veracity of human generated referring expressions in order to better understand the most successful strategies for directing people's gaze.

KEYWORDS: psycholinguistics, referring expressions, surface realisation, wayfinding, urban

1. Psycholinguistics and Referring Expressions

Landmarks are one aspect of the environment frequently referenced, as they assist in forming mental representations of space (Hirtle and Heidorn 1993, Tversky 1993), and in way-finding tasks (Werner *et al.* 1997, Lovelace *et al.* 1999, Caduff and Timpf 2008, Winter *et al.* 2008, Duckham *et al.* 2010). Landmarks are defined as identifiable features in an environment, whose saliency may be calculated by comparing scores for particular attributes (e.g. their size) and identifying those which deviate from the mean (Raubal and Winter 2002, Elias 2003a, Elias and Brenner 2004). These are the buildings unlikely to be confused with others, either which appear very different to their surroundings (e.g. churches) or are well known major international brands (e.g. Starbucks). The focus of this paper is not on modelling landmarks. Instead its focus is on i) determining what governs choice of the characteristics that allow landmark identification in 'vista space' (Montello 1993), and 2) how those characteristics are formed into a string of words that constitute a referring expression – a process called 'surface realization' (Jurafsky and Martin 2008).

1.1. Referring Expressions and common ground

Referring expressions (RE) can be 1) optimally informative (e.g. 'the small apple' – Figure 1a), 2) under-informative ('the apple') or 3) over/hyper- informative ('the small green apple'). There are various reasons why subjects, when asked to write referring expressions, might create these different forms. They may fail to undertake a full visual scan of the image and fail to see a need to further differentiate. But probably more important than this, the creation of a referring expression depends on a shared conceptualisation between the subject describing the object, and the viewer who interprets and is thus able to locate the object in the scene; this is referred to as 'common ground' (Horton and Keysar 1996). This is very pertinent in the context of the urban (Figure 1b) where reality can be conceptualised (and so described) at very different levels of granularity. Where there is thought to be little common ground or uncertainty between the subject and viewer, we might expect the referring expression to be hyper informative, and therefore contain information that may essentially be redundant.

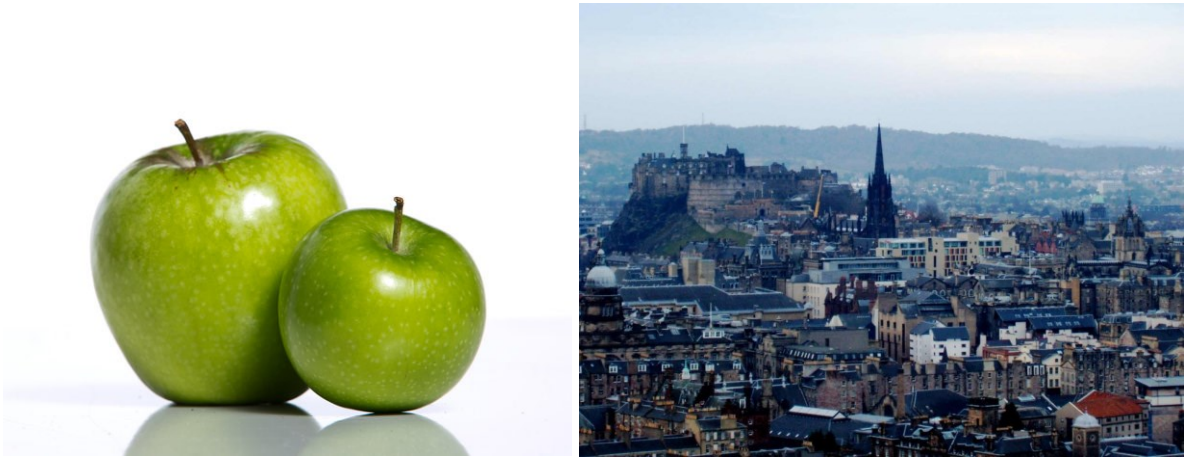


Figure 1: Simple and complex worlds: a) two apples and b) an urban vista

1.2. Influence of Vista, Relatum and Distractors on Referring Expression Generation

The context of this research is the growing interest in dialogue-only based interaction in which the user is both hands free and eyes free to explore the environment as they move through it (Bartie and Mackaness 2006; Mackaness et al. 2014). In a context of only having spoken text as the description, the question arises: 'how best do we describe urban objects that are in the field of view?'. Too brief and we risk uncertainty; too verbose and the cognitive effort is unnecessarily high. To that end we draw inspiration from Grice who wrote 'make your contribution such as it is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange' (Grice 1975) and from which we can apply the maxims of *quantity, quality, relation and manner* as useful frameworks by which to assess 'the perfect description'.

The task of referring to something is easy if its uniqueness is readily apparent; 'distractor' is the term given to objects of the same class or form, and their presence requires disambiguation (eg the need to distinguish between the two apples in Figure 1a). We see in Figure 1b) there is a large number of distractors. What constitutes a distractor in this urban vista is complex; buildings are of similar shape, colour and size (eg the large beige buildings), or they could be deemed as distractors because they perform a similar function (eg churches). We might therefore envisage that longer RE descriptions are required to disambiguate.

Figure 1b also makes clear that some objects are far more salient than others – unambiguous in their form (eg 'the castle'). It is common to use such features as an anchor by which we might describe the (less salient) object, and thus differentiate it from its distractors (eg 'The church immediately right of the castle'). Here the castle acts as a *relatum.* We note that the relatum is acting to reduce the area of search in the scene (the solution space) – provided of course, that the subject and viewer have the same conceptual understanding of castle (common ground). We also note that the larger the vista, the greater the likelihood of distractors. From perusal of the literature in psycholinguistics (the processes by which we understand utterances) (Aitchison 2011) and work in qualitative spatial reasoning (Freska 1991) we can identify the various ways in which referring expressions can be combined both to reduce the solution space, and differentiate between distractors (Table 1). One can readily imagine a large permutation in these choices in order to direct someone's gaze.

Table 1 Ways of reducing the solution space or referring to objects in an urban scene

Form of reference	**Ambition**	**Example**
Absolute distance	Reducing search	In the *far distance*…
Egocentric description (self to object)	Reducing search	To your *left* you will see..
Allocentric description (object to object)	Reducing search	The library is *to the right of* the museum
Colour, colour tone, texture, size	Description	The *large* house with the *dark blue* door…
Architectural age	Description	..the *Victorian* looking house
Cardinality	Description	The *south* facing windows…
Landmark brand	Description	..between *MacDonalds* and *Subway*
Landmark type	Description	The *library*…, the *church*…
Via relatum	Description	Two doors down from the *Fire station*….
Relatum type <network, region>	Description	On the right side of the *river*…, beyond the *park*…
Composition by form	Description	..comprising *steps* leading to a *set of columns*
Composition by shape	Description	..large *block* with a *pointy* top
(superlative) adjectives, proper nouns	Description	…the *taller* of the two *grand* towers on the Houses of Parliament
Topological, container descriptions	Description	..immediately *next to* the pub, which is *in* the park

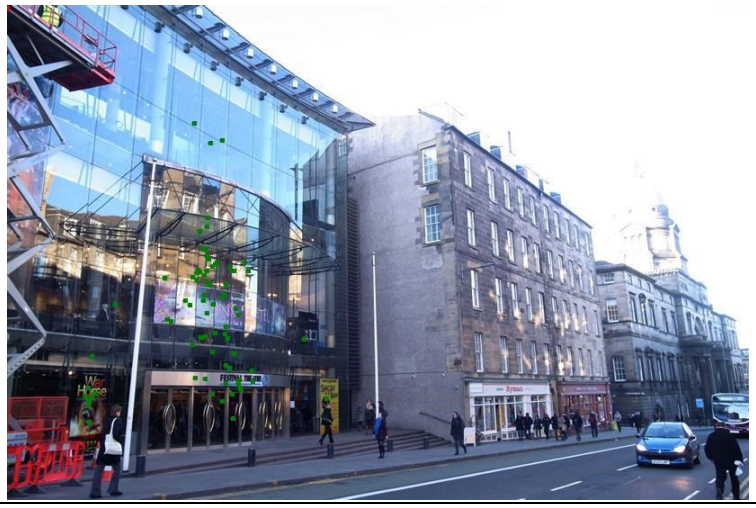
2. The Experiment

A web based internet based experiment was set up to identify the most common strategies and types of natural language phrases used to describe the location of an urban feature. We wished to understand 1) how subjects dealt with distractors, 2) reduced the solution space, 3) whether there were any patterns in the ordering or choice of variables, 4) what are the implications in the design of databases necessary to support to automatic generation of RE.

Subjects were asked to write descriptions for each of five images chosen randomly from a choice of 32 images, which varied widely in vista. A press of a button temporarily placed an outline over the building in question. Subjects were asked to provide a text based description sufficient for another person to be able to identify the highlighted feature. In a second phase of the experiment, subjects were presented with a second set of random images together with textual descriptions (provided via previous subjects). Based on the description, subjects were asked to identify the feature (by clicking on the object in the image). Subjects could record if they were uncertain or the description was ambiguous. The images with descriptors were presented to at least three viewers. This provided a means of assessing the veracity of a description. The experiment was first promoted via Crowdflower but with mixed results. Greater success came from promotion at conference, via Facebook and to subjects involved in a previous experiment (Mackaness et al. 2014). Subjects were incentivised by an opportunity to win Amazon vouchers. The number of participants grew quickly and within a month close to 200 participants provided a total of 800 annotations distributed over the 32 images.

3. Results

Figure 2 is an example of an image, with a list beneath of a subset of the descriptions that successfully led viewers to identify the target (shown by the green dots).



Large, modern glass fronted building, butted up against traditional Victorian terrace, slightly set back from road, and with facing bowed frontage.
The target is the Festival Theatre on North Bridge. It is a large glass fronted modern building slightly set-back from the road.
Just look at the first building from the left, the one with really big and nice glass walls.
A large modern building with a totally glass front.
A rather square-shaped, glass walled building, which shows no resemblance to its environment due to its complete different, modern style.
Festival Theatre. The glass-fronted building with obvious posters advertising shows.
Festival Theatre - large glass fronted building with theatre posters in the windows. To the left of Rymans, as you're looking at it.

Figure 2: Image with a subset of successful descriptors illustrating the breadth of techniques used.

We observe that most of the descriptors are hyper informative (despite the absence of distractors). For this reason, and contrary to expectations, across the images we could not discern a relationship between vista and length of descriptor. In the case of Figure 2, we surmise that subjects might feel this object is not prototypically 'theatre' and this has led to hyper informative descriptions. The issue of granularity is very apparent - subjects utilise detail where it is discernable. Where the object reveals less detail (eg Figure 3), the subject is forced to use alternate strategies.

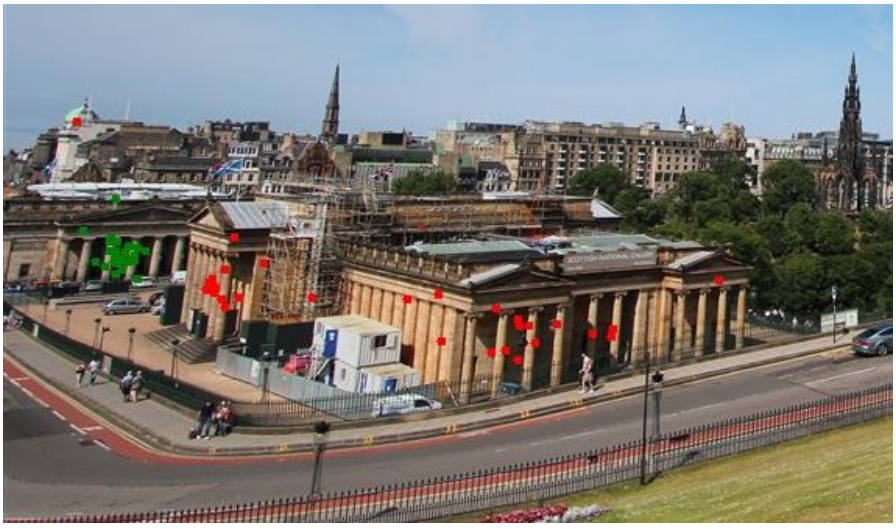


Figure 3: Cropped image of an urban vista showing viewers who selected the object (green) or distractors (red dots)

In Figure 3 the target is a grand building with columns and two distractors. Among the very few successful descriptors was one that anticipated the confusion and likely fixation upon the prominent

central building. Their solution was to direct the gaze away from the distractor ('Not the first building with the stone pillars but the one behind it'), the other was to use the distractor as a relatum ('the second building with the columns, the farthest away one').

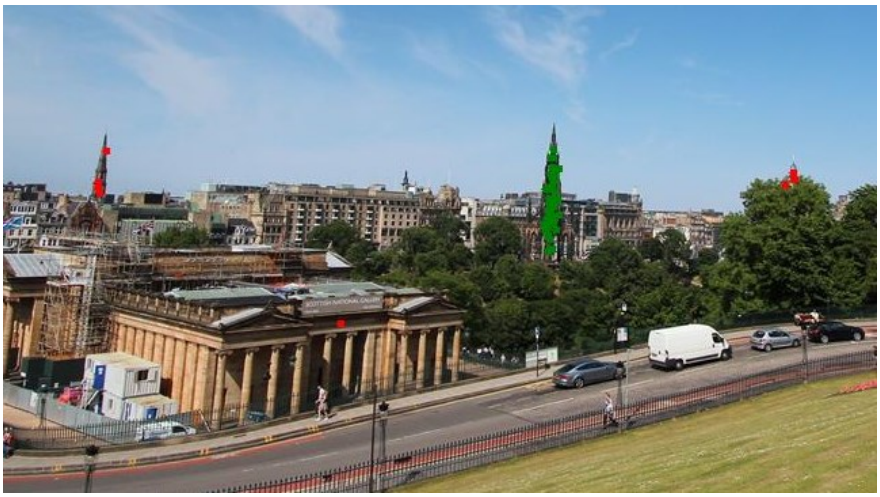


Figure 4: A monument that is 'cathedral like'.

Figure 4 is interesting because it has a large solution space and a number of distractors. Of the successful descriptors, 60% used superlatives (dominating, taller, enourmous, emblematic). Optimally 'monument' proved sufficient. 20% used their knowledge of the city to name the monument. 60% used the trees as a container relatum (eg 'the spire among the trees'). The term 'spire' was used by 50% of the subjects – but ironically this *created* distractors since there are other spires in the image, (whereas it is the only *monument*). This required additional disambiguation and a lengthier descriptor.

3.1 Failed Descriptors

A review was also made of descriptions which were deficient. Some were ascribed to poor English, were puerile or directed the viewer to the wrong target. Some used terms that lay outside common ground (eg 'mansard roof with dormer', 'triangular pediment', 'neo-classical'). Some used names requiring local knowledge (eg 'Story telling Centre'), and others linked the description to events or people (eg 'where J K Rowling writes'). Most frequent were instances where the subject failed to discriminate between the target and other discriminators. In some cases the relatum had distractors which meant the viewer searched in the wrong part of the image (eg 'second block to the right of the church', where the image contained two churches).

4. Conclusion

This research contributes to research in generating referring expressions (GRE). The web based experiments did not reveal a gender bias, nor a preferred scan direction (vertical or horizontal), nor a relationship between vista (complexity) and length of description. It did reveal how a mix of strategies were used to reduce the solution space, and direct the user's gaze.

Some might argue that the experiment is of poor design given its multi variate complexity. We would counter that the simple scene analysis typical of psycholinguistic experiments (eg Baltaretu et al 2013; Hanna et al 2003) has little relevance to the generation of RE in urban vistas. We argue that the experiment shows that a critical step prior to surface realisation is scene analysis in which both relatum and distractors are identified. Applying Grice's norm depends on 1) a deep understanding of the context (objects in the vista), and 2) an understanding of the level of detail discernible to the naked eye. Identifying appropriate relatum requires assessment of its saliency (since relatum may have their own distractors!) Determining whether something is a distractor is fraught with difficulty, since the granularity of the feature will govern the viewer's perception of whether it is a distractor or not. We would argue that this work is of use as a corpus to the psycholinguistic community.

The implications for a database that supports automatic generation of referring expressions is intriguing. It would appear that multi scale (or multi resolution) representations (Burghardt et al. 2014) are required of each potential candidate object in order to cope with both 'close up' RE and 'vista based' RE, as well as RE in between!

Acknowledgements
We are very grateful for EU funding FP7/2007-13 via the SpaceBook project (No 270019). Our thanks too to the participants in the web based experiments.

Biography
William Mackaness is a senior lecturer at The University of Edinburgh in the School of GeoSciences. Phil Bartie is a lecturer in the Biological and Environmental Sciences at University of Stirling.

References

Abella A, Kender J.R. (1999) From images to sentences via spatial relations. Proceedings of the Integration of Speech and Image Understanding.

Aitchison, J (2011) The Articulate Mamman: An Introduction to Psycholinguistics. Routledge, London.

Bartie, P. and Mackaness, W.A. (2006) Development of a speech-based augmented reality system to support exploration of cityscape. *Transactions in GIS* 10: 63-86

Baltaretu, A.A. Krahmer E.J. Maes, A. (2013) Factors influencing the choice of relatum in referring expressions generation: animacy vs. position. *Proceedings of the CogSci workshop on the production of referring expressions PRE-CogSci 2013*.

Burghardt, D. Duchêne, C. and Mackaness, W.A. 2014 Abstracting Geographic Information in a Data Rich World Methodologies and Applications of Map Generalisation. Springer.

Caduff D, Timpf S (2008) On the assessment of landmark salience for human navigation. Cognitive Processing 9: 249-267

Duckham M, Winter S, Robinson M (2010) Including landmarks in routing instructions. Journal of Location-Based Services 4 28-52

Elias B, Brenner C (2004) Automatic generation and application of landmarks in navigation data sets. IN Fisher, PF (Ed.) Developments in Spatial Data Handling. Springer, Berlin

Elsner, M., Rohde, H. Clarke, A.D.F. (2014) Information Structure Prediction for Visual-world Referring Expressions. *EACL2014* April 26-30th Gothenburg, Sweden. http://aclweb.org/anthology/E/E14/E14-1055.pdf

Freska, C. (1991) Qualitative Spatial Reasoning in *Cognitive and Linguistic Aspects of Geographic Space*, D.M. Mark & A.U. Frank (eds.),361-372.

Grice, P (1975). "Logic and conversation". In Cole, P.; Morgan, J. *Syntax and semantics*. 3: Speech acts. New York: Academic Press. pp. 41–58.

Hanna,J.E., Tanenhaus, M.K. and Trueswell J.C. 2003 The effects of common ground and perspective on domains of referential interpretation *Journal of Memory and Language* 49: 43-61

Hirtle SC, Heidorn PB (1993) The structure of cognitive maps: Representations and processes. Behavior and Environment: Psychological and Geographical Approaches: 170-192

Horton, W.S., and Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, 59, 91-117.

Jackendoff, R. (1992) Languages of the Mind, Cambridge, MA: MIT Press.

Jurafsky, D. and Martin J.H. (2008) Speech and Language processing, Prentice Hall.

Lovelace K.L., Hegarty M., Montello D.R. (1999) Elements of good route directions in familiar and unfamiliar environments. IN Freksa, C, Mark, D (Eds.) Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science. Springer Berlin / Heidelberg.

Mackaness, W.A. Bartie, P. Dalmas, T., Janarthanam, S., Lemon, O.,Liu X., Webber B., (2014) Talk the Walk and Walk the talk: Design, Implementation and Evaluation of a Spoken Dialogue System for Route Following and City Learning, *Annual Conference of the Association of American Geographers*, Tampa Florida, 7-13 April.

Mackaness, W.A. Bartie, P. and Sanchez-Rodilla Espeso, C. (2014) Understanding Information Requirements in 'Text only' Pedestrian Wayfinding Systems, *GIScience 2014*, Vienna.

Montello D (1993) Scale and multiple psychologies of space. Spatial Information Theory A Theoretical Basis for GIS: 312-321

Moratz, R. and Tenbrink, T. (2006) Spatial Reference in Linguistic human-robot interaction: iterative, empirically supported development of a model of projective relations *Spatial Cognition and Computation 6(1), 63-107.*

Raubal M, and Winter S (2002) Enriching wayfinding instructions with local landmarks IN Egenhofer, MJ, Mark, DM (Eds.) Second International Conference GIScience. Springer, Boulder, USA

Richter, K-F and Winter S. (2014) Landmarks: GIScience for Intelligent Services

Tversky B (1993) Cognitive maps, cognitive collages, and spatial mental models. IN Frank, AU, Campari, I (Eds.) Spatial Information Theory: A Theoretical Basis for GIS. Italy, Springer-Verlag

Werner S, Krieg-Bruckner B, Mallot HA, Schweizer K, Freksa C (1997) Spatial cognition: The role of landmark, route, and survey knowledge in human and robot navigation. IN Jarke, M (Ed.) Informatik '97 GI Jahrestagung. Berlin, Heidelberg, New York. Springer

Winter S, Tomko M, Elias B, Sester M (2008) Landmark hierarchies in context. Environment and Planning B: Planning and Design 35: 381 – 398