Place as Location Categories: Learning from Language

Clare Davies[1] and Thora Tenbrink[2]

1. Dept. of Psychology, University of Winchester, Winchester SO22 4NR, UK. clare.davies@winchester.ac.uk
2. School of Linguistics and English Language, Bangor University, Bangor LL57 2DG, UK. t.tenbrink@bangor.ac.uk

Abstract How do people refer to places in their environment, and to what extent do the underlying spatial concepts correspond to officially defined regions? We exemplify some types of evidence that may help to determine local vernacular place concepts. The output of LSA on a web-scraped text corpus was compared with mapping and linguistic data from a pilot experiment, to see how localities within the same geographic area tended to be clustered, how far the spatial geography is similarly distorted, and how far participants' verbal protocols revealed a tendency to group places together (and how). Finally, we list some challenges for future triangulation of such data sources, in deriving vernacular place data.

1. Introduction

The human tendency to name and store knowledge of places that are regions, i.e., larger than a single landmark or point, may be considered as a cognitively efficient means of categorizing known locations in space. This is true both at the scale of single point locations experienced through wayfinding, and at higher hierarchical levels such as grouping localities within cities, and thence into larger geographic regions. Categorization allows assumptions, reasoning and linguistic references to be applied across a range of category members all at once, saving effort in processing and communication [1].

Officially (i.e., administratively) defined places tend to have rigid, non-overlapping boundaries, with a degree of hierarchy. However, the usage of place names and place-based reasoning by their human inhabitants, even within relatively modern and planned New World cities, is much messier. The key challenge for so-called 'vernacular geography' is to capture people's fuzzy, context-changeable and possibly spatially distorted understanding of such places, so that the typically intentional extent of a given toponym can be modeled, and so that locations or lo-

[1] (Author for correspondence)

calities which are seen by local people to 'belong' together (or not) can be identified even where their groupings lack a recognized toponym.

In this paper, we discuss ways of addressing this challenge through analyzing natural language data mentioning local places. After defining the key aspects of human cognitive categories and applying them to vernacular places, two relevant data sets will be compared: data from a pilot study in which participants were recorded talking through a mapping task for localities in their home area, and the outputs of LSA performed on a web-scraped text corpus mentioning the same set of places. Ideally, in language data, place-as-category information may be extractable not only from the co-occurrences of place names (toponyms) within sentences, but also from additional verbal cues implying clustering of localities (villages and suburbs) into groupings which may, but may not always, have a collective toponym applied to them.

The category information implied by such verbal data can also be triangulated with that from further sources, and synthesized into a composite model of the common cognitive groupings of localities in a given area. In the future, learning 'classifier' algorithms may be employable to 'learn' the local vernacular geography from a variety of sources, as available. First, however, we need to improve our understanding of how and when different phrases can be taken as indicating spatial 'place' categorization - and when they cannot.

2. Places as Categories of Locations

In what sense can places be viewed as locational categories? Arguably, any place larger than a single point must contain a collection of individual locations - be they navigational landmarks and intersections, or a string of villages along the shore of a large water body. Montello [2] argued that cognitively, regions reflect the general human tendency to organize knowledge categorically, trying to minimize within-category variation and maximize between-category differences – often to the point of stereotyping or over-generalization. As Montello pointed out, this tendency apparently aids cognitive efficiency and avoids spurious precision in our assumptions and speech.

However, half a century of research into categorical cognition has gone way beyond this general observation. Categories in human cognition have a range of well-established properties, emerging from several decades of research, which in turn have specific implications for our understanding of place. Such properties include fuzziness [3], graded membership (often depending on ‘ideal’ comparison rather than ‘typicality’ – see [4] and [5]), and classification using characteristic but not necessarily defining (necessary and sufficient) features. These kinds of properties may help us to identify suitable machine-learning classifiers for building place knowledge based on human-sourced data, since they would in this case need to be cognitively plausible rather than factually accurate.

A few studies have shown evidence that categorical thinking about locations within places can also be influenced in similar ways to experiments on semantic categorization (e.g., [5]). Meanwhile, a body of work mainly focused at what Montello [6] defined as figural-scale spaces [7], and figural-scale representations of geographic spaces [8, 9] has demonstrated what might be thought of as 'category errors' in spatial memory and thinking. Both in adults and children, locations and shapes of dots, lines and geometric forms tend to be mentally simplified and distorted to more regular or distinctly clustered patterns [7,8]. Similarly, we see similar spatial distortions in mental representations of geographic locations: individual items may be clustered together more distinctively, along straighter lines and with broader separations between clusters, than in physical reality [10,11].

If this is how people remember and reason about locations – such that their grouping into places distorts the space into one akin to the more semantic 'mapping' of categories in non-spatial domains – then it would be useful to know whether we might see the same patterns and tendencies show up in different sources of (vernacular) data about the same set of places. For a given geographic area, we should expect that the locational place category memberships obtained from linguistic data will indicate similar patterns to the results of an experimental mapping task, and to corpus data scraped from the web or social media and reduced to a 'semantic map'. We may later be able to triangulate such information sources together [12], to build predictive models of place grouping and toponym referents. Furthermore, if spatial regions can be viewed as a special case of general cognitive categorizing, then we may be able to apply to models of 'place' many of the findings and models from half a century of cognitive science research on, and models of, semantic memory and reasoning.

The next section will compare such example outcomes from two very different data sources: one set re-analyzed from a previous project in the Southampton area of southern England, and another collected as a pilot human-subjects experiment.

3. Comparing sample data sources: pilot evidence

3.1 Web-sourced text corpus

The example data shown in this section was previously extracted and analyzed as part of a project presented at COSIT'13 [13]. The project, as previously reported, was attempting to replicate work by Max Louwerse [14], demonstrating that the geographic positions of named places could be replicated via latent semantic analysis (LSA, [15][2]) on a corpus of non-georeferenced online text that mentioned

[2] LSA was adopted here because of Louwerse's previous success, and because other methods tend to presume groupings of points exist from the start; we did not.

them. The text consisted of web pages 'scraped' from the internet via a corpus builder, and the pages were not selected as 'spatial' descriptions at all. They merely had to mention at least one of the identified locality names together with 'Hampshire' (the enclosing administrative county – to minimize false positives from identical toponyms elsewhere), and within the context of continuous text rather than a list.

Following Louwerse's method as far as he had specified, the corpus was processed using an implementation of LSA, and the resulting matrix of associations among the toponyms was subjected to MDS to reduce the dimensionality to two. The final 'map' of the toponyms was then geometrically transformed using an affine transformation, and the result compared to the true geographic locations, using Tobler's bidimensional regression technique [16]. The whole process was run iteratively, eventually optimizing the final r^2 value in the bidimensional regression to around 0.8. Therefore, the final map derived from the original webpage corpus was as close to overall topographic accuracy as possible.

As discussed in [13] and illustrated in Figure 1, this 'optimized' map still showed certain patterns of distortion compared to the localities' true locations:

1. A preservation, but also a broadening, of the central geographic divide that exists within this area due to Southampton Water – an unbridged sea inlet approximately 2-3km wide (see Figure 1).
2. Exaggerated clustering of geographically close localities, which in reality are much more evenly distributed through the space. Thus neighboring localities were moved closer together, and further away from other clusters.
3. Some degree of 'cardinalizing' of spatial directions: Southampton Water seemed rotated to a north-south axis, and localities along its irregular shoreline were arranged in straighter linear configurations than in reality.

The question we now ask about these distortions is whether they might correspond to a genuine cognitive geography, in the minds of local people such as those who (usually) wrote the online texts which were scraped for the corpus. Otherwise, the above could simply reflect limitations and flaws in the computational analysis method, or the lack of genuine spatial locational information inherent in the online texts themselves. Thus for present purposes, this data serves as a basis for comparison with the more directly human-sourced pilot data presented below.

3.2 Pilot empirical data: dot-placement task

To begin to 'ground-truth' the above data, by comparing it to individual local human perceptions of the geography of the same area, a pilot empirical study has been performed. This involved a small field experiment in which, after briefing, participants had to arrange and place labeled foam counters onto a blank sheet of white card, cut to match the proportions of the area in question. Each counter bore

a number, and had a label attached giving the name of the locality it represented. Participants were encouraged to talk aloud during the task, to provide a verbal account of their decision processes while placing the counters. When they declared themselves finished (without placing any completely unknown names, as guessing was discouraged), the 'map' was photographed.

To compare the relative locational patterns, the pilot data from eight participants was transformed to the same scale as the above LSA data and the 'true' coordinates of the set of localities. Figure 1 combines the true locations of each locality (hollow circles) with the mean coordinates of each locality's placement (black diamonds) in the mapping task, across the eight pilot participants (all long-term residents of, and tested in, Hythe – a village shown near the center of the map), and with the LSA-derived locations ('+' symbols) discussed earlier. For simplicity, cartographic detail is excluded apart from the approximate shoreline of Southampton Water. The lack of points across the center-left and towards bottom right reflect part of the New Forest National Park and Southampton Water, respectively.

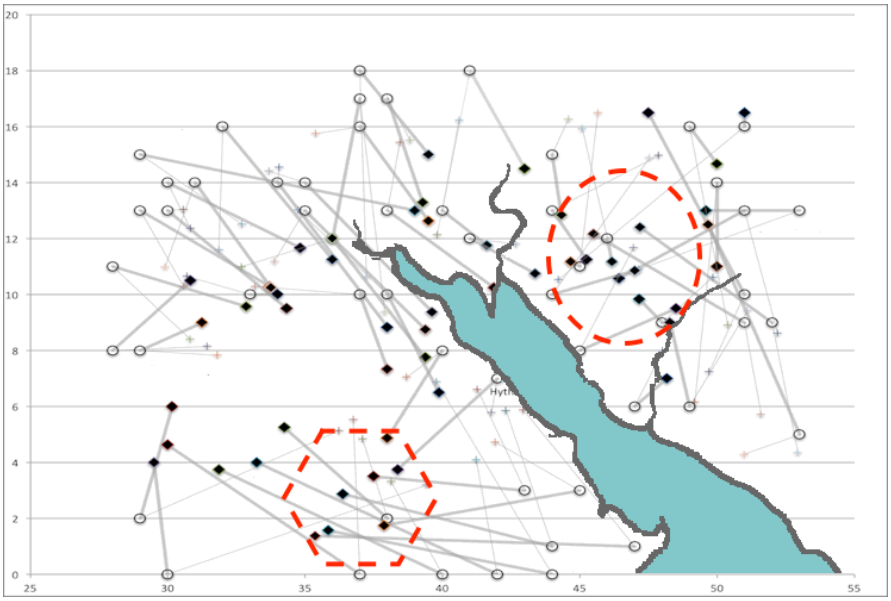

Fig. 1. Configuration of Southampton area localities: true locations (open circles), mean placement by pilot study participants (black diamonds), and coordinates extracted from LSA on web-sourced text corpus (fainter + signs).

It will again be noted that relative to the original distribution of localities (the circles), both the task- and LSA-based locations are more closely clustered. For discussion purposes, two of the task-based clusters are tentatively outlined with red dashes in Figure 1, representing:

1. Hexagon: a string of seven settlements (including Hythe) on or near the western shore of Southampton Water, known locally (though not in any map or gazetteer) as 'the Waterside'. (Note that the fainter lines in Figure 1 also show the same group of places clustering at a slightly different location in the LSA-based map.)
2. Oval: a similarly tight clustering of eleven eastern Southampton suburbs on the other side of Southampton Water.

The second cluster is also close to the placement of the city of Southampton itself, and to its western suburbs of Shirley and Millbrook – in other words, the city's urban geography (less well known to Hythe residents) is effectively shrunk. The clusters effectively move the two shores further apart, emphasizing the role of (bridgeless) Southampton Water in dividing the area.

It appears that while to some extent the clustering is similar to that found in the LSA data, the relative locations and memberships of the clusters have partly shifted in some cases, although this may simply reflect this small and hence unreliable pilot sample. The geography of the area in general is again clearly distorted – with Southampton Water not only widened but also rotated almost 90 degrees – and this reflects the tendency by most participants to leave a vertical north-south space for the inlet, rather than its true northwest-southeast orientation. This presumably reflects the tendency to rotate and simplify axes and coastlines, noted by many previous cognitive studies [8, 11]. Thus this data, and possibly the LSA data, does seem to reflect known cognitive biases in spatial representation. More importantly, it clearly begins to reveal the clustering of places at this scale by local inhabitants.

2.3 Pilot linguistic data

As mentioned earlier, participants in the pilot study were encouraged to provide simultaneous verbal protocols while laying out the counters. If the clustering and distortions in the layout reflect a categorizing tendency within people's stored mental geography, rather than ignorance or lazy shortcutting in performing the spatial task, then we would expect to see this also reflected in the way participants described the locations in question. Although the transcripts have not been formally (let alone quantitatively) analyzed as yet, it is easy to find hints of categorical thinking about groupings of places, using various words for such groupings as shown in the examples below. Note that often, the groupings are initially linear along major roads or shorelines, but still ending up closer together than their real-life positions. (Author's italics.)

P01: “so mostly *stacked up* along the waterfront, south-west of Southampton … right by Fawley, Dibden and Marchwood y'know getting stacked up…”
P02: “Everything’s in *groups*. Groups because I know, if I've travelled around that area I'll recognize the names…”
P03: “Marchwood I've heard of and I think of that as being sort of on the way to Totton so I'd probably *put them together*… I'd probably put that with Bucklers Hard in a New Foresty *part of the world*… These are all places that I think of as being *along the A326* [road] although some of them aren't exactly.”
P04: “Winsor is somewhere *in amongst this lot* I think… Durley is over with *this little batch*.”
P05: “… have to imagine the Southampton water going down there and I'll try and put *all of these uh these foreign east of the water places*…”
P07: “So what I'm doing I'm now looking for *places in the Waterside* as I can use it as *a lateral line* up to Southampton… Just *bunching* um Brockenhurst Lyndhurst and Beaulieu up closer together… So I'll be using *the line of the M27* [highway] to get a lot of the *places along there*…”
P08: “Pooksgreen that can be *part of the Waterside*… Exbury mm just need to go here in a *little clump* this can go there with *the foresty ones*… have to move *the Waterside* over a bit… Bartley that can go over with *the Winsor Cadnam lot*… Calmore can *snuggle in there with* Totton and Testwood.”

3. Discussion and Challenges

The above data is being presented mainly to stimulate discussion; clearly there is a long way to go before reliable empirical and linguistic data can be triangulated with web-sourced data to enable derivation of local vernacular geography. Some key questions arise, not least the extent to which the apparent groupings of places in both the task and verbal data were an artifact of the counter-placing task (given the sheer number of counters – 55 – that needed to be placed somewhere on the card). Even within the small pilot sample, it was also evident that different strategies might be applied – e.g., participant 06 gave no indication of clustering places during the task performance. This corresponds to earlier findings indicating different spatial strategies across individuals (e.g., cluster-based vs. trajectory-based navigation strategies in [17]).

Nevertheless, the tendency to simplify spatial memory via categorical encoding has already been demonstrated at various other spatial scales [e.g., 7-11, 18], so it would be surprising if people did not also tend to group localities in this way, notwithstanding their direct experience of them as navigated regions in themselves (unlike the data points used in most previous studies). And indeed, the parallels between our different data sets in this respect were striking. Although the actual locations on the map did not necessarily coincide, the cognitive biases were clearly present across the board, in terms of clustering, simplifying and ‘shrinking’ the

available space – even though at least half of the space was highly familiar to participants from an immersed, egocentric perspective.

Data such as this, when collected from wider samples, may help us to model why and how place categories form in familiar spaces (as opposed to those only experienced as haptic representations), and how far those categories match the 'messy' aspects of human categories mentioned at the start of this paper. It appears that toponyms are not essential for clustering to occur – we can cope with spatial categories which are effectively nameless – so under what circumstances do toponyms get applied consistently enough, by enough local residents, to become established as verifiable vernacular geography?

Scale is also a consideration. Here, unusually in spatial cognition studies, we were considering the scale above that of an urban streetscape or university campus but below national level; the places that were apparently being grouped into categories were suburban and village settlements or localities, themselves each already a collection of locations which afford the individual 'vistas' or 'reference frames' discussed by work such as [19]. Those in turn, of course, tend to represent a collection of different points (and potential viewpoints) within a single scene or 'vista' space. In this study we were considering a relatively high level of the place hierarchy – an area covering approximately 432km^2. To what extent do different principles and heuristics apply at different scales, for the formation of place clusters or categories? Will these be reflected in linguistic utterances about them?

This reminds us of still further questions, concerning the role of linguistic data in trying to identify vernacular place categories. In the above pilot study, a wide range of phrases was used to indicate groupings of places – making it difficult to imagine the use of any kind of automatic language parsing to help to identify them. Even so, trawling data for similar 'grouping' hints in people's discussions of local places – even if it could never be exhaustive - might in future help to augment a cruder, more co-occurrence-based approach to toponym groupings.

References

1. Hahn U, Ramscar M (2001) Similarity and Categorization. Oxford University Press, Oxford
2. Montello DR (2003) Regions in geography: Process and content. In: M Duckham, M F Goodchild, M F Worboys (eds) Foundations of geographic information science (pp 173-189). Taylor & Francis, London
3. Hampton JA (2007) Typicality, graded membership and vagueness. Cognitive Science 31: 355--384
4. Barsalou LW (1985) Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. Journal of Experimental Psychology: Learning Memory and Cognition 11: 629--654
5. Davies C (2009) Are Places Concepts? Familarity and Expertise Effects in Neighborhood Cognition. In: Spatial Information Theory: 9th International Conference COSIT 2009, Aber Wrac'h, France, September 21-25 2009. LNCS Vol. 5756: 36—50. Springer, Berlin

6. Montello D (1993) Scale and Multiple Psychologies of Space In: Frank AU Campari I(eds) Spatial Information Theory: A Theoretical Basis for GIS, Proceedings of COSIT '93: 312—321. Springer-Verlag, Berlin
7. Newcombe N, Huttenlocher J, Sandberg E, Lie E, Johnson S (1999) What do misestimations and asymmetries in spatial judgment indicate about spatial representation? Journal of Experimental Psychology: Learning Memory and Cognition 25 (4): 986--996
8. Tversky B (1981) Distortions in memory for maps. Cognitive Psychology 13: 407--433
9. Friedman A (2009) The role of categories and spatial cuing in global-scale location estimates. Journal of Experimental Psychology: Learning Memory and Cognition 35(1): 94--112
10. Hirtle SC, Jonides J (1985) Evidence of hierarchies in cognitive maps. Memory and Cognition 13: 208-–217
11. Lloyd R, Heivly C (1987) Systematic distortions in urban cognitive maps. Annals of the Association of American Geographers 77(2): 191--207
12. Gao S, Janowicz K, Montello DR, Hu Y, Yang J-A, McKenzie G, Yan B (2017) A data-synthesis-driven method for detecting and extracting vague cognitive regions. International Journal of Geographical Information Science 31(6): 1245--1271
13. Davies C (2013) Reading Geography between the Lines: Extracting Local Place Knowledge from Text. In: Spatial Information Theory: 11th International Conference COSIT 2013, Scarborough, UK, September 2-6 2013: Proceedings. LNCS Vol 8116: 320—337. Springer, Berlin
14. Louwerse MM, Zwaan RA (2009) Language encodes geographic space. Cognitive Science 33 : 51--73
15. Landauer TK, McNamara DS, Dennis S, Kintsch W (2007) Handbook of Latent Semantic Analysis. Lawrence Erlbaum, Mahwah, NJ.
16. Tobler W (1994) Bidimensional Regression. Geographical Analysis 26: 186–212
17. Tenbrink T, Wiener J (2009) The verbalization of multiple strategies in a variant of the traveling salesperson problem. Cognitive Processing 10(2): 143--161
18. Lansdale MW (1998) Modeling memory for absolute location. Psychological Review 105 (2): 351--378
19. Meilinger T, Riecke BE, Bülthoff HH (2014) Local and global reference frames for environmental spaces. The Quarterly Journal of Experimental Psychology *67*(3): 542--569