What Do You Mean You're *in* Trafalgar Square? Comparing Distance Thresholds for Geospatial Prepositions

Niloofar Aflaki[1]
Massey Geoinformatics Collaboratory, Massey University, Auckland, New Zealand

Kristin Stock
Massey Geoinformatics Collaboratory, Massey University, Auckland, New Zealand

Christopher B. Jones
School of Computer Science and Informatics, Cardiff University, UK

Hans Guesgen
Massey Geoinformatics Collaboratory, Massey University, Auckland, New Zealand

Jeremy Morley
Ordnance Survey, Southampton, UK

Yukio Fukuzawa
School of Natural and Computational Sciences, Massey University, Auckland, New Zealand

Abstract

Natural language location descriptions frequently describe object locations relative to other objects (*the house near the river*). Geospatial prepositions (e.g.*near*) are a key element of these descriptions, and the distances associated with proximity, adjacency and topological prepositions are thought to depend on the context of a specific scene. When referring to the context, we include consideration of properties of the relatum such as its feature type, size and associated image schema. In this paper, we extract spatial descriptions from the Google search engine for nine prepositions across three locations, compare their acceptance thresholds (the distances at which different prepositions are acceptable), and study variations in different contexts using cumulative graphs and scatter plots. Our results show that adjacency prepositions *next to* and *adjacent to* are used for a large range of distances, in contrast to *beside*; and that topological prepositions *in*, *at* and *on* can all be used to indicate proximity as well as containment and collocation. We also found that reference object image schema influences the selection of geospatial prepositions such as *near* and *in*.

2012 ACM Subject Classification Computing methodologies → Natural language processing

Keywords and phrases contextual factors, spatial descriptions, acceptance model, spatial template, applicability model, geospatial prepositions

Digital Object Identifier 10.4230/LIPIcs.COSIT.2022.1

Funding This work is partly funded through by an Ordnance Survey (UK) PhD scholarship.

1 Introduction

In natural language location descriptions, people tend to describe their location or that of a point of interest (POI) using relative relation terms (e.g. *the house beside the park* describes the location of the house relative to the park). These types of location descriptions contain three essential elements: the locatum (the object for which the location is being described);

[1] Corresponding author

© Niloofar Aflaki, Kristin Stock, Christopher B. Jones, Hans Guesgen, Jeremy Morley, and Yukio Fukuzawa;
licensed under Creative Commons License CC-BY 4.0

15th International Conference on Spatial Information Theory (COSIT 2022).
Editors: Toru Ishikawa, Sara Irina Fabrikant, and Stephan Winter; Article No. 1; pp. 1:1–1:14
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

the relatum (used as a reference location for describing a locatum) and the spatial relation term (which specifies the relation in space between the locatum and relatum) [35] (*the house* *<locatum> beside <spatial relation term> the park <relatum>*).

Relative spatial descriptions are a common method for describing location in human communication, and accurate automated interpretation of these kinds of expressions can be of critical importance for many applications. For example during emergency events, they may describe the location of stranded people or dangerous conditions as in *there is a fire in the house on Victoria street, next to the Coffee Club cafe* [39, 13]. The availability of large amounts of text on web sites, blogs and social media motivates the development of methods to automatically interpret and generate such natural language relative location descriptions.

Most of the previous work on georeferencing relative spatial descriptions has focused on toponym recognition and disambiguation [20, 21, 17, 16, 18] without taking into account the role of spatial relation terms. For example, in a description such as *behind the Shell building*, consideration of the preposition *behind* improves the accuracy of the georeference that would be obtained if only the place name were used. To consider geospatial prepositions[2], it is necessary to understand the locations (relative to the relatum) in which a given spatial relation term may validly be used (e.g., how near does a locatum have to be to a relatum for *near* to be an appropriate spatial relation term?). To address this kind of question, a number of models have been developed for specific spatial prepositions, known as acceptance models, applicability models or spatial templates [26, 9, 31, 2, 11, 37, 32, 42, 6, 28, 3, 4]. These models are often probabilistic or predictive, describing areas in which a given preposition is highly suitable, compared to others where it may be borderline.

We address two gaps in the previous research. Firstly, previous work has mostly focused on the task of developing models for small numbers of individual prepositions in isolation. Here, we compare the acceptance thresholds (the distances at which a preposition is acceptable) for different prepositions in the English language in order to study their semantic similarities and differences. Secondly, the importance of contextual factors on location interpretation has been emphasised in a number of previous works [12, 15, 27, 38, 24, 41, 8, 28], but the comparative impact of context on prepositions has not been widely studied empirically. We address these gaps by comparing the use of the nine prepositions across three well-known landmarks in London, UK (Trafalgar Square, Buckingham Palace, and Hyde Park) taking account of context with particular reference to the nature of the relatum and its associated image schema. We address two research questions of: **RQ1**: How do distances between relata and locata that are acceptable differ between geospatial prepositions? **RQ2**: How important is context in the use of geospatial prepositions?

In Section 2, we discuss relevant previous work, while Section 3 defines the data extraction method; Section 4 presents the results and Section 5 provides a discussion and findings.

2 Previous work

2.1 Acceptance models

Spatial acceptance models define the areas in which a given preposition may be applied, relative to a relatum, and have been investigated for several purposes, including location prediction [2, 4], selection of an appropriate preposition for a description (language generation) [6, 28] and georeferencing [9, 3]. Chang et al. [2] and Yu and Siskind [42] used

[2] We consider a spatial preposition to be *geospatial* if the relatum is a geographical object.

acceptance models to draw a spatial scene in 3D using textual descriptions and to find objects in videos in an indoor environment respectively. Malinowski and Fritz [22] and Lan et al. [19] used deep learning and other machine learning models (CNN and latent ranking SVM) to retrieve specific objects in image configurations, relying on spatial acceptance models. However, they did not compare individual spatial prepositions or consider contextual factors, and the studies were image-based without consideration of geographic space. In a series of studies Hall et al. [10, 9, 11] used location descriptions from the Geograph website and human subjects experiments to quantify the distances and angles for projective and proximity spatial relations, and to create spatial templates for the purpose of generating and interpreting natural language photo captions. The scale varied between urban and rural locations and corresponds to the environmental and geographical spaces of Montello [25]. A study of the distances associated with the relation *near* was conducted in Derungs and Purves [5] based on evidence of n-grams mined from the web. They considered relations between places that were either cities or points of interest, and hence environmental and geographical spaces, and found the distances were typically between nearest neighbours and random. Our work considers a relatively wide range of spatial prepositions and focuses on the environmental scale using data scraped from web page sources.

2.2 Effects of contextual factors on spatial preposition use

The importance of context in selection and interpretation of spatial prepositions has been well documented. Herskovits [12] identified geometric configuration, use types, salience, relevance, tolerance and typicality as important in determining whether a preposition would apply in a given situation or not. Tyler and Evans [36] counted context and other elements of spatial descriptions such as the locatum and relatum as important factors for understanding spatial prepositions such as *over*. Stock and Yousaf [34] used a Linguistically Augmented Geospatial Ontology (LAGO) to extract five contextual factors for locata and relata and calculated the similarity between them using WordNet [40], showing that the consideration of contextual factors improves the ability to identify semantically similar descriptions. Wallgrun et al. [38] used a similar method to us to extract content from internet search engines, and demonstrated the variation in distances associated with three proximal spatial relations according to the mode of transport. Collell et al. [4] used spatial templates to predict the location of objects in photos using contextual factors such as embeddings (which describe linguistic context) and size of locatum. However, their focus was on spatial relations in the form of verbs and they did not address prepositions. Stock and Hall [33] identified a broad range of factors that affect the interpretation of spatial descriptions including: proximity and locatum and relatum characteristics such as liquid/solid and image schema. However, they did not explicitly describe the ways that context influences the interpretation of spatial prepositions or the distance between locatum and relatum.

3 Data extraction method

We used Google searches to extract descriptions from the World Wide Web that contained three elements: locatum, geospatial preposition and relatum. We used place names for the locatum and relatum from OpenStreeMap[3], identified coordinates of the relatum and locatum used with a specific preposition and calculated the distance between them for

[3] `https://www.openstreetmap.org/`

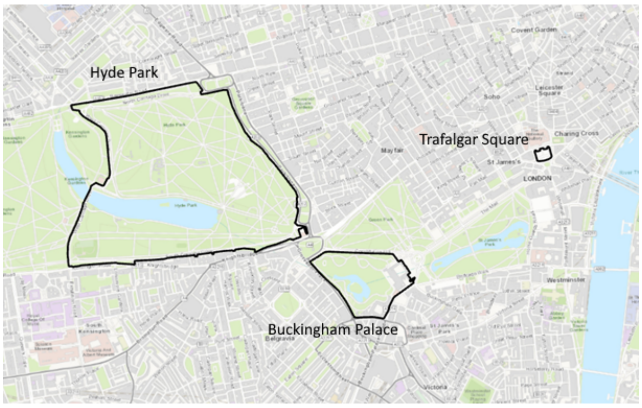


Figure 1 Locations of Hyde Park, Buckingham Palace and Trafalgar Square in London.

further analysis. We use the frequency of mentions of a locatum with a particular geospatial preposition-relatum combination as a proxy for the applicability of that geospatial preposition. For example, a search for *Green Park next to Buckingham Palace* returned a count of 83 mentions (which we refer to as frequency). We consider that this frequency of use indicates that the *next to* preposition is more acceptable for the Green Park-Buckingham Palace locatum-relatum pair than for some other pair of place names with a lower frequency.

In order to compare the influence of context, we selected three popular tourist attractions in the London area (Figure 1) as relata with a variety of scales and feature types: Trafalgar Square (area: $18040m^2$, perimeter: $954m$), Hyde Park (area: $1388013m^2$, perimeter: $5629m$), and Buckingham Palace (the building and its grounds)(area: $201240m^2$, perimeter: $1997m$). We used the following steps to extract the data for each relatum:

1. We used the OpenStreetMap export service to extract all places in the general area of the three relata using a bounding box that covered a large section of central London.
2. From the set returned in Step 1, we identified those features that had centroids within a specified distance of the centroids of each relatum. A maximum distance of 2km was used for Trafalgar Square and Buckingham Palace, and 3km for the much larger Hyde Park. These distances were selected to retrieve a manageable number of locata but with the aim of encompassing typical extents of acceptable use of the prepositions. Our results indicate that the selected distance ranges were sufficient in most cases (see Section 4.2). We only extracted point and polygon geometries and defer consideration of linear objects to a later study.
3. We manually checked and excluded place names returned from Step 2 that had multiple instances (for example, *McDonald's* has multiple branches across the London area) in Google Maps, to avoid ambiguity regarding the coordinate location of the locata.
4. After Step 3, we had around 800 locata for each relatum. Google search counts were used to identify the 100 most frequently mentioned places as candidate locata for each relatum (though it is acknowledged that these counts might be approximate due to Google's search algorithm).
5. We generated query triples combining each of the 100 locata for a given relatum, each of the prepositions and the relatum itself, enclosed with quotation marks (e.g. *"National Gallery near Trafalgar Square"*), using Python's Beautiful Soup library [29] to run a query for each triple and scrape the URL and excerpt from the retrieved page.
6. We ran a version of the previous step in which a wildcard character was included before the preposition, to accommodate the common presence of verbs.

7. We manually reviewed the output returned from the previous steps in order to remove mentions that were repetitive, non-spatial or outside the geographical area. We also examined outliers and removed invalid expressions manually (e.g. expressions that referred to buildings that had moved since the text was written, which occurred particularly for historical documents online).
8. We combined the frequencies of mentions for searches with and without the wildcard character.
9. We retrieved the geometry for each locatum and relatum from OpenStreetMap, along with the category (e.g., tourism), type (e.g., attraction) and centroid coordinates (if a polygon).
10. Finally, we calculated the shortest distance between the location of each locatum and relatum using point geometries (for points) or boundary geometries (for polygons).

This analysis process resulted in a total of 1970, 1523 and 1746 mentions (across all locata and all nine prepositions) for Buckingham Palace, Hyde Park and Trafalgar Square respectively.

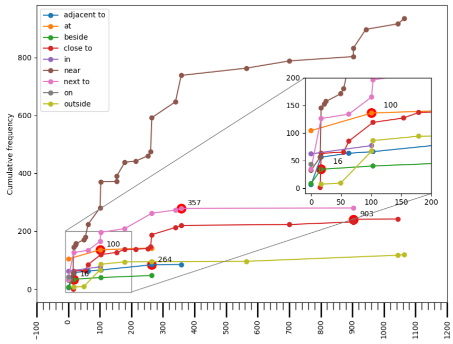
4 Results

4.1 RQ1: How do acceptable distances between relata and locata differ between prepositions?

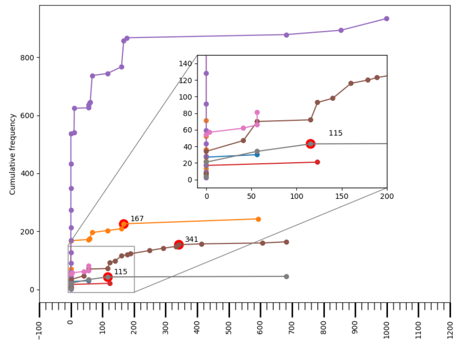
Figure 2 shows the cumulative frequency graphs [14] for each preposition for each of the three sites. Each coloured line is a preposition, as labelled in the legend. The points on the graphs represent locata and are positioned on the x axis using the distance between the relatum and locatum geometries. The vertical axis shows cumulative frequencies, being the total of all mentions of a given locatum plus all locata at shorter (closer) distances. We use cumulative frequency graphs because they provide a clearer picture of the behaviour of each preposition in comparison to raw frequency graphs in which individual locata can obscure the visualisation. Note that the point at which each curve flattens is the distance beyond which there are very few new mentions, so we consider this to be the acceptance threshold for each preposition.

On the graphs in Figure 2, acceptance thresholds for each preposition for each relatum are marked with large red dots (Figure 2 (a-c)). They are identified as follows. Starting from the right-most point on each line for a given preposition in the graph, we move progressively leftward, point by point, to identify the first point for which the slope of the edge connecting that point to the point to its right exceeds 5°, and consider this point the acceptance threshold. If the right-most edge (between the last two points) exceeds 5°, we consider that we have insufficient data to identify the acceptance threshold. If none of the edges for a given preposition has a slope exceeding 5° (they are all relatively flat), we calculate the average slope for all edges for the preposition line, and repeat the above process to find the right-most point for which the slope of the edge connecting that point to the point to its right exceeds the average slope. The figure of 5° was selected through a process of trial and error with a range of other angles, with 5° best identifying the point at which the lines flatten consistently across all of the prepositions.

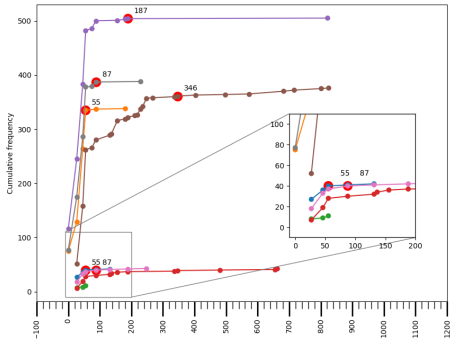
In addition to the graphs for each relatum, we present an aggregated cumulative frequency graph Figure 2(d), in which we sum the data for all relata, and adjust frequency values to account for varying site popularity. Some relata are more popular than others (i.e., attract more mentions in social media in total). So, if more popular sites are not adjusted for popularity, their values will have a disproportionately large influence on the shape of the graph that combines the results of all relata. We therefore scale down the mention count for



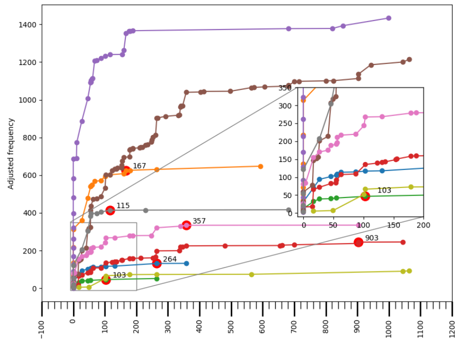
(a) Buckingham Palace.



(b) Hyde Park.



(c) Trafalgar Square.



(d) d) All relata aggregated.

Figure 2 Cumulative frequency graphs for each relatum and aggregated (Euclidean distances on x axis and frequencies on y axis).

these more popular sites so that all relata have an equal total adjusted number of mentions, this being equal to the minimum total across all relata (in this case it is for Hyde Park, which has 1523 mentions). The reason for doing this rather than normalising (adjusting to values between 0 and 1) is that it still gives some indication of the number of mentions in absolute terms (i.e. it is clear if there are very few mentions) (Equation 1).

$$adjfreq(r_i, l_i) = \frac{freq(r_i, l_i)}{\sum_{i=1}^{n} freq(r_i, l_i..n)} * min(\sum_{i=1}^{n} freq(r_i..n, l_i..n)) \quad (1)$$

$freq(r_i, l_i)$ indicates the frequency of a given preposition for $locatum_i(l_i)$ and $relatum_i(r_i)$, $\sum_{i=1}^{n} freq(r_i, l_i..n)$ indicates the sum of the frequencies of the prepositions across all locata for $relatum_i$ and $min(\sum_{i=1}^{n} freq(r_i..n, l_i..n))$ indicates the sum of the frequencies of the prepositions across all the locata for the relatum that has the minimum sum of the preposition frequencies (in this case Hyde Park). The numbers of mentions for Trafalgar Square and Buckingham Palace were higher than those for Hyde Park, and were adjusted down accordingly.

The data show some clear patterns across the prepositions, with the range in acceptance thresholds for each site summarised in Figure 3. We were unable to identify an aggregate threshold for *in* and *near* because the aggregated data did not reach a point of flattening. The *close to* preposition has the highest acceptance threshold, with (903m) for Buckingham Palace and for the aggregated data, but there is insufficient data to establish a threshold

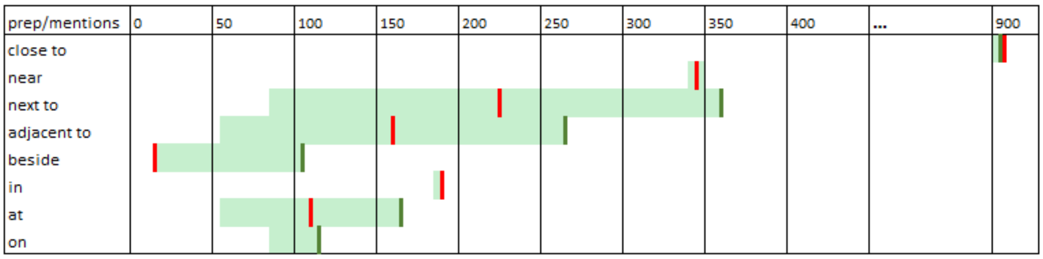


Figure 3 Acceptance threshold ranges for each preposition (red vertical line is mean, green vertical line is aggregate).

for the other two relata (*close to* is a relatively infrequently used preposition, so we have few mentions in the data). The *near* preposition also has a high threshold: 341m for Hyde Park and 347m for Trafalgar Square. Furthermore, we see *near* being used infrequently for much greater distances than the threshold (700-800m for Hyde Park and Trafalgar Square). The acceptance threshold for Buckingham Palace (and for the aggregated data) is much higher (>1100m), and the graph does not level off for this site, indicating that the acceptance threshold is beyond our last data point.

Of the prepositions that convey notions of adjacency, *next to* has the highest threshold for the aggregate data (357m). However, the range in thresholds between 358m for Buckingham Palace and 87m for Trafalgar Square has some substantial overlap with the range for *adjacent to* (265m to 55m, with an aggregate data threshold of 264m). *Beside* has a much smaller threshold, being 103m for the aggregate data, and 17m for Buckingham Palace (*beside* did not appear in our data for Hyde Park, and only infrequently, with distances up to 50m, for Trafalgar Square). From this evidence, we postulate that *beside* is typically limited to much closer locations than *next to* and *adjacent to*, both of which are used for locations within a closer range than the proximity prepositions *near* and *close to*, but more data is needed to confirm this. *Outside* only appeared for the Buckingham Palace site, but no acceptance threshold could be determined as the slope between its rightmost two points exceeds 5°.

Moving to the containment and collocation prepositions *in*, *at* and *on*; surprisingly, the acceptance thresholds for *in* appear to be large, being beyond our last data point for the aggregate data and for Hyde Park, and to a lesser extent, for Trafalgar Square (187m), even though, given that our distances are measured boundary to boundary, we might expect distances of zero (the locatum inside the relatum). For Buckingham Palace, only two locata were used with the *in* preposition. The first one is within its boundary (The Royal Mews) and the second is 100m away (Victoria Memorial). While the latter is located on the site of Buckingham Palace, it was not within the boundary geometry we extracted from OpenStreetMap. The Hyde Park data is affected by the location of the neighbouring Kensington Gardens, which people sometimes refer to as Hyde Park. For example, the description *Princess Diana Memorial Playground in Hyde Park* appears, but the playground is on the east side of Kensington Gardens, 900m outside the boundary of Hyde Park. In the case of Trafalgar Square, the most distant locata within the acceptance threshold are Her Majesty's Theatre (187m) and the Nigerian High Commission (180m). We expect that the reason for using *in Trafalgar Square* for these two locations is that Trafalgar Square is a well-known landmark in the area. It appears that in natural language location descriptions, the geographic boundaries of well-known landmarks may be "stretched", but more work is needed to validate this. We also see an unusual outlier for Trafalgar Square: the Methodist

Central Hall. It is marked with a dashed line on the graph as we suspect that it is an error. The source description reads "*16/09/2015 – The largest air raid shelter in England was at the Methodist Central Hall in Trafalgar Square which could hold 2,000 people each night.*" (Westminster Reporter, September 2015, Issue 120, page 21). Methodist Central Hall is 821m from Trafalgar Square and other documentation indicates it was associated with a large air raid shelter, but we have no evidence that either was in Trafalgar Square.

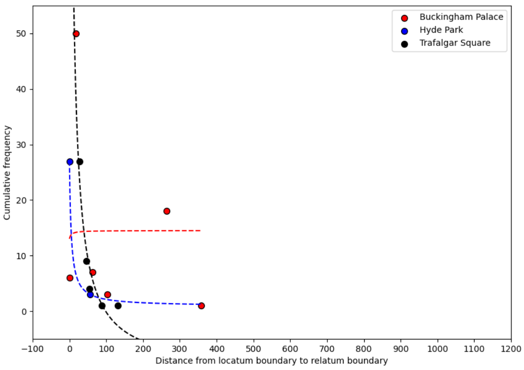
The *at* preposition has acceptance thresholds in a similar range to the *in* preposition, but with the lowest acceptance threshold of the three topological prepositions, being for Trafalgar Square. The *on* preposition has thresholds in a similar range (87m-116m) with a threshold for the aggregate of 115m, being the lowest for the three containment and collocation prepositions.

4.2 RQ2: How important is context in the use of geospatial prepositions?

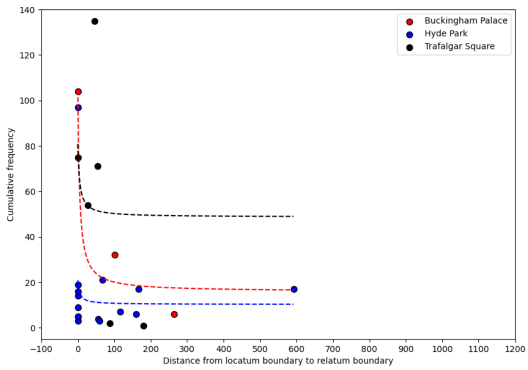
Our three relata were specifically selected for their difference in size and feature type (Figure 1) to enable investigation of contextual differences that relate to the nature of the relatum. Figure 4(a-h) compares the three relata for each of the prepositions using scatter plots and regression lines for each relatum. We use a reciprocal, linear regression equation to plot the regression line (y=1/x), and we refer to these regression lines as acceptance profiles, as they show the distances at which the preposition is highly acceptable, and those at which it becomes less so. This provides more information than the acceptance threshold alone. We do not plot the *outside* preposition, as we only identified its use with the Buckingham Palace relatum, and in RQ2 our focus is on comparison of the context.

Several of the prepositions show clear similarity across all three sites, including *next to* and *close to*. The curves across the three relata for *next to* are very similar, the main difference being in frequency of mentions, which is discussed further below. The ranges of the data points vary for the three sites, with Buckingham Palace having low frequency mentions for more distant locata, while in contrast Hyde Park only uses *next to* with locata that are relatively close (up to 56m). For smaller distances, the *adjacent to* and *beside* preposition graphs are similar to those of *next to*. Both are used with Buckingham Palace for distances up to 250m, but the most frequent uses across the other two sites are less than 100m. We see a very similar pattern for *close to*, with Buckingham Palace attracting mentions out to approximately 1km, Trafalgar Square to 660m (albeit very low frequency) and Hyde Park to 123m. We consider it likely that the larger distances associated with Buckingham Palace are influenced by the ambiguity in the specific size/area of the relata: the entire grounds, or only the palace building itself. In this analysis, we used the entire grounds, since public access to the grounds and palace is limited, so it is less likely that mentions would refer only to the palace. Our comparative analysis confirmed this, with calculations based on the use of the palace geometry alone returning less consistent results than for the combined palace and grounds.

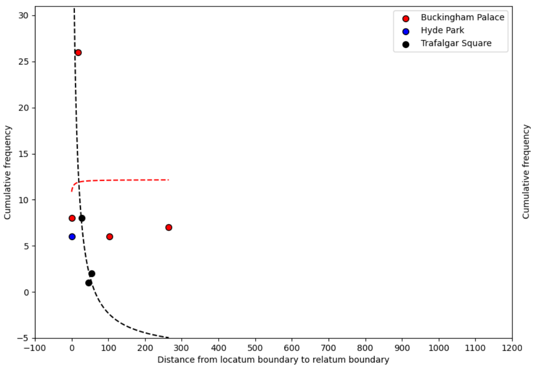
Despite the differences among the three sites discussed above, the highest frequency uses of *close to* and *next to* (the point at which the steeper sections of the graph level out being less than 100m in both cases) are consistent across all three sites, suggesting that context has limited impact on the most common uses of these prepositions. While the *near* graph is similar to that of *close to*, it does show some variation between the three relata: namely the absence of high frequency use at very low distances for Buckingham Palace and Hyde Park. It appears that while *near* may be used for Trafalgar Square for locations very close to (or at) the Square, this does not apply to the other two relata, in contrast to *close to*, which



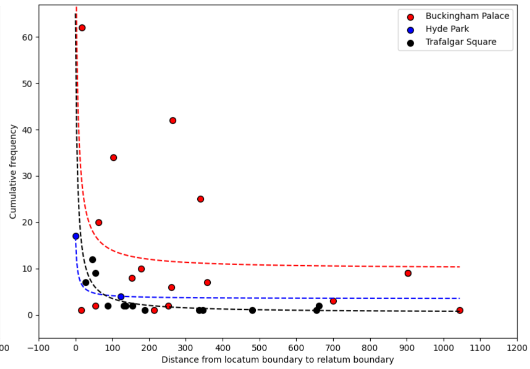
(a) adjacent to.



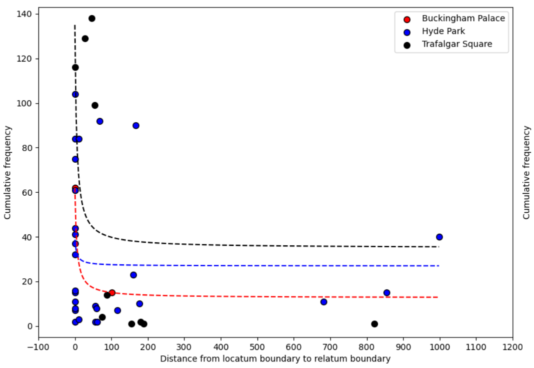
(b) at.



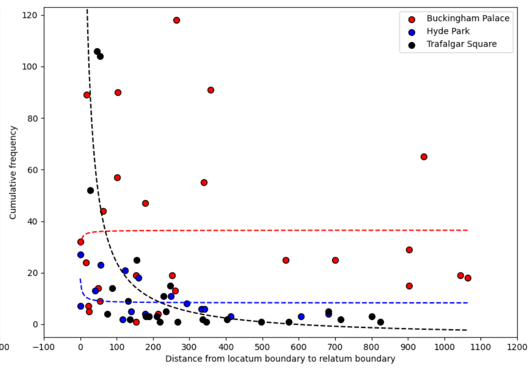
(c) beside.



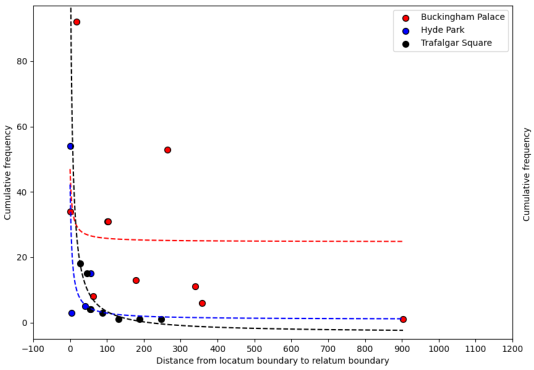
(d) close to.



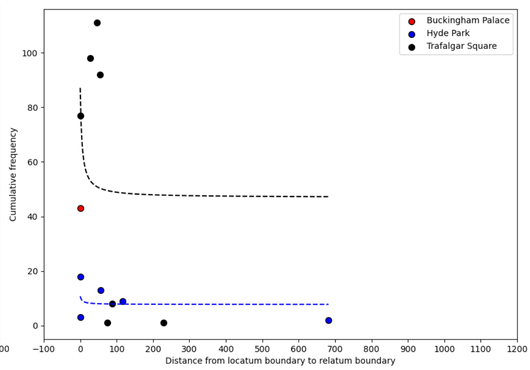
(e) in.



(f) near.



(g) next to.



(h) on.

Figure 4 Scatter plot and regression lines for the frequency of each preposition across all three sites.

is applied at very small distances for all three relata. Trafalgar Square is a smaller feature, with arguably vaguer boundaries, than the other two relata, which may explain the more liberal use of *near* in that case.

The *in* preposition shows similar patterns across all three sites. Hyde Park and Trafalgar Square have a much greater range of mentions (going up to 1km), but all of the curves flatten at a distance within approximately 50m of the relata. It is interesting to note that, counterintuitively, the *in* preposition is used relatively frequently with locata much further from the relatum than the *next to* and *close to* prepositions, and clearly for objects that are well outside the boundary of the relatum. For example, "*Baglioni Hotel in Hyde Park*" is 700 metres away from the boundary of Hyde Park.

Like *in*, the *at* preposition shows flattening at distances very close to the relatum, but this distance is greater for Buckingham Palace than for the other two relata. This may be due to the closed nature of Buckingham Palace (public access is strictly controlled, being limited in timing, volume of visitors and area of access, and requiring payment) compared to the other sites. Thus, the description *I'm at Buckingham Palace* could mean that the speaker is outside the Palace gates, while this is less likely (but still possible) for the other two relata. *At* is used at a much greater distance for Hyde Park than for the other two relata, but this may be related to the Kensington Gardens effect described above.

Use of the *on* preposition is much more frequent for Trafalgar Square than for Hyde Park and is only used once for Buckingham Palace in our data set, with distance zero. This is likely the result of image schema, with squares and plazas being more frequently associated with a platform schema than parks or palaces ([23, 24]). However, we do not see a similar pattern for the *in* and *at* prepositions, which are commonly used for parks and similar types of objects. Trafalgar Square is frequently used with *in*, *at* and *on*, suggesting that a range of different image schemata are suitable for this feature type, while parks and palaces are more limited.

Across all of the prepositions, we see much greater variation between relata at the outer extremes of acceptability of the prepositions. That is, many of the prepositions studied are used less frequently for quite large distances for some sites more than others; while the most frequent uses are much more uniform across the sites, despite the differences in size, feature type and level of urban construction among the relata. Generally, the proximity (*near*, *close to*) and adjacency (*next to*, *adjacent to*, *beside*) prepositions are more frequently used for greater distances for Buckingham Palace than the other two relata; while *in*, *on* and *at* are most frequently used for Trafalgar Square across all distances. Like *in*, the *at* preposition shows flattening at distances very close to the relatum, but this distance is greater for Buckingham Palace than for the other two relata.

5 Discussion

5.1 Acceptance thresholds (RQ1)

Our findings show that among the proximity/adjacency prepositions, *near* and *close to* have the highest acceptance threshold distances from the relatum with *near* > 1100m and *close to* = 903m. *Next to* and *adjacent to* were used for distances between 55-358m, and the smallest threshold for proximity/adjacency prepositions was for *beside* (17m for the Buckingham Palace) but there was not enough data on the other two sites to confirm this finding. Carlson and Covey [1] ran a human subjects experiment to estimate the distance associated with some spatial prepositions such as *next to*, *beside* and *near*. Similar to our findings, their research showed that *beside* and *next to* are not associated with relatum size and these prepositions specify smaller distances than *near*. Fisher and Orf [7] also reviewed the interpretation of *near* and *close* in a university campus area and found that people did not use these terms for buildings that were very close to the student centre (relatum), but instead those that were further away. Hall et al.'s [11] density models of spatial prepositions such as *near*, *between*, *at the corner*, *at* and *next to* also show a larger distance for *near* and smaller distances for *at* and *next to*. Skoumas et al. [32] also confirmed that the use of *near* is not restricted only to locations close to the relatum, while *at* and *next to* were used for areas close to the relatum.

5.2 Contextual factors (RQ2)

The analysis in Section 4.2 highlights a number of observations regarding contextual variations in the use of spatial prepositions. Firstly, we note that *near* is used less frequently for locata very close to the relatum for two of the three relata, the exception being Trafalgar Square. This 'doughnut effect' was less evident in the analysis of Hall et al. [9], but was identified by Fisher and Orf [7] who claimed that this might be due to the similarity in place names or functions. More research is needed to confirm this.

Secondly, we noted in Section 4.2 the likely impact of image schema on the use of the containment and collocation prepositions. This has been identified by other researchers, who identified the use of the *on* preposition when a platform schema is used, or *in* for a container schema ([23, 24]). However, our results identify a variation in image schemata applied to different feature types. Some of our relata were strictly subject to a single image schema (e.g., Hyde Park with the container schema, indicated by the use of the *in* preposition in preference to *at* or *on*), while others were more promiscuous (notably Trafalgar Square, which uses all three of these prepositions liberally, suggesting platform, container and possibly other image schemata such as link are appropriate).

Thirdly, we note that the *outside* preposition is only used with Buckingham Palace in our data (see Figure 2(a)). While the *outside* preposition would normally be associated with the container image schema, we also see low frequency use of *in* (also associated with the container image schema) with Buckingham Palace compared to the other two relata. However, this may be due to the access limitations reducing the frequency of mentions for Buckingham Palace. In contrast, *in* is the most frequently used containment or collocation preposition used for Hyde Park, but *outside* is not used for this relatum in our data. This suggests that *outside* requires a stronger form of containment than *in*, with Buckingham Palace being a stronger container than Hyde Park.

In addition to the influence of generalisable characteristics of different relata on the acceptance profiles and thresholds of prepositions, our data shows that individual contexts can influence the use of prepositions, as in the case of Hyde Park and the likely influence of the neighbouring Kensington Gardens. This suggests that general models of preposition applicability, even if they are able to incorporate a rich range of contextual factors such as feature type, accessibility or image schema, are still likely to be limited in accuracy, as they are unable to capture these individual nuances. Also in the Hyde Park/Kensington Garden example, the influence of familiarity on the use of prepositions and the associated selection of locata to describe a location, reported in [41], is confirmed by our research. Thus people appear to use the reference objects whose names they are more familiar with, and this may influence the acceptance profiles and thresholds of prepositions in specific contexts.

These results also show that while there are distances that are acceptable for a given preposition across all three of our relata, which we might refer to as "ranges of agreement", there are also outer extremes of those ranges that are only used in certain circumstances, for only one or two relata, and depending on context. Both Hyde Park and Buckingham Palace have data points at much greater distances than Trafalgar Square, usually with low frequency (for example, the *next to* preposition is used at around 900m for Buckingham Palace), but with greater frequency at much lower distances for all three relata. We thus consider it likely that the acceptance areas for geospatial prepositions follow prototype theory [30] in having a range of exemplars, and outliers, which are at least partly determined by context.

5.3 Locatum popularity

In addition to the impact of relata on the selection of appropriate geospatial prepositions, we acknowledge the impact of locatum popularity on our dataset. It is likely that people chose the most popular or salient locata when describing a scene, in favour of less noticeable objects. While this may result in bias, since we are looking at the same three relata for all prepositions, the comparison between the relata is still valid, and we see that for a given locatum, a mixture of different prepositions is selected, rather than a single preposition. It would however be of interest in future work to examine more closely the possible influence of variation in popularity of the locata as well as variation in their properties such as size and type.

6 Conclusion

In this paper, we investigated the acceptable distances for a set of prepositions used in geospatial situations, and the impact of context on those distances. We used Google searches to extract locatum, preposition and relatum triples for three sites in London (Trafalgar Square, Buckingham Palace, and Hyde Park) and analysed the frequency of mentions of specific locata with each preposition and relatum. Our experiment led to a number of findings. Firstly, proximity/adjacency geospatial prepositions such as *near* and *close to* are used for larger distances than *adjacent to*, *next to* and *beside*. *Adjacent to* and *next to* are used for relatively large distance ranges, while *beside* is confined to small distances. Also, *in*, *on* and *at* are not only used for locata that are inside the relata, but sometimes when there is a short distance between locata or relata (depending on the context). Secondly, degree of adherence to a particular image schema varies depending on feature type: some use a single schema, while others use a variety depending on context. Thirdly, the use of the *outside* preposition relies on a notion of strength of containment which may depend on other contextual factors such as accessibility.

It is acknowledged that this study is limited in being confined to three relata and to scales relating only to an urban space. Thus future work is needed to consider a wider range of relata within different geo-spatial contexts. However there are challenges in such work with regard to obtaining adequate volumes of good quality data for individual spatial prepositions when used with multiple occurrences of specific combinations of named locatum and relatum as was done here. As part of future studies it would also be of interest to investigate relationships between locatum feature type and acceptance distance, as well as the relationship between image-schema and feature types and flexibility in the use of image-schema by a given feature type.

References

1 L. A. Carlson and E. S. Covey. How far is near? inferring distance from spatial descriptions. *Language and Cognitive Processes*, 20(5):617–631, 2005.
2 A. Chang, M. Savva, and C. D. Manning. Learning spatial knowledge for text to 3d scene generation. In *EMNLP*, pages 2028–2038, 2014.

3 H. Chen, S. Winter, and M. Vasardani. Georeferencing places from collective human descriptions using place graphs. *Journal of Spatial Information Science*, 17:31–62, 2018.

4 G. Collell, L. Van Gool, and M.-F. Moens. Acquiring common sense spatial knowledge through implicit spatial templates. In *Proc. AAAI 2018*, 2018.

5 Curdin Derungs and Ross S. Purves. Mining nearness relations from an n-grams web corpus in geographical space. *Spatial Cognition & Computation*, 16(4):301–322, 2016. doi:10.1080/13875868.2016.1246553.

6 S. Du, X. Wang, C. Feng, and X. Zhang. Classifying natural-language spatial relation terms with random forest algorithm. *Int. J. Geographical Information Science*, 31(3):542–568, 2017.

7 P. F. Fisher and T. M. Orf. An investigation of the meaning of near and close on a university campus. *Computers, Environment and Urban Systems*, 15(1-2):23–35, 1991.

8 N. Gronau, M. Neta, and M. Bar. Integrated contextual representation for objects' identities and their locations. *Journal of Cognitive Neuroscience*, 20(3):371–388, 2008.

9 M. Hall, P. Smart, and C.B. Jones. Interpreting spatial language in image captions. *Cognitive Processing*, 12(1):67–94, 2011.

10 M. M. Hall and C. B. Jones. Quantifying spatial prepositions: an experimental study. In *Proc. 16th ACM SIGSPATIAL*, pages 1–4, 2008.

11 M.M. Hall, C.B. Jones, and P. Smart. Spatial natural language generation for location description in photo captions. In *COSIT*, pages 196–223. Springer, 2015.

12 Annette Herskovits. Semantics and pragmatics of locative expressions. *Cognitive science*, 9(3):341–378, 1985.

13 Y. Hu and J. Wang. How do people describe locations during a natural disaster: an analysis of tweets from hurricane harvey. *arXiv preprint*, 2020. arXiv:2009.12914.

14 Arthur J Jelinek. Use of the cumulative graph in temporal ordering. *American Antiquity*, 28(2):241–243, 1962.

15 M. Johnson. *The body in the mind: The bodily basis of meaning, imagination, and reason.* University of Chicago Press, 2013.

16 E. Kamalloo and D. Rafiei. A coherent unsupervised model for toponym resolution. In *Proc, WWW Conference*, pages 1287–1296, 2018.

17 Morteza Karimzadeh. Performance evaluation measures for toponym resolution. In *Proc. GIR Workshop*, pages 1–2, 2016.

18 T. Kew, A. Shaitarova, I. Meraner, J. Goldzycher, S. Clematide, and M. Volk. Geotagging a diachronic corpus of alpine texts: Comparing distinct approaches to toponym recognition. In *Proc. Workshop on Language Technology for Digital Historical Archives*, pages 11–18, 2019.

19 Ti. Lan, W. Yang, Y. Wang, and G. Mori. Image retrieval with structured object queries using latent ranking svm. In *European Conf. Computer Vision*, pages 129–142. Springer, 2012.

20 Jochen L Leidner. *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names.* Universal-Publishers, 2008.

21 M. D Lieberman and H. Samet. Adaptive context features for toponym resolution in streaming news. In *Proc. ACM SIGIR*, pages 731–740, 2012.

22 M. Malinowski and M. Fritz. A pooling approach to modelling spatial relations for image retrieval and annotation. *arXiv preprint*, 2014. arXiv:1411.5190.

23 D.M. Mark. Cognitive image-schemata for geographic information: Relations to user views and gis interfaces. In *GIS/LIS*, volume 89 (2), pages 551–560, 1989.

24 D.M. Mark and A.U. Frank. Experiential and formal models of geographic space. *Environment and Planning B: Planning and Design*, 23(1):3–24, 1996.

25 Daniel R. Montello. Scale and multiple psychologies of space. In Andrew U. Frank and Irene Campari, editors, *Spatial Information Theory A Theoretical Basis for GIS*, pages 312–321, Berlin, Heidelberg, 1993. Springer Berlin Heidelberg.

26 R. Moratz and T. Tenbrink. 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, 2006.

27 D.G. Morrow and H.H. Clark. Interpreting words in spatial descriptions. *Language and Cognitive Processes*, 3(4):275–291, 1988.

28 G. Platonov and L. Schubert. Computational models for spatial prepositions. In *Proc. Workshop on Spatial Language Understanding*, pages 21–30, 2018.

29 Leonard Richardson. Beautiful soup documentation. *April*, 2007.

30 E. Rosch. Cognitive representations of semantic categories. *J. Experimental Psychology*, 104(3):192, 1975.

31 S. Schockaert, M. De Cock, and E. E. Kerre. Location approximation for local search services using natural language hints. *International Journal of Geographical Information Science*, 22(3):315–336, 2008. doi:10.1080/13658810701626277.

32 G. Skoumas, D. Pfoser, A. Kyrillidis, and T. Sellis. Location estimation using crowdsourced spatial relations. *ACM Trans. Spatial Algorithms and Systems*, 2(2):1–23, 2016.

33 K. Stock and M. Hall. The role of context in the interpretation of natural language location descriptions. In *COSIT*, pages 245–254. Springer, 2017.

34 K. Stock and J. Yousaf. Context-aware automated interpretation of elaborate natural language descriptions of location through learning from empirical data. *Int. J. Geographical Information Science*, 32(6):1087–1116, 2018.

35 L. Talmy. How language structures space. In *Spatial Orientation*, pages 225–282. Springer, 1983.

36 A. Tyler and V. Evans. *The semantics of English prepositions: Spatial scenes, embodied meaning, and cognition.* Cambridge University Press, 2003.

37 M. Vasardani, L.F. Stirling, and S. Winter. The preposition at from a spatial language, cognition, and information systems perspective. *Semantics and Pragmatics*, 10:3, 2017.

38 Jan Oliver Wallgrün, Alexander Klippel, and Timothy Baldwin. Building a corpus of spatial relational expressions extracted from web documents. In *Proceedings of the 8th workshop on geographic information retrieval*, pages 1–8, 2014.

39 D. Wu and Y. Cui. Disaster early warning and damage assessment analysis using social media data and geo-location information. *Decision support systems*, 111:48–59, 2018.

40 Z. Wu and M. Palmer. Verb semantics and lexical selection. *arXiv preprint*, 1994. arXiv:cmp-lg/9406033.

41 X. Yao and J.-C. Thill. How far is too far?–a statistical approach to context-contingent proximity modeling. *Trans. in GIS*, 9(2):157–178, 2005.

42 H. Yu and J.M. Siskind. Sentence directed video object codiscovery. *Int. J. Computer Vision*, 124(3):312–334, 2017.