International Journal of Geographical Information Science
Vol. 00, No. 00, Month 200x, 1–31

RESEARCH ARTICLE

Generating geographical location descriptions with spatial templates: a salient toponym driven approach

Mark M. Hall[a] and Christopher B. Jones[b*]

[a]*School of Computing and Communications, The Open University, UK*
[b]*School of Computer Science and Informatics, Cardiff University, UK*

(*Received 00 Month 200x; final version received 00 Month 200x*)

Natural language descriptions of geographical location are used very frequently in daily life and there is a motivation to create systems that generate such descriptions automatically, particularly for purposes such as documentation of where events have taken place, where a person is located, where photos were taken and where plants and animals are located. Typically these location descriptions combine references to named geographical features with often vague spatial relational terms, such as *near*, *north of* and *at* that relate locations to the features. Here we describe a system for generating location descriptions, in the context of photo captioning, that combines spatial templates, that model the applicability of different spatial relations relative to a reference location, with toponyms in the vicinity of the described location that are selected according to aspects of salience. The toponyms are retrieved from a gazetteer service based on OpenStreetMap for which we create a consistent hierarchical feature classification scheme to facilitate selection of toponyms according to the distinctiveness of their feature types (in addition to other aspects of salience). The advantages of the approach are demonstrated in a user study, relative to an existing state of the art system and to other baseline approaches that include manually created captions and the automated methods of two widely used photo captioning systems.

Keywords: Spatial Natural Language; Referral expressions; Locative expressions; Salience; Landmarks; Spatial Preposition Applicability Models; Vagueness; Natural Language Generation; Spatial Templates; Gazetteers; OpenStreetMap

1. Introduction

The use of spatial language relating to geographic space is commonplace in everyday speech and arises whenever we need to describe the location of things or explain how to

[*] Corresponding author. Email: jonescb2@cardiff.ac.uk

ISSN: 1365-8816 print/ISSN 1362-3087 online

DOI: 10.1080/1365881YYxxxxxxxx
http://www.informaworld.com

navigate to a location. Given the ubiquity of spatial language there is a motivation to develop automated systems that use spatial natural language to describe the geo-location of objects, such as vehicles, photos, animals and people, particularly when their location can be defined quantitatively with coordinates such as from GPS devices. Generating spatial language automatically poses a challenge however due to the often vague and context dependent variation in its use and interpretation, depending on factors such as the size and category of the things that are referred to and the associated activities or tasks (Talmy 2000, Tyler and Evans 2003, Coventry and Garrod 2004, Stock and Yousaf 2018). At present the most common commercial automation of spatial language is in mobile and desktop navigation systems that produce instructions to direct travel on road networks. In practice that language is usually confined to a small vocabulary of phrases that are relatively unambiguous, mostly concerning taking turns at junctions (though there has been research on developing richer descriptions especially those that make more use of landmarks (Dethlefs *et al.* 2011)). There are fewer examples of systems that automatically generate 'static' descriptions of locations. One example of systems that do include 'static' descriptions is photo sharing apps, which often include automatically generated descriptions of where a photo was taken. The form of the descriptions produced in such systems is however typically quite limited, being confined to expressions of the form *at X* or *near X*, possibly in combination with a comma separated geographic hierarchy of place names (see for example Geograph[1]). Here we present a system that, given as input the coordinates of a point location, exploits contextual geo-data to generate richer forms of natural language location descriptions that combine a variety of spatial relations with the possibility of multiple reference landmarks. In doing so we present methods that have the potential to be used not just in the employed case study of photo captioning but in multiple other applications where it would be helpful for geographical coordinates to be translated to natural language descriptions of location. These could include describing the location of an injured person based on their mobile device coordinates; the location of remote robotic devices; assistance on where to find bikes or cars in shared vehicle schemes; and in navigational tools that help a user to understand their present or future location, such as in self driving cars, for pedestrians, for sight-impaired people, or in locations with limited visibility.

Locational, or *locative*, expressions typically associate the location being described with a reference location by means of a spatial relation (Herskovits 1986, Talmy 2000), as for example "the car on the high street", in which the car is the located object (LO) and the high street the reference object (RO), associated with the spatial relation of *on*. Alternatively the located object can be implicit as when describing a scene simply as "on the high street". Often such spatial relations are composed into more complex expressions, as in "on the high street, near the church, in Chipping Norton". In the present work we generate expressions that can have multiple spatial relations and reference objects. Our application is that of photo captioning in which the located object is implicitly the camera location. The approach is multi-faceted in that it is comprised of the three main tasks, of selecting suitable toponyms (to serve as reference objects), modelling the applicability of a set of spatial relations in order to choose the most appropriate spatial relation for a given spatial configuration, and generating a complete natural language expression composed of one or more spatial relationships. The approach only uses the geo-tagged photo's location data as its input and does not require any image analysis. Selection of appropriate reference toponyms is governed by a procedure to rank toponym

[1] http://www.geograph.org.uk/

salience based on factors that include their proximity to the photo location, their distinctiveness with regard to their name and feature type, and their popularity as reflected in frequency of use on an existing photo sharing web site. We implement a gazetteer service to find candidate names that employs OpenStreetMap[1] (OSM) as the source but, to overcome the somewhat arbitrary feature classifications scheme in OSM, we introduce a consistent hierarchical feature classification to which OSM names are mapped. Selection of appropriate spatial relations adopts a spatial template approach (Schirra 1993, Logan and Sadler 1996, Mukerjee *et al.* 2000) in which each spatial relation is represented by a density field (template) that is anchored on the reference toponym. An individual spatial relation is preferred when the density field value at the photo location is higher for the selected spatial template than for alternative spatial relations. The structure of the composed location descriptions combines simple patterns of the form identified by Hall *et al.* (2011) in their analysis of photo caption language. We evaluate and demonstrate the benefits of the presented method using a human subject experiment that compares our method with an existing state of the art approach as well as with manually and automatically generated expressions from the Flickr[2] and the Geograph photo sharing websites.

The work reported in this paper investigates the following two hypothesis:

H1 Complex, natural-sounding location descriptions can be generated fully automatically, outperforming current automatic methods.

H2 While human-generated descriptions of photo locations are typically short, more detail is preferred by photo caption users.

The rest of the paper is structured as follows: section 2 covers background and related work relevant to our contribution; section 3 describes our methods, of retrieving toponyms including an explanation of the creation of a feature classification hierarchy using OpenStreetMap data; computing toponym salience; and generation of the natural language location description with spatial templates. Our human-subject evaluation is presented in section 4, while section 5 presents conclusions.

2. Related work

2.1. *Spatial language*

The form and components of location descriptions (referred to also as locative or locational expressions) have been described in several studies including Herskovits (1985), Levinson (2003), Talmy (1983) and Tenbrink (2011). Location descriptions are commonly regarded as composed of triple structures of a located object (or trajector or locatum), a spatial relation and a reference object (or landmark or relatum), but there are several possible frames of reference to which the expressions can conform and which are reflected in the choice and interpretation of spatial relation. Particular types of frames of reference, described by Levinson (2003), are *relative*, *absolute* and *intrinsic*. In the *relative* frame of reference a location is defined relative to the location of a reference object, for example “near Paris” in the case of a proximal (distance-based) spatial relation, or “to the left of the church” in the case of a directional (projective) relation that here is from the point of view of the observer. Spatial relations that depend upon an

[1] https://www.openstreetmap.org/
[2] https://www.flickr.com/

observer, such as the latter directional relation, can be referred to as *deictic*. In the *absolute* frame of reference, some property of the environment is used to provide orientation, with the most common case being the the cardinal directions based on the earth's axis. An example would be "east of Paris". The *intrinsic* frame of reference uses properties of the related object to define direction. Thus a person or a house for example can be regarded as having a front, i.e. the direction that they face, giving rise to an example such as "in front of the house" where "in front" is relative to the orientation of the house (as opposed to that of an observer). As Tenbrink (2011) has pointed out, the frames of reference described by Levinson (2003) are generally regarded as applicable to external spatial relations in which the located and reference objects are separate from each other. Cardinal directions are quite often used to refer to parts of a region, e.g. "in the north of Paris" which can be described as an *internal* frame of reference, as opposed to "north of Paris" which would be an external relation (Tenbrink (2011). In the current paper we generate location descriptions that use *relative* and *absolute* but not *intrinsic* frames of reference. All relations are also *external*, i.e. separate located and reference objects. The location of the photo is always treated as the located object and is therefore implicit in the description. We do not employ *deictic* or *intrinsic* locations relations as we do not maintain data on the orientation of the camera (although some cameras do provide that possibility) or on the intrinsic directional properties of reference objects (though again in certain cases it might perhaps be possible to deduce that information).

2.2. *Applicability of spatial relations*

A key aspect of the method adopted here to generate location descriptions is the use of applicability models of individual spatial relations that are used to decide which spatial relation should be applied in a given situation. The approach is analogous to that of (Schirra 1993) who used a set of density fields for spatial prepositions including *in*, *near*, *left of*, *at* and *in front of* to select a particular spatial relation based on anchoring the density field on the reference object and determining the value of the field at the location of the located object. The spatial relations that provide the highest values of their density field were used in the natural language description. The application was that of describing a football game and the density fields were generated procedurally without specific empirical evidence for their form, whereas in our work they are based on human subject studies and on evidence of actual uses in the context of photo captioning.

The principle of selecting a spatial relation on the basis of its level of applicability can be found in several subsequent works, for example that of Gapp (1995) who studied projective (directional) relations which were modelled on the basis of human-subject experiments that determined the nature of an angular decay function relative to a prototypical direction. Note that the concept of deviation from a prototypical relation was introduced in Herskovits (1985) but was not accompanied by methods to generate specific models. Human subject experiments were used in Mukerjee *et al.* (2000) to create density field models of the relations of *between*, *near*, *far*, and *to the right of* in which linear regression functions were fitted to trends in the empirical data. The models were used in the interpretation of spatial language. Another early example of human subject experiments to create density fields, though without applying them, can be found in Logan and Sadler (1996) who introduced the concept of a spatial template which models applicability, and for which a density field would be an approach to its representation.

These latter examples relate to very localised, so-called table top space (as opposed to our interest in geographic space), and there are many examples of subsequent work at the same very localised scale in the context of robotics. An example is Kelleher and Costello (2009) who adapted the spatial template concept to model proximity with respect to salience of landmark objects as well as distance to them. The applicability of a spatial relation at a given location relative to a particular landmark then depends on proximity to other candidate landmarks, referred to as distractor objects. Our approach is analogous in that when describing a location we take account of all candidate landmarks in the vicinity and select them based both on proximity and other aspects of salience. In the context of robotics the applicability fields have been modelled in various ways including for example human-subject based assertions of relations between multiple generated configurations of table top items, and enhanced datasets based on these observations (Kunze *et al.* 2014), and an interactive environment in which a robot uses human feedback progressively to learn models of spatial relations that it cannot recognise (Spranger and Steels 2015) (see also Spranger *et al.* (2016), Spranger and Pauw (2012)). The tutor learning approach of Spranger and Steels (2015) was also applied to learning grammatical constructs including noun phrases relating to spatial locational expressions, with the intention to understand complete sentences. An example of the use of spatial templates for image retrieval tasks that employ spatial relations is found in Malinowski and Fritz (2015), while the concept of spatial templates underlies the work of Collell *et al.* (2018) that uses word embeddings to estimate the location of objects as inferred by natural language descriptions that employ words of action.

Examples of the use of spatial templates in a geographical context are found in Hall *et al.* (2011), in which they use them to infer the locations referred to by vague spatial relations, and Hall *et al.* (2015) in which they are used in a system to generate natural language location descriptions describing the locations at which photos were taken. Given a photo location and a set of neighbouring named geographic features Hall *et al.* (2015) use the spatial templates (referred to there as field models) to select the spatial relation that is most applicable given a combination of candidate neighbouring, and salience-ranked, toponyms to which the spatial relation could refer. The templates were created from data obtained in human subject experiments relating separately to urban and rural contexts in which participants were are asked to select, and rate the level of applicability of, the most appropriate spatial relation to describe a location relative to a reference location. The spatial relations considered were *north of*, *south of*, *east of*, *west of*, and *near* for the rural experiment, and *near*, *north of*, *next to*, *at*, *at the corner* and *between* for the urban experiment. In the rural experiment kriging was used to interpolate data to create a density field (i.e. the spatial template), while in the urban experiment linear spline functions were fitted to distance-based values of applicability and assumed, for *next to*, *at*, and *at the corner*, to apply equally in all directions relative to the reference location. Due to data sparseness, the urban models of cardinal directions were scaled versions of the rural models. In the present experiments we obtained and applied the applicability models, i.e. spatial templates, used in Hall *et al.* (2015) which facilitated a more direct comparison with our methods in our evaluation experiments.

2.3. *Referral expressions*

The process of generating location descriptions, as presented here, is a particular case of generating referral expressions, for which a widely adopted approach is described by Dale and Reiter (1995). They present an incremental algorithm that is designed to ensure that the referred object is distinguished with descriptive attributes from any other objects in the context with which it might be confused. The algorithm iteratively selects values of attributes that are most discriminative, in that no other value of the attribute would eliminate more potentially ambiguous objects. An adaption of Dale and Reiter (1995) that addresses the problem of determining the salience of objects is presented in Krahmer and Theune (2002). Our approach differs somewhat from the latter studies in that, while we seek to create a description that is precise, there is no sense of a discourse in which objects might previously have been mentioned, and our objects are geographical, for which we present specifically adapted salience criteria.

Several aspects of salience of geographical objects have been identified, notably distinguishing between visual, structural and cognitive (Sorrows and Hirtle 1999), with specific visual, semantic and structural properties being proposed for example in Raubal and Winter (2002). Some of these properties are motivated by the context of wayfinding and navigation. These are highlighted in Lynch (1960) which considers more wide ranging aspects of perception of urban environments and identifies key elements of path, landmark, edge, node and district. Several of these latter structural elements are particularly relevant to our task of static location description, in which for example named buildings and streets, and junctions between streets, can be essential in providing accurate description of a specific static location in combination with the regional context provided by districts. In addition to the need to refer to specific objects that are coincident or very close to the location to be described, it is important to retain distinguishing properties, as mentioned with regard to the approach of Dale and Reiter (1995), but also emphasised in the accounts of geographic salience (see also Golledge (1993)). Our salience measures therefore emphasise proximity for purposes of accuracy, along with distinctiveness as measured here with respect to names and to geographic feature types. We also introduce an element of perceptual salience by paying attention to whether users of the photo sharing web site Flickr have chosen to select a given place name or landmark.

As in some robotics applications of referral expressions (Kelleher and Kruijff 2006) the spatial relationships play a key role here in distinguishing the located object (in association with salience). We select spatial relationships on the basis of those that are most applicable to the configuration of located and reference objects and, similar to the approach presented in (Hall *et al.* 2015), we generate prepositional phrases that are concatenated to create complete location descriptions. Our approach differs significantly from Hall *et al.* (2015) in that we choose the most appropriate toponyms (from a set of candidate toponyms) to act as reference locations, before then determining the most applicable spatial relations, as opposed to what we refer to as the preposition-driven approach of Hall *et al.* (2015) that selects spatial relations before toponyms. Notably a limitation identified in their evaluation was the fact that the algorithm attempts to find a single toponym that provides both spatial accuracy and high toponym salience. This latter system is similar in its objectives to the approach that we adopt here and we treat it as one of several baselines in our experimental evaluation.

3. Location Description Generation System

The Location Description Generation System (LDGS) we present here consists of two main components: the gazetteer, which provides the geo-data that underpins the location descriptions (LD) and the Location Description Generator (LDG), described in the next section, which generates the LDs based on the geo-data provided by the gazetteer. Figure 1 shows the components and data-flows within the system.

INSERT FIGURE 1 images/system_overview.png HERE

3.1. *Gazetteer*

The gazetteer that provides all geo-data for the LDGS is built on top of the OpenStreetMap (OSM) data. It is only loosely coupled to the LDG and the LDG can also use other gazetteers. The gazetteer provides the following pieces of data for a given pair of co-ordinates f:

- urban / rural distinction - whether f is an urban or rural location;
- containment toponyms - the hierarchy of containment toponyms within which f lies;
- proximal toponyms - toponyms that lie within a specific distance of f;

For both the containment and proximal toponyms the gazetteer provides classification functionality and for the proximal toponyms salience values are calculated.

Due to the loose coupling, it is possible to use the LDG with a different gazetteer. This alternative gazetteer would have to be able to provide the three pieces of data listed above, prior to computation of the salience values. Additionally, the alternative gazetteer's classification would have to be mapped into our gazetteer's classification structure.

3.1.1. Urban / rural distinction

To distinguish urban and rural locations, the gazetteer uses a heuristic based on the OSM geo-data. If there is a feature in the OSM data that lies within a given threshold distance from f and is classified as a building, then the location is classified as urban, otherwise it is classified as rural. This is a very liberal definition of urban and is driven by the requirements of the LDG. The LDG uses separate spatial templates (density field models) for urban and rural areas, but, importantly, the urban spatial templates are based on buildings and can thus be used anywhere there is a building available as the reference object. Due to this flexibility, the gazetteer indicates urban whenever there is the possibility of a building serving as a reference object, for which the urban spatial templates are appropriate.

The threshold distance is configurable, but for the LD generation scenario, a threshold of 400 metres was chosen. It refers to the shortest distance between the feature geometry and the point location f. The threshold choice is based upon the maximum distance used in the urban spatial preposition applicability models (i.e. the spatial templates) that are described in sections 2.2 and 3.2.

3.1.2. Containment toponyms

The gazetteer first generates the containment toponyms, before finding proximal toponyms, as the proximal toponyms' salience calculations require at least one containment toponym.

The containment toponyms are selected from the OSM data by retrieving all polygon areas that contain f. As it cannot be guaranteed that each containment toponym's area is fully contained within the larger containment toponyms (for example a national park can have parts in various counties), the containment toponyms are instead sorted by their size to create the toponym hierarchy. This is appropriate for the LD generation task, as the containment toponyms are essentially ordered by how precisely they describe the location f.

Table 1. Example OSM Data for the Wales Millennium Centre (https://www.openstreetmap.org/way/26584146), showing the generic and instance-specific keys used to define the meta-data

Generic Keys	Value	Instance-specific Keys	Value
amenity	theatre	addr:city	Cardiff
building	yes	name	Wales Millennium Centre
		name:cy	Canolfan Mileniwm Cymru
		wikidata	Q2631977

3.1.3. Proximal toponyms

The proximal toponyms are selected from the OSM data by retrieving all toponyms for which the shortest distance between the toponym and the location f is less than a given threshold distance. Due to the nature of OSM, proximal toponyms will be a mixture of point, linear, and polygonal features.

For proximal toponyms the gazetteer provides an optional filtering step. When filtering is enabled, all proximal toponyms that have a name that also occurs in the list of containment toponyms are removed from the proximal toponyms. This is to enable avoidance of repetition of a name in generated location descriptions. Filtering is only based on the name, as in OSM the same feature can be represented with different geometry types, most frequently as both a point and a polygon (e.g. for cities).

3.1.4. Classification

OSM defines a feature's attributes via sets of key-value pairs (Table 1). Some of these specify attributes are specific to that instance (its name, address, wikipage, ...), while others describe the type of feature (building, railway, ...). A challenge in using OSM is that due to its crowdsourced nature there is a vast range of different attributes and values for describing feature types and often features of fundamentally the same type are described using different attribute-value combinations.

The classification layer in our gazetteer addresses two aspects of this issue. It implements a hierarchical classification structure that provides higher levels of feature type abstraction and it maps different, but semantically equivalent, feature type definitions in the OSM data into a single feature type in the classification system.

3.1.4.1. Classification hierarchy generation. The classification hierarchy was built using a mixed top-down and bottom-up strategy. In an initial top-down step, the very high-level structure was created by using the CORINE land-use/land-cover classification system[1] to define the top-level feature types.

Next, using a bottom-up approach, specific feature types were extracted from OSM. The scale of different feature types and the manual nature of the bottom-up process means that it was impossible to initially create feature types in the classification hierarchy for all OSM feature types. Instead we focused on modelling the most common feature types first. To determine the most common feature types we processed the OSM UK data-set. For each feature in the data-set we removed those attributes that defined instance-specific aspects (such as *name: Cardiff Central* that refers to a particular train station), leaving only those that defined generic feature type information (such as *building: train*

[1] https://www.eea.europa.eu/data-and-maps/data/clc-2006-raster-4

station). These were then sorted by frequency and the top 1000 were included in the classification hierarchy.

For each of the top 1000 attribute pairs we manually identified the type of feature and created the appropriate feature type in the classification hierarchy. In order to link the feature type entry to the OSM data, the set of key-value pairs used in the OSM data is stored with the feature type in the classification hierarchy. If the manually identified feature type already exists in the classification hierarchy, then the corresponding OSM key-value pairs are added to the list of those already stored with the feature type. This approach enables us to map the large amount of variation in how features are described in OSM into a much more manageable set of feature types.

In the final step the feature types created in the previous step are added under the appropriate CORINE top-level feature types. The CORINE feature types are very high-level, thus additional mid-level feature types were added between the CORINE types and the specific feature types. For example CORINE has a feature type "Road and Rail Network and associated land". Below this we added two feature types "Road" and "Rail" and below these we placed the specific feature types identified in the OSM data, for example the feature type "Railway station".

This process enabled us to create a classification hierarchy that is both specific enough to represent the detail available in the OSM data, but at the same time provides generic feature types for those aspects of the LD generation that do not require that level of detail.

3.1.4.2. OSM feature type mapping. After retrieving containment and proximal features, each feature is assigned a feature type from the classification hierarchy. To do this, the feature's key-value attributes are filtered using the same set of rules as when creating the classification hierarchy. The remaining key-value pairs are then compared to all the key-value sets stored with each of the feature types in the classification hierarchy. If the key-value set stored with a feature type is a sub-set of the key-value pairs of the feature, then the feature type is marked as a candidate feature type.

If after processing all feature types there are multiple candidate feature types, then the following set of rules are used to determine the final feature type. First, the candidate feature types that are deepest in the hierarchy are selected, as these are more specific and thus more precise descriptors of the feature. If there are multiple candidate feature types at the same depth, then the feature type with the largest number of key-value pairs is used. Again, this is because this is the more specific feature type and thus preferred. Finally if there are still multiple candidates, then one is selected randomly. For performance reasons the classification is then cached with the OSM data.

When the gazetteer returns proximal and containment toponyms, for each toponym its feature type and also all its ancestor feature types are returned. The advantage of this is that down-stream processes, such as the LDG, can use any level of abstraction in the feature type hierarchy, without having to themselves load and process the classification hierarchy.

3.1.5. Salience calculation

The final processing step the gazetteer provides is to calculate the salience of the identified proximal toponyms. Five types of salience are calculated to support a wide range of requirements: Flickr, Name, Feature, Distance, and Error Distance. The first three will be referred to as feature-related saliences, as they derive from the features themselves, while the Distance and Error Distance metrics are referred to as spatial-related saliences, as they derive from the spatial information.

All salience metrics produce values in the range $[0, 1]$, with 0 the least salient and 1 the most salient. The salience values are calculated relative to the set of identified proximal toponyms and can only be used to compare between those. This means that even if two image locations f_1 and f_2 produce partly overlapping sets of proximal toponyms, the toponyms that appear in both sets may have different salience values and no comparisons can be made between the salience values of the two sets.

3.1.5.1. Flickr salience. The Flickr salience is used to distinguish landmarks from other toponyms, based on the heuristic that well-known toponyms are photographed and labelled more frequently. To determine a toponym's Flickr salience, the gazetteer queries Flickr's API for photographs within a given threshold of the candidate toponym's location that are tagged with the same name as that toponym. The number of these photographs is then cached with the OSM data for performance reasons. For urban areas the threshold was set at 400 metres and for rural areas 3 kilometres. The distances were based upon the maximum distances used in the urban or rural spatial preposition acceptability models (see Hall *et al.* (2015)).

The absolute photo counts are normalised to the $[0, 1]$ range. There is one exception to this and that is features that relate to transportation (bus stops, train stations, ...). These often share the name of landmarks, but are less likely to be the focus of the photographs and should thus not share in the actual landmark's salience. For these feature types the Flickr salience is always set to 0. This very simple solution obviously does not provide the correct salience value for those transportation features that are themselves very salient and well-known landmarks (for example the Grand Central Terminal in New York). However, it does not prevent them from being selected and, as they are the exception, overall the solution produces less incorrect salience values.

3.1.5.2. Name & feature salience. For many toponyms the Flickr salience is 0, as they have never been photographed, making a distinction between these toponyms difficult. To address this the gazetteer calculates two further salience metrics based on the heuristic that the more unique a toponym is, the more salient it is. These two salience metrics also provide a fallback if the Flickr data were to become temporarily or permanently unavailable in the future.

To calculate the name salience the gazetteer retrieves all toponyms P that are contained in the smallest containment toponym and have the same name as the toponym that the salience is being calculated for. The toponym's salience is then calculated as $\frac{1}{|P|}$. Thus it has a salience of 1 if it is unique and the salience approaches 0 the more toponyms there are with that name. For example the Wales Millennium Centre is unique in Cardiff and thus gets a value of 1, while the nearby Sainsbury's supermarket (part of a chain) gets a value of 0.09, as there are a total of 11 Sainsbury's supermarkets in Cardiff.

The same process is performed for the feature salience, measuring the salience the same way by retrieving all toponyms in the smallest containment toponym that have the same feature type as the toponym for which salience is being calculated.

3.1.5.3. Distance & error distance salience. Two distance salience values are calculated. The main distance salience s_{dist} is calculated as the shortest distance from the co-ordinates f to the toponym t and represents the heuristic that a toponym that is closer is a more precise description of the location. The salience values are normalised to $[0, 1]$ using equation 1 (d_t is the distance from f for the toponym t; D is the set of distances for all proximal topnyms). This gives the closest toponym a salience value of 1, and the most distant toponym a salience value of 0.

$$s_{dist}(t) = \frac{d_t - min(D)}{max(D) - min(D)} \tag{1}$$

The second distance metric, the error distance, is calculated as the distance from f to the most distant vertex of the toponym t. This models the heuristic that large features, even if close, also contain a large number of points that do not describe the location f well. Common examples for this case are rivers and lakes. Values are normalised to $[0, 1]$ using equation 2 (ed_t is the error distance from f for the toponym t; ED is the set of error distances for all proximal topnyms).

$$s_{errdist}(t) = 1 - \frac{ed_t - min(ED)}{max(D) - min(ED)} \tag{2}$$

3.2. *Generating location descriptions*

To generate the location descriptions the toponyms provided by the gazetteer are combined with the preposition-specific spatial templates and the resulting spatial configurations turned into a natural language representation. In this section we present our novel algorithm for performing the combination. It will be referred to as the *toponym-driven* algorithm, as it starts with picking good toponyms and only then adds the spatial preposition. Note that the spatial templates that model the applicability of spatial prepositions are the same ones as were used by Hall *et al.* (2015) as described in the related work section. They represent the spatial relations of *north of*, *south of*, *east of*, *west of*, *near*, *next to*, *at*, *at the corner* and *between*. Different scaled versions of the models are used according to whether the context is determined to be urban or rural, using the approach described in section 3.1.1.

3.2.1. Creating and selecting spatial templates

We describe here the process of creating the spatial templates. An illustration of the creation of a spatial template from point-referenced data is given in Figure 2, where the points represent the locations of examples of the use of the spatial relation ‘north of’ relative to a location that is at X in the figure. The density field on the right is obtained by applying kernel density estimation to the data points on the left. Its field strength at a given location can be regarded as a measure of the applicability of the spatial relation at that point relative to a reference toponym at position X. In Figure 3 we illustrate the process of selecting a spatial template that is applicable for given combinations of photo location and candidate toponym location. We see that a reasonably high density field value is obtained when a ‘north of’ template is anchored on toponym T4 but, for example, a very weak (or zero) value is obtained when it is anchored on T3. Thus ‘on T4’ could be selected to describe the location of P. A similar process is illustrated for street spatial templates in Figure 4, where stronger field values for photo position P are obtained for the spatial relations ‘on R4’ and ‘between R1 and R3’ both of which can be regarded as applicable. Note that the ‘between’ template is anchored on a pair of adjacent point-referenced junctions (such as J2 and J4 in the figure) and adjacent road names are then selected for the location description. Note also that the 'on' template is anchored on a line rather than a point. The 'at the corner' template is anchored on a single point corresponding to a junction.

INSERT FIGURE 2 HERE images/SpatialTemplates1.png

INSERT FIGURE 3 HERE images/SpatialTemplatesSelectingNorth.png

INSERT FIGURE 4 HERE images/SpatialTemplatedStreets.png

3.2.2. The location description algorithm

The algorithm does not directly generate natural language location descriptions, instead it generates an intermediate representation, which is then rendered into natural language. The advantage of this is that the intermediate representation can be rendered into multiple languages and that changes to the location description algorithm do not impact the language rendering, enabling future extensions and application to other domains.

INSERT FIGURE 5 images/caption_pattern.png HERE

The overall location description is based on the same patterns as those identified in Hall *et al.* (2011) and is shown in Figure 5. Algorithm **1** shows a high-level abstraction of the captioning algorithm. The proximal toponyms P and containment toponyms C are fetched from the gazetteer and the list of spatial relations R initialised depending on whether the photo location is rural or urban (1 - 7). The spatial relations R are ordered by the size of the area they describe, in order of increasing size. If a road element can be generated, then it is added to the location description LD and the list of proximal toponyms is sorted by feature-salience, otherwise P is sorted by spatial-salience (8 - 13). The main loop (14 - 20) runs while there is at least one proximal toponym and one spatial relation that could still be used. It initially generates a selected proximal toponym SP and spatial relation SR and adds these to the location description. Then it re-ranks the toponyms P by their feature-salience and filters out all toponyms that are less salient than the selected toponym SP. Finally the list of spatial relations R is filtered (19), removing all spatial relations that describe areas smaller than SR. After generating the proximal elements, the location description is completed by adding a sub-set of the containment toponyms (21). We will now present each of these steps in detail.

Algorithm 1:

```
1: $P, C \leftarrow from\_gazetteer()$
2: $LD \leftarrow []$
3: if rural then
4:     $R \leftarrow [northof, southof, eastof, westof, near]$
5: else
6:     $R \leftarrow [at, nextto, near]$
7: end if
```

```
if can_generate_road_element(P) then
    $LD \leftarrow LD + road\_element(P)$
    $P \leftarrow rank\_by\_feature\_salience(P)$
else
    $P \leftarrow rank\_by\_spatial\_salience(P)$
end if
while $|P| > 0 \wedge |R| > 0$ do
    $SR, SP \leftarrow generate\_proximal\_element(P, R)$
    $LD \leftarrow LD + (SR, SP)$
    $P \leftarrow rank\_by\_feature\_salience(P)$
    $P \leftarrow filter\_less\_salient(P, SP)$
    $R \leftarrow filter\_spatial\_relations(R, SR)$
end while
$LD \leftarrow LD + select\_containment\_elements(C)$
```

3.2.3. Road elements

The road element generation logic initially checks whether there is a junction that could be used with the *at the corner* spatial-preposition model and, if not, attempts to find a road for which *on* is applicable. To determine whether an *at-the-corner* element can be generated, the algorithm first generates all intersections between all roads returned by the gazetteer that have distinct names. The check for distinct names is required due to the OSM data splitting linear features into multiple features to improve performance when rendering the data to a map. After generating all intersections, the *at-the-corner* model is instantiated for each intersection and the applicability at the image location calculated from the spatial template. Note that an applicability value for a particular spatial relation is obtained by anchoring its spatial template (density model) on the reference toponym location and finding the density value at the photo location. If the applicability is $\geq \frac{2}{3}$ then the intersection is used to generate an *at-the-corner* element. The limit of $\frac{2}{3}$ was chosen as the applicability data was acquired on a 9-point scale in Hall *et al.* (2015) and applicability values $\geq \frac{2}{3}$ represents applicability ratings of 7 or higher, indicating high agreement that the location is 'at the corner of' the junction. If multiple junctions are applicable, then the one with the highest applicability value is chosen.

If no *at-the-corner* element was generated, then the algorithm will attempt to generate an *on* element. As the road features in OSM are represented as linear features with no width, the algorithm checks whether there is a road within $10m$ of the image location and if so, generates an *on* element for that road. The limit of $10m$ was determined empirically and is derived from the average width of a two-lane road with pavement in the UK, including a buffer based on testing with a number of commercial GPS receivers.

3.2.4. Proximal description generation

Our algorithm can generate multiple proximal elements, focusing first on spatial salience and then on feature salience. To achieve this, the algorithm uses five salience metrics (see 3.1.5) provided by the gazetteer to rank all toponyms P. To calculate an overall salience for each toponym, the individual salience values are combined using a weighted sum. By varying the weights, the ranking can be weighted more towards the spatial or feature salience values.

Three sets of weights have been determined (Table 2). When ranking with a focus of spatial salience, urban and rural contexts are distinguished, in that the Error Distance is less important in the rural context due to the generally larger distances between features

Table 2. Salience weights used by the proximal description generation. For the *Spatial salience*, 60% of the weighting is on the spatial salience metrics *Distance* and *Error Distance*, while for the *Feature salience* 90% of the weighting is on the feature salience metrics (*Flickr*, *Name*, and *Type*). For the *Spatial salience* a distinction is made between the *urban* and *rural* contexts, as in the *rural* context the distances between features and the image location are generally larger, thus the *Error Distance* weight is lower as there is less need to down-weight larger toponyms.

Mode	**Context**	**Distance**	**Error**	**Flickr**	**Name**	**Type**
Spatial salience	Urban	.45	.15	.3	.05	.05
	Rural	.55	.05	.3	.05	.05
Feature salience	Urban & Rural	.1	0	.8	.05	.05

and image location. For the feature salience no distinction between urban and rural is made.

As can be seen in Algorithm **1** the proximal toponyms are initially ranked using the spatial salience weights ($rank\_by\_spatial\_salience$) unless the location description contains a road element, which provides spatial accuracy, in which case they are ranked using the feature salience weights ($rank\_by\_feature\_salience$). The highest ranked toponym is selected and combined with the most specific spatial preposition that has an applicability score for the toponym and the image location $\geq \frac{2}{3}$. If the highest ranked toponym has no applicable spatial preposition, the next toponym is tested and so on until all toponyms have been tested or one has an applicable spatial preposition ($generate\_proximal\_element$). After finding a toponym / spatial-preposition pair, a proximal caption element is generated for that pair and added to the location description.

Next the toponyms P are re-ranked using the salience weights ($rank\_by\_feature\_salience$). Then all toponyms that are ranked below the selected proximal toponym SP are filtered out ($filter\_less\_salient$). Similarly all spatial prepositions that describe areas smaller than the selected spatial preposition SR are also filtered ($filter\_spatial\_relations$). This ensures that each element in the location description is more generic than the previous one. The process of adding toponym / spatial-preposition pairs is repeated until either there are no further toponyms P or spatial relations R.

The effect of this algorithm is that initially a toponym will be selected that balances spatial and feature salience, but where, if none can be found that satisfies both, the preference is for spatial salience. Then, in the second step, if the first toponym has a low feature salience a second toponym can be added that provides a highly feature-salient reference point. Using this approach the location description will contain one or more toponyms that provide both local spatial accuracy and also a higher recognition factor for people less familiar with the local area.

The preposition applicability models (i.e. spatial templates) from Hall *et al.* (2015) that are used here were constructed using point-referenced data. However, to incorporate the line and polygon data that is becoming more widely available from OSM, when line or polygon features are used with a preposition applicability model, the point on the feature that is nearest to the photo location is used to access the spatial template (and hence retrieve the applicability value at that location in the density field).

3.2.5. Containment description generation

Testing during the algorithm's development indicated that the containment hierarchies associated with proximal toponyms frequently contained more information than necessary ('City of London, London, England') or contained duplicate information ('Cardiff,

Cardiff, Wales‘). To avoid this three heuristics were developed that aim to create as concise a location description as possible. These are applied in the following order:

Duplicate containment areas Where two containment areas have the same name, for example in the case of a city and administrative area, the larger of the two areas is filtered out. The smaller location is kept as it provides more spatial accuracy, while having the same salience as the larger area.

Minimal size difference Where a containment area is less than 25% larger than the next smaller area, the larger area is filtered, based on the assumption that the larger area adds little information to the location accuracy.

Uniqueness filtering Where the first containment area is unique within the third, the second containment area is filtered. This heuristic is applied repeatedly until either the condition does not hold or the number of containment toponyms has been reduced to three. The exception is if the first containment toponym is a national park or the first is an administrative district and the second a county, in which case the number of containment toponyms can be reduced to two. These exceptions handle the cases of national parks and large cities, which are unique within the nation.

Together these will result in containment hierarchies that generally contain between two and three levels and are as concise as possible.

3.2.6. Language generation

The location description LD generated by the main algorithm (Algorithm **1**) consists of a list of caption elements, each a combination of a spatial preposition type and one or two toponyms. The final step in the generation of the location descriptions is to convert these into cohesive natural language expressions. For each caption element the appropriate English-language spatial preposition representation is generated as shown in Table 3, with the exception of *in*, where further processing is applied. If the toponym used in the caption element is a water feature, beach, bridge, road, or path then ‘on’ is generated instead of ‘in’ and for locations contained in buildings ‘at’ is used. Additionally if the English-language representation generated for the previous element was ‘in’ or a comma ‘,’ and the current representation is also ‘in’, then a comma ‘,’ is output instead (‘in Edinburgh, Scotland’).

When rendering a toponym, the generator has to determine whether the toponym has to be prefixed with the definite article ‘the’. This uses a heuristic approach based on English grammar rules, extended with some OSM-specific rules (see Table 4). The rules are applied greedily in the order shown, with a default behaviour of using the definite article.

The use of the intermediate representation also makes it possible to generate location descriptions in other languages. However, there are differences in how spatial language is used in different languages (Grabowski and Miller 2000, Beller *et al.* 2005). Thus, in order to generate LDs in other languages, it would be necessary to create language-specific spatial-preposition models, as well as additional language generation modules for each language that should be supported.

Table 3. Mappings from the various spatial preposition types used in the location descriptions into English-language representations. For the spatial preposition type *in* further processing is applied to select one of the four possible English-language representations (see section 3.2.6)

Spatial Preposition Type	**English Language Representation**
northof	'north of'
southof	'south of'
eastof	'east of'
westof	'west of'
near	'near'
at	'at'
nextto	'next to'
at-the-corner	'at the corner of'
on	'on'
in	'in', 'on', 'at', ','

Table 4. Heuristic rules for the use of 'the' with a toponym. Column 2 indicates whether for a toponym that matches the rule in column 1, the toponym is prefixed with 'the'. Rules are applied in the given order. If no rule applies, then by default 'the' is generated for the toponym.

Rule	**Include 'the'**	**Example**
Already starts with 'the'	No	at the Butcher's Arms
Is a populated place	No	in Cardiff
Includes 'kingdom', 'states', 'republic', or 'borough'	Yes	in the United Kingdom
Is a ceremonial county	No	in Cheshire
Is an instance of flowing water	Yes	next to the Thames
Is a natural feature	No	near Pen-Y-Fan
Is a reservoir	No	at Pontsticill Reservoir
Is a road with just a number	Yes	on the A475
Is a road	No	on Queen Street
Is a commercial or educational building	No	next to Oxford University
Is an airport terminal or railway station	No	at Cardiff Central
Is a park	No	in Bute Park

4. Evaluation

To determine whether the *toponym-driven* location description algorithm achieves its goal of generating more flexible, natural-sounding location descriptions, it has been evaluated using a set of captions drawn from Flickr and Geograph. The evaluation uses the complete natural-language descriptions generated by our system. The algorithm is compared against human-created location descriptions, location descriptions generated using the system described in Hall *et al.* (2015), and the automatically generated location descriptions provided by both Flickr and Geograph.

4.1. *Data set*

To ensure that the evaluation results are reliable, an evaluation data-set has been created by drawing test images from both Flickr and Geograph. The two sources represent different types of images, with the Geograph images focusing very heavily on location, while the Flickr images represent location descriptions from a more every-day type of spatial language use.

To generate the test data-set, ten test areas were selected in the UK, five rural (Snowdonia, Lake District, Cotswolds, Peak District, Dartmoor) and five urban (Edinburgh, Oxford, London, Cardiff, Liverpool) with a wide geographic spread. For each test area both Flickr and Geograph were searched for all images that contained at least one of the spatial prepositions that the algorithms use: near, at, next to, north, south, east, west. This was done to ensure that the spatial language used is comparable between the human- and algorithm-generated location descriptions. An additional manual filter was applied to the Flickr images to remove images where the image co-ordinates differed significantly from the location description in the captions. This was necessary as the accuracy of the location meta-data in Flickr is very variable, sometimes off by hundreds of metres. A Flickr image was only included in the test data-set if a manual analysis of the image content indicated that it was highly likely the image was within 100m of the location specified by the image coordinates.

From the resulting set of 162 images, for each test area one Flickr and one Geograph image were selected randomly. One rural location was discarded as its human-generated caption referred to a toponym for which no spatial data could be found. This location was replaced with another randomly selected image.

For each test location the gazetteer was used to fetch all proximal and containment features and to identify all feature types that were not represented in the gazetteer's classification scheme. These were then added to ensure that all features in the test locations could be mapped to types and thus are available to the location description algorithms.

For each test location four location descriptions were included in the evaluation: our novel *toponym-driven* location descriptions, the *human-generated* location descriptions, *preposition-driven* location descriptions using the method described in Hall *et al.* (2015), and an automatic *baseline*. The human-generated location descriptions were created by selecting those parts of the original image captions that focused on the location description. The caption parts that describe the content of the image were removed as they represent an element of the caption that the location description algorithms cannot recreate. The human-generated descriptions represent the target quality that the automatic location description algorithms should achieve. For the *preposition-driven* location descriptions, place names were obtained from our OSM gazetteer. The *baseline* location descriptions were sourced from Flickr and Geograph. Both provide basic, automatically generated location descriptions for each image that represent the minimum that our algorithms need to achieve. In the case of Flickr the location description created automatically by Flickr is the given location and its containment hierarchy, while Geograph seems to select a nearby toponym and adds the containment hierarchy. Tables 5 and 6 show the location descriptions generated for the Flickr and Geograph evaluation locations.

4.2. *Evaluation Procedure*

The evaluation used human participants to evaluate the captions using a web-based platform mixing standard survey and crowd-sourcing techniques.

Participants initially confirmed their informed consent[1] and then had to provide demographic information (age, gender, language skills, education, employment status). These will be used to characterise the participant cohort. Participants were then given a set of instructions before being shown the main evaluation interface.

[1] The study was approved by the Research Ethics Committee of the Faculty of Arts and Sciences at Edgehill University, UK. Participants signed an electronic consent form that informed them of their freedom to withdraw from the study, anonymity in storage and use of the data, and giving permission for the research team members to use anonymised responses in research publications.

Table 5. Location Descriptions generated for the Flickr evaluation locations. # is the location identifier. *Toponym-driven* are generated using the novel algorithm, *Human* are taken from the original Flickr captions, *Preposition-driven* are generated using Hall *et al.* (2015), and the *Baseline* is taken from the automatic location description on Flickr.

#	Context	Toponym-driven	Human	Preposition-driven	Baseline
FR1	*Rural*	Near Guiting Power in Cotswold, England	Near Guiting Power	Near Guiting Power in The Cotswolds, Gloucestershire, Great Britain	Guiting Power, England, United Kingdom
FR2		Near Ladybower Reservoir in High Peak	Near Derwent Reservoir	Near Fairholmes in Peak District National Park, Derbyshire, Great Britain	Derwent Moors, England, United Kingdom
FR3		At Eastwrey Barton in Teignbridge District	Near Lustleigh Cleave	Near Sanduck Wood in Dartmoor National Park, Devon, Great Britain	Lustleigh, England, United Kingdo
FR4		Near Llyn Peris in the Snowdonia National Park	Near Llanberis	Near Llyn Peris in Snowdonia National Park, Gwynedd, Great Britain	Pentre Castell, Wales, England
FR5		Near Helm Crag in South Lakeland, Cumbria	Near Grasmere	Near Helm Crag in Lake District National Park, Cumbria, Great Britain	Grasmere, England, United Kingdom
FU1	*Urban*	On Wapping next to the Salthouse Dock in Liverpool	Near Albert Dock	On Wapping near Merseyside Police Headquarters in Liverpool, Great Britain	Albert Dock, Liverpool, England
FU2		Next to Staircase 12 in Second Quad at Jesus College in Oxford	At Jesus College	At Jesus College in Oxford, Oxfordshire, Great Britain	Oxford England, United Kingdom
FU3		Next to Mermaid Quay on the Roald Dahl Plass in Cardiff	At Cardiff Bay	At Demiro's in Cardiff, Great Britain	Cardiff Bay, Cardiff, Wales
FU4		At Edinburgh Castle in the City Centre & Leith	At Edinburgh Castle	Next to St Margaret's Chapel in City of Edinburgh, Great Britain	West End, Edinburgh, Scotland
FU5		On Belvedere Road next to Jubilee Gardens in the London Borough of Lambeth, London	Next to the London Eye	On Belvedere Road next to Jubilee Gardens in London Borough of Lambeth, London, Great Britain	South Bank, London, England

INSERT FIGURE 6 images/evaluation_map.png HERE

Table 6. Location Descriptions generated for the Geograph evaluation locations. # is the location identifier. *Toponym-driven* are generated using the novel algorithm, *Human* are taken from the original Geograph captions, *Preposition-driven* are generated using the method of Hall *et al.* (2015), and the *Baseline* is taken from the automatic location description on Geograph.

#	Category	Toponym-driven	Human	Preposition-driven	Baseline
GR1	*Rural*	Near Llyn Newydd in Ffestiniog Community, Gwynedd	West of Llyn Newydd	Near Power Station in Gwynedd, Great Britain	Near to Thywbryfdir, Gwynedd, Great Britain
GR2		Near the Blelham Beck in South Lakeland, Cumbria	Near Blelham Tarn	Near Wray Castle in Lake District National Park, Cumbria, Great Britain	Near to High Wray, Cumbria, Great Britain
GR3		On Main Road near The Bramwell Memorial Instittue in Taddington Parish, Derbyshire Dales	East of Taddington	On Main Road near The Bramwell Institute in Peak District National Park, Derbyshire, Great Britain	Near to Taddington, Derbyshire, Great Britain
GR4		Near Lakehead Hill in Dartmoor Forest Parish, Devon	B3212 near Postbridge	Near Powder Mills in Dartmoor National Park, Devon, Great Britain	Near to Bellever, Devon, Great Britain
GR5		Near Chedworth in Cotswold, England	Near Chedworth	Near Chedworth in The Cotswolds, Gloucestershire, Great Britain	Near to Chedworth, Gloucestershire, Great Britain
GU1	*Urban*	Near the Volunteer's Walk in Old Town, City of Edinburgh	On the Parade Ground at Holyrood	In City of Edinburgh, Great Britain	Near to Edinburgh, Great Britain
GU2		Near Albert Dock in Liverpool	Near Albert Dock, Liverpool	At Legacy Sculpture in Liverpool, Great Britain	Near to Birkenhead, Wirral, Great Britain
GU3		Next to Queen Alexandra Dock in the Port of Cardiff, Cardiff	At Queen Alexandra Dock	Near Shed D in Cardiff, Great Britain	Near to Penarth, The Vale of Glamorgan/Bro Morgannwg, Great Britain
GU4		On The Plain next to the St Clement's in Oxford, Oxfordshire	Near Magdalen Bridge	At the corner of The Plain and Magdalen Bridge in Oxford, Oxfordshire, Great Britain	Near to Oxford, Oxfordshire, Great Britain
GU5		At Trafalgar Square in London	At Trafalgar Square, London	At Trafalgar Square in City of Westminster, London, Great Britain	Near to City of London, Great Britain

The main evaluation interface consisted of two unlabelled maps onto which the photo location and all features mentioned in at least one of the location descriptions were drawn, as shown in Figure 6. By only showing the features mentioned in the location descriptions, participants are focused on evaluating the four descriptions, rather than on what other location description they might have used. For the same reasons the participants were not shown the photos. To give participants a good overview over each test location, the two maps use different zoom levels. The left-hand map always showed a more zoomed-in map, while the right-hand map showed a zoomed-out overview. Different zoom levels were used for the *urban* and *rural* contexts, but within each context the same zoom levels were used for all locations. The maps were static, with no interactions possible. Below each map participants were asked to rate each caption on a 5-point Likert-like scale for the questions:

(1) How accurately do each of the following four location descriptions describe the location of the photograph marked on the map?
(2) How natural is the language in each of the following four descriptions of the photograph's location?

Additionally participants were asked to pick which of the four location descriptions they would use to describe the photograph's location.

Participants were always shown five randomly selected test locations and were required to evaluate a minimum of five locations, but could judge all 20 locations, in sets of 5, if they so wished. Each set of locations to evaluate was selected randomly from the 20 available locations. The system automatically ensures that the random selection does not show a single participant the same location more than once and that overall all locations are judged almost the same number of times (due to the unpredictable nature of online experiments, there is minor variation in the number of judgements per location).

At the end of the experiment participants could optionally provide an e-mail address to be included in a 50 pound Amazon voucher raffle as an incentive.

4.3. *Participants*

Participants were recruited from staff and students at Edge Hill University and Cardiff University via e-mails and posts on announcement message boards. Staff at a national mapping agency, the Ordnance Survey, were also invited via e-mail. The invitation provided a link to the online experiment, indicated the expected minimum duration of 10 minutes, and the optional 50 pound Amazon voucher raffle.

In total 261 participants were recruited, with 120 completing the experiment. 9 participants were filtered, as they spent less than 20 seconds per set of five locations they were evaluating, resulting in a total of 111 participants that form the basis of the analysis.

Of these 70 were female and 41 male. Approximately half of the participants were undergraduate students (54) and 24 were undertaking further study. 69 classified themselves as students and 39 as employed (3 unemployed participants). The majority of the participants were native English speakers (91) with a further 6 who were not native speakers, but primarily spoke English at home. The age distribution is skewed towards the 18 - 25 range (60), but shows an otherwise good spread across the higher age groups as well (25-35: 18, 35-45: 12, 45-55: 12, 55-65: 9).

4.4. *Results*

Table 7 shows the preference counts for all location descriptions. The captions obtained in our experiments and in the human and automated baselines for each of these locations

Table 7. Preference Counts for each of the 20 evaluation locations. # is the location identifier. 'Human' and 'Baseline' refer to the manually and automatically generated captions in Flickr (for FR locations) and Geograph (for GR locations). The variation in the total number of judgements is due to the crowdsourcing nature of the experiment and the participant filtering process. The final row shows the number of locations for which the approach is clearly preferred, defined as where the description with the highest number of preferences has at least three more votes than any other description. Fractions do not always add up to 1 due to rounding errors.

#	Toponym-driven	Human	Preposition-driven	Baseline	Total
FR1	15 (.38)	4 (.1)	**16 (.41)**	4 (.1)	39
FR2	4 (.11)	12 (.32)	**17 (.46)**	4 (.11)	37
FR3	**25 (.63)**	1 (.03)	13 (.33)	1 (.03)	40
FR4	15 (.38)	6 (.15)	**16 (.41)**	2 (.05)	39
FR5	9 (.24)	2 (.05)	**25 (.66)**	2 (.05)	38
FU1	**20 (.54)**	3 (.08)	11 (.3)	3 (.08)	37
FU2	**14 (.38)**	9 (.24)	11 (.3)	3 (.08)	37
FU3	**19 (.51)**	6 (.16)	7 (.19)	5 (.14)	37
FU4	7 (.18)	**22 (.56)**	9 (.23)	1 (.03)	39
FU5	**16 (.43)**	12 (.32)	9 (.24)	0 (0)	37
GR1	**21 (.58)**	5 (.14)	5 (.14)	5 (.14)	36
GR2	14 (.38)	3 (.08)	**16 (.43)**	4 (.11)	37
GR3	11 (.31)	5 (.14)	**14 (.39)**	6 (.17)	36
GR4	**21 (.53)**	9 (.23)	7 (.18)	3 (.08)	40
GR5	12 (.3)	6 (.15)	**19 (.48)**	3 (.08)	40
GU1	10 (.27)	**27 (.69)**	2 (.05)	0 (0)	39
GU2	16 (.42)	4 (.11)	**17 (.45)**	1 (.03)	38
GU3	**19 (.56)**	13 (.38)	2 (.06)	0 (0)	34
GU4	10 (.24)	13 (.32)	**16 (.39)**	2 (.05)	41
GU5	**15 (.38)**	14 (.36)	9 (.23)	1 (.3)	39
Pref	*8 (.53)*	*2 (.13)*	*5 (.33)*	*0 (0)*	*15*

are given in Tables 5 and 6. The results are in-line with previous studies on spatial languages, showing a large amount of variation between participants' spatial language preferences (Fisher and Orf (1991), Robinson (2000), Worboys (2001), Schockaert *et al.* (2008)), and that almost any location description will be preferred by somebody. Only for three locations (FU5, GU1 & GU3) is there a location description that no participant preferred. At the same time for three-quarters of the locations there is a preference[1] for one location description over the others. Of these 15 locations with a preference, for eight locations the *toponym-driven* location description is preferred, for five it is the *preposition-driven*, and for two the *human* location description. For no location is the automatic *baseline* of either Flickr or Geograph preferred. Together with the results from the naturalness evaluation (Table 8) this validates our hypothesis H1 that the focus on toponyms will lead to an improved caption quality. An example that illustrates strong preference for the toponym-driven caption over the preposition-driven caption is location FU3 in which the caption for the toponym-driven method is *Next to Mermaid Quay on the Roald Dahl Plass in Cardiff* while the caption for the preposition-driven method is *At Demiro's in Cardiff, Great Britain*, in which it appears that participants might have found reference to the local landmarks of *Mermaid Quay* and *Roald Dahl Plass* preferable to simply *Demiro's*. Note that the values reported in Tables 8 and 9 are the median and, in square brackets, modal values of the 5-point Likert scale scores (in which 5 is a high score while 1 is low), along with inter-quartile ranges and counts respectively.

[1] We define a preference as where the most popular description has at least three more votes than any other.

Table 8. Naturalness results for the evaluation location descriptions. The results are reported 'median (inter-quartile-range) [mode (count)]'. The values range from 1 - lowest score to 5 - highest rating. Values in bold indicate the highest rating for that location. The final row shows the number of location descriptions for which the approach has the highest median score.

#	**Toponym-driven**	**Human**	**Preposition-driven**	**Baseline**
FR1	4.0 (1.0) [4 (18)]	**5.0** (1.5) [5 (20)]	4.0 (2.0) [5 (14)]	3.0 (2.0) [3 (14)]
FR2	**4.0** (2.0) [5 (14)]	**4.0** (1.0) [4 (17)]	**4.0** (2.0) [4 (13)]	3.0 (1.0) [3 (12)]
FR3	**4.5** (1.0) [5 (20)]	4.0 (2.0) [5 (15)]	4.0 (1.0) [4 (17)]	4.0 (1.25) [4 (13)]
FR4	**5.0** (1.0) [5 (20)]	4.0 (2.0) [5 (19)]	4.0 (2.0) [3 (15)]	4.0 (3.0) [5 (12)]
FR5	**4.0** (1.0) [4 (18)]	**4.0** (2.0) [5 (15)]	**4.0** (2.0) [5 (18)]	3.0 (1.0) [3 (11)]
FU1	**4.0** (1.0) [4 (16)]	**4.0** (2.0) [5 (16)]	**4.0** (1.0) [3 (12)]	**4.0** (1.0) [4 (12)]
FU2	4.0 (2.0) [4 (14)]	**5.0** (1.0) [5 (20)]	4.0 (1.0) [4 (16)]	3.0 (1.0) [3 (12)]
FU3	**5.0** (1.0) [5 (21)]	4.0 (2.0) [5 (17)]	4.0 (2.0) [5 (13)]	4.0 (2.0) [4 (13)]
FU4	3.0 (2.0) [2 (12)]	**5.0** (1.0) [5 (26)]	4.0 (2.0) [4 (12)]	3.0 (2.0) [3 (10)]
FU5	4.0 (2.0) [5 (12)]	**5.0** (1.0) [5 (21)]	3.0 (3.0) [3 (12)]	4.0 (1.0) [4 (16)]
GR1	4.0 (2.0) [5 (14)]	**4.5** (1.0) [5 (18)]	3.0 (1.25) [3 (16)]	3.0 (1.5) [3 (11)]
GR2	4.0 (1.0) [5 (15)]	**5.0** (1.0) [5 (19)]	4.0 (2.0) [4 (14)]	4.0 (1.0) [4 (14)]
GR3	4.0 (2.0) [4 (11)]	**5.0** (2.0) [5 (19)]	4.0 (2.0) [5 (15)]	4.0 (2.0) [4 (13)]
GR4	**4.0** (1.0) [5 (19)]	3.5 (2.0) [4 (11)]	**4.0** (2.0) [5 (17)]	**4.0** (1.25) [3 (13)]
GR5	**4.0** (2.25) [5 (13)]	**4.0** (2.0) [5 (17)]	3.0 (2.0) [3 (13)]	**4.0** (1.0) [4 (19)]
GU1	4.0 (1.0) [4 (15)]	**5.0** (1.0) [5 (28)]	3.0 (1.5) [3 (13)]	3.0 (1.5) [4 (12)]
GU2	**5.0** (1.0) [5 (23)]	4.0 (1.0) [4 (17)]	4.0 (2.0) [5 (16)]	2.5 (1.75) [2 (14)]
GU3	4.0 (1.75) [4 (14)]	**5.0** (1.0) [5 (23)]	3.0 (2.0) [3 (11)]	3.0 (1.0) [2 (12)]
GU4	4.0 (1.0) [4 (19)]	**5.0** (1.0) [5 (21)]	4.0 (1.0) [4 (18)]	3.0 (2.0) [3 (16)]
GU5	**5.0** (1.0) [5 (22)]	4.0 (1.0) [5 (18)]	3.0 (1.0) [3 (12)]	3.0 (2.0) [3 (14)]
Highest	*10*	*14*	*4*	*3*

Notably the novel *toponym-driven* algorithm generates location descriptions that are regarded as more natural than those produced by the *preposition-driven* algorithm. As Table 8 shows, ten of the *toponym-driven* location descriptions receive the highest naturalness scores, while for the *preposition-driven* algorithm it is only four. At the same time naturalness is one area where the *human* location descriptions still outperform all other approaches, with 14 location descriptions receiving the highest ratings. It may be noted that while in Hall *et al.* (2015) the *preposition-driven* algorithm generated location descriptions that were evaluated as weaker than the human location descriptions, in this experiment the *preposition-driven* algorithm outperforms the *human* baseline used here with regard to people's preferences (as does the toponym-driven algorithm). It should also be remarked that the evaluation presented here used location descriptions written by the photographers for the purpose of captioning. They thus provide a more realistic comparison than the bespoke, spatial-language-focused baseline captions that were created specifically for and used in the evaluation of Hall *et al.* (2015).

The evaluation also validates our second hypothesis H2 that humans prefer more detailed location descriptions. Overall for 17 of the 20 locations the preference is for one or other of the toponym-driven and preposition-driven automatic location descriptions, which are uniformly longer and more detailed than the *human* ones. Examples that illustrate this preference are provided for location descriptions FR1 and GR5. In FR1 the toponym and preposition driven captions are *Near Guiting Power in Cotswold, England* and *Near Guiting Power in The Cotswolds, Gloucestershire, Great Britain*, while the human-subject created caption is *Near Guiting Power.* In GR5 the toponym and preposition driven captions are *Near Chedworth in Cotswold, England* and *Near Chedworth in The Cotswolds, Gloucestershire, Great Britain* while the human-subect caption is simply *Near Chedworth.* For both locations one or both of the automated location descriptions have the same proximal toponym and preposition as the human location description, thus the preference for the automated location descriptions can be attributed to the additional containment detail. While we can only speculate on the reason the human location descriptions are shorter, we believe this might be due to the trade-off between effort needed to write the location descriptions and the minimum level of detail that is required.

Notably, there is uncertainty in these the results about what is the appropriate level of detail in the containment hierarchy. The extra detail compared to the *human* location descriptions is preferred, but the optimal level of detail is unclear. As described earlier the *toponym-driven* algorithm applies strong filtering heuristics to the captions and thus always produces containment hierarchies that are shorter than the *preposition-driven* ones. In FR1 and FR4 there is a minimal preference for the *preposition-driven* location description (one extra preference each), but the *toponym-driven* location description is seen as more natural. For GR5 there is a clear preference for the *preposition-driven* algorithm, but rather than the additional detail this might be due to the less natural sounding 'Cotswold' in the *toponym-driven* location description, compared with 'The Cotswolds' in the *preposition-driven* location description (note that Cotswold and The Cotswolds are different areas, both present in OSM). This interpretation is supported by GU5, where the additional detail in the *preposition-driven* location description is not preferred, possibly because London on its own is such a strong reference point. Clearly more study is needed here to identify exactly what level of detail is required in what context.

Exceptions to the preference for the two automatically generated location descriptions are captions FU4 and GU1. In the case of FU4 it seems that considering the prominence of the proximal toponym (Edinburgh Castle), the automatic choice of containment toponym (City Centre & Leith) creates an unnatural sounding caption, as is also apparent from the naturalness scores (Table 8). For GU1 the toponym used in the human location description (Parade Ground) is a vernacular name that does not exist in the gazetteer, resulting in poorer quality automatically generated location descriptions. Adding vernacular toponyms to the gazetteer should thus be a focus for future work.

The results also provide some insight into what makes a location description 'natural'. In GU2 and GU5 the *toponym-driven* and *human* location descriptions are almost exactly the same, except that the *toponym-driven* location description uses 'in', rather than ','. The use of 'in' leads to a higher naturalness score in both cases, again making it likely that the use of ',' in the human location descriptions is primarily due to reduced time effort it needs, rather than a preference.

The evaluation also assessed the accuracy of the location descriptions and as Table 9 shows, both the two automatic algorithms and the human baseline perform very well. For the toponym-driven location descriptions the only exception is FR2. For that location, the larger size of 'Ladybower Reservoir' and its proximity to a major tourist route through the Peak District leads to a significantly higher Flickr salience score than for 'Derwent Reservoir'. This higher Flickr salience outweighs the fact that the photo location is significantly closer to 'Derwent Reservoir'. Potentially it is necessary to modify how the Flickr salience is calculated for very large features to compensate for this, something that also needs to be considered for future work.

Additionally, the preposition-driven location descriptions are judged to be more accurate for 7 of the test locations, while the toponym-driven algorithm only outperforms the preposition-driven descriptions in 4 cases. As the location descriptions in Table 6 show, the preposition-driven location descriptions always end with 'Great Britain'. Whether

Table 9. Accuracy results for the evaluation location descriptions. The results are reported as 'median (inter-quartile-range) [mode (count)]'. The values range from 1, the lowest score, to 5 the highest. Values in bold indicate the highest rating for that location. The final row shows the number of location descriptions for which the approach has the highest median score.

#	**Toponym-driven**	**Human**	**Preposition-driven**	**Baseline**
FR1	**5.0** (1.0) [5 (20)]	4.0 (2.0) [5 (14)]	**5.0** (1.0) [5 (22)]	4.0 (1.0) [4 (16)]
FR2	3.0 (1.0) [3 (12)]	**4.0** (2.0) [5 (13)]	**4.0** (1.0) [4 (16)]	3.0 (1.0) [2 (13)]
FR3	**5.0** (1.0) [5 (25)]	2.5 (1.0) [2 (14)]	4.0 (1.0) [4 (19)]	2.0 (1.25) [2 (15)]
FR4	4.0 (1.0) [5 (19)]	3.0 (1.0) [3 (15)]	**5.0** (1.0) [5 (21)]	3.0 (2.0) [3 (14)]
FR5	4.0 (2.0) [4 (15)]	3.0 (2.0) [3 (13)]	**5.0** (1.0) [5 (25)]	2.0 (1.75) [2 (13)]
FU1	**5.0** (1.0) [5 (23)]	3.0 (2.0) [2 (10)]	**5.0** (1.0) [5 (21)]	2.0 (3.0) [1 (11)]
FU2	**5.0** (1.0) [5 (25)]	4.0 (2.0) [5 (17)]	4.0 (1.0) [4 (17)]	3.0 (2.0) [2 (13)]
FU3	**4.0** (1.0) [5 (17)]	**4.0** (1.0) [4 (13)]	**4.0** (2.0) [4 (13)]	**4.0** (2.0) [4 (14)]
FU4	4.0 (2.0) [4 (14)]	**5.0** (1.0) [5 (24)]	**5.0** (2.0) [5 (20)]	2.0 (2.0) [1 (14)]
FU5	**5.0** (0.0) [5 (29)]	4.0 (2.0) [4 (15)]	**5.0** (0.0) [5 (28)]	3.0 (2.0) [3 (13)]
GR1	**5.0** (1.0) [5 (19)]	4.0 (1.0) [4 (14)]	3.0 (1.0) [3 (15)]	3.0 (2.0) [2 (13)]
GR2	**4.0** (1.0) [4 (19)]	**4.0** (1.0) [4 (16)]	**4.0** (1.0) [4 (18)]	3.0 (2.0) [2 (11)]
GR3	4.0 (1.0) [5 (16)]	3.0 (2.25) [2 (9)]	**5.0** (1.0) [5 (23)]	4.0 (2.0) [4 (12)]
GR4	**4.0** (1.0) [4 (18)]	3.0 (1.0) [3 (16)]	**4.0** (1.25) [3 (16)]	3.0 (1.25) [2 (14)]
GR5	**4.0** (1.0) [4 (14)]	3.0 (2.0) [3 (14)]	**4.0** (1.25) [5 (19)]	4.0 (0.25) [4 (21)]
GU1	4.0 (1.0) [4 (22)]	**5.0** (1.0) [5 (28)]	3.0 (2.5) [1 (10)]	3.0 (2.0) [3 (13)]
GU2	4.0 (1.0) [5 (17)]	4.0 (1.0) [4 (18)]	**5.0** (1.0) [5 (26)]	2.0 (2.0) [2 (15)]
GU3	**5.0** (1.0) [5 (21)]	4.5 (1.75) [5 (17)]	3.0 (2.0) [3 (9)]	2.0 (1.75) [1 (16)]
GU4	4.0 (1.0) [4 (19)]	4.0 (2.0) [4 (16)]	**5.0** (1.0) [5 (24)]	3.0 (2.0) [2 (13)]
GU5	**5.0** (1.0) [5 (22)]	**5.0** (1.0) [5 (24)]	**5.0** (0.0) [5 (30)]	2.0 (2.0) [2 (13)]
Highest	*12*	*6*	*15*	*1*

this is sufficient to account for the differences remains for future work.

5. Conclusions

Location descriptions are used to provide geographic context in a range of scenarios, from photo sharing to social media or travel sites. The work presented here shows that it is possible to fully automatically generate complex, natural-sounding location descriptions that largely outperform current methods and are preferred by users. In particular our novel *toponym-driven* algorithm creates highly natural sounding location descriptions that balance accuracy, detail, and naturalness and are consistently preferred to the automatic methods currently in use on sites such as Flickr and Geograph. Our algorithm also produces an intermediate, relatively language-neutral representation of the location description, that will in future work allow us to generate location descriptions in other languages.

The evaluation results support our second hypothesis that people prefer location descriptions with more detail. At the same time there remains an open question about the appropriate level of detail, particularly when generating containment hierarchies. How many levels it should have, how the actual toponyms in the containment hierarchy influence the number of levels, and whether there is a context factor at work all need to be the subject of further study.

One limitation of our algorithm is that the salience calculation depends heavily on data taken from Flickr. This obviously biases our algorithm's view of "salience" towards that held by the kind of people who make their photographs available on Flickr. However, since the aim in the current study is to create location descriptions for photographs, we believe that this bias has no negative impact. If the algorithm were to be used to generate location descriptions for other contexts, then this would have to be revisited.

With regard to future work, the evaluation suggests that the inclusion of vernacular place names could improve the perceived caption quality. Enriching gazetteers with vernacular place names is not a new goal (e.g. Jones *et al.* (2008), Twaroch *et al.* (2008a,b, 2019), Schockaert (2011), Vasardani *et al.* (2013), Cunha and Martins (2014)), but it appears reasonable to suppose that their presence might be expected to improve the naturalness of automatically generated location descriptions.

The methods that we have used here can be regarded as limited by the fact that they only use the location of the photo in combination with associated geo-data. It can be envisaged that richer descriptions of the field of view of the camera could be produced by making use of directional meta-data obtained from the camera, while information about the subject of the photo could be obtained with the application of computer vision methods to identify features in the photo image.

Finally, our algorithm makes use of heuristics and empirically determined limits. Further work is needed to investigate whether and how these need to be adapted in order for the algorithm to be used in other contexts and languages.

5.1. *Data and codes availability statement*

The data and codes that support the findings of this study are available with the identifier `https://doi.org/10.21954/ou.rd.12876938.v1`.

Acknowledgements

Some of the material here is based upon work supported originally by the European Community in the TRIPOD project (FP6 cr n°045335).

References

Beller, S., Bender, A., and Bannardo, G., 2005. Spatial frames of reference for temporal relations: A conceptual analysis in English, German, and Tongan. *In*: *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 27.

Collell, G., Gool, L.V., and Moens, M.F., 2018. Acquiring Common Sense Spatial Knowledge Through Implicit Spatial Templates. *In*: *Thirty-Second AAAI Conference on Artificial Intelligence*, Apr..

Coventry, K. and Garrod, S., 2004. *Saying, seeing, and acting: The psychological semantics of spatial prepositions.* Psychology Press.

Cunha, E. and Martins, B., 2014. Using one-class classifiers and multiple kernel learning for defining imprecise geographic regions. *International Journal of Geographical Information Science*, 28 (11), 2220–2241.

Dale, R. and Reiter, E., 1995. Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions. *Cognitive Science*, 19 (2), 233–263.

Dethlefs, N., *et al.*, 2011. Generation of Adaptive Route Descriptions in Urban Environments. *Spatial Cognition & Computation*, 11 (2), 153–177.

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

Gapp, K., 1995. Angle, distance, shape and their relationship to projective relations. *In*: *Proceedings of the 17th Annual Conference of the Cognitive Science Society* Psychology Press, 112–117.

Golledge, R., 1993. Chapter 2 Geographical Perspectives on Spatial Cognition. *In*: T. Gärling and R. Golledge, eds. *Behavior and Environment: Psychological and Geographical Approaches* Elsevier Science Publishers B.V.

Grabowski, J. and Miller, G., 2000. Factors Affecting the Use of Dimensional Prepositions in German and American English: Object Orientation, Social Context, and Prepositional Pattern. *Journal of Psycholinguistic Research*, 29 (5), 517–553.

Hall, M.M., Smart, P., and Jones, C.B., 2011. Interpreting Spatial Language in Image Captions. *Cognitive Processing*, 12 (1), 67–94.

Hall, M.M., Jones, C.B., and Smart, P., 2015. Spatial Natural Language Generation for Location Description in Photo Captions. *In*: S.I. Fabrikant, M. Raubal, M. Bertolotto, C. Davies, S. Freundschuh and S. Bell, eds. *Spatial Information Theory.*, Vol. 9368 of *Lecture Notes in Computer Science* Springer International Publishing, 196–223.

Herskovits, A., 1985. Semantics and pragmatics of locative expressions. *Cognitive Science: A Multidisciplinary Journal*, 9 (3), 341–378.

Herskovits, A., 1986. *Language and Spatial Cognition: An Interdisciplinary Study of Prepositions in English.* Cambridge University Press.

Jones, C.B., *et al.*, 2008. Modelling vague places with knowledge from the Web. *Int. J. Geogr. Inf. Sci.*, 22 (10), 1045–1065.

Kelleher, J. and Costello, F., 2009. Applying computational models of spatial prepositions to visually situated dialog.. *Computational Linguistics*, 35 (2), 271–306.

Kelleher, J.D. and Kruijff, G.J.M., 2006. Incremental Generation of Spatial Referring Expressions in Situated Dialog. *In*: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Jul.. Sydney, Australia: Association for Computational Linguistics, 1041–1048.

Krahmer, E. and Theune, M., 2002. Efficient context-sensitive generation of referring expressions. No. 143 Lecture Notes, *In*: K. van Deemter and R. Kibble, eds. *Information Sharing: Reference and Presupposition in Language Generation and Interpretation, Center for the Study of Language and Information-Lecture Notes.* CSLI Publications, 223–264.

Kunze, L., Burbridge, C., and Hawes, N., 2014. Bootstrapping Probabilistic Models of Qualitative Spatial Relations for Active Visual Object Search. *In*: *AAAI Spring Symposium 2014 on Qualitative Representations for Robots*, March, 24–26. Association for the Advancement of Artificial Intelligence.

Levinson, S., 2003. *Space in language and cognition: Explorations in cognitive diversity.* Cambridge: CUP.

Logan, G. and Sadler, D., 1996. A computational analysis of the apprehension of spatial relations. *Language and space*, 493–529.

Lynch, K., 1960. *The Image Of The City.* MIT Press.

Malinowski, M. and Fritz, M., 2015. A Pooling Approach to Modelling Spatial Relations for Image Retrieval and Annotation. *arXiv:1411.5190 [cs]* ArXiv: 1411.5190.

Mukerjee, A., *et al.*, 2000. Conceptual description of visual scenes from linguistic models. *Image and Vision Computing*, 18 (2), 173–187.

Raubal, M. and Winter, S., 2002. Enriching Wayfinding Instructions with Local Landmarks. *In*: *International conference on geographic information science* Springer, 243–259.

Robinson, V., 2000. Individual and multipersonal fuzzy spatial relations acquired using human–machine interaction. *Fuzzy Sets and Systems*, 113 (1), 133–145.

Schirra, J., 1993. A contribution to reference semantics of spatial prepositions: The visualization problem and its solution in VITRA. *The Semantics of prepositions: from mental processing to natural language processing*, p. 471.

Schockaert, S., de Cock, M., and Kerre, E., 2008. Location Approximation for Local Search Services using Natural Language Hints. *International Journal of Geographical Information Science*, 22 (3), 315–336.

Schockaert, S., 2011. Vague Regions in Geographic Information Retrieval. *SIGSPATIAL Special*, 3 (2), 24–28.

Sorrows, M. and Hirtle, S., 1999. The Nature of Landmarks for Real and Electronic Spaces.. *In*: C. Freksa and D.M. Mark, eds. *Spatial Information Theory..*, Vol. 1661 of *Lecture Notes in Computer Science* Springer Science & Business Media, 37–50.

Spranger, M. and Pauw, S., 2012. Dealing with Perceptual Deviation: Vague Semantics for Spatial Language and Quantification. *In*: L. Steels and M. Hild, eds. *Language Grounding in Robots*. Boston, MA: Springer US, 173–192.

Spranger, M. and Steels, L., 2015. Co-Acquisition of Syntax and Semantics: An Investigation in Spatial Language. *In*: *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, Buenos Aires, Argentina AAAI Press, p. 1909–1915.

Spranger, M., Suchan, J., and Bhatt, M., 2016. Robust Natural Language Processing: Combining Reasoning, Cognitive Semantics, and Construction Grammar for Spatial Language. *In*: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16 Event-place: New York, New York, USA AAAI Press, 2908–2914.

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

Talmy, L., 1983. How Language Structures Space. *In*: H.L. Pick and L.P. Acredolo, eds. *Spatial Orientation: Theory, Research, and Application*. Boston, MA: Springer US, 225–282.

Talmy, L., 2000. *Toward a Cognitive Semantics: Concept Structuring Systems*. The MIT Press.

Tenbrink, T., 2011. Reference frames of space and time in language. *Journal of Pragmatics*, 43, 704–722.

Twaroch, F.A., *et al.*, 2019. Investigating behavioural and computational approaches for defining imprecise regions. *Spatial Cogn. Comput.*, 19 (2), 146–171.

Twaroch, F.A., Jones, C.B., and Abdelmoty, A.I., 2008a. Acquisition of a vernacular gazetteer from web sources. *In*: S. Boll, C.B. Jones, E. Kansa, P. Kishor, M. Naaman, R. Purves, A. Scharl and E. Wilde, eds. *Proceedings of the First International Workshop on Location and the Web, LocWeb 2008, Beijing, China, April 22, 2008*, Vol. 300 of *ACM International Conference Proceeding Series* ACM, 61–64.

Twaroch, F.A., Jones, C.B., and Abdelmoty, A.I., 2008b. Acquisition of Vernacular Place Names from Web Sources. *In*: I. King and R. Baeza-Yates, eds. *Weaving Services and People on the World Wide Web* Springer, 195–214.

Tyler, A. and Evans, V., 2003. *The Semantics of English Prepositions: Spatial Scenes, Embodied Meaning, and Cognition*. Cambridge University Press.

Vasardani, M., Winter, S., and Richter, K.F., 2013. Locating place names from place descriptions. *International Journal of Geographical Information Science*, 27 (12), 2509–2532.

Worboys, M., 2001. Nearness relations in environmental space. *International Journal of Geographic Information Science*, 15 (7), 633–651.

Files associated with each figure are as follows:

FIGURE 1 images/system_overview.png

FIGURE 2 images/SpatialTemplates1.png

FIGURE 3 images/SpatialTemplatesSelectingNorth.png

FIGURE 4 images/SpatialTemplatedStreets.png

FIGURE 5 images/caption_pattern.png

FIGURE 6. images/evaluation_map.png

The Figure captions are on the next page

Figure 1. Overview of the main components of the Location Description Generation System: the Gazetteer on the left and the Location Description Generator on the right. OpenStreetMap (OSM) data are used to retrieve proximal and containment toponyms, which are then classified. The proximal toponyms are then further used to distinguish between urban and rural locations and salience is calculated for them. Proximal and containment toponyms, together with the urban/rural distinction are then passed to the Location Description Generator, which first generates abstract Location Descriptions, that are then turned into an actual natural language representation in the final step.

Figure 2. Creating a spatial template. The data points on the left (a) represent examples of locations at which an expression of the form “north of X” has been used when the reference location is positioned relatively at the location marked X, where in practice X is a place or feature name. The data points in the example were obtained from Geograph captions. On the right (b) is the spatial template or density field generated from the points using kernel density estimation.

Figure 3. Selecting a spatial template relative to a toponym. The location marked P is a photo location to be described, while the locations marked T1 to T5 are candidate toponyms in the vicinity. A spatial template under consideration (here that for “north of”) is anchored at each of the toponym locations and the strength of the density field at the photo location is determined. In the two example anchorings in the figure, P has a fairly strong density field value when the template is anchored on toponym T4 but a weak value for T3. Thus T4 is a reasonable candidate toponym for an expression of the form “P north of T4”.

Figure 4. Selecting a spatial template relative to a street toponym. The location marked P is a photo location to be described. The lines marked R1 to R5 are candidate named streets in the vicinity, and J1 to J5 are street junctions. Of the three spatial relation templates under consideration here for describing the location of P, the “on” template is anchored on each of the roads, while the “at the corner” template is anchored at each corner and the “between” template is anchored between pairs of adjacent junctions. For each anchored candidate template, the density field value is determined at the location of the photo. In the example relatively high density values are found for “on” anchored on R4, and for “between” anchored between J2 and J4 (and hence streets R1 and R3), while for the anchoring of “at the corner” there is no strong field value at any junction. Thus both the “on R4” and the “between R1 and R3” can be regarded as applicable spatial relations.

Figure 5. Basic caption pattern used for generating the location descriptions. The caption starts with an optional road element, followed by zero or more proximal elements, and it ends with one or more containment elements.

Figure 6. Example interface used in the evaluation experiment. The left-hand map shows a zoomed-in view, while the right-hand map gives a wider overview. Only features mentioned in the location descriptions were labelled. For this example the descriptions were “Near Albert Dock” (human), “Albert Dock, Liverpool, England” (Flickr baseline), “On Wapping next to the Salthouse Dock in Liverpool” (toponym-driven) and “On Wapping near Merseyside Police Headquarters in Liverpool, Great Britain” (preposition-driven).