Context-sensitive interpretation of natural language location descriptions

A thesis submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy in Information Technology at

Massey University, Auckland, New Zealand

Niloofar Aflaki

January 2022

Abstract

People frequently describe the locations of objects using natural language. Location descriptions may be either structured, such as *26 Victoria Street, Auckland,* or unstructured. Relative location descriptions (e.g., *building near Sky Tower*) are a common form of unstructured location description, and use qualitative terms to describe the location of one object relative to another (e.g., *near, close to, in, next to*). Understanding the meaning of these terms is easy for humans, but much more difficult for machines since the terms are inherently vague and context sensitive.

In this thesis, we study the semantics (or meaning) of qualitative, geospatial relation terms, specifically **geospatial prepositions**. Prepositions are one of the most common forms of geospatial relation term, and they are commonly used to describe the location of objects in the geographic (geospatial) environment, such as rivers, mountains, buildings, and towns. A thorough understanding of the semantics of geospatial relation terms is important because it enables more accurate automated georeferencing of text location descriptions than use of place names only. Location descriptions that use geospatial prepositions are found in social media, web sites, blogs, and academic reports, and georeferencing can allow mapping of health, disaster and biological data that is currently inaccessible to the public. Such descriptions have unstructured format, so, their analysis is not straightforward.

The specific research questions that we address are:

RQ1. Which geospatial prepositions (or groups of prepositions) and senses are semantically similar?

RQ2. Is the role of context important in the interpretation of location descriptions?

RQ3. Is the object distance associated with geospatial prepositions across a range of geospatial scenes and scales accurately predictable using machine learning methods?

RQ4. Is human annotation a reliable form of annotation for the analysis of location descriptions?

To address RQ1, we determine the nature and degree of similarity among geospatial prepositions by analysing data collected with a human subjects experiment, using clustering, extensional mapping and t-stochastic neighbour embedding (t-SNE) plots to form a semantic similarity matrix. In addition to calculating similarity scores among prepositions, we identify the senses of three groups of geospatial prepositions using Venn diagrams, t-sne plots and density-based clustering, and define the relationships between the senses. Furthermore, we use two text mining approaches to identify the degree of similarity among geospatial prepositions: bag of words and GloVe embeddings. By using these methods and further analysis, we identify semantically similar groups of geospatial prepositions including: 1- *beside, close to, near, next to, outside* and *adjacent to;* 2- *across, over* and *through* and 3- *beyond, past, by* and *off.* The prepositions within these groups also share senses. *Through* is recognised as a specialisation of both *across* and *over.* Proximity and adjacency prepositions also have similar senses that express orientation and overlapping relations. *Past, off* and *by* share a proximal sense but *beyond* has a different sense from these, representing *on the other side.* Another finding is the more frequent use of the preposition *close to* for pairs of linear objects than *near*, which is used more frequently for non-linear ones. Also, *next to* is used to describe proximity more than touching (in contrast to other prepositions like *adjacent to*). Our application of text mining to identify semantically similar prepositions confirms that a geospatial corpus (NCGL) provides a better representation of the semantics of geospatial prepositions than a general corpus. Also, we found that GloVe embeddings provide adequate semantic similarity measures for more specialised geospatial prepositions, but less so for those that have more generalised applications and multiple senses.

We explore the role of context (RQ2) by studying three sites that vary in size, nature, and context in London: *Trafalgar Square, Buckingham Palace,* and *Hyde Park.* We use the Google search engine to extract location descriptions that contain these three sites with 9 different geospatial prepositions (*in, on, at, next to, close to, adjacent to, near, beside, outside)* and calculate their acceptance profiles (the profile of the use of a preposition at different distances from the reference object) and acceptance thresholds (maximum distance from a reference object at which a preposition can acceptably be used). We use these to compare prepositions, and to explore the influence of different contexts. Our results show that *near, in* and *outside* are used for larger distances, while *beside, adjacent to* and *at* are used for smaller distances. Also, the acceptance threshold for *close to* is higher than for other proximity/adjacency prepositions such as *next to, adjacent to* and *beside.* The acceptance threshold of *next to* is larger than *adjacent to,* which confirms the findings in Chapter 2 which identifies *next to* describing a proximity rather than *touching* spatial relation*.* We also found that relatum characteristics such as image schema affect the use of prepositions such as *in, on* and *at.*

We address RQ3 by developing a machine learning regression model (using the SMOReg algorithm) to predict the distance associated with use of geospatial prepositions in specific expressions. We incorporate a wide range of input variables including the similarity matrix of geospatial prepositions (RQ1); preposition senses; semantic information in the form of embeddings; characteristics of the located and reference objects in the expression including their liquidity/solidity, scale and geometry type and contextual factors such as the density of features of different types in the surrounding area. We evaluate the model on two different datasets with 25% improvement against the best baseline respectively.

Finally, we consider the importance of annotation of geospatial location descriptions (RQ4). As annotated data is essential for the successful study of automated interpretation of natural language descriptions, we study the impact and accuracy of human annotation on different geospatial elements. Agreement scores show that human annotators can annotate geospatial relation terms (e.g., geospatial prepositions) with higher agreement than other geospatial elements.

This thesis advances understanding of the semantics of geospatial prepositions, particularly considering their semantic similarity and the impact of context on their interpretation. We quantify the semantic similarity of a set of 24 geospatial prepositions; identify senses and the relationships among them for 13 geospatial prepositions; compare the acceptance thresholds of 9 geospatial prepositions and describe the influence of context on them; and demonstrate that richer semantic and contextual

information can be incorporated in predictive models to interpret relative geospatial location descriptions more accurately.

Acknowledgements

All Ph.D. journeys have a story, a colourful story with many ups and downs. Since I was a teenager, I dreamed about it, and four years ago, this dream came true. I became a Ph.D. student, and to be honest, I enjoyed every moment of it.

Deciding on moving to a place like New Zealand and being far away from my loved ones was not an easy decision, but I made it to follow my heart. New Zealand was a peaceful home for me, and I became stronger and more confident every day.

First of all, from the bottom of my heart, I would like to thank my primary supervisor, Dr. Kristin Stock, for her dedicated support, guidance, and encouragement throughout the process. You taught me many things and, most importantly, supported me along this journey. Indeed, if you were not here, I would not have been able to complete this research. Furthermore, I would like to thank my co-supervisor, Mr. Jeremy Morley, in Ordnance Survey Company (UK), that funded my Ph.D. study in New Zealand and gave me this opportunity to follow my dreams. I want to say thank you to two other amazing and supportive people, my co-supervisors, Prof. Hans Guesgen, Prof. Chris Jones, for all your thoughtful recommendations and feedback on my research. I would also like to thank my examiners Prof Stephan Winter, Dr. Benjamin Adams, and Dr. Napoleon Reyes for their feedback on my work and their great support and encouragement.

I'm also very grateful to all the friends in my lab (Azadeh, Rahila, Jinjin) and staff at Massey University (SNCS) (Annette, Sue, Dianne, and all the other friendly staff in our campus) for your significant role in the quality of my Ph.D. life.

I would also like to thank my dad, mom, and brother for your smiles on our video calls that enlightened my heart every time. It's been three years that we couldn't meet, but I always feel your unconditional love and support, even when we are this far.

Last but not least, thanks to my husband. You were the only person who I could talk when I was confused and exhausted. Thanks for your patience, positivity, and support that gave me the strength to do my best every day to accomplish this research.

Papers

This thesis contains 7 chapters (including the introduction and conclusion), and 5 chapters are research papers. Two of these papers have been published, one is submitted and is under review, and 2 other papers are ready for submission.

Chapter	Paper title	Conference/journal
Chapter 2, manuscript 1	***Aflaki, N.*** *Stock, K. Jones, C.B. Guesgen, H. Morley, J. An Empirical Study of the Semantic Similarity of Geospatial Prepositions and their Senses*	Submitted to the *Spatial cognition and computation* journal
Chapter 3, manuscript 2	***Aflaki, N****., Jones, C., & Stock, K. (2019, September 18). Mining the Semantic Similarity of Spatial Relations from Text. Geocomputation 2019 - Adventures in GeoComputation, Queenstown, NZ.*	Presented in *Geocomputation* Conference (Queenstown 2019)
Chapter 4, manuscript 3	***Aflaki, N.*** *Stock, K. Jones, C.B. Guesgen, H. Morley, J. Geospatial preposition acceptance thresholds and the role of context*	Ready for submission
Chapter 5, manuscript 4	***Aflaki, N.*** *Stock, K. Jones, C.B. Guesgen, H. Morley, J. A Machine learning model to predict the distance between geospatial locations using contextual factors*	Ready for submission
Chapter 6, manuscript 5	***Aflaki, N****., Russell, S., & Stock, K. (2018). Challenges in Creating an Annotated Set of Geospatial Natural Language Descriptions (Short Paper). 10th International Conference on Geographic Information Science (GIScience 2018).*	Presented in *GIScience* Conference (Melbourne 2018)

Beside the above-mentioned papers in this thesis, I contributed to two other papers during my Ph.D. study.

1- Egorova, E., **Aflaki, N**., Fagundes, C. and Stock, K. (2019). Cross-corpora Analysis of Spatial Language: The Case of Fictive Motion. *Conference on Spatial Information Theory: COSIT 2019,* 10-13 September, Regensburg, Germany.

2- Stock, K., Jones, C. B., Russell, S., Radke, M., Das, P., & **Aflaki, N**. (2021). Detecting geospatial location descriptions in natural language text. *International Journal of Geographical Information Science*, 1-38.

Glossary of technical terms

Abstraction: the process of reduction to a set of basic features by eliminating or omitting characteristics from it.

Antonymy: two words are antonyms if they have opposite meanings (e.g., arrive and depart).

Axial structure: structure of the main axis of an object, determining which parts of the object are referred to with projective relations (front, back, left, right, top, bottom).

BNC (British National Corpus): BNC is a 100-million-word textual collection including written and spoken English samples collected from a variety of sources.

Boundedness: focuses on internal fragmentation of a quantity.

Conceptual neighbourhood graphs: graphs that define the semantic similarity between topological relations.

Degree of extension: point-like form, bounded extent, or unbounded extent of an object.

Dividedness: whether an object is divided to multiple parts or not.

Elongation: a formula that identifies the length of an object divided by its width (how long an object is).

Force dynamics: the nature of objects and the goal of the statement, as well as the force dynamics that happen between them (e.g., whether objects push against one another). These have an impact on how language is applied.

Frame of reference: the conceptual frame in which a location/an object is being viewed or described (egocentric, relative, absolute).

Geometry type: the shape of an object in geographical space (e.g., line, polygon, point).

Human Intelligence Tasks (HIT): a HIT is a specific, self-contained online job that a Amazon Mechanical Turk (a job brokering web site) Worker can complete by working on it, submitting an answer, and receiving a reward.

Hypernym: a word is a hypernym of another word, if it has a more general concept (e.g., colour is the hypernym of blue).

Image schema: the schema (e.g., platform, container) that is used to conceptualisation an object, and that influences language selection. The geometry of an object, its location, and the trajectory of mobility are all represented schematically in image schemas.

Liquidity/solidity: specifies if a location has a liquid characteristics or solid (e.g., river: liquid, building: solid)

Locatum/trajector: the object whose location is being described in a relative location description.

Metric spatial relations: metric relationships are based on the distance between two items.

NCGL (The Nottingham Corpus of Geospatial Language): a geospatial corpus which contains geospatial descriptions collected from news websites, geographic websites, tourism, and local history websites.

Pattern of distribution: include the pattern of distribution of matter through space or of activity through time.

Pearson correlation: Pearson product-moment correlation coefficient (also known as the Pearson correlation coefficient) is a measurement of a linear relationship between two variables.

Plexity: The state of being made up of a specific number of items (e.g., uniplex, duplex).

Probability density fields: providing a likelihood measure in the form of a continuous field to indicate whether a geographic relation term fits in a given context.

Projective relations: the relations which determine the location of an object relative to a framework that is projected on the scene (e.g., left, right, behind).

Proximity/proximal prepositions: prepositions that indicate object closeness.

Relatum/landmark: a reference object for the locatum, used with a spatial relation term (e.g. a preposition).

Schematization: to attain cognitive competence, a representation is purposely simplified beyond technical needs which is called schematization.

Senses of geospatial prepositions: the meanings associated with geospatial prepositions. Each geospatial preposition can have more than one sense (meaning).

SMO regression (Sequential Minimal Optimization): is a more efficient method of solving the Support Vector Machine (SVM) training optimisation problem than typical QP (quadratic programming) solvers.

Spatial qualifier: A word or set of words that adds more information to a spatial relation term.

Spatial relations: identify the relationships between locatum/relatum.

Spatial specifier: a word of set of words that describes particular subparts of a feature.

tf-idf (term frequency, inverse document frequency): tf-idf is a statistical measure for assessing the relevance of a word to a document in a set of documents.

Topological relations: topological relations are relations that do not change with rotation, translation, and scaling. The basic topological relationships are connectivity, adjacency, and enclosure.

Toponym recognition: locating geographic place name references in text.

t-SNE (t-distributed stochastic neighbour embedding): a statistical method to reduce the dimension and visualise high dimensional data in 2D or 3D map.

Word embeddings: word embeddings are vectors that show the semantics of words for text analysis based on their collocation with other words across many instances.

Table of contents

Table of figures

Table of tables

Chapter 1 - Introduction

1.1 Motivation and problem statement

Location and route data is widely used in modern society, saved in databases for easy information retrieval. We can ask applications such as Google Maps to find a specific address such as "*126 Queen Street, Auckland, New Zealand*" and be provided with corresponding geographic coordinates (e.g., latitude and longitude). Location data of this kind is quick and easy to use.

However, dealing with natural language place and route descriptions is much more difficult. When people describe a location, especially in an emergency such as a traffic accident or violent crime, they do not normally provide an exact address. Instead, they may describe the location using qualitative terms. For example, *the accident site is on Queen Street, near the Noel Leeming store.* Determining coordinates for (georeferencing) these kinds of descriptions is not a simple task for a machine and requires interpretation of qualitative terms to receive an accurate result. However, the ability to georeference location descriptions would be useful for a wide range of types of text, including the descriptions people provide during emergency situations; social media text such as that used on Flickr[1] or Twitter[2] (Kelm et al. 2013; Landwehr & Carley 2014), textual descriptions for the task of search and rescue (Doherty et al. 2011); text documents describing objects in historical sites (Rupp et al. 2013; Alex et al. 2015), textual data describing the location of specimens (Beaman and Conn 2003; Hill et al. 2009; van Erp et al. 2015), or reports of the starting point for specific disease like flu (Achrekar et al. 2011).

To georeference a relative location description, first the elements of a description should be identified. The main elements of relative location descriptions, that describe the location of one object relative to the secondary object, are the locatum (the object whose location is

[1] https://www.flickr.com/

[2] https://twitter.com/home

being described), the relatum (a reference object for the locatum) and spatial relation terms (that specify the relationship in space between the locatum and relatum) (Talmy 1983). For example, *a house (locatum) on (spatial relation term in a form of a preposition) the A25 road (place name relatum).* Following identification of these elements, their characteristics and all the contextual factors in the described spatial scene can be considered in georeferencing tasks.

Georeferencing of place names has been addressed by many researchers and there are a number of existing applications and mapping tools (Google maps[3], OSM[4], Stanford NER[5] and SpaCy[6]) that can complete this task (Leidner 2008; Karimzadeh 2016; Cardoso et al. 2019). However, the problem of georeferencing text documents is not limited to specific place names (toponym resolution/recognition). Many textual documents contain qualitative spatial relation terms followed by place names or without any place names. For instance, a description such as *the building near Victoria Embankment Gardens* contains the place name *Victoria Embankment Gardens,* but the description really refers to *the building* which is related to the *Victoria Embankment Gardens* by the spatial relation term *near*. This term could specify any place in the vicinity of *Victoria Embankment Gardens,* but its location is still unclear and needs further processing for the location of the building to be identified. While humans are often able to work with such uncertainty and interpret the meanings of expressions like this based on knowledge of the scene or other cues, automated methods for georeferencing require the semantics of these spatial relation terms to be understood, and of particular importance is accurate determination of the locations associated with spatial relation terms.

[3] https://www.google.com/maps

[4] https://www.openstreetmap.org/

[5] https://nlp.stanford.edu/software/CRF-NER.html

[6] https://spacy.io/

In addition to better understanding the semantics of, and developing georeferencing methods for, spatial relation descriptions, one possibility is to find semantically similar spatial relation terms, as this could reduce the cost of automated georeferencing and make it faster. As an example, if we need to georeference the description (1) *the building near Victoria Embankment Gardens*, and already know the georeference for two other descriptions: *(2) building close to Victoria Embankment Gardens* and (3) *a building on Victoria Embankment Gardens,* we can conclude that description 1 is likely to have a similar location to description 2, but not to description 3. Knowledge of the semantic similarity among pairs of spatial relation terms can enable us to do this and is particularly useful for machine learning methods for automated georeferencing, in which interpretations of descriptions are learnt from other similar descriptions.

The similarity between spatial descriptions that may aid in automated georeferencing is not limited to their meaning. Many other factors can contribute to the similarity between spatial descriptions and may also be helpful in interpretation. These may include characteristics of the relatum and locatum, including feature types, size, image schema, axial structure, and scale. Other contextual factors that describe aspects of the surrounding scene may also be useful. Identifying these factors and using them in the process of georeferencing spatial descriptions can increase the accuracy of georeferencing and may also reduce the time taken in cases that require quick response such as emergency incidents like violent crime, earthquake damage, traffic accidents or fires.

We will address the questions mentioned in this section through this thesis. In this chapter, we summarise previous works in the area of georeferencing and research gaps in Section 1.2, and in Section 1.3 we identify the research goals, objectives and research questions that we cover in this thesis. Section 1.4 lists manuscripts and the methods we used in each chapter. In section 1.5 we visualise the prepositions we work with in each chapter and Section 1.6 presents a table of data used in each manuscript/chapter. 1.7 summarises our contributions in each chapter and Section 1.8 is the outline of this thesis.

1.2 Background and research gaps

Spatial data includes not only coordinate-based information described with maps, diagrams, databases, or models to specify the location of objects that are referenced in location descriptions, but also other forms of information describing locations, movements and routes (Lautenschütz et al. 2006; Blaylock et al. 2009).

Spatial language has been defined as a relation between a cognitive model of space and language (Talmy 1983; Landau and Jackendoff 1993; Emmorey et al. 1993; Hayward and Tarr 1995; Richardson et al. 2001; Munnich and Landau 2003; Steels and Loetzsch 2006; Chatterjee 2008), and studied by many researchers with goals such as toponym resolution/recognition (Leidner 2008; Karimzadeh 2016; Cardoso et al. 2019), locating objects or places in text descriptions or documents (Landau and Jackendoff 1993; Coventry et al. 2009), studying the meaning and use of spatial relations (Bitters 2009; Coventry and Garrod 2004; Hois et al 2009; Kemmerer 2006; Levinson and Meira 2003; Retz-Schmidt 1988, Zwarts 2005, Zwarts and Winter 2000, Tenbrink 2008) and spatial topic modelling (Long et al. 2012; Joseph et al. 2012; Lim et al. 2017). Landau and Jackendoff (1993) defined the identification of objects; spatial search and navigation tasks as the three main areas of spatial knowledge. They claimed that because we are human, we can easily use language to describe the location of spatial objects and incidents around us and highlighted the key role that language plays in human spatial cognition, assisting us in talking about our perception of events.

A number of researchers have explored the nature of the objects involved in relative location descriptions (the relatum and locatum). For example, Landau and Jackendoff (1993) conducted experiments that show that objects that are described as nouns or noun phrases are easily described using axes, volume, surface information. Talmy (1983), in his early work, named the locatum and relatum as figure and ground and specified characteristics of them. Talmy advanced on his earlier work by identifying additional characteristics of the locatum, having simpler geometry, being more salient and more dependent on the relatum. On the other hand, the relatum is more likely to have complex geometry, be more familiar, more stable, and more independent (Talmy 2000). For example, '*the car is near the post office*' is a much more common order of objects than '*the post office is near the car',* which is not factually incorrect, but unusual (Olivier and Gapp 1998). In relative location descriptions, the locatum and relatum can be place names (Tezuka and Tanaka 2005; Rose-Redwood et al. 2010), geographic feature types such as rivers, buildings, roads (Usery 2020; Ying et al. 2011; Richter et al. 2012) or moving objects such as vehicles, people, or weather events (Talmy 1975, 1983, 2000; Richter et al. 2012). Talmy has pointed out that the geometry of the locatum and relatum is an important factor in choosing the right preposition to relate them. For instance, in the expression *the board lay across the river,* the geometry and axes of the locatum and relatum let us use *across* instead of other prepositions (Talmy 1983).

Core to a location description is a term or terms that determine the position of the locatum relative to the relatum, and these are most often spatial prepositions (Zelinsky-Wibbelt 1990; Vasardani et al. 2013; Khan et al. 2013; Kim et al. 2016), although other words or groups of words (including verbs, adverbs) may also be used to describe relative location, and have more generally been referred to as spatial relation terms or spatial indicators (Kordjamshidi et al. 2011; Vasardani et al. 2013; Khan et al. 2013; Kim et al. 2016).

Accurate georeferencing relies on an understanding of the semantics of spatial prepositions, the ways in which they are used and the spatial locations that specific spatial relations describe, a topic that has been addressed by a number of researchers (Bitters 2009; Coventry and Garrod 2004; Hois et al 2009; Kemmerer 2006; Levinson and Meira 2003; Retz-Schmidt 1988; Zwarts 2005; Zwarts and Winter 2000; Tenbrink 2008). A consideration of the semantic similarity among spatial prepositions is useful for tasks such as language learning, automated translation, as well as machine learning methods for georeferencing. However, previous work has only addressed the semantic similarity of a limited number of prepositions (or other spatial relation terms) or considered them only in a limited context such as *road park (*Mark and Egenhofer 1994). Furthermore, previous work has mainly used human subjects experiments to measure the semantic similarity of spatial prepositions. Recent automated text processing work, including the application of word embeddings (Mikolov et al. 2013) shows some promise as a tool for determining semantic similarity, but to the best of our knowledge, there was no previous work using this text mining technique to measure the semantic similarity among spatial prepositions using word embeddings considering their context or Global Vector embeddings (GloVe).

In addition, in some cases, the same spatial preposition may be used to describe different kinds of spatial configuration (e.g., *the house across the road; houses across the country*), referred to as senses. Some authors have studied the senses of spatial prepositions that influence the interpretation of spatial descriptions (Cooper 1968; Leech 1970; Bennett 1972; Miller and Johnson-Laird 1976; Talmy 1983; Lakoff 2008), but this work is limited in scope, focussing on a small number of spatial prepositions (Tyler and Evans 2003; Coventry and Garrod 2004), and not considering the semantic similarity of, or relations between, different senses of spatial prepositions.

While early models of the areas referred to by spatial prepositions and other spatial relation terms were adopted from the field of Qualitative Spatial Reasoning (QSR) and mapped to natural language spatial relation terms (Schwering 2007), more recently qualitative interpretations of spatial relation terms (prepositions or other elements such as verbs) have been defined as acceptance models such as density fields (Hall and Jones 2008; Hall et al. 2011; Hall et al. 2015). These fields describe the areas within which the use of a given preposition is acceptable, and are extracted from large volumes of expressions that use the preposition concerned, collected by either human subjects experiments or crowdsourced web sites (e.g. Geograph[7]) (Lan et al. 2012; Hall et al. 2011; Hall et al. 2015). However, these works largely aggregate the use of prepositions across many contexts to create a generalized model and do not explore the contextual variations in acceptance models. Furthermore, they provide only limited comparison of the acceptance models of different prepositions.

Moving on from generalised acceptance models, predictive models have been developed to determine the location that corresponds to a particular location description (including spatial preposition), often using machine learning. These models incorporate a range of features and

[7] https://www.geograph.org.uk/

learn the interpretation of a spatial preposition from a training set in order to estimate the location of objects, georeference locations, predict distances or retrieve objects in images (Chang et al. 2014; Chen et al. 2018; Collell et al. 2018). Thus far, there are very few works that have applied predictive models to the interpretation of geographic location described by location descriptions, with these kinds of models more commonly being applied to image retrieval (e.g. to provide automated methods to find the photo with *the boy on the horse*) (Lan et al. 2012), or robotics, in blocks world or indoor environments (Moratz and Tenbrink 2006). Furthermore, the range of contextual features included in these models has been limited in the previous work. For example, Bisk et al. (2018) and Collell et al. (2018) included locatum and relatum size, and embeddings indicating feature type to predict the location of objects in a spatial scene or in images, and Stock and Yousaf (2018) incorporated a range of characteristics of locatum and relatum in their model, but incorporation of a broader range of characteristics, and more general contextual information about the environment, have been addressed only in very limited ways (Chen et al. 2018; Novel et al. 2020).

Finally, much work in the area of geospatial natural language processing relies on human annotation as ground truth (Stock and Yousaf 2018; Kordjamshidi et al. 2011; Hois et al 2009). While some studies calculate inter-annotator agreement, the reliability of human annotators in identifying, parsing, and tagging location descriptions has not been established.

1.3 Research goals, objectives, and questions

This thesis addresses the twin goals of advancing the understanding of English geospatial prepositions and automating their interpretation.

We break down these goals into several research objectives to address the research gaps as follows:

- To understand the semantic similarity of geospatial prepositions and their senses using a human subjects study.
- To understand the semantic similarity of geospatial prepositions using text mining methods.

- To review the effect of contextual factors and distance between the locatum and relatum and understand the use of specific prepositions in different contexts.
- To predict the distance between the locatum and relatum using contextual information and extract the most important factors that influence this prediction.
- To understand the reliability of human subjects in annotating geospatial language.

The specific research questions that we will address are:

RQ1. Which geospatial prepositions (or groups of prepositions) and senses are semantically similar?

RQ2. Is the role of context important in the interpretation of location descriptions?

RQ3. Is the object distance associated with geospatial prepositions across a range of spatial scenes and scales accurately predictable using machine learning methods?

RQ4. Is human annotation a reliable form of annotation for the analysis of location descriptions?

We present this thesis by publication so that each chapter describes a step towards the end goals of the project. Table 1.1 shows the research questions, the objectives we mentioned here, and the manuscripts that address each objective and research question.

1.4 Manuscripts and methods

This thesis combines five different manuscripts in order to progressively advance the goals of increasing the level of understanding of geospatial prepositions and automating their interpretation. Prepositions are geospatial when they are being used in a geospatial context. A context is geospatial when it's happening in an outdoor environment in an open geographic space (Radke et al. 2019; Stock et al. 2021).

Table 1.1. Research questions, objectives, and papers

Research question	Objective	Manuscript
RQ1: Which geospatial prepositions (or groups of prepositions) and senses are semantically similar?	•To understand the semantic similarity of geospatial prepositions and their senses using a human subjects study	Manuscript 1 (Chapter 2): *An Empirical study of the semantic similarity of geospatial prepositions and their senses* -Measures the semantic similarity among 24 geospatial prepositions and describes the senses of three groups of them
	•To understand the semantic similarity of geospatial prepositions using text mining methods •To evaluate how well text mining methods could be used to determine the semantic similarity of geospatial prepositions	Manuscript 2 (Chapter 3): *Mining the semantic similarity of spatial relations from text* -Measures the semantic similarity among 25 geospatial prepositions using text mining methods including bag of words and word embeddings and compares the results to a human subjects experiment that is a ground truth data
RQ2: Is the role of context important in the interpretation of location descriptions?	•To review the effect of contextual factors and distance between the locatum and relatum and understand the use of specific prepositions within different contexts	Manuscript 3 (Chapter 4): *Spatial preposition acceptance thresholds and the role of context* -Identifies the impact of the contextual factors of the relatum on choosing prepositions in a spatial scene. -compares the distances at which each preposition is acceptable
RQ3: Is the object distance associated with geospatial prepositions across a range of spatial scenes and scales accurately predictable using machine learning methods?	•To predict the distance between the locatum and relatum using contextual information and extract the most important factors that influence this prediction	Manuscript 4 (Chapter 5): *Machine learning model to predict the distance between geospatial locations using contextual factors* -Predicts the distance between the locatum and relatum in a given expression using contextual factors, word embeddings of geographic features, the semantic similarity of geospatial prepositions and environmental factors
RQ4: Is human annotation a reliable form of annotation for the analysis of location descriptions?	•To understand the reliability of human subjects in annotating geospatial language	Manuscript 5 (Chapter 6): *Challenges in creating an annotated set of geospatial natural language descriptions* -Creates an annotated dataset using six different spatial language elements, as marked by specific labels that use an annotation scheme) and evaluates the degree to which annotators can consistently identify the elements

1.4.1 Manuscript 1 Chapter 2

The first research paper in Chapter 2 describes the semantic similarity between 24 geospatial prepositions using an empirical study. For this research, a human subjects experiment was implemented using diagrammatic representations following Stock and Yousaf (2018) to measure the degree and nature of the semantic similarity among geospatial prepositions using clustering, t-stochastic neighbour embedding (t-SNE) (Maaten and Hinton 2008) and extensional mapping (Levinson and Meira 2003). Then, the senses of three groups of geospatial prepositions are identified using Venn diagrams, t-SNE and density-based clustering (Ester et al. 1996). The data used were 720 geospatial expressions extracted from the Nottingham Corpus of Geospatial Language (NCGL) created by Stock et al. (2013) and the Manaaki Whenua - Landcare Research Specimen Collection data, consisting of four different data sets (soils[8], flora[9], terrestrial invertebrates[10] and fungi[11]), including specimen types and collection locations in the form of natural language descriptions.

1.4.2 Manuscript 2 Chapter 3

The second research paper in Chapter 3 defines the semantic similarity between 25 geospatial prepositions (included a ternary preposition *between* in addition to the 24 prepositions studied in Chapter 2) using text mining methods, including the bag of words model and GloVe embeddings (Pennington et al. 2014) using two different datasets, the British National Corpus (BNC[12]) and NCGL. The bag of words model consists of a vector with a normalised measure of word frequency values for the 1000 most frequent words across the expressions that use each of the geospatial prepositions.

[8] https://soils.landcareresearch.co.nz/soil-data/national-soils-data-repository-and-the-national-soils-database/

[9] https://www.landcareresearch.co.nz/resources/collections/allan-herbarium

[10] https://www.landcareresearch.co.nz/resources/collections/nzac

[11] https://www.landcareresearch.co.nz/resources/collections/pdd

[12] http://www.natcorp.ox.ac.uk/

GloVe embeddings are a dimension-reduced vector representation pre-trained on a very large corpus. We use both of these methods to calculate the semantic similarity between each pair of geospatial prepositions using the cosine similarity. Then, the results are compared to the previous human subjects experiment conducted by Stock and Yousaf (2018), and the Pearson correlation coefficient is calculated between the human subjects data and three methods separately.

1.4.3 Manuscript 3 Chapter 4

This Manuscript in Chapter 4 studies the distance acceptance thresholds for two subsets of 24 geospatial prepositions (9 geospatial prepositions in total). Then, explores the role of context on the use of geospatial prepositions by measuring the frequency of mentions of specific locatum-spatial preposition-relatum triples and the distance between the locatum and relatum. A web scraping method was used to extract spatial expressions from Google containing six proximity and adjacency prepositions *(next to, close to, beside, adjacent to, outside, near)* and three topological prepositions *(in, on, at)* for three relata (*Trafalgar Square, Buckingham Palace and Hyde Park)* all based in London, each with 100 surrounding unique place names as locata.

1.4.4 Manuscript 4 Chapter 5

Chapter 5 uses the preposition and sense semantic similarities extracted from the empirical study (Chapter 2) as an input for prediction of distances indicated by geospatial prepositions in different spatial contexts using a machine learning method. In addition to preposition and sense semantic similarity, characteristics of the relatum and locatum such as geometry type, image schema, liquidity/solidity and semantic similarity and contextual information such as object density are used as input for a regression model to predict the distance between two locations described in a spatial scene. Two datasets were used in this study: a data set containing 690 geospatial expressions extracted from Geograph[13] and Foursquare[14] and a set of about 7400 expressions from a previous study (Morris 2020).

1.4.5 Manuscript 5 Chapter 6

The last Manuscript in Chapter 6 explores the reliability and consistency of human annotation of geospatial language. It identifies, describes, and explains an annotation scheme consisting of six elements and an experiment in which four annotators tagged location descriptions. Inter-annotator agreement was calculated and the reliability of annotation between annotators and between different elements were compared to evaluate human annotation as a method for ground truthing geospatial natural language processing methods.

1.5 Spatial prepositions addressed in each chapter

This thesis addresses geospatial prepositions from a number of different angles, and each manuscript uses a particular set of geospatial prepositions to answer the research questions. Figure 1.1 shows the prepositions used in each chapter of this thesis.

We only considered English in our analysis of geospatial prepositions and associated elements like locatum and relatum. For examples of studies of spatial preposition use in other languages, see (Cuyckens 1991; Takenobu et al. 2005; Marchi Fagundes et al., 2021). Furthermore, we did not consider different English dialects. The data that we used in Chapters 2-5 are collected from a range of different sources such as Google search engine, Geograph or Foursquare, and were not limited to only one dialect, as our focus was on generic characteristics of geospatial prepositions that occur across dialects, and dialect detection can be difficult in the kinds of sources we used.

[13] https://www.geograph.org.uk/

[14] https://foursquare.com/

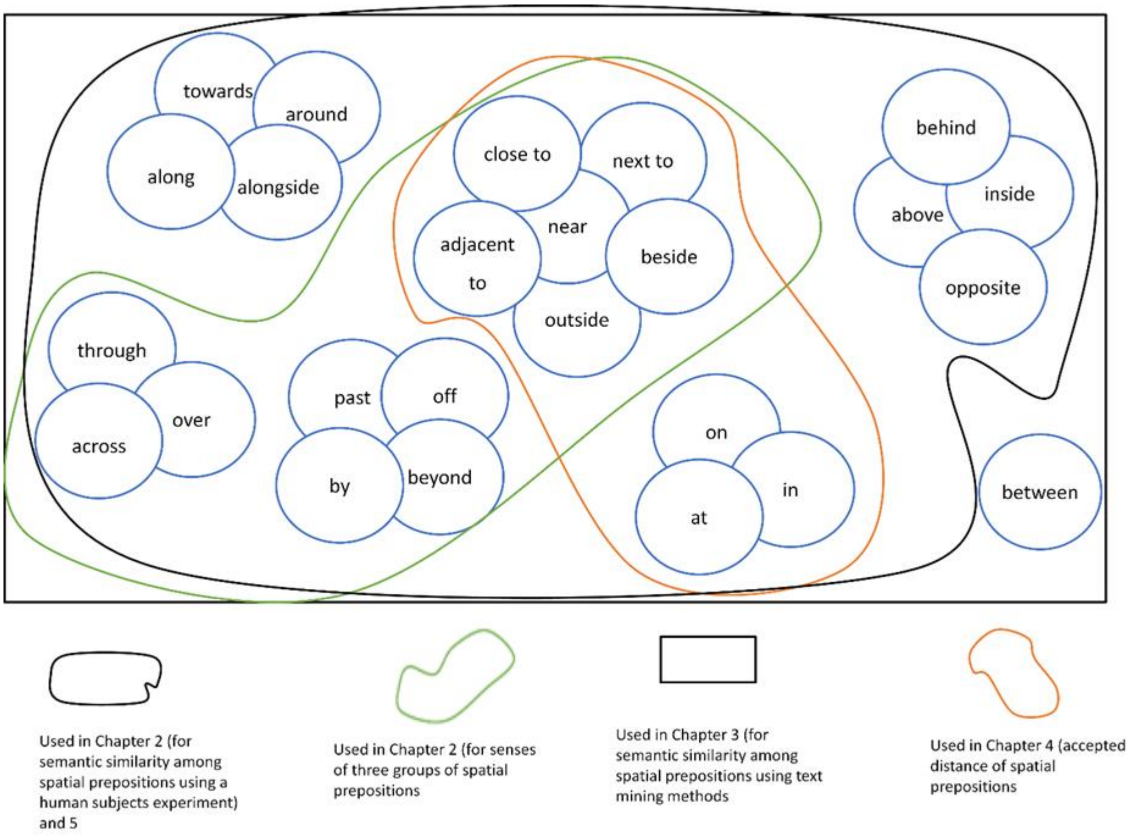

towards
around
along
alongside
close to
next to
near
adjacent to
beside
outside
behind
above
inside
opposite
through
over
across
past
off
by
beyond
on
in
at
between
Used in Chapter 2 (for semantic similarity among spatial prepositions using a human subjects experiment) and 5
Used in Chapter 2 (for senses of three groups of spatial prepositions
Used in Chapter 3 (for semantic similarity among spatial prepositions using text mining methods
Used in Chapter 4 (accepted distance of spatial prepositions

Figure 1.1. Prepositions used in each chapter of thesis

1.6 Table of the data for each manuscript

Table 1.2 presents the datasets we used for each manuscript.

1.7 Contribution in each chapter

The scientific contributions and the research questions for each individual paper, which contribute to the wider thesis research questions, are as follows.

Chapter 2 two research questions have been defined for this chapter:

RQ1. Which geospatial prepositions are semantically similar to each other across a range of geospatial contexts, and what is the degree and nature of that similarity?

Table 1.2. Datasets used for each manuscript

Manuscript	Dataset
Manuscript 1 Chapter 2	720 geospatial expressions extracted from: ○ The Nottingham Corpus of Geospatial Language[15] (NCGL) (Stock et al. 2013) ○ The Manaaki Whenua - Landcare Research Specimen Collection data, consisting of four different data sets (soils[16], flora[17], terrestrial invertebrates[18] and fungi[19])
Manuscript 2 Chapter 3	-NCGL (Stock et al. 2013) -BNC data[20] (British National Corpus)
Manuscript 3 Chapter 4	-5000 spatial expressions ○ Google
Manuscript 4 Chapter 5	-690 geospatial expressions from (London area TQ3080): ○ Geograph[21] ○ Foursquare[22] -7300 expressions from: ○ Geograph (all over the UK)
Manuscript 5 Chapter 6	-1000 sentences from: ○ NCGL ○ where am I (Stock et al. 2015) - The Manaaki Whenua - Landcare Research Specimen Collection data,

[15] http://geospatiallanguage.massey.ac.nz/ncglindex.htm

[16] https://soils.landcareresearch.co.nz/soil-data/national-soils-data-repository-and-the-national-soils-database/

[17] https://www.landcareresearch.co.nz/resources/collections/allan-herbarium

[18] https://www.landcareresearch.co.nz/resources/collections/nzac

[19] https://www.landcareresearch.co.nz/resources/collections/pdd

[20] http://www.natcorp.ox.ac.uk

[21] https://www.geograph.org.uk/

[22] https://www.foursquare.com/

To answer the first research question, we studied the semantics of 24 geospatial prepositions using the data collected from a human subjects experiment and defined a matrix of the semantic similarity among these geospatial prepositions.

The output is an important contribution of this chapter, as no one has previously measured the semantic similarity of the wide range of geospatial prepositions. The groups of semantically similar prepositions we identified in this chapter can be used in future research to identify semantically similar spatial descriptions, as existing semantic networks such as WordNet do not adequately cover prepositions.

RQ2. How are the semantics of similar geospatial prepositions and their senses related to each other?

To answer this second question, we studied the senses of three subsets of the larger group of 24 prepositions (13 in total), using both quantitative and qualitative approaches. Our contribution here was to define the senses of these 13 prepositions, and the relationships between them. Senses of geospatial prepositions are important in understanding their semantics. Two spatial descriptions may be different at first glance because they use different geospatial prepositions, but individual senses of the geospatial prepositions may be similar. Conversely, prepositions may appear similar, but in particular expressions, their semantics may be different due to the senses being used.

Chapter 3 the research question of this chapter is the same as the first research question of the previous chapter, i.e., RQ1. However, this chapter uses a different methodology to determine semantic similarity. Instead of a human subjects experiment, this chapter studies the accuracy of text mining in measuring the semantic similarity among geospatial prepositions, addressing the research question:

Which geospatial prepositions are semantically similar to each other across a range of geospatial contexts, and what is the degree and nature of that similarity?

To answer this question, we used a Bag of Words (BoW) (expressing context by representing the frequency of words around the geospatial prepositions) and GloVe embeddings (only embeddings of geospatial prepositions) on two corpora (one geospatial and one non-geospatial) to calculate the semantic similarity between pairs of spatial relations. We are not aware of any previous works using text mining methods to measure the semantic similarity of geospatial prepositions, and in particular, we address geospatial uses of these prepositions. This work indicates that the Pearson correlation between the human subject experiment is higher than the BNC which is a general corpus. So, by using a text mining method with geospatial corpus, we can identify the semantic similarity between all 25 geospatial prepositions with the correlation 47%, and if we only analyse the geospatial prepositions that have less descriptions, mainly because they only have spatial sense like *opposite* and *beyond,* the correlation will increase to 76%.

Another contribution of this chapter was an analysis of the parts of speech words that occur frequently with each geospatial preposition, showing that the geospatial prepositions *adjacent to, beside and next to* are more likely to co-occur (to be in a same spatial description as) with nouns, while *above, in and off* more frequently co-occur with other prepositions. However, in the case of *on,* we see more *adjectives* in the given spatial descriptions. This contribution is also important for further analysis of geospatial prepositions and understanding of their context.

Chapter 4 addresses two research questions:

RQ1. How do distances between relata and locata that are acceptable differ between prepositions?
RQ2. How important is context in the use of geospatial prepositions?

The contribution of this chapter was to identify and compare the acceptance thresholds (distances at which a given preposition is acceptable) for 9 qualitative, non-directional geospatial prepositions using frequency graphs with data extracted from the Google search engine, for three different sites (relata) in Central London. The results show that *near* and *close to* have the highest distance acceptance thresholds, that proximity prepositions vary widely in their thresholds (*next to* being high and *beside* being used mainly for very short distances) and that the preposition *in* is frequently used for locata that are not physically inside the relatum, but nearby. Contextual factors such as size, image-schema, popularity, and accessibility also influence the acceptance thresholds of some geospatial prepositions.

Chapter 5 uses the information gained from the three previous chapters to build a machine learning model using a range of features to predict the distance between the locatum and relatum indicated by a geospatial preposition, in a spatial scene. We defined two specific research questions for this chapter:

RQ1. How accurately can we predict distance using machine learning regression methods?

RQ2. How important are specific model features in the success of that prediction?

Our contribution was a model that used a much wider range of features in our model than previous work, including:

- Characteristics of the relatum and locatum, such as:
    - Information about feature type, using both broad and detailed classifications, and word embeddings
    - Liquidity/solidity, scale, geometry type and image schema
    - Elongation
- Semantic information about the prepositions:
    - Semantic similarity data for geospatial prepositions obtained from Chapter 2 and Chapter 3
    - Data about geospatial preposition sense
- Information about the wider context:
    - GloVe embeddings of the whole expression
    - Density of objects in the area, and of specific types of objects.

The novelty of this research is the usage of contextual factors as well as geospatial prepositions characteristics to predict the distance between two locations. The methodology was tested on two different geospatial corpora to predict the distance, achieving an average distance prediction with 93.5% of distances being predicted within

50m of their correct locations, and improvements of 15-38% over the best baseline, for different experiments.

Another contribution of this chapter was to identify the importance of variables such as the scale and geometry type of the locatum and relatum, as well as semantic information about the geospatial prepositions in the accuracy of the distance predictions.

Chapter 6 sheds some light on the effectiveness of human manual annotation, a topic that has only been addressed in a limited way before. This chapter shows that agreement between annotators varies by individuals and by the kinds of language elements that are being annotated. Annotators were most consistent in their annotation of the *spatial relation term (including prepositions), locatum (trajector)* and *relatum* (landmark), with average inter-annotator agreement ranging from 0.53 (for locatum) to 0.64 for spatial relation and 62% for relatum. Chapter 6 shows a lower agreement accuracy between human annotators who are not experts in this field, although they had some training through our experiments. Human annotation is widely relied upon as a method for ground truthing automated natural language processing, including for spatial language, and this chapter shows the significant role of annotators and how they can contribute to the body of knowledge in this field, as further work on the context and characteristics of spatial elements is highly dependent on the accuracy of annotations.

1.8 Chapter outline

This thesis by publication has seven chapters. It consists of an introduction chapter, five chapters that each contain a publication and a conclusion chapter.

Chapter 1 introduces the research and its importance, some of the previous literature, research gaps, research goals and questions, then summarises the main methods of each manuscript and the datasets used in each.

Chapter 2 and ***Chapter 3*** answer the first research question, measuring the semantic similarity of spatial relation terms (we focused on geospatial prepositions) using a human subjects study (Chapter 2) and text mining methods (Chapter 3).

Chapter 2 also investigates senses of geospatial prepositions, and the relationships between them. These studies help us to understand more about geospatial prepositions, their semantic similarities and those of their senses and are also used in our method for automated prediction of distances associated with geospatial prepositions (Chapter 5).

In ***Chapter 4*** we describe a web scraping approach to extract geospatial expressions from Google and analyse how contextual factors and the distance between the locatum and relatum influence the selection of geospatial prepositions. This chapter sheds more light on the use of geospatial prepositions in text location descriptions as we did not limit user selection to specific options but extracted available data and analysed it based on the spatial scene elements.

Chapter 5 presents a method to predict the distance between the locatum and relatum in a spatial description using the features belonging to its locatum, relatum, and preposition and the surrounding environment. These features include the embeddings of the geographic feature types; contextual factors such as building density and the semantic similarities we measured in Chapter 2 .

Chapter 6 measures the success of human annotation of spatial language by defining a scheme of spatial elements including, *trajector (locatum), landmark (relatum), spatial relation (such as preposition), location and movement verb, spatial specifier, and spatial qualifier.* We asked four annotators to annotate spatial expressions with these elements following training and compared the results to determine consistency. The best agreement was for elements such as spatial relation, trajector (locatum) and landmark (relatum).

The findings, contributions, conclusion, and future work are discussed in ***Chapter 7*** .

Chapter 2 - An Empirical study of the semantic similarity of geospatial prepositions and their senses

ABSTRACT

Spatial prepositions have been studied in some detail from multiple disciplinary perspectives. However, neither the semantic similarity of these prepositions, nor the relationships between the multiple senses of different geospatial prepositions, are well understood. In an empirical study of 24 geospatial prepositions, we identify the degree and nature of semantic similarity and extract senses for three semantically similar groups of prepositions using t-SNE, DBSCAN clustering, and Venn diagrams. We validate the work by manual annotation with another data set. We find nuances in meaning among proximity and adjacency prepositions, like the use of *close to* instead of *near* for pairs of lines, and the importance of proximity over contact for the *next to* preposition, in contrast to other adjacency prepositions.

2.1 Introduction

The locations of objects on the earth are commonly described using natural language in human speech and written documents. Locations may be identified using place names, but may also be described with relative location expressions, consisting of a spatial preposition and a reference object (Herskovits 1986). For example, the expression *I am near the cinema* describes the speaker's location (*near*) relative to a cinema. In this case, the preposition *near* does not describe a precise, specific location. *Near* could refer to a location in any direction within a short distance of the cinema. The distance specified by *near* is vague, and likely to depend on the context (Purves et al. 2007).

Spatial prepositions are a key element of relative location descriptions, and a clear understanding of their meaning (semantics) and applicability in different contexts is key to the study of location language but is far from straightforward. Spatial prepositions are vague, and they might have different interpretations in different contexts (Landau and Jackendoff 1993). In addition to their vagueness, spatial prepositions often have multiple senses and contexts of use (Talmy 1983; Coventry and Garrod 2004; Tyler and Evans 2003). They are known to be among the most difficult kinds of words for second-language learners to use correctly (Chodorow et al. 2010), and spatial prepositions are often used metaphorically to apply to other situations (for example, *I am at the end of my tether*) (Coventry and Garrod 2004). In addition to the inherent interest in the study of spatial prepositions for our understanding of human language use, a clear understanding of the semantics of spatial prepositions in different situations is crucial for advancing the effective methods for automated georeferencing and generation of location descriptions. Such automation has multiple applications, including natural language spatial querying; georeferencing of social media, blogs, reports, and archives, automated georeferencing of emergency calls and natural language support for navigation (Chen et al. 2019; Al-Olimat 2019; Hu and Wang 2020).

An important element in understanding the semantics of geospatial prepositions and their senses is the consideration of semantic similarity. The semantics of concepts are often understood through their relations with other words (Bittner et al. 2005; Sánchez et al. 2012), and if we know which geospatial prepositions and/or geospatial preposition senses are synonymous or nearly synonymous, we can better understand their meaning based on their senses of the interpretation of their synonyms. This knowledge can also be applied in automated natural language processing methods, as it enables us to learn correct interpretations from other semantically similar expressions. For example, *the restaurant next to the Auckland Harbour Bridge* and *the restaurant beside the Auckland Harbour Bridge* describe the same location, and awareness of this similarity may be useful for machine learning tasks, or for ontology-based information retrieval. Semantic similarity has long been an essential element for many information retrieval problems, including web search (Hliaoutakis et al. 2006), and for tools like WordNet (Fellbaum 1998), which is built on semantic relations.

Researchers have investigated the semantics of spatial prepositions in some detail (e.g., Talmy 1983; Coventry and Garrod 2004; Tyler and Evans 2003; Herskovits 1985), exploring the different contexts of use, and describing their senses (Talmy 1983; Herskovits 1986; Coventry 1999; Tyler and Evans 2003; Coventry and Garrod 2004; Tenbrink 2008). However, much of this work focusses on spatial prepositions and/or their senses individually, rather than addressing the semantic similarity between them. A number of formal, mathematical models have been developed to enable rule-based calculation of the physical configurations in which specific spatial relations occur (Freeman 1975; Clementini et al. 1994), but these works focus on the definition of spatial relations on a theoretical level, not natural language spatial prepositions, and do not take context into account. Some work has addressed the problem of mapping spatial relations to the natural language prepositions that are used to describe them, and explored the semantic similarity of different spatial prepositions, but these works largely focus on a single contextual situation (road and park, with different spatial relation terms), rather than developing more broadly applicable models, and do not address different senses of spatial prepositions (Mark and Egenhofer 1994, Mark et al. 1995; Shariff et al. 1998, Du et al. 2017, Schwering 2007). A third strand of investigation of spatial prepositions comes from the computational linguistics and computer science fields, in which methods for automated interpretation of spatial prepositions include applicability models, or spatial templates (Zenasni et al. 2015; Hall et al. 2015; Collell et al. 2018). These works provide a picture of the operation of some spatial prepositions, but they do not address semantic similarity or individual senses.

In this chapter, we address these gaps in the previous literature and pursue two research questions:

RQ1: Which geospatial prepositions are semantically similar to each other across a range of geospatial contexts, and what is the degree and nature of that similarity?

RQ2: How are the semantics of similar geospatial prepositions and their senses related to each other?

We address these research questions by studying the semantics of 24 spatial relation prepositions and their senses using empirical data from a human subjects experiment. Our focus is particularly on the geospatial context, in which these geospatial prepositions are used to describe situations in geographic, environmental or some cases of vista space, in Montello's typology (Montello 1993; Stock et al. 2021). We asked respondents to match 720 expressions to the diagrams (from a set of 55) that best reflect their meaning. From the analysis of the human subjects data, we make two main contributions. Firstly, we study geospatial preposition semantics using both quantitative and qualitative approaches. Using a quantitative approach, we identify groups of semantically similar geospatial prepositions using clustering and t-distributed stochastic neighbour embedding (t-SNE), contrasting the groupings of similar prepositions to the typologies and groupings of prepositions that have been proposed thus far. Then, using a qualitative approach (although based on our quantitative data), we explore the aspects of similarity and difference within and between groups of prepositions using extensional maps.

In our second contribution, we explore the senses of three groups of semantically similar geospatial prepositions, again using a combination of qualitative and quantitative approaches. We apply density-based clustering (DBSCAN) to the x, y coordinates for each individual expression that were determined using t-SNE. We then examine the clusters using Venn diagrams to isolate individual senses and the relationships between them using a manual approach. We do not attempt to build sense networks that show the ways in which senses may have been abstracted from other senses of a particular preposition like Tyler and Evans (2003) and Lakoff (2008). Our focus is rather on identifying the senses used in *geospatial* natural language, and the relationships between the senses of different prepositions. We are particularly interested in geospatial natural language because of the applications of semantic similarity work on the problem of georeferencing. We combine computational and manual methods to explore the semantic similarity of specific prepositions and their senses, and do not attempt to define an automated approach to the extraction of senses.

The structure of this chapter is as follows. Section 2.2 describes related work addressing the spatial prepositions and the similarity between them, Section 2.3 describes the method used for the human subjects experiment, Section 2.4 describes the data collection process, and Section 2.5 describes the analysis applied to the data to represent the semantics of the geospatial prepositions. Section 2.6 analyses the semantic similarity of the geospatial prepositions using qualitative and quantitative methods and discusses the results. Section 2.7 analyses the senses of three subgroups of geospatial prepositions and discusses the results. Section 2.8 presents future work and the conclusion.

2.2 Literature review

2.2.1 The Semantics of spatial prepositions

The main elements of a relative spatial description are the locatum (the object being located), the relatum (the reference object) and the spatial relation term, which describes the position of the locatum relative to the relatum (Lehmann 1983, Taylor and Evans 2003; Quirk et al. 1985). Spatial relation terms are commonly prepositions (Talmy 1983; Retz-Schmidt 1988) but may alternatively (or as well as) consist of other parts of speech such as verbs, adverbs (Kordjamshidi et al. 2011). Prepositions may specify the geometric configuration of the relatum relative to the locatum, as well as shape, magnitude, and orientation (Talmy 1983; Dirven 1993).

Experimental work has demonstrated the importance of context in the selection of spatial prepositions to describe a scene (Coventry 1999), and their selection and use may be influenced by space schematisation, idealisation, image schema and abstraction. For example, in the expression *a bar inside the hotel*, the spatial preposition *inside* may indicate that *bar* is smaller than *hotel*, *hotel* has a volume geometry and both objects have locative characteristics (Herskovits 1980; Herskovits 1985; Talmy 1983; Vorwerg and Rickheit 1998; Zwarts 1997; Zelinsky-Wibbelt 1993), although note that the application of these aspects depends on the specific situation and perspective of the observer. Other aspects that may impact on the semantics of prepositions include frame of reference, which may be object-centred (intrinsic), viewer-centred (relative) or environment-centred (absolute). Stock and Hall (2017) identified a broad range of factors that affect the interpretation of spatial descriptions including: *proximity*, *visibility, immediacy, object shape, contact, centrality, physical containment, projection, convergence, collinearity, vertical contact, surroundedness, termination,* and some locatum and relatum characteristics such as *liquid/solid* (Lautenschütz et al. 2006); *image schema* which is mainly connected to the spatial preposition (Lakoff and Johnson 1980; Mark and Frank 1996); *axial structure or axes* that describes whether objects have *back, side, top, bottom, left and right* (Landau and Jackendoff 1993); *perspectival mode* which is a subcategory of perspective defined by Talmy (2000) as *stationary* or *moving.* Other factors that are subcategories of configurational structure have been described in Talmy (2000), including: *boundedness* (focuses on internal fragmentation of a quantity); *dividedness; quantity; plexity* (referring to single or multiple objects); *pattern of distribution* (which may include the pattern of distribution of matter through space or of activity through time) and *degree of extension* (point, bounded extent or unbounded extent). *Geometry type* (point, line, and polygon shape of locatum and relatum) (Landau and Jackendoff 1993); *scale* (the size of spatial elements) (Lautenschütz et al. 2006) and *force dynamics* (Coventry and Garrod 2004) are also known as important contextual factors that impact the interpretation of a spatial description. While these different aspects of the semantics of spatial prepositions have been studied in some detail, particularly by linguists and cognitive scientists, investigation of the semantic similarity and relatedness between spatial prepositions is more limited.

2.2.2 Semantic similarity

Semantic similarity is a subset of the general idea of semantic relatedness, which includes any kind of relation between concepts. A vast range of different kinds of semantic relations between objects have been defined, including contrasts (e.g., antonyms, incompatibilities); case/syntactic/syntagmatic relations (e.g., agent-action), part-whole relations and causality (Chaffin and Herrmann 1984; Ballatore et al. 2014; Budanitsky and Hirst 2006).

Definitions of semantic similarity vary, with Chaffin and Herrmann (1984) including synonymity (car-auto); attributional similarity (have the same salient attributes); dimensional similarity (smile-laugh) and necessary attribution (lemon-sour); Ballatore et al (2014) including synonymity, hypernymity or hyponymity (e.g. house *is a kind of* building) and Miller and Charles (1991) defining semantic similarity in terms of substitutability (whether terms can be used in place of one another without changing meaning, or in a weaker form, truth value). Several criticisms of definitions of similarity have been proposed (Goodman 1972), but the notion of semantic similarity nevertheless plays a key role in many information retrieval and querying tasks.

Much of the work on semantic similarity has focussed on objects, rather than relations, and methods for determining semantic similarity have considered the presence of shared or similar attributes, relations (e.g., analogy) or affordances (Turney 2006; Ballatore et al. 2014; Janowicz and Raubal 2007; Hahn and Chater 1997); proximity in space; correspondence between objects; or number of transformations needed to change one object into another (Goldstone and Son 2005). Janowicz et al. (2011) provide a comprehensive review of the semantics of similarity, describing a range of approaches to the measurement of similarity in the context of geographic information retrieval, and identifying the benefits of each. Ontology-based approaches, which formally specify the semantics of concepts using their attributes and relations, have been used to identify semantically similar objects, and have been applied to geographic concepts (river, mountain, forest) by a number of researchers (Rodríguez and Egenhofer 2004). Initiatives such as WordNet define a range of different types of relations to assist in the automation of semantic processing (Pedersen et al. 2004). Another common approach to determining the semantic relationship between objects (or types of objects) uses word context in natural language, assuming that similarity in the terms that appear near words in text corpora indicates that they are semantically similar (Rubenstein and Goodenough 1965, Agirre et al. 2009; Wang et al. 2020). However, text-based approaches more accurately describe semantic relatedness than semantic similarity, as they do not account for situations such as antonymy (Budanitsky and Hirst 2006; Miller and Charles 1991; Ballatore et al. 2014).

2.2.3 Semantic similarity of spatial prepositions

Despite extensive investigation into the notion of semantic similarity, application of the concept in the context of spatial prepositions is more limited. Several researchers have addressed the semantics of spatial prepositions by attempting to categorise them, indicating some level of semantic similarity or relatedness (e.g., adjacency and proximity) (Bitters 2009; Coventry and Garrod 2004; Hois et al 2009; Kemmerer 2006; Levinson and Meira 2003; Retz-Schmidt 1988; Zwarts 2005; Zwarts and Winter 2000; Tenbrink 2008). However, many of these studies cover only a subset of spatial relation terms, and there is little consensus among schemes (e.g., *beside* can be classified as projective or proximal) (Retz-Schmidt 1988, Zwarts and Winter 2000, Coventry and Garrod 2004). Other classes contain prepositions that are related in some way but are not semantically similar (e.g., the class of topological prepositions includes various types of connection or containment (e.g., *contains, outside, overlaps*) (Kemmerer 2006; Levinson and Meira 2003). Similarly, the class of projective relations contains relations that rely on projected axes (e.g., *left, right, in front, behind*) (Coventry and Garrod 2004; Kemmerer 2006), but would not be considered semantically similar for many purposes.

Theoretical work by Bitters (2009) describes equivalent and synonymous relations for the spatial preposition *near*, equivalents being *near to, nearby, close, close to*, and *nigh*, and synonyms being *adjacent, adjacent to, beside, by, alongside,* and *next to*. However, the focus of this work is to identify the frequency of use of prepositions with particular feature type pairs, and the semantic equivalence and synonymous relations are not experimentally verified. In a quantitative approach, Schwering (2007) defines a semantic similarity measure between pairs of 15 natural language spatial terms, combining Shariff et al's (1998) mapping from natural language terms to topological and metric relations with Mark and Egenhofer's (1994) conceptual neighbourhood graphs that define the semantic similarity between topological relations. They test their measure with a human subjects experiment, identifying three groupings of semantically similar terms (broadly representing containment, intersection and near/avoid/bypass). However, they experiment only with road and park as locatum and relatum respectively, and do not consider a wider range of situations. Du et al. (2017) develop a random forest classifier to predict spatial relation from a sketch also using Shariff et al's (1998) parameters. To aid prediction success, they identify sets of five and seven groups of semantically similar prepositions (from a set of 69) using three methods: human judgement with a sketch drawing task; examination of a confusion matrix to identify misclassification (and thus likely similarity) and average distance between vectors of features. Their groups roughly correspond to: *starts* and *ends in; alongness/enclosure; leads up to; containment; crosses/overlaps; goes into* and *near*. However, their similarity assessment is relatively course-grained, with some groups containing a wide range of terms, and is again confined to the road + park context only. Stock (2008) demonstrates an approach to determine semantic similarity of spatial relations using a restricted natural language called Natural Semantic Metalanguage, but investigates only the *intersects, next to, on* and *contains* spatial relation terms in a theoretical treatment.

2.2.4 Spatial preposition senses

It is common for words to have multiple meanings in natural language generally, and spatial prepositions are no exception. Several spatial prepositions are known to be used to describe multiple, different spatial configurations (e.g., the preposition *on* in *the cup is on the table* and *the key is on the chain*) (Coventry and Garrod 2004). These different meanings of the same preposition are referred to as senses. In some cases, the same word is used to refer to objects or concepts that appear to have no semantic connection (homonyms) (e.g., the word *bank* can be used to describe a geographic feature or a financial institution) (Lakoff 2008), but in the case of spatial prepositions, senses are commonly thought to be connected through some underlying principle (polysemes) (Tyler and Evans 2003). Principles of support and location control have been posited as playing this role for the *on* and *in* prepositions respectively (Coventry and Garrod 2004). Lakoff (2008) describes connections between senses as being defined by metaphors and image schemas and shows how multiple senses are connected for the spatial preposition *over*. Herskovits (1986) cites contiguity, attachment, and support, but also identifies other factors and exceptions in different cases, rather than a single organising principle.

Senses of spatial prepositions have been studied and enumerated by several researchers (Cooper 1968; Leech 1970; Bennett 1972; Miller and Johnson-Laird 1976; Talmy 1983; Lakoff 2008), and application of the specific senses of prepositions have been shown to be influenced by the surrounding context (Dahlmeier et al. 2009). In the Preposition Project (PP) Litkowski and Hargraves (2005) define senses based on dictionary definitions. Cannesson and Saint-Dizier (2002) discuss the difference in senses based on the characteristics of the noun and verbs in the context. Cooper (1968) defines senses based on a semantic marker that is a specification of a concept, defining different concepts and interpretations. To disambiguate senses, Dahlmeier et al. (2009) and Tratz and Hovy (2009) designed a classifier and trained it on an annotated data to get the annotations of senses for test data prepositions. While this work has investigated senses, work on the semantic similarity of senses is limited.

In addition to studying distinct senses, researchers have investigated the means by which senses are related to each other (e.g., through metaphor). Herskovits (1986) refers to use types that describe variations on the ideal meaning of a preposition, and the 'stretching' of prepositions to apply in different situations. How then, do we define a distinct sense? Tyler and Evans (2003) propose two criteria. Firstly, "it must contain additional meaning not apparent in other senses associated with a particular form" (pp. 42-43). Secondly, "there must be instances of the sense that are context independent, that is, in which the distinct sense could not be inferred from another sense and the context in which it occurs" (p.43). As an example, they mentioned "*Joan nailed a board over the hole in the ceiling*". This meets the first criterion, by describing a sense of *over* that is distinct from the more standard sense meaning 'on top of' conveying the idea of *covering*; and the second criterion in that this different sense cannot be extracted from the context in which it occurs. They distinguish uses of a preposition that meet these two criteria, and thus count as distinct senses, as those that are "conventionalised in semantic memory" (p.45), in contrast to other uses that are the result of inference and "produced on-line for the purposes of understanding" (p.45). They acknowledge that these criteria are strict, and that agreement about how fine-grained sense distinctions should be has not been agreed on, and also discuss the notion of a primary sense, which they define as the most prototypical, which can be identified through empirical means (from language studies) and linguistic means (the earliest use, role in the semantic network relative to other senses, inclusion in composite words, participation in contrast sets with other prepositions (e.g. *above/below*) and ability to be substituted for related senses) (Tyler and Evans 2003; Langacker 1987).

While in previous work, the semantics of many common spatial prepositions and their senses has been explored, limited attention has been given to the semantic similarity of spatial prepositions and senses, except in a narrow range of situations (e.g., road + park). In the next section we explain the human subjects experiment that forms the basis of our determination of semantic similarity of geospatial prepositions and their senses, across a range of different contextual situations.

2.3 Method

Our method for studying geospatial prepositions and their senses has its theoretical foundations in Gärdenfors' conceptual spaces, in which the semantics of an object can be described by its position in a multidimensional vector space whose axes are defined by quality dimensions, and the distance between objects in that vector space can be used to determine semantic similarity (Gärdenfors 2004). We create a conceptual space in which objects are geospatial prepositions and their senses, and we use 55 geometric configuration diagrams, based on Stock's (2014) Geometric Configuration Ontology, to represent each quality dimension. Values for each quality dimension for a given preposition are determined by respondents' assessments of how well each geometric configuration diagram fits a range of expressions using the preposition. We use 30 expressions for each preposition in order to incorporate a range of different contextual situations (explained in Section 2.3.2), as the interpretation of spatial relations is acknowledged to be highly influenced by context (Coventry and Garrod 2004). By using a range of different expressions for each preposition, we explore the aspects of preposition semantics that are generic in different situations, as well as different preposition senses.

Like a number of previous researchers (Mark and Egenhofer 1994, Levinson and Meira 2003, Coventry 1999, Stock and Yousaf 2018), we use a diagram matching task, in which respondents select diagrams that match each expression and rate the degree of agreement on a Likert scale. While grouping and pairwise comparison tasks are common alternatives to diagram matching methods for determining semantic similarity (e.g., Miller and Charles 1991; Chaffin and Herrmann 1984; Mark and Egenhofer 1994), we consider them less useful for gaining a clear understanding of the specific meanings of geospatial prepositions and their senses because we are interested in exploring the use of prepositions in different contexts, and in the range of different ways that prepositions are used, aspects that can be highlighted through the diagram matching approach. Drawing tasks have also been used in the study of geospatial prepositions (Shariff et al. 1998), but unlike many studies that focus on a single expression (for example, *the road crosses the park*), we study prepositions across many different contexts, and we considered that it would be difficult to obtain comparable diagrams across such a range of situations, when the experiment is not based on a limited number of expressions. Employing the results of our diagram matching experiment, we apply several methods to determine semantic similarity, including clustering, t-distributed stochastic neighbour embedding (t-SNE) (Section 2.6.1), as well as qualitative methods (Section 2.6.2).

2.3.1 Selection of geospatial prepositions

There are about 80 spatial prepositions and some of them have multiple senses. Out of these 80, we investigate the semantics of 24 frequently used geospatial prepositions. These prepositions were identified by extracting geospatial expressions from the Geograph[23] and Foursquare[24] websites. Geograph aims to crowd-source geographically representative photos and associated captions, descriptions, and locations for every square kilometre of Great Britain and Ireland. Foursquare is a social networking application and website that contains attractions and user reviews. We extracted descriptions and comments from both sites in the central London area (specifically, the TQ 3080 map tile on the British National Grid) and excluded any descriptions that did not include place names or location information, resulting in 890 geospatial descriptions. From these descriptions, we manually identified geospatial prepositions as those that described either the location or movement of a geographic object/place. For instance, we excluded the expression *a cat behind the table* as it does not refer to a

[23] https://www.geograph.org.uk/

[24] https://foursquare.com/city-guide

named place, but we include *the bridge over the Thames River*. We excluded the geospatial prepositions *to* and *from* because their interpretations are based on the verbs that they are collocated with (e.g., *the road goes to the church; the ferry came to the island*), and ternary prepositions (e.g. between). This process resulted in 700 expressions with 24 geospatial prepositions. The final list consisted of twenty-one single word prepositions (*above, across, along, alongside, around, at, behind, beside, beyond, by, in, inside, near, off, on, opposite, over, outside, past, through, towards*) and three prepositional phrases (*adjacent to, close to, and next to*). Figure 2.1 shows the frequency of expressions for each preposition.

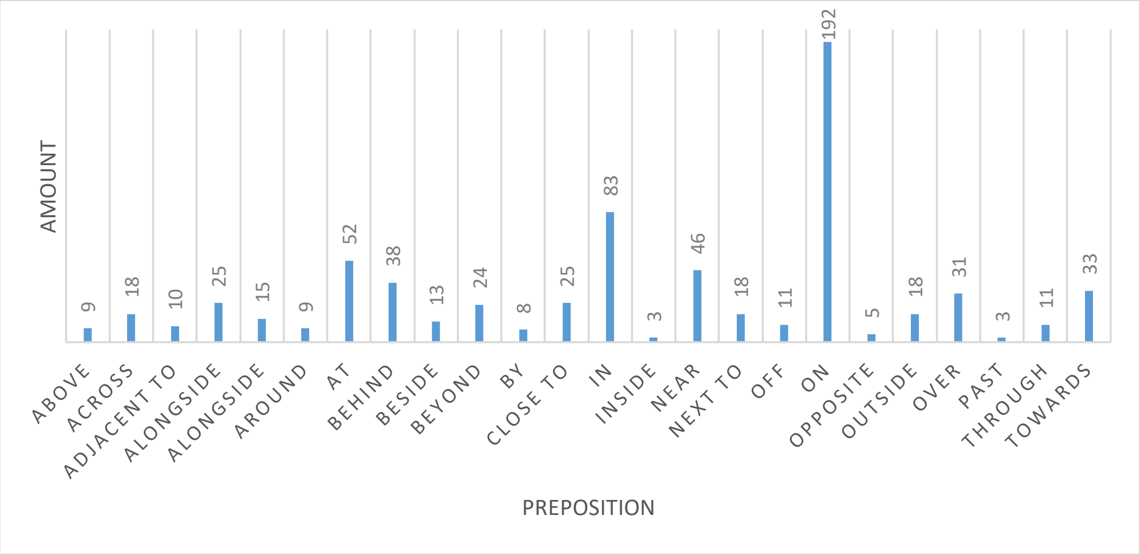


Figure 2.1. Number of prepositions in 700 geospatial expressions

2.3.2 Selection of descriptions

Having selected 24 frequently appearing geospatial prepositions, we randomly selected 30 expressions for each preposition from two other data sets (Table 2.1):

1. The Manaaki Whenua - Landcare Research Specimen Collection data, consisting of four different data sets (soils[25], flora[26], terrestrial invertebrates[27] and fungi[28]), including specimen types and collection locations in the form of natural language descriptions.
2. The Nottingham Corpus of Geospatial Language[29] (NCGL) (Stock et al. 2013), consisting of around 11,000 geospatial expressions collected from 46 websites with content such as news, travel, tourism.

Table 2.1. The properties of each dataset

Dataset	Number of expressions	Number of tokens	Example
Landcare collection locality data	132,954	237,936	"Beside Lake Wairarapa 1 km north of Burling's Stream."
NCGL	10,147	812,145	"At the crossroads by the church, turn right down the hill down Trent Lane."

From these expressions, we manually extracted the relatum and locatum for each preposition in each of the 720 expressions. Many of the expressions were complex, involving other elements (e.g., adjectives, adverbs), but these additional elements were disregarded. Specific place names were replaced with the relevant geographic feature type to avoid bias specific to particular locations. For instance, the first example in Table 2.1 becomes *"beside the lake, 1km north of stream".* We did not use any automation process to identify the grammar and extract these elements.

[25] https://soils.landcareresearch.co.nz/soil-data/national-soils-data-repository-and-the-national-soils-database/

[26] https://www.landcareresearch.co.nz/resources/collections/allan-herbarium

[27] https://www.landcareresearch.co.nz/resources/collections/nzac

[28] https://www.landcareresearch.co.nz/resources/collections/pdd

[29] http://geospatiallanguage.massey.ac.nz/ncglindex.htm

2.4 Data collection

We collected assessments of the semantics of each expression from respondents using Amazon Mechanical Turk[30], a platform for crowdsourcing responses to Human Intelligence Tasks (HITs) that has been used in a range of research projects (Schnoebelen and Kuperman 2010; Mason and Suri 2012). We created a separate HIT for each of our 720 expressions, and Mechanical Turk Workers were paid US$0.1 to complete each HIT. Workers could complete as many or as few HITs as they liked but could only complete a given HIT once. We collected 30 responses (from 30 different respondents) for each of our 720 expressions (30 expressions for each of the 24 geospatial prepositions), in order to ensure that the results were not biased by responses of one, or a small number of respondents.

Each HIT page contained introductory instructions (see Appendix B - Full instructions for experiment), an explanation of geospatial prepositions, and an ethical statement. The research was conducted in accordance with the Massey University Code of Ethical Conduct for Research, Teaching and Evaluations involving Human Participants, and low-risk ethical approval was obtained from Massey University Ethics Committee prior to the commencement of data collection[31]. For each HIT, we showed respondents an expression and asked them to select diagrams from a set of 55 (see Figure 2.2) derived from the Geometric Configuration Ontology (GCO) (Stock 2014) (Appendix A - Geometric configurations Stock (2014)), with the inclusion of more than one diagram for some GCO concepts to reflect different geometry types, in line with the two basic models of representation of place as regions and vectors (Zwarts 2017). The GCO provides a comprehensive ontology of different geometry configurations extracted from the literature and text analysis, and includes topology, distance, linear orientation,

[30] https://requester.mturk.com/

[31] Ethical Approval Number 4000021526

horizontal projective orientation, direction, adjacency, collocation, and object parthood. The diagrams depict the locatum (in red) and the relatum (in blue) and include spatial relations that are relative to the position of the observer (projective, egocentric frame of reference) (Diagrams 1-10) and cardinal direction relations (absolute frame of reference) (Diagrams 11-26). The observer was represented by a stick figure while the direction of North was represented by an arrow labelled with the word ‘North’. Several diagrams reflect multiple kinds of spatial relations (e.g., Diagram 53 depicts the topological *contains* relation and a parthood *centre of* relation).

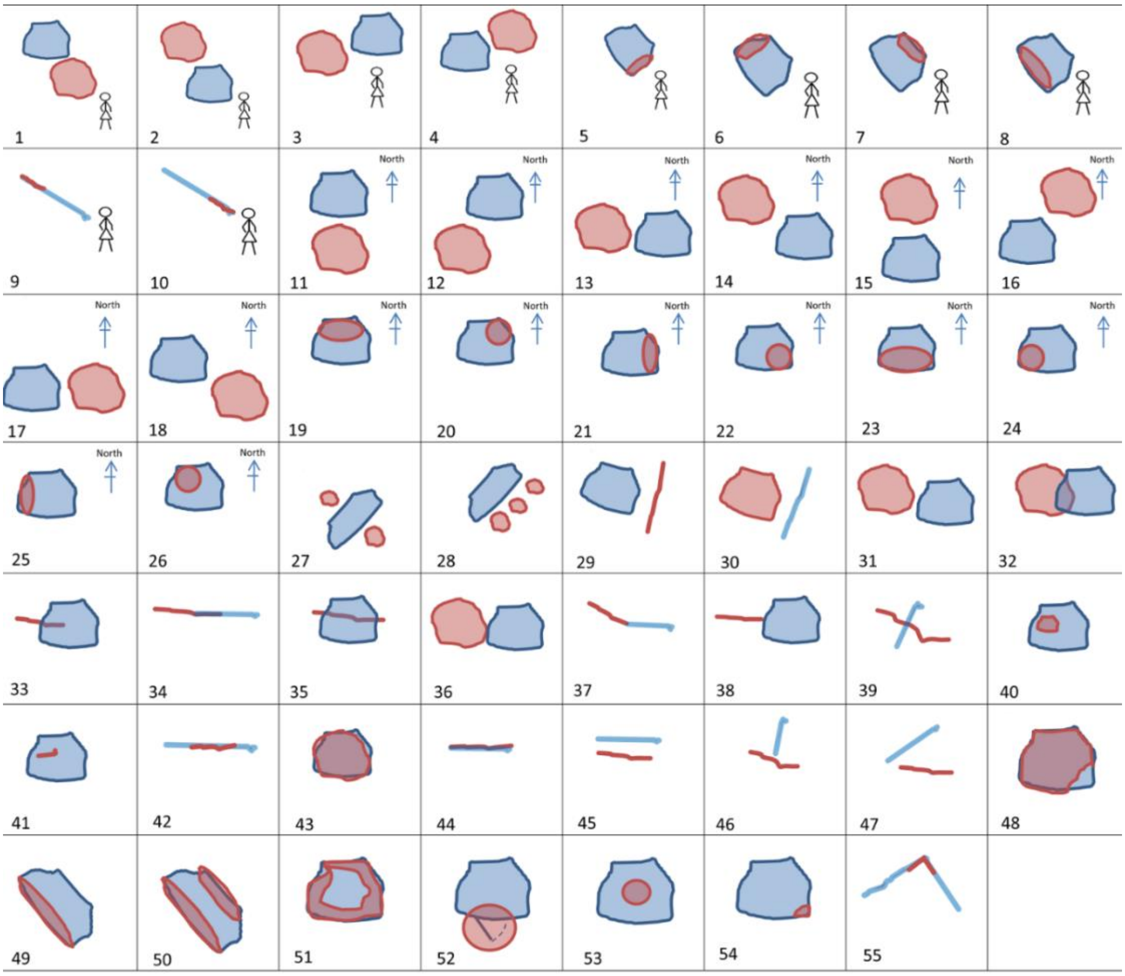


Figure 2.2. Diagrams of the human subject experiment for the HIT task

The diagrams intentionally omit contextual information (e.g., scale, location of other objects in the scene). This is because our goal is to focus on the semantics of spatial relations and their senses that occur across a range of different situations, relata and locata, rather than through a single relatum-locatum pair (Egenhofer and Shariff 1998; Shariff et al. 1998; Mark and Egenhofer 1994), or a specific aspect of context (e.g., Tenbrink 2008). We acknowledge that this approach excludes a deeper level of understanding of contextual aspects of geospatial preposition semantics but leave this for later work.

We asked respondents to select at least one and no more than 3 diagrams for each expression (in case a single diagram didn't exactly reflect the expression and additional diagrams were needed), and to specify closeness of match from a half-Likert[32] scale with options: *agree somewhat, agree,* and *strongly agree.* To remove bias created by the order of the diagrams in the experimental stimulus, we produced 100 different diagram matrices, each containing the same diagrams, but in different orders (changing the order of diagrams in Figure 2.2). Each of the 720 HITs was sequentially allocated one of the 100 diagram matrices.

The experiment was restricted to fluent English speakers through self-selection (workers were asked to proceed only if they met this criteria), since prepositions (and not least geospatial prepositions) are one of the more difficult aspects of English for learners to obtain (Bitchener et al. 2005; De Felice and Pulman 2008).

2.5 Analysis

From the 21600 HITs (30 responses x 30 expressions x 24 geospatial prepositions), 956 blank HITs were submitted. These responses were rejected (Mechanical Turk provides the option to accept or reject responses before payment) and we re-ran the rejected

[32] The negative half of the scale is removed because respondents are asked to select diagrams that they consider do reflect the expression.

expressions to get new responses. The total number of respondents was 921 and the majority completed fewer than 21 HITs (Figure 2.3).

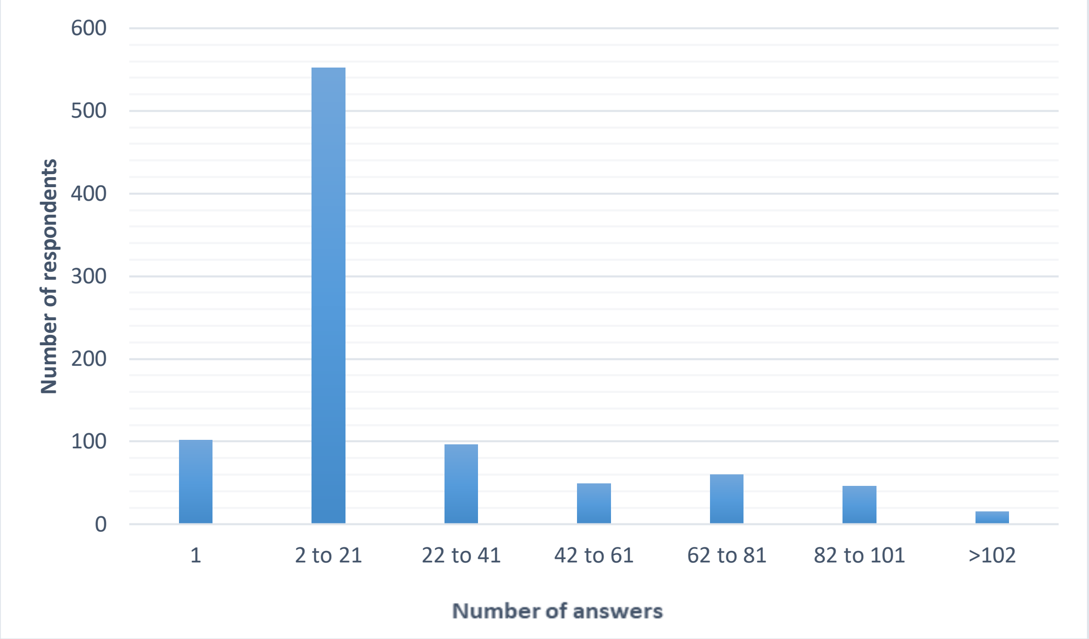


Figure 2.3. Number of respondents and number of HITs completed

We calculated a total agreement score for each expression – diagram combination using the following formula (Equation 2.1):

Total agreement score $_{\text{expression, diagram}}$ = $(\sum_{k=0}^{n} response_k \; * \; weight_k)/n$ Equation 2.1

We assigned a weight to each response: 0.5 for "*agree somewhat*", 0.75 for "*agree*" and 1 for "*strongly agree*". $Response_k$ specifies an individual response and has a value of 1 (for each respondent who selected the diagram concerned), $weight_k$ is the weight of that response and *n* represents the total number of responses for the given description. We produced a 55-dimension vector (one number for each diagram representing the average weighted agreement across all respondents with the diagram for that expression) for each expression. We refer to these vectors as *expression diagram vectors*.

Previous studies have shown that although Mechanical Turk can be a cheap and fast platform for collecting data, sometimes the quality of data may not be at the level that requesters expect (Schnoebelen and Kuperman 2010; Mason and Suri 2012). When computing the *Total agreement score,* we average across all 30 responses for a given expression in order to reduce the effects of outliers amongst respondents, and we further removed noise from the vectors by considering only average values that were equal to or greater than 0.1 (all average values for a dimension below 0.1 were set to zero). Very low numbers for a given diagram in an expression diagram vector suggest that only one or two people selected the diagram, and therefore it does not reflect a common view across all, or even most, respondents.

We then produced a single diagram vector for each geospatial preposition by calculating an average score for each diagram across all 30 expressions that contained the geospatial preposition. We refer to these vectors as *preposition diagram vectors*.

2.6 Semantic similarity of geospatial prepositions

In this Section, we use the results from our experiment to explore the semantics of geospatial prepositions and their similarity. We firstly apply quantitative techniques (clustering and t-SNE) to identify groupings of geospatial prepositions and discuss the results from this process. We then study the prepositions using qualitative methods, with an extensional map.

2.6.1 Quantitative analysis, results, and discussion

We apply clustering to the preposition diagram vectors in order to identify groups of semantically similar geospatial prepositions, following the assumption that respondents will select similar diagrams for geospatial prepositions that have similar meaning. We applied several different clustering configurations in order to identify the dominant groupings robustly, as follows:

- We applied two clustering techniques: Agglomerative Hierarchical Clustering (AHC) and K-means (Johnson 1967; Hartigan and Wong 1979).
- We applied the techniques to both the preposition diagram vectors and a modified form of the vectors, in which only the top three diagram values in each preposition diagram vector were retained, and all other values were set to zero (this eliminates all but the most dominant selections), because the top three values show the most frequently chosen diagrams for that specific expression, and thus carry more information than other small values that may be outliers.
- We applied these techniques with different numbers of clusters (3, 5, 7, 9 and 11).

We then calculated the co-occurrence between pairs of prepositions as the percentage of configurations in which they appear in the same cluster, across all of these different clustering configurations (20 in total – 5 x 2 x 2) in order to ensure that our groupings of semantically similar prepositions are not influenced by a particular clustering configuration, using the following formula (Equation 2.2):

$$co-occurrence_{x,y} = \frac{number\ of\ configurations\ in\ which\ x\ and\ y\ are\ in\ the\ same\ cluster}{total\ number\ of\ configurations} \qquad \text{Equation 2.2}$$

We created a co-occurrence matrix representing the pairwise co-occurrence of the prepositions and plot this data on a t-distributed stochastic neighbour embedding (t-SNE) plot (Figure 2.4). T-SNE plots are able to express the similarity between multi-dimensional non-linear vectors in two-dimensional space (Maaten and Hinton 2008).

The t-SNE plot shows several interesting groupings. Unsurprisingly, *in* and *inside* are grouped together. While there are differences in the way these prepositions are used (e.g, *I live in the street* makes sense, while *I live inside the street* is unlikely), there are significant overlaps that suggest this common positioning in the reduced dimension space. Several adjacency and proximity prepositions are grouped together (*next to, near, adjacent to),* while *beside* and *close to* are together, but some distance from the other proximity and adjacency relations.

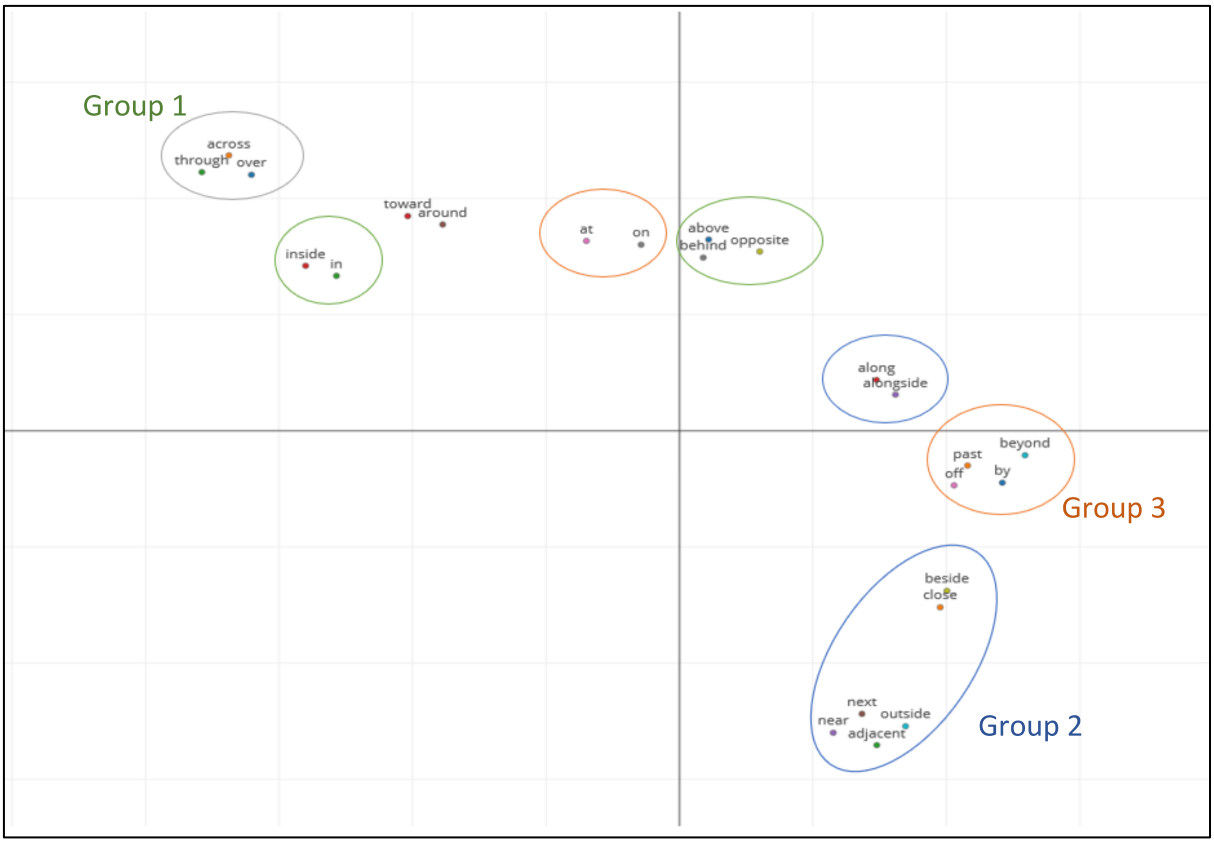


Figure 2.4: t-SNE plot of preposition co-occurrence matrix

The groupings do not reflect the distinction between proximity (*near, close to*) and adjacency (*beside, next to, adjacent to*) that has been identified in preposition typologies (Bitters 2009). While Zwarts and Winter (2000) and Retz-Schmidt (1988) class *beside* as a projective relation, Coventry and Garrod (2004) class it as a proximity relation, consistent with its position in Figure 2.4 with the *close to* relation. Interestingly, *outside* is grouped with *next to, near* and *adjacent to,* although it is not commonly presented as either an adjacency or proximity relation, but rather a topological or containment relation (in that it would typically be considered to refer the situation in which the locatum is external to the containing relatum) (Bitters 2009).

Past, beyond, off and *by* are grouped together in the t-SNE plot. While *by* might be considered more akin to the adjacency and/or proximity relations, the similarity between the four prepositions is further confirmed by the Pearson Product Moment

Correlation Coefficients between the preposition diagram vectors as shown in Appendix C - Similarity matrix of geospatial prepositions, in which the similarity of *by* with *off* and *past* is 0.95 and with *beyond* is 0.7 (Figure 2.5 shows the shaded similarity matrix).

Across, through and *over* are close together in the plot. Although the relations expressed by these prepositions might vary if viewed in three-dimensional space, because our diagrams only depict plan view, there is significant overlap in the diagrams selected.

Above, behind, and *opposite* are also close to each other in the plot, even though they appear to be semantically very different. As for the *across, through* and *over* group, this grouping may be affected by the absence of the three-dimensional view in our diagram, and the tendency for respondents to select diagrams in which one object is above the other, even though the diagrams are not intended to depict the vertical dimension. Thus Diagram 2 was highly scored for both *above* and *behind*. While it is intended to reflect the *behind* relation, given the position of the objects relative to the observer, some respondents also applied it to the *above* preposition. We consider the specific diagrams selected for each preposition and explore these aspects in more detail in the next Section.

2.6.2 Qualitative analysis, results, and discussion

Following Levinson and Meira (2003), we present an extensional map (Figure 2.6) of the three diagrams with the highest agreement for each preposition. Extensional maps are used to highlight the findings of diagram matching experiments and depict groups of diagrams that are most frequently selected for a given linguistic expression (in our case prepositions). Diagrams are positioned on the extensional map in a way that facilitates display of groups of similar diagrams (i.e., diagrams used for the same preposition are grouped together on the map), and most importantly for our work, enables comparison of the semantics of individual prepositions.

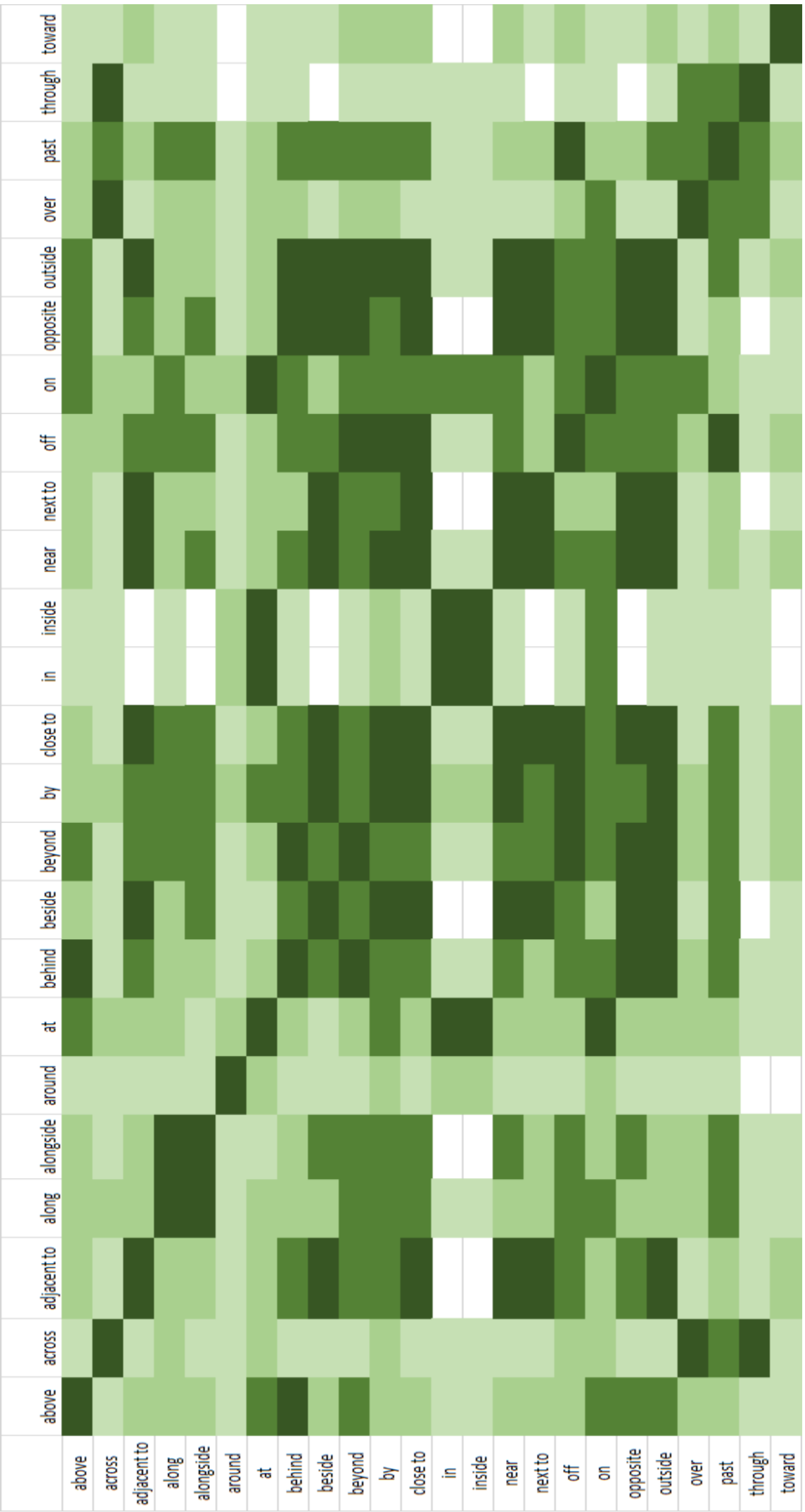


Figure 2.5. Shaded similarity matrix of the prepositions- from darkest (highest similarity) to lightest (low similarity)

The extensional map of our experimental results further elucidates some of the groupings shown in the t-SNE plot. It is important to note that while the t-SNE plot incorporates the full set of average diagram vectors for a preposition, and position on the plot can be influenced by diagrams that have lower agreement scores, the extensional map only shows the three most highly scored diagrams, so gives a more general view of the similarities of the prepositions. Nevertheless, it highlights the explicit distinctions between those views, which is informative.

In the extensional map, *in, inside* and *at* all share the same highest scoring diagrams (Diagrams 26, 40 and 53) those that indicate containment, with greater or lesser degree of centrality in the relatum. As in the t-SNE plot, the proximity and adjacency relations in the extensional map form two distinct groups, but these do not coincide with the distinction between proximity and adjacency. *Beside, by* and *close to* all have the same three highest scoring diagrams, one of which indicates two objects touching, and the other two of which depict a linear object near a polygon object. In contrast, all three highest scoring diagrams for *adjacent to, near* and *next to* show two polygons, in one case touching (which overlaps with those for the *beside, by* and *close to* group) and the other two near each other. This suggests that *beside, by* and *close to* are more appropriate for linear objects than polygonal ones, where *adjacent to, near* and *next to* might be preferred. *Outside*, which was grouped with *adjacent to, near* and *next to* in the t-SNE plot, shares two highly scored diagrams with each of the other two groups, and those groups include linear objects as well as touching and near polygons, indicating more general semantics. *Past, beyond, off* and *by*, which are grouped together in the t-SNE plot, all share the same two highly scored diagrams (Diagrams 29 and 30), as well as one other which they do not share (*past*: Diagram 35, *beyond*: Diagram 2, *off*: Diagram 38 and *by*: Diagram 36). They are the same two diagrams that are included in the top three for *beside* and *close to*: a polygon and a linear object near each other.

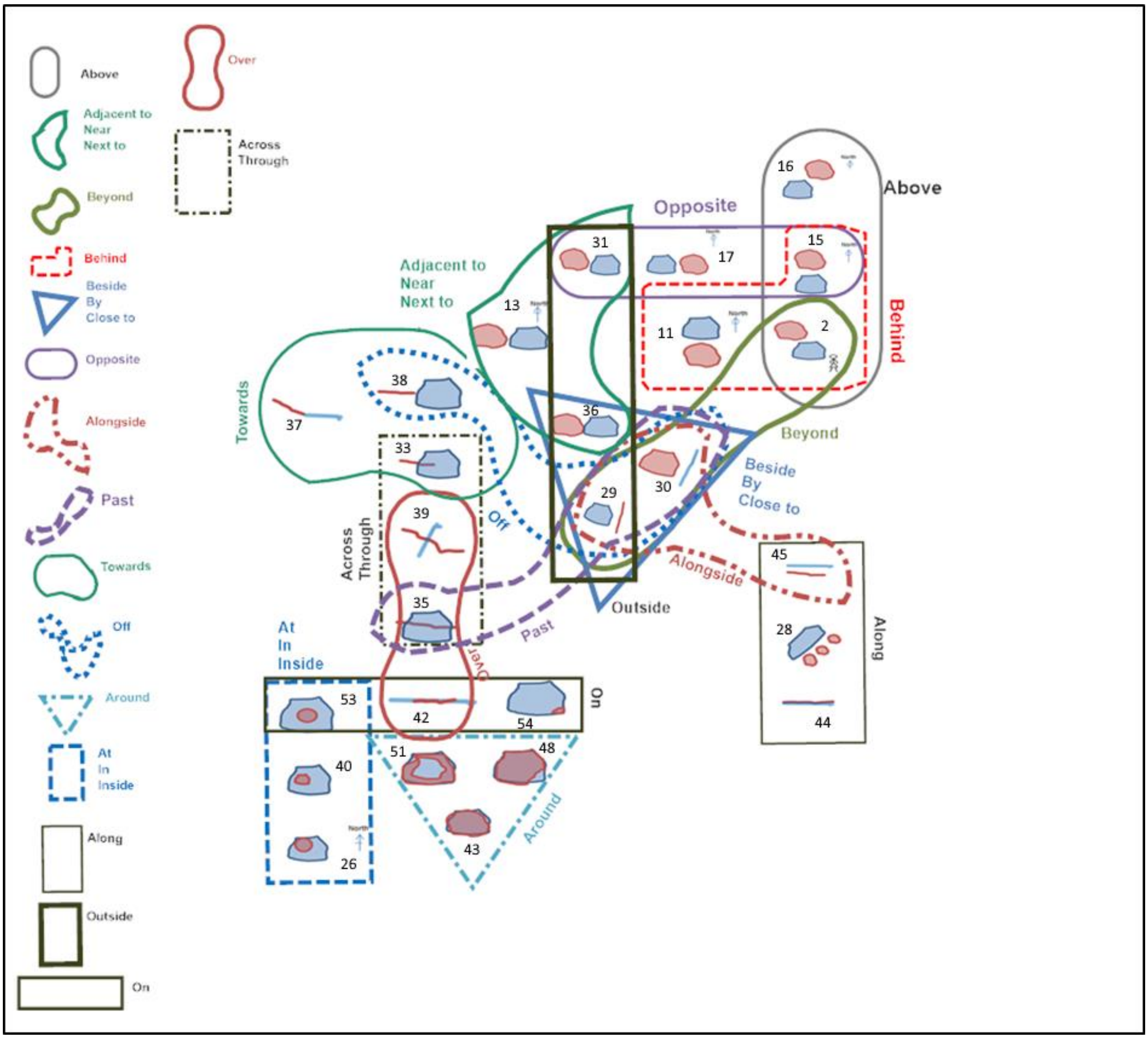


Figure 2.6: Extensional map of geospatial prepositions

These prepositions thus clearly have some shared semantics, while also some additional aspects of meaning that are independent of the others. In the case of *beyond*, this additional diagram is a projective relation, indicating one object behind another, relative to the observer, and is also shared with *behind*. *Past* includes a diagram showing a linear locatum over a polygonal relatum, and all three of its diagrams combine linear and polygon objects. *Off* also includes a third diagram involving linear and polygon objects, with the linear locatum outside and leading up to the edge of the polygonal relatum.

Across, through and *over* were very closely clustered in the t-SNE plot, while in the extensional map, *across* and *through* share the same three highly scored diagrams and *over* shares two of those, with one different diagram. *Across*, *through* and *over* all share a diagram involving two crossing lines, as well as one in which a linear locatum crosses a polygonal relatum. *Across* and *through* (but not *over*) also share a diagram with a linear locatum going into and stopping in the middle of a polygonal relatum. The extra diagram that is highly scored for *over* involves two overlaid lines, one inside the other. It should be noted that our diagrams are only in plan view, so three-dimensional diagrams are not available, even though they may be more suitable for prepositions like *over*, and this may affect the results.

The *above*, *behind,* and *opposite* group from the t-SNE plot is not visible in the extensional map, with the three prepositions only sharing one diagram. *Above* and *behind* share two diagrams, but it is possible that this is because of a mistaken identification of the diagrams concerned as a view from the side, rather than from above, in the case of the *above* preposition. All three of the diagrams selected for *above* show the locatum object geometrically above the relatum object in the diagram (i.e., further up the page), but when the diagrams are interpreted in plan view, they do not reflect the *above* relation. Instead, in plan view, diagrams that show one object inside another may be considered the most accurate depiction of the *above* relation.

The *above, behind* and *opposite* prepositions also reveal the tendency for respondents to ignore the intended meaning of the north arrow in the diagrams. Diagrams 11 to 26 include a north arrow and were intended to show cardinal direction spatial relations (north of, south of) from the original Geometric Configuration Ontology (Stock 2014). Cardinal directions were not included in our set of 24 spatial relations, although a small number of our expressions (14 expressions) did include cardinal direction references in other parts of speech (e.g., *a kitchen on the north side of the town*). In any case, respondents appeared to ignore the north arrows, and see the diagrams as if only the objects themselves appeared, in contrast to Diagrams 1 to 10, which included an observer to reflect spatial relations that were relative to the observer's position (the projective relations), for which the selection of diagrams did appear to take the existence of an observer into account.

It is clear from the extensional maps that for some geospatial prepositions, the three most highly scored diagrams include different kinds of spatial configurations. For example, the top three diagrams for the *around* preposition include one in which the entire locatum covers the relatum, and another in which it is only around the edges of the relatum. Some of these selections of different diagrams suggest different senses of the geospatial preposition. In the next section, we explore preposition senses in more detail.

2.7 Geospatial preposition senses

In this section, we focus on three groups of prepositions that were shown to be semantically similar in the previous section (Figure 2.4):

- *across, through* and *over*;
- proximity and adjacency: *beside, close to, near, next to, outside* and *adjacent to* and
- *past, off, beyond* and *by*.

Again, we combine quantitative and qualitative approaches to study individual geospatial prepositions and their senses, using the diagram vectors and applying Tyler and Evans' (2003) criteria for identification of distinct senses. Within each group, we present our findings, validating them with explanation and examples in the tradition of Talmy (1983), Tyler and Evans (2003) and Herskovits (1985) and relating them to the previous literature. We then further validate our findings using manual classification.

2.7.1 Qualitative and quantitative analysis method

In this section, we interpret the prepositions, their similarity, and their senses using both qualitative and quantitative means. We first apply t-distributed stochastic neighbour embedding (t-SNE) to the expression diagram vectors (in contrast to the preposition diagram vectors that were used in Section 2.6.1), reducing them to x, y coordinates in two-dimensional space. We then apply density-based clustering (DBSCAN) (Ester et al. 1996) to the t-SNE coordinates for the expressions for each geospatial preposition, to identify clusters of expressions that have similar agreement score profiles across the 55 diagrams. We consider each of these clusters a candidate sense for the preposition concerned. We perform manual, qualitative analysis on these clusters using Venn diagrams for each preposition to study the semantics of prepositions and identify their senses. The Venn diagrams[33] allow us to identify which aspects of the semantics of the prepositions (represented by the highly scored diagrams) are shared across all senses (in the section of the Venn diagram where the clusters intersect). The Venn diagrams also clearly identify the aspects of the semantics of each cluster that are distinct to that cluster, as required by Tyler and Evans' (2003) first criteria for a distinct sense (see Section 2.2.4). To address Tyler and Evans' second criterion, which specifies that instances of a sense must not be capable of being inferred from the context they appear, we consider three kinds of similarity between diagrams that may invalidate a given cluster as a separate sense (see Appendix E - Venn diagrams for *across, through* and *over*, Appendix F - Venn diagrams for adjacency/proximity prepositions, and Appendix G - Venn diagrams for *by, past, off, beyond* for examples):

- semantic similarity, determined from the semantic similarity matrix in Appendix C - Similarity matrix of geospatial prepositions;
- representations of the same relation with different geometric types, determined from our mapping from the GCO ontology to diagrams, in which some GCO concepts were mapped to multiple diagrams with different geometry types and

[33] We always use the prefix Venn when referring to these to avoid confusion with the geometric configuration diagrams used in our experiments, which also appear within the wider Venn diagrams

- representations of the same relation with different plurality (one diagram depicts a single object while another depicts multiple objects, but the diagrams are otherwise identical).

While distinct senses may be invalidated by other kinds of similarity than these three (since Tyler and Evans' second criterion is not clearly specified), we consider that these give an indication of clearly similar clusters that do not qualify as distinct senses, and during our manual study of each sense, we require a clearly different semantic intent for each sense and discuss equivocal cases.

Figure 2.7 shows the Venn diagram for *across* (Appendix E - Venn diagrams for *across, through* and *over* , Appendix F - Venn diagrams for adjacency/proximity prepositions, and Appendix G - Venn diagrams for *by, past, off, beyond* show the Venn diagrams for each of the three groups of prepositions discussed in this Section). Each Venn diagram shows the six most highly scored (by maximum total agreement score for any expression within the cluster) diagrams for each cluster, scaled by maximum total agreement score. Diagrams that appear in more than one cluster are scaled for the highest maximum total agreement score, and the maximum total agreement scores for all clusters are shown as vertical bars beside the diagram, colour coded for the cluster. For example, in the Venn diagram for *across* (Figure 2.7), Diagram 35 had the highest maximum total agreement score in cluster 1 (green), with much lower scores in clusters 2 and 3, indicated by the smaller blue and orange bars. The lines between diagrams represent the types of semantic similarity discussed above:

- solid lines indicating semantic similarity are weighted by degree of similarity (Diagram 48 is more semantically similar to diagram 43 than 51, based on the results of our human subjects experiment);
- dashed lines indicating the same spatial relation represented with different geometry types (for example, Diagrams 35 and 39 are both representations of an

overlaps/crosses spatial relations, but in one case the relatum is a line, while in the other it is a polygon) and

- dot-dash lines indicating the same spatial relation with different plurality.

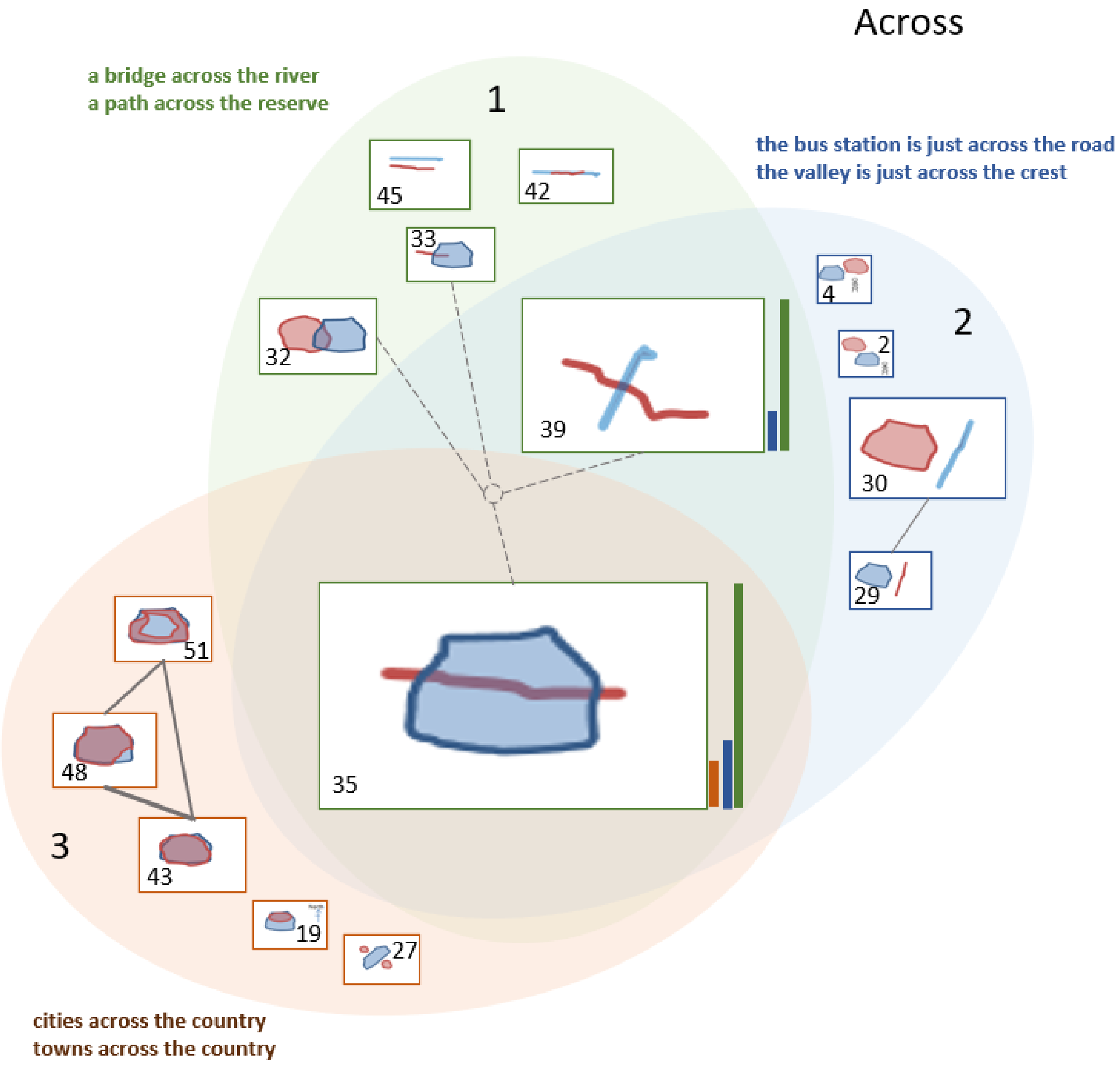

Across
a bridge across the river
a path across the reserve
1
the bus station is just across the road
the valley is just across the crest
2
3
cities across the country
towns across the country
45
42
33
32
39
4
2
30
29
51
48
35
43
19
27

Figure 2.7. Appendix E (a) (Venn diagrams for across)

In cases in which diagrams is sufficiently highly scored to be among the top six and thus appear in the Venn diagram, but that on closer examination has gained that high score based on use with only one expression, we exclude it from the analysis (shown without borders in the Venn diagrams). Such cases are normally due to other aspects of the expression than the original preposition (e.g., referred to part of a relatum) and are considered outliers (e.g., *A doorway close to the head of the north-western staircase*). We also consider that the intersecting section of the Venn diagrams may be used as guidance as to the primary sense of a preposition, given that Tyler and Evans (2003) view the primary sense of a preposition as its prototypical use, and the intersecting portion of the Venn diagram indicates a 'central' meaning of the preposition, but further research is required to verify this.

We also show example expressions from each cluster to assist in analysis of the differences between the kinds of expressions. Appendix D - Summary of sense extraction, summarises the extraction of the senses from the Venn diagrams for all the prepositions.

2.7.2 *Across, through* and *over*

All three of these prepositions have a sense that indicates an overlapping relation between the located and referenced objects. In addition to this sense, we identify two other senses for *across*, one in which there is a third object between the observer and locatum, and the observer is often implied (e.g., *the bus station is just across the road* [from me]) (see Appendix D - Summary of sense extraction). A third sense indicates a relation in which multiple locata appear throughout different areas of the relatum (e.g., *cities across the country*). The previous literature mainly refers to the first, and most dominant (given its role in the intersecting part of the Venn diagram) of these senses (Cooper 1968; Landau and Jackendoff 1993; Lindstromberg 2010). Cooper (1968) also identifies a sense that has some similarities with our second sense (e.g., *the town across the river*), but specifies that "x is located in the space which is contiguous with the distal boundary of y" (p.19).

The *through* preposition (Appendix E - Venn diagrams for *across, through* and *over*(b)) has only one sense, which it shares with *across*. We thus consider that *through* is a specialisation of *across*, being semantically similar to *across* sense 1, but not encompassing the semantics of senses 2 and 3. Expressions in across clusters 2 and 3 in which *through* is substituted for *across* make little sense (*the bus station through the road*) or alter the semantics of the expression (*the valley is just through the crest*). The preposition *through* has not been widely studied, although Dirven (1993) describes spatial and non-spatial senses of through. In the spatial context, the focus of this work is that *through* is used in movements in a 2D or 3D enclosure (e.g., channel, tunnel or surface).

The *over* preposition also shares the overlapping sense with *through* and *across*, as identified by a number of other researchers (Cooper 1968; Brugman and Lakoff 1988; Mackenzie 1992; Tyler and Evans 2001;2003; Lakoff 2008; Kreitzer 1997). The second sense combines the overlapping relation with varying degrees of linear alignment between relatum and locatum and was identified by Lindstromberg (2010). Our third sense places a greater emphasis on verticality, with diagrams such as *the tower over part of the bay* reflecting a meaning that is more akin to *above* than *across* and *through*, a sense that has been identified by other researchers (Bennett 1975; Brugman and Lakoff 1988; Lakoff 2008; Kreitzer 1997). It must be pointed out that only 2-dimensional diagrams were available to respondents, and that these are limited in their ability to represent some uses of *over*, given that they represent a survey perspective (from above) (Taylor and Tversky 1996). A final sense that has been described for *over* but that was not identified in our research describes the case in which the locatum is on the other side of the relatum (*e.g., "Arlington is over the river from Georgetowns, Tyler and Evans 2003, page 48")* (Tyler and Evans 2003; Geeraerts and Cuyckens 2007; Lakoff 2008; Lindstromberg 2010), and is like our second sense for *across*.

Figure 2.8 illustrates the senses of the prepositions in the *across*, *through* and *over* group, and the relationships between them, highlighting the common overlapping sense across all three prepositions that was also identified by Kreitzer (1997). The overlapping sense frequently has a dynamic component of transition relative to the reference object.

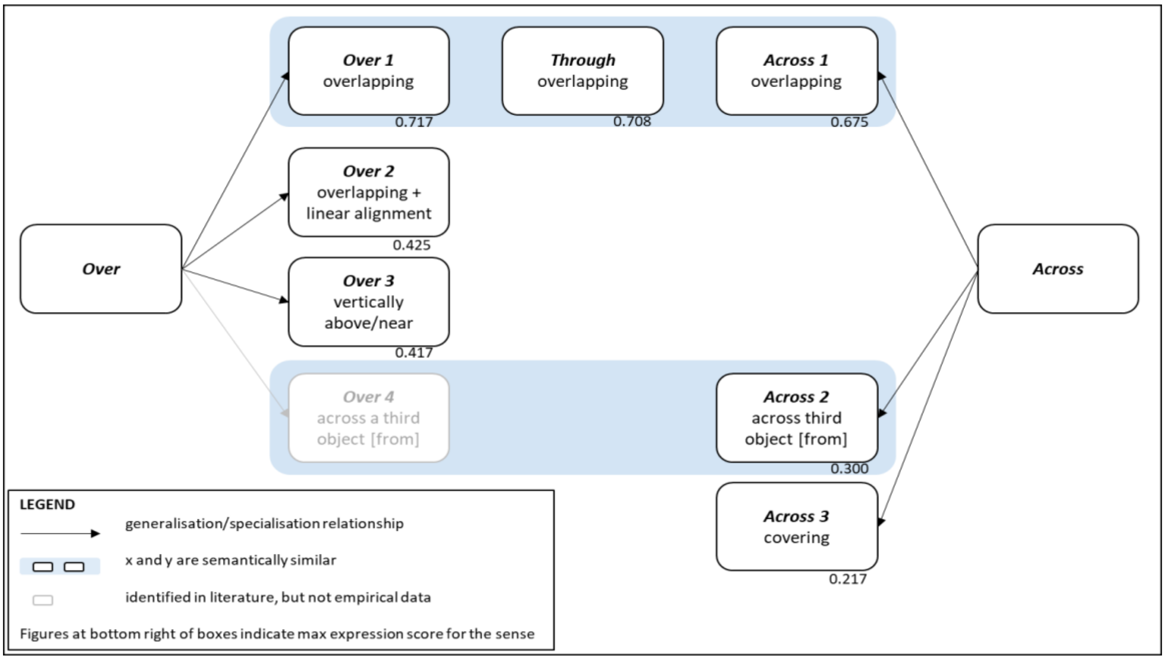

Over 1
overlapping
0.717
Through
overlapping
0.708
Across 1
overlapping
0.675
Over 2
overlapping + linear alignment
0.425
Over
Across
Over 3
vertically above/near
0.417
Over 4
across a third object [from]
Across 2
across third object [from]
0.300
Across 3
covering
0.217
LEGEND
generalisation/specialisation relationship
x and y are semantically similar
identified in literature, but not empirical data
Figures at bottom right of boxes indicate max expression score for the sense

Figure 2.8. Senses of across, through and over prepositions

2.7.3 Adjacency and proximity prepositions

Six prepositions that relate to adjacency and proximity were grouped together in Figure 2.4, and the Venn diagrams for these are presented in Appendix F - Venn diagrams for adjacency/proximity prepositions, and the senses and the relationships between them are summarised in Figure 2.9.

We identify two senses for the *adjacent* preposition: one describing spatial proximity, and another describing the overlap relation. The more dominant touching or proximal sense reflects the sense of *adjacent* identified by Klien and Lutz (2005) in their analysis of Wordnet definitions. There can be some debate about whether the second sense (overlapping) is merely a stretching of the proximity sense (such 'stretched' semantics are described by Herskovitz (1986)) to accommodate vague boundaries. Expressions for which diagram 32 was selected include *land adjacent to the mountain*, and *a wetland adjacent to the avenue*.

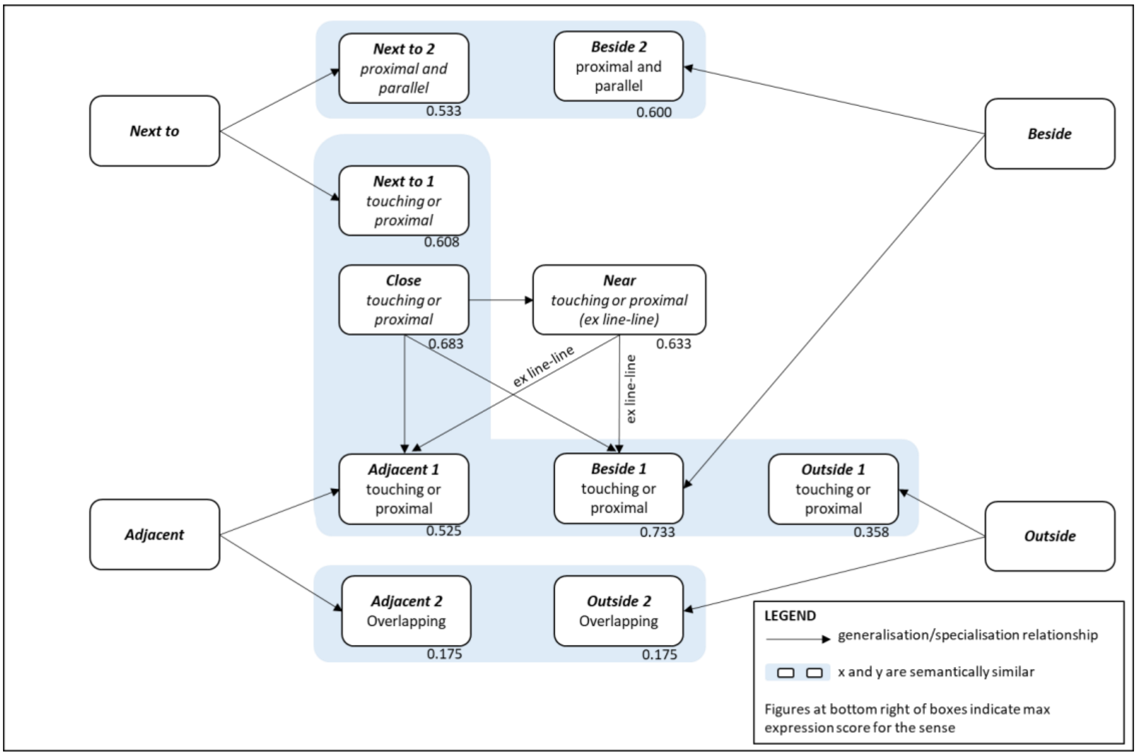


Figure 2.9. Senses of adjacency and proximity prepositions

A similar overlapping sense was identified for the *outside* preposition (Appendix D - Summary of sense extraction), and thus we have included this as a sense of both prepositions, but it should be noted that it is weaker than the other senses, as the maximum scores given by respondents for the overlapping diagrams are much lower. The shared senses, absence of additional senses for either preposition and close positioning in Figure 2.4 confirms the semantic similarity of *adjacent* and *outside*.

The touching or proximal sense is also shared by *beside, close to* and *next to*. *Near* has a similar sense (like *close to*, *near* has only one sense), but interestingly *near* only used the sense for polygon-polygon and line-polygon pairs. The other prepositions also use touching or proximal sense for line-line pairs. In order to confirm this finding, we examined the expressions, and noted that all of the *near* expressions in the data set (randomly extracted from the NCGL) involved polygon objects (with another polygon or a line). We further confirm this by randomly selecting a larger sample of 174 expressions using *near (87 expressions)* and *close to (87 expressions)* from Geograph, and manually identifying the geometry types of the locatum and relatum using the Linguistically Augmented Geospatial Ontology (Stock and Yousaf 2018), which identifies geometry types for a range of geographic feature types. The results showed that 31% of *close to* expressions referred to line-line feature type pairs, in contrast to 3% of *near* expressions. Figure 2.10 shows this distribution.

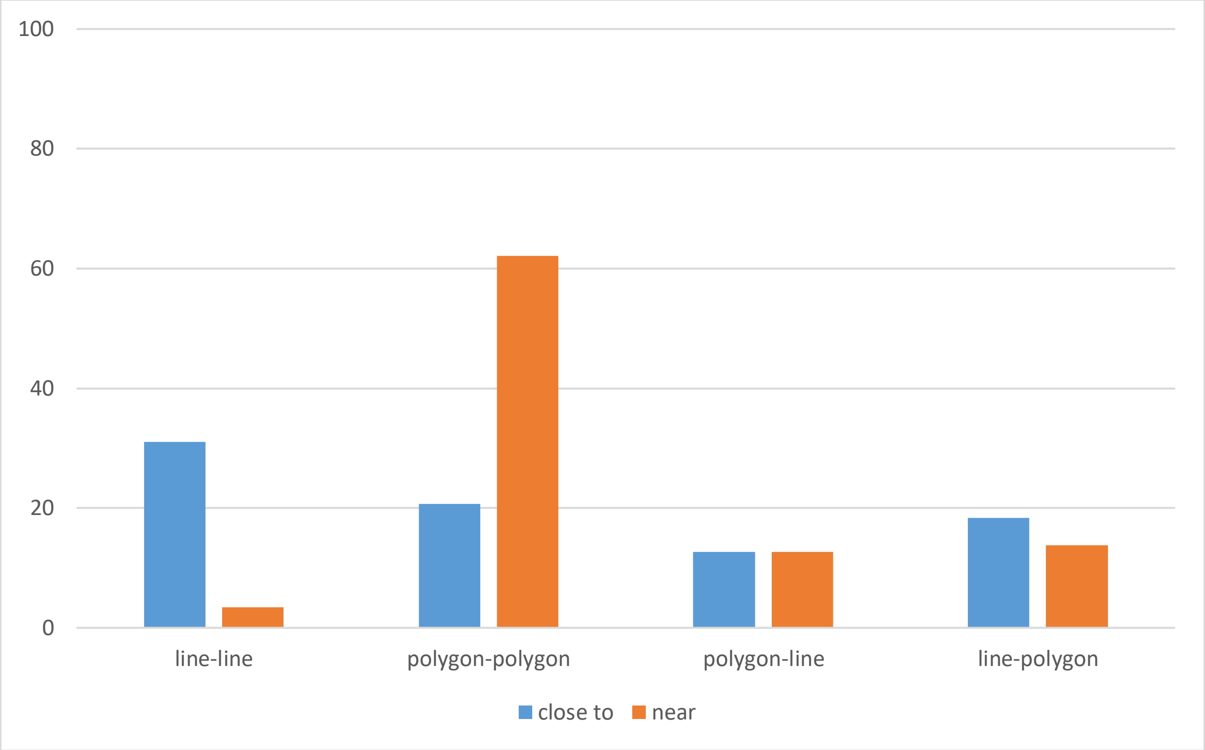


Figure 2.10. The distribution of close to and near prepositions in the expression

An additional sense that was evident for *next to* and *beside* was the proximal and parallel sense, which was used for pairs of linear objects, rows of multiple objects in a line, or sides of a larger polygon object.

Another interesting observation was the relative importance of the proximal and touching aspects of this group of prepositions. Figure 2.11 compares the maximum expression scores for diagrams that depict a touching relation *vs* those that depict a proximal relation. It is unsurprising that proximity is more important than touching for *close to* and *near*, and that touching is more important for *adjacent*. However, *next to* is more similar to *close to* and *near* in that proximity is more important than touching, and *beside* gives equal scores to both.

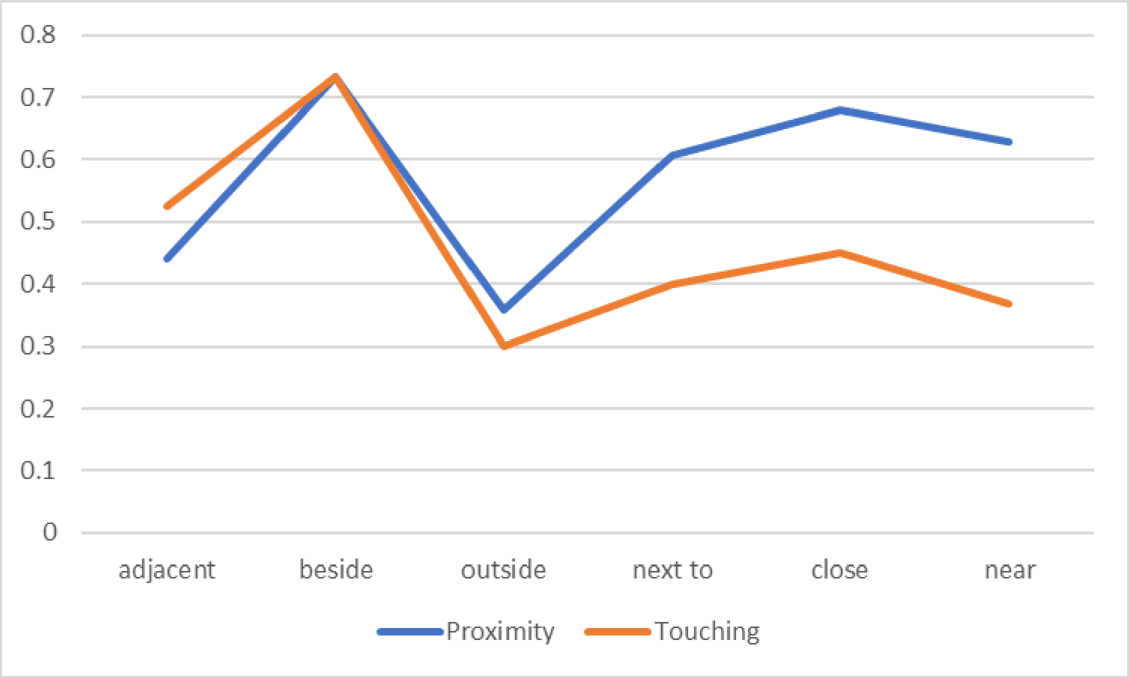


Figure 2.11. Max scores for proximity and touching diagrams

It must be acknowledged that our method does not capture the importance of the vertical elements of the adjacency prepositions identified in the literature (Herskovits 1980; Lautenschütz et al. 2006; Lindstromberg 2010), since we work only with diagrams in plan/survey view. However, the previous literature confirms the role of proximity and the possibility of contact (Mackenzie 1992; Zwarts 1997; Saint-Dizier 2006; Lindstromberg 2010), without identifying the nuances and inter-relationships shown in Figure 2.9.

2.7.4 *Past, off, beyond* and *by*

The third group of prepositions that we examine in more detail also captures varying kinds of proximity, with some additional semantics for particular senses. The Venn diagrams are presented in Appendix G - Venn diagrams for *by, past, off, beyond*, and the senses and relationships between them are summarised in Figure 2.12.

Off, past and *by* all have a sense that conveys proximity. This sense for *by* is particularly used in expressions involving '*by the side of*' (e.g. *a house by the side of the lake*), and has been identified by multiple researchers for linear objects (Cooper 1968; Mackenzie 1992; Landau and Jackendoff 1993; Lindstromberg 2010).

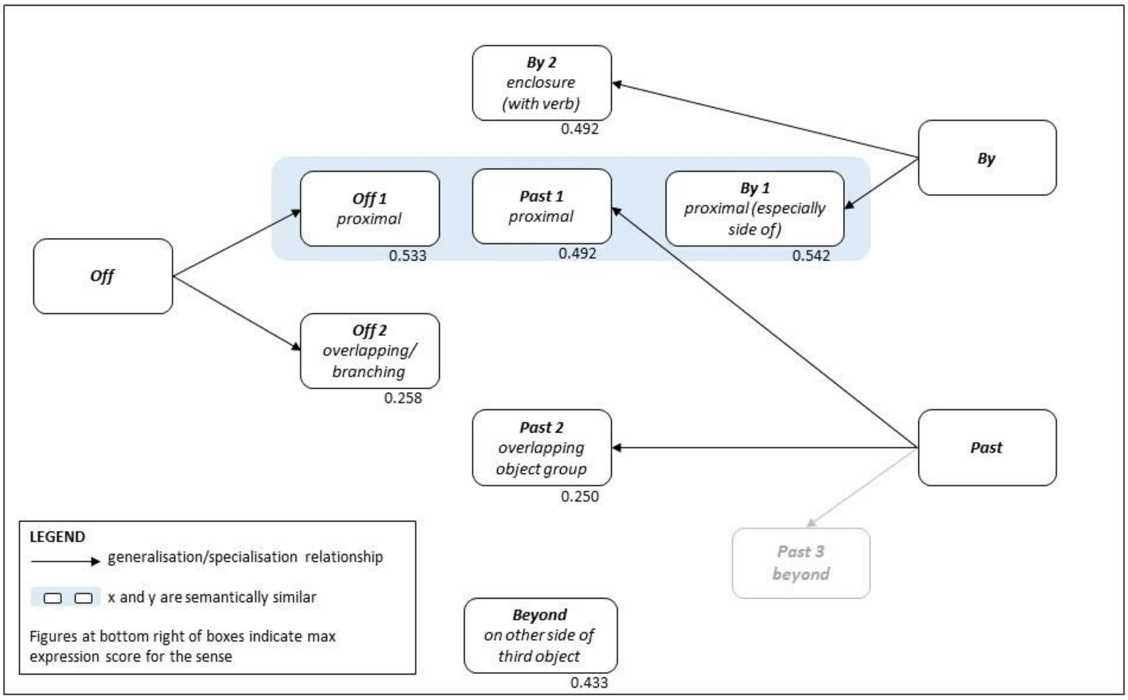

By 2
enclosure
(with verb)
0.492
By
Off 1
proximal
0.533
Past 1
proximal
0.492
By 1
proximal (especially side of)
0.542
Off
Off 2
overlapping/ branching
0.258
Past 2
overlapping object group
0.250
Past
Past 3
beyond
Beyond
on other side of third object
0.433
LEGEND
generalisation/specialisation relationship
x and y are semantically similar
Figures at bottom right of boxes indicate max expression score for the sense

Figure 2.12. Senses of off, by, past and beyond prepositions

Our data also identifies an additional sense that has been discussed by Hois and Kutz (2008), in which particular verbs combine with the preposition to indicate enclosure (*a field bounded by the canal, the platform is surrounded by a ditch*). The previous literature identifies the first sense of *off* shown in our data (Cooper 1968; Landau and Jackendoff 1993; Lindstromberg 2010). Our second sense of *off* is used for pairs of linear features in various relative orientations, and indicates a branching or veering configuration, sometimes combined with a verb (*the avenue off the main road, the path leading off the track*). In addition to the proximal sense, *past* also includes a sense in which the located object overlaps a reference object that is a group, conveying the notion of travelling through that group (*a walk past the buildings, a river past the villages*). This is similar to the sense described by Lindstromberg (2010), but our data mostly confined this sense to grouped objects. Lindstromberg (2010) also identified a sense of *past* that was similar to *beyond*, which we did not observe in our data, possibly because it is a less common use of *past* and did not appear in our sample of 30 expressions. Finally, *beyond* has only one, distinct sense and is thus different from the other three prepositions. That sense is similar to the third sense of *across* and indicates an object on the other side of some reference object from the observer (*a chapel beyond the river*). This extends the semantics of beyond described in the previous literature, which mainly focusses on distance (objects that are far away) (Cooper 1968; Mackenzie 1992; Landau and Jackendoff 1993; Mackenzie 2003; Lindstromberg 2010). We postulate that *beyond* is close to *past, off* and *by* in Figure 2.4 mainly because some respondents selected diagrams 29 and 30, rather than the diagrams that depicted the observer. While this group of prepositions appear close to each other due to common semantics mainly related to the proximal sense, they also have additional senses that clarify the nature of their semantic variation.

2.7.5 Validation of the senses

In addition to comparison with the senses identified in the literature, we validate the senses extracted above in two ways. Firstly, we validate the repeatability of the manual sense extraction process. The candidate and main supervisor independently extracted the senses for Group 3 (*past, off, beyond* and *by*) using the method described in Section 2.7 and the resulting senses were compared. Both independently produced the same senses for all four prepositions using the Venn diagram methodology.

Secondly, we validate the senses by classifying additional data using our senses to identify gaps and/or ambiguities. Two other co-supervisors, who were not involved in the sense identification step, classified a sample of 100 expressions involving each of the

13 prepositions for which we extracted senses. Four of the prepositions were excluded as we only identified one sense for them (*close to, beyond, near* and *through*). The annotators were given a description of the senses (the right most column in Appendix D - Summary of sense extraction) and asked to classify the expressions into each of the senses, with the addition of two other classes: non-spatial use (for uses of the prepositions in a non-spatial sense, as these are excluded from our work here) or other sense (a sense that is not included in the set we have extracted here), and to identify any ambiguous cases. The latter two classes validate our set of senses by determining

(1) *completeness:* identifying any senses that are found in the sample of expressions but were not identified by our approach; and
(2) *distinctness:* identifying cases in which the sense classification was ambiguous, suggesting that our senses are not sufficiently distinct or well defined.

The sample of 100 expressions for each preposition was randomly selected from the combined set of the NCGL and Landcare corpora, excluding the expressions that had been extracted and used in the main experiment. In the case of *adjacent to and beside* there were insufficient expressions, so additional expressions were sourced from Geograph[34], a photo posting web site that includes photo captions and descriptions in which geospatial prepositions often appear. For each of the lower-frequency two prepositions, we conducted a manual search using the geospatial preposition in Geograph's search images function, and manually extracted the first 75 for *adjacent to* and 67 for *beside* expressions (142 in total being the number needed to achieve a total of 100 together with expressions already obtained from NCGL and Landcare corpora) that contained each respective preposition and that included both a locatum and a

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

relatum (some captions in Geograph have an implied locatum, and these were excluded).

The 100 expressions for each preposition were divided among the two annotators with an overlap of 22 expressions to check inter-annotator agreement (each annotator classified the shared 22 expressions plus half of the remaining 78). Following annotation, we calculated the inter-annotator agreement for the 22 shared expressions, achieving an average agreement score across all eight of prepositions of 86%, with a range between 72% (*by*) and 100% (*next to*). The *past* preposition had much lower agreement (50%), in part because an additional sense was identified by one annotator (see below).

2.7.5.1 Completeness of senses

Across the nine senses, only one additional sense was identified by the annotators that had not emerged from our analysis, for the *past* preposition, with a *beyond/after* sense. For example:

- *I'm standing one street from Long Bay College past the roundabout on the right next to the giveaway sign.*
- *I am standing at the first driveway past the side street on the right side of the road as you face downhill…*

This additional sense was identified by Lindstromberg (2010) as discussed in Section 2.7.4, but not found in the 30 expressions that were used for our experiment (and that were a different set of expressions from those used for the validation), due to its low frequency of use (9 expressions out of the 61 expressions included in the validation).

2.7.5.2 Distinctness of senses

We asked the annotators to identify expressions that were of ambiguous class, with a view to determining the distinctness of our set of senses. 6 expressions for the *by* preposition were marked as ambiguous across both annotators. For example, in the expression below, *traversed by* is the ambiguous case that was not identified by our experiment:

- *The Chesterfield canal here passes through the ridge of ground, that is traversed by the road to the north, by means of a tunnel some 270 yards in length and 15 feet in breadth and height.*

Furthermore, 2 expressions for the *off* preposition were identified as ambiguous. For example:

- *The bus ride across the Pyrenean Mountain passes into Andorra is spectacular, although a new tunnel cuts off part of the original road over the pass.*

The ambiguity is mainly due to the verbs that accompany the prepositions (e.g., c*uts off, traversed by*, *crossed by)* conveying a different meaning than the uses of *by* and *off* with verbs in our experiment, in which most of the expressions used *by* with verbs of boundedness (e.g., *surrounded by, flanked by*).

2.7.5.3 Frequency of senses

Figure 2.13 shows the frequency of each sense for the eight validated prepositions, using all 100 expressions and averaging across annotators for the overlapping portions of the sample. As can be seen, most prepositions have a clearly dominant sense, along with other sense/s that are much less frequent.

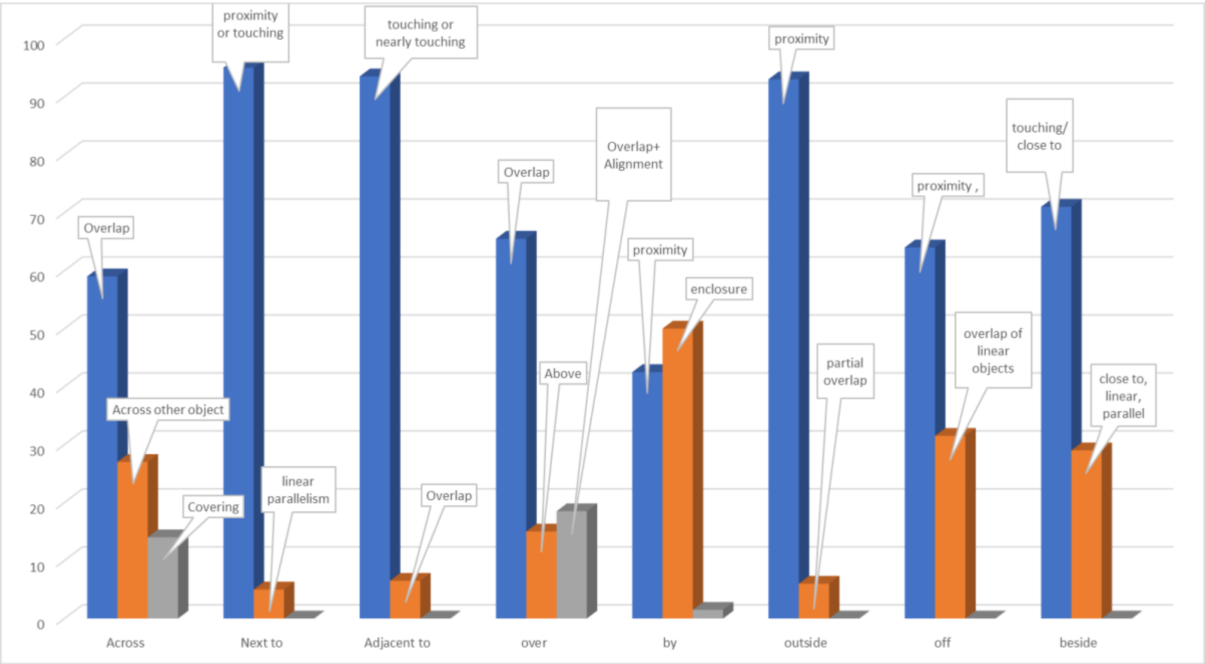


Figure 2.13. Average distribution of each senses

2.8 Conclusion

In this chapter, we used a human subject experiment with 720 expressions across 24 spatial relations and multiple geospatial contexts, in order to study the semantic similarity among spatial relations and their senses. We identified groups of semantically similar prepositions using t-SNE and studied the nature of differences between the prepositions using an extensional map to address Research Question 1. Groups that were particularly similar included *across, through* and *over*; the proximity and adjacency prepositions (*beside, close to, near, next to, outside* and *adjacent to*) and *past, off, beyond* and *by*. We then studied the senses of these three groups of similar geospatial prepositions, identifying the senses and the semantic relations between them using Venn diagrams to address Research Question 2. We validated this work though comparison to previous literature and manual annotation. We found that *through* is a specialization of *across* and *over*, sharing only one of their senses; and that the adjacency and proximity prepositions share a complex network of senses. While these were centered on proximity and touching relations, overlap, orientation and geometry type were also relevant for some senses. The senses of *past*, *off* and *by* were similarly overlapping, while the single sense of *beyond* was distinct. Our results further showed that:

- The *near* preposition is rarely used for line-line relations, with *close to* being preferred to describe proximity in this case.
- The *next to* preposition is used to describe proximity more than immediate adjacency (touching), in contrast to *adjacent*, which more frequently requires a touching relation.

We acknowledge that this analysis provides one perspective on the semantics of the geospatial prepositions: a perspective mediated by the experimental method used. The diagrams were deliberately designed to be context neutral in order to study the generic semantics of geospatial prepositions across a range of contextual situations (although geometry type is an exception to this given that it is a key component of diagrammatic elicitation methods), but the importance of context in the application of geospatial prepositions in specific geographic situations is acknowledged (Talmy 1983; Landau and Jackendoff 1993; Schwering 2007). Future work to build on these findings by exploring specific aspects of context (e.g., image schema, scale) is needed, particularly to identify the degree to which these contextual aspects affect semantic similarity. The focus of our work on two dimensional (survey view) diagrams is another potential limitation, particularly when applied to prepositions that have a clear vertical component (e.g., *above*). Future work using three dimensional diagrams is appropriate to address the semantic similarity of these prepositions and their senses in particular.

In the next chapter, we will investigate the semantic similarity among geospatial prepositions, using a text mining method. We use the human subjects experiment data as a ground truth for the semantic similarity information we extract from the text.

Chapter 3 - Mining the semantic similarity of spatial prepositions from text

ABSTRACT

Spatial prepositions are one of the most important components in a location description, conveying information about proximity, direction, adjacency and topology among other things. However, despite being studied for many years, the semantics of spatial prepositions are still not well understood, particularly given that the use of spatial prepositions can vary with context.

Chapter 2 has shown that we can understand the semantic similarity among geospatial prepositions by analysing human subjects data. In this chapter, we take one step further and investigate whether it is possible to mine the semantics of spatial prepositions from text, particularly focusing on semantic similarity, but also exploring the extraction of richer semantic information about the relationships between spatial prepositions, with the long-term goal of moving towards the automation of the interpretation and generation of locative expressions. We test three similarity methods, including a bag of words technique, with both general and geospatial corpora, and using word embeddings. We compare the results to ground truth data from human subjects experiments.

3.1 Introduction

Spatial language is an essential aspect of communicating information about geographical locations whether in speech or in textual documents. The main distinctive component of such language is the use of words that describe spatial relationships between the location or object to be described and one or more reference objects, as in *I am standing in front of the cinema*. A major challenge in geographical information retrieval is the automated interpretation of locative expressions such as this, which is essential for translation of natural language expressions into georeferenced locations, allowing information about the location of people, objects or events in text documents to be located for applications such as emergency response or navigation. A related challenge is the automated generation of spatial language to provide descriptions of locations and navigational instructions. Many of the words that are used to communicate individual spatial relations are prepositions, though other parts of speech, such as verbs, can also play an important role. A widely acknowledged characteristic of prepositions used in a spatial sense is that they are often vague and overloaded in meaning, in that a single word, such as *at* might imply different interpretations of the corresponding geometric configuration to which they refer (Landau and Jackendoff 1993). Thus, *at* can mean *inside,* or *next to*, or *just outside* of a reference object. Therefore, some spatial preposition can be used interchangeably, while others, such as *beneath* and *above*, have much more specific meanings. Successful automated interpretation and generation of geo-spatial language depends on understanding factors including the geometric configuration to which a spatial preposition refers in a given context and the semantic relationships, such as synonyms and hypernyms, between spatial relational terms. In this chapter, as a step towards the creation of a semantic network of spatial prepositions, we present the results of some experiments to determine the degree of similarity between natural language spatial prepositions.

We investigate the use of three different text-based approaches to determine the semantic similarity between spatial preposition terms. As Firth (1957) mentioned: "You shall know a word by the company it keeps". Two of our approaches use a bag of words method, where a bag of words is a vector of frequencies of occurrence of the words in the document collection (explained further below). The first approach employs a generic corpus (the British National Corpus) and the second a corpus that contains geospatial language (the Nottingham Corpus of Geospatial Language (Stock et al. 2013). The third approach uses GloVe (Global Vectors) embeddings from the Stanford Common Crawl[35] which is a vector space representation of terms obtained using an unsupervised learning algorithm (Pennington et al. 2014). An embedding of an individual word is a reduced

[35] https://nlp.stanford.edu/projects/glove/

dimensionality representation of the co-occurrence of other words with that word. In the two bag of words approaches, the bag of words is a vector containing a dimension for every word used in a document collection, and the values of the vector are a function of the frequency of use of the respective word in the context of the represented spatial preposition, using tf-idf, which attaches more weight to words that are specific to the spatial preposition and less common throughout the document collection. Similarity between a pair of spatial prepositions is then measured by the cosine similarity between their vectors. In the case of the GloVe embeddings, we calculate the cosine similarity between GloVe embedding vectors for the spatial prepositions concerned.

We evaluate the similarity values by comparing each matrix of derived similarity values with a matrix of similarities that was created using the results of human subject experiments to measure the extent to which each of the spatial preposition terms correspond to each of a number of geometric configurations representing a variety of possible spatial prepositions. In addition to reporting the results of this evaluation we highlight a number of observations of the degree to which particular spatial prepositions were found, using these methods, to be similar to many other spatial prepositions, and hence of a generic nature, or different from most other spatial prepositions, and hence more specific in their meaning and usage.

In the remainder if the chapter we review related work in Section 3.2, before describing in Section 3.3 the methods applied. In Section 3.4 we present the results and their evaluation. The chapter concludes in Section 3.5 with a summary of the conclusions and a discussion of future work.

3.2 Literature review

As we discussed in depth in Section 2.2, spatial language is often regarded as serving the purpose of locating objects and places in space (Landau and Jackendoff 1993; Coventry and Garrod 2004).

The idea of exploiting semantic networks in the context of natural language processing is well established. Their potential for disambiguation was recognised in Au (2010) who proposed a semantic network of words, giving examples of the use of informs and is-a links. Fellbaum (1998) proposed some semantic and lexical relations, or factors, that are influential in creating a semantic network of verbs. Some of these factors are entailment, hyponymy, and opposition. WordNet (Miller et al. 1990) also provides a rich semantic network for some parts of speech such as nouns, but its support for semantic relations between terms, such as prepositions, that serve as spatial prepositions is very limited.

In recent years there has been interest in using vector space representations, to infer semantic relations between words. Word embeddings provide such a vector space representation, which can be regarded as form of conceptual space as proposed by Gardenfors (2004), mentioned in Section 2.3. In word embeddings the dimensions correspond to meanings associated with the words that have been mapped to the respective dimension by a dimensionality reduction procedure. It was demonstrated in Mikolov et al. (2013) that cosine distances between word embeddings represent vector offsets that correspond to semantic relations (especially analogy) between the represented words. Subsequent studies (Fu et al. 2014; Attia et al. 2016) have also exploited word embeddings to determine semantic relations (e.g., synonyms, hypernyms) between words. In our work we present an investigations of the use of word embeddings to measure similarity between spatial prepositions as well as investigating similarity between the textual contexts of spatial prepositions represented in bag of word vectors.

3.3 Method

In order to test the ability of text mining approaches to determine the semantic similarity of spatial prepositions, we compared three methods. Using these methods, we tested 25 spatial prepositions, 22 of which were single word prepositions, and the other 3 of which were prepositional phrases (*next to, adjacent to, close to*). The set of prepositions was selected from content obtained from the Geograph [36] and Foursquare web sites[37]. We manually identified the spatial prepositions present in a sample of 1010 expressions from these two sources from central London (780 expressions from Geograph and 230 expressions from Foursquare), excluding spatial prepositions that rely on verbs for their spatial interpretation. For example, prepositions like *to* and *from* usually require a verb for complete interpretation (e.g., *the road comes from the city centre*), and were thus excluded.

3.3.1 Method 1: bag of words with BNC

In the first method, we extract eight-word windows (four words on either side of the spatial prepositions) and build a bag of words model that contains the tf-idf value for the most frequently appearing 1000 words across all of the spatial prepositions. We tested windows of two words and eight words (on either side). For two words the window was so small that we lost some useful information. Eight words was too much, and sometimes went beyond one sentence into the next one. We chose four to keep all necessary information around the preposition and to avoid losing important text or going beyond one sentence. We then calculate the tf-idf sum for each word-spatial preposition pair (summing across all expressions that include the spatial preposition concerned), and thus producing a vector for each spatial preposition, with each value in the vector being the sum of tf-idf values for one of the words in the bag. We then calculate the cosine similarity between pairs of vectors, to establish a measure of the semantic similarity between the corresponding spatial prepositions. We used the British National Corpus (BNC) [38] to create the model, searching only for the spatial prepositions, and did not distinguish between spatial and non-spatial senses. This means that some of the expressions included for a given spatial preposition are likely to contain non spatial senses (e.g., *not all children in the family are gifted* or *in the period between the*

[36] https://www.geograph.org.uk/

[37] https://foursquare.com/city-guide

[38] http://www.natcorp.ox.ac.uk/

first European landings and the First World War...). Metaphoric, figurative, and temporal uses of spatial preposition words and phrases in text are common, and these are included in the bag of words alongside everything else.

3.3.2 Method 2: bag of words with NCGL

Method 2 is very similar to Method 1: bag of words with BNC, except that is uses a geospatial corpus, rather than a general corpus, and thus we aim to reduce some of the non-spatial senses of the spatial relation words and phrases. The Nottingham Geospatial Corpus of Geospatial Language (Stock et al. 2013) contains 10,146 expressions (sentences or paragraphs), each of which contains geospatial content, including at least a location reference and a spatial preposition (i.e., only a place name is not sufficient for addition to the Nottingham Corpus). The content of the Nottingham Corpus was harvested from a 46 different web sites, from a range of domains (e.g., local history, tourism, news).

We performed the same steps using the bag of words approach as for Method 1, producing a second set of similarity measures between the 25 spatial prepositions. Since the Nottingham Corpus only includes geospatial expressions, the incidence of non-spatial uses of the spatial prepositions is likely to be much lower than for the BNC. However, given that the Nottingham Corpus contains some complex expressions, occasional non-spatial senses are still likely to occur. For example, the following expression includes a temporal sense of *at*: *This is known as Stony balk and was at one time a paved way across the field*. However, these non-spatial senses are in the minority.

3.3.3 Method 3: GloVe embeddings

The third method uses the published GloVe embeddings from the Stanford Common Crawl[39] (Pennington et al. 2014). We extracted vectors for each of the 25 spatial prepositions from the 840B token, 2.2M vocabulary, cased, 300-dimension vector data

[39] https://nlp.stanford.edu/projects/glove/

set. For the three spatial prepositions that consist of two words (*close to, next to* and *adjacent to*), we used only the first word, as the data set did not include embedding vectors for bigrams. We tested the use of hyphenated bigrams, which do appear in the GloVe dataset, but these provided negative cosine similarities, in contrast to every other word in the matrix, and thus were not thought to be representative of the bigram spatial prepositions phrases concerned, so were excluded. We calculated the cosine similarity between pairs of embedding vectors to create a third similarity matrix.

3.3.4 Human subjects data

Data from a human subjects experiment (described in more detail in Stock and Yousaf (2018)) was used to calculate similarity between pairs of the 25 spatial prepositions for comparison with the similarity determined using the three methods described above. Human subjects were presented with a series of natural language expressions, each of which contained one of the 25 spatial prepositions, in the context of a particular pair of geographic features (locatum and relatum). The expressions were randomly selected from instances of the selected 25 spatial prepositions in the Nottingham Corpus of Geospatial Language, and then in some cases simplified to exclude non-spatial adjectives and create expressions conforming to a standard construction as described in Stock et al. (2013).

Alongside the expression, respondents were also presented with a matrix of diagrams, each showing a particular geometric configuration between two objects, indicating one of 50 different spatial relations (Appendix A - Geometric configurations Stock (2014)). To avoid overloading the respondents with many diagrams, the diagrams were divided into subsets so that each respondent was presented with only 16 diagrams for each expression. The 16 diagrams were randomly selected from the full set, ensuring that diagrams from the same class of spatial relation (e.g., topological, projective) were included. Figure 3.1 contains an example stimulus. Respondents were asked to select between 1 and 3 diagrams that best reflected the expression, and to rate the degree to which those 1 to 3 diagrams fitted the expression using a half-Likert scale (*agree* *somewhat, agree* and *agree strongly*). This approach was designed to force respondents to select the diagram/s that best matched the expression, and then indicate the degree of match.

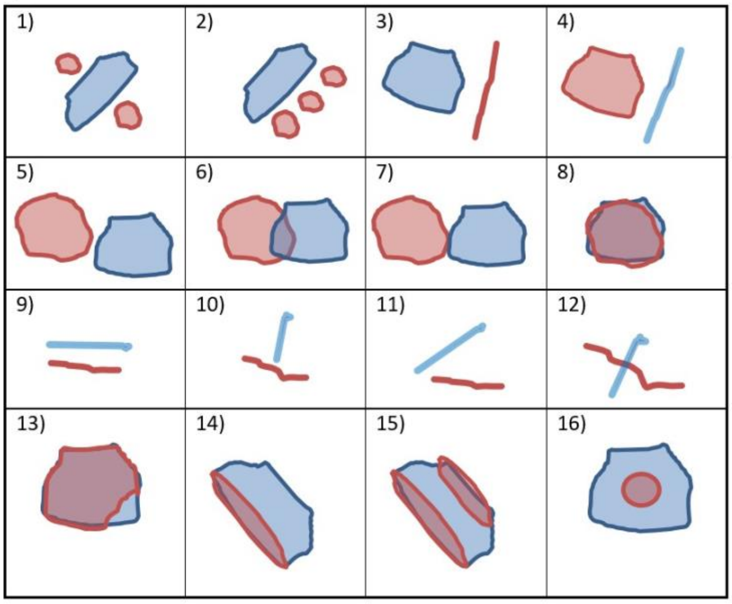


Figure 3.1. Example stimulus for human subjects experiment

In total, 1882 expressions were scored, with each respondent scoring 20 randomly selected expressions, each expression being scored by between 21 and 36 respondents recruited from Survey Monkey Audience. Following the experiment, a score was calculated for each of the 50 spatial relations, using Equation 3.1, where response k represents each individual response, which is multiplied by the weight, depending on the selection of the respondent for the given expression-spatial relation combination, and n is the total number of responses for the expression.

$$GCOscore_{expression,spatial\ relation} = \left(\sum\nolimits_{k=0}^{n} response_k weight_k\right)/n \qquad \text{Equation 3.1}$$

Weights were applied to each response (0.5, 0.75 and 1 for agree somewhat, agree and agree strongly respectively). The score for each expression-spatial relation combination was then calculated as the mean of all individual responses across all diagrams that depicted the relation.

We then created a single vector for each spatial preposition by calculating the mean of the values across all expressions that used the term. Table 3.1 shows the number of expressions that were used to calculate the mean, for each spatial preposition, and as can be seen, there are wide variations in the number of expressions that were used to calculate the mean vectors, and some spatial prepositions have very few (or only one) expressions. Therefore, those spatial prepositions that are included in many are likely to represent a broader range of contexts than those that are included in only one expression. This issue and its implications are discussed further in Section 3.4.1.

Table 3.1. Frequency of spatial preposition terms in corpora

Spatial preposition	Nottingham corpus	BNC	Human subjects
beyond	46	782	1
opposite	68	408	1
close to	54	360	1
between	368	11178	2
toward	24	272	2
behind	56	828	3
off	245	1418	4
past	131	1729	8
outside	95	955	10
inside	49	522	11
near	518	526	13
adjacent	18	101	15
alongside	32	208	16
around	262	2266	19
over	413	6027	19
beside	23	40	20
next to	99	83	56
by	1325	39248	67
through	567	6876	84
along	411	1127	95
at	2259	21223	196
on	2507	36313	302
in	5185	8999	327

3.4 Results

3.4.1 How well do the three methods match the human subjects experiments, and which method matches most closely?

Our first analysis considers how well the text mining methods presented match the human subjects experiments.

Table 3.2 presents the Pearson Product Moment Correlation Coefficient for each of the three methods when compared with the human subjects experiments (and between methods 2 and 3), calculated using the lower triangular half of the diagonally symmetrical matrix. As shown in Table 3.1, the numbers of expressions included in the mean calculations for each spatial preposition vary widely, and we tested the inclusion of different subsets of spatial prepositions by expression frequency, to determine whether the lower number of expressions produced poorer correlations, given that spatial prepositions with few expressions would be expected to represent a smaller number of different contexts and therefore be less representative. Unexpectedly, the reverse was true, with the spatial prepositions with most expressions showing lower correlation between the text mined methods and the human subjects experiments, with moderate correlation (as defined in (Hinkle et al. 1988)) for spatial prepositions with fewer than 50 expressions in the human subjects comparison set, and high correlation with fewer than 15 expressions (around half of the spatial prepositions). This decreased correlation may be due to noise resulting from the multiple uses and meanings of expressions in many different contexts and situations, and therefore may have been matched to different spatial prepositions by respondents. We can see in the Nottingham corpus the results for all spatial prepositions are higher, due to the fact that most of the spatial prepositions appeared in spatial or geospatial senses. Notably prepositions that have the largest numbers of expressions are the most general, with a broad range of applications in different contexts (especially *in, at* and *on*), while most of those with fewer expressions, and higher correlations, are more specific spatial prepositions (e.g., *opposite, between and beyond*, although *close to* might be considered a counter example, that may be considered to have a general meaning, but fewer expressions in the human subjects data set).

Of the three methods, the bag of words (BoW) method using the Nottingham Corpus (Method 2: bag of words with NCGL) provided the best results, with GloVe (Method 3: GloVe embeddings) slightly poorer and the BoW using the BNC (Method 1: bag of words with BNC) noticeably worse.

Table 3.2. Pearson product moment correlation coefficients

Comparison	All spatial prepositions	Prepositions with <100 expressions	<50 expressions	<20 expressions	<15 expressions	<10 expressions	<5 expressions
Number of spatial prepositions	25	22	18	16	12	9	8
BNC/ human subjects	0.285	0.248	0.331	0.491	0.569	0.666	0.727
Nottingham/ human subjects	**0.468**	**0.482**	**0.517**	**0.596**	**0.716**	**0.746**	0.761
GloVe/ human subjects	0.45	0.434	0.515	0.556	0.648	0.707	**0.77**
Nottingham/ GloVe	0.701	0.763	0.818	0.835	0.825	0.889	0.902

The only distinction between Methods 1 and 2 is the corpus from which the context words (the four words on either side of the spatial preposition) were selected, and an additional potentially confounding factor in Method 1: bag of words with BNC, is that multiple senses of the spatial prepositions are likely to be included in the data, while for Method 2: bag of words with NCGL, non-geospatial senses are likely to be relatively infrequent. Since Method 3 also uses generic text but nevertheless provides a clear improvement over the BoW approach, we might expect that the use of embeddings trained on a geospatial rather than a generic corpus would result in additional improvements. This is an area for future work. Given that Method 2: bag of words with NCGL, produced the best results, our subsequent analysis focuses on the data produced using that Method.

3.4.2 Do some spatial prepositions correlate better with human subjects experiments than others?

In Figure 3.2, a matrix of cosine similarity between specific pairs of spatial prepositions using Method 2, four spatial prepositions show high similarity to each other: *at, in, on* and *by*. In addition to these specific, strong pairwise similarities, the sum of the cosine similarities between these four spatial prepositions and all others are also higher than the sums for other spatial prepositions (see Figure 3.3, which presents all geospatial prepositions in order of their total similarity, being the sum of cosine similarities with all other prepositions. For example, the sum of the total cosine similarity of the spatial preposition *on* and all the other 24 spatial prepositions is 15.7). At the other end of the spectrum, *alongside, beside* and *toward* have particularly low sum of similarity. Thus, there is a trend for the more general spatial prepositions, that can be used in different contexts and could often be substituted with more specific spatial prepositions, to have higher total correlation with other spatial prepositions. These spatial prepositions at the top of the list are relations of proximity, collocation, and containment, while some more specific relations appear further down the list. Surprisingly, there are some spatial prepositions (*close to, beside, next to*), that might reasonably be expected to appear higher up the list, and be more similar to other spatial prepositions (e.g. *near* to *close to*).

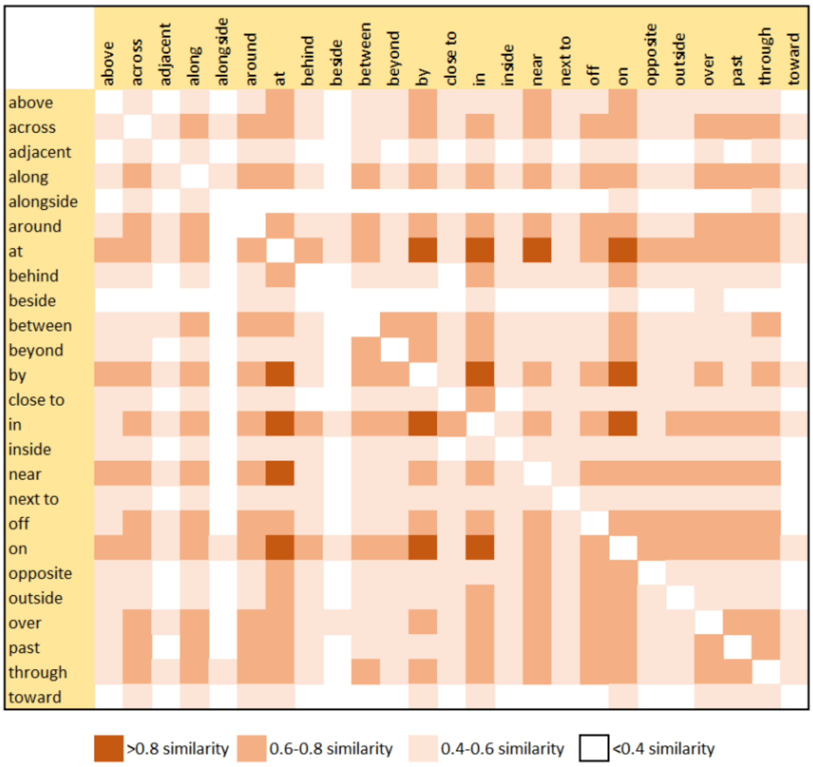


Figure 3.2. Matrix of spatial preposition cosine similarity for method

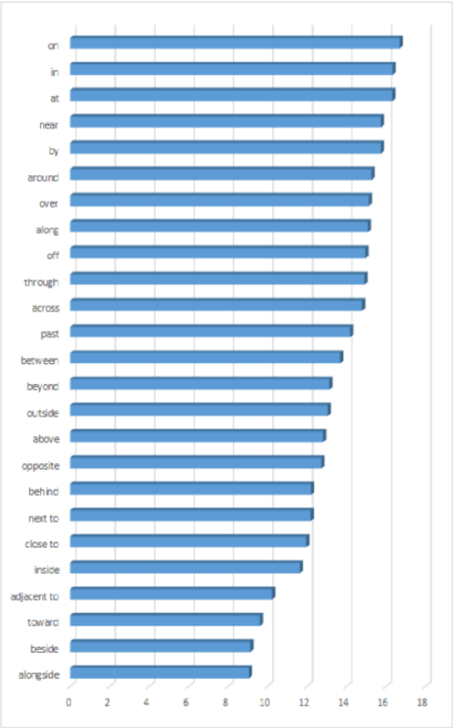


Figure 3.3. Total cosine similarity for each spatial preposition with all others

3.4.3 Do we see clusters among the spatial prepositions?

To answer this question, we used unsupervised clustering techniques to see whether meaningful groups of spatial prepositions could be extracted from the text. The following dendrogram (Figure 3.4) shows the clusters among spatial prepositions using Agglomerative Hierarchical Clustering. To calculate the distance among clusters, the complete linkage agglomeration method was selected which clustered the spatial prepositions in a similar manner to human subject spatial prepositions' similarity. Other methods such as Average and Ward were tested, but they produced sparse clusters that appeared less effective than those from the complete linkage method. The reason might be that the complete linkage method can perform well on dissimilar and distinct clusters and is sensitive to outliers (Schütze et al. 2008). The dendrogram groups together the more general relations (*in, at, on*), discussed in Section 3.4.1. *alongside* and *beside* also appear together, but some other relations that might be considered similar (e.g., *adjacent*, *next to*) do not. However, *next to* is grouped with *outside*, and in some contexts, this similarity is likely to be valid (e.g., *I am outside the post office,* and *I am next to the post office*). Another collection of adjacency/proximity relations (*around, near, by*) appear together in another group. Spatial prepositions that are commonly used in route directions (e.g., *through, across, along, past*) also appear together.

The dendrogram identifies some particular sub-groupings of spatial prepositions, but also highlights the often ambiguous and context-sensitive nature of spatial prepositions, and it may be necessary for a more sophisticated semantic similarity measure to consider different senses of some commonly overloaded spatial prepositions.

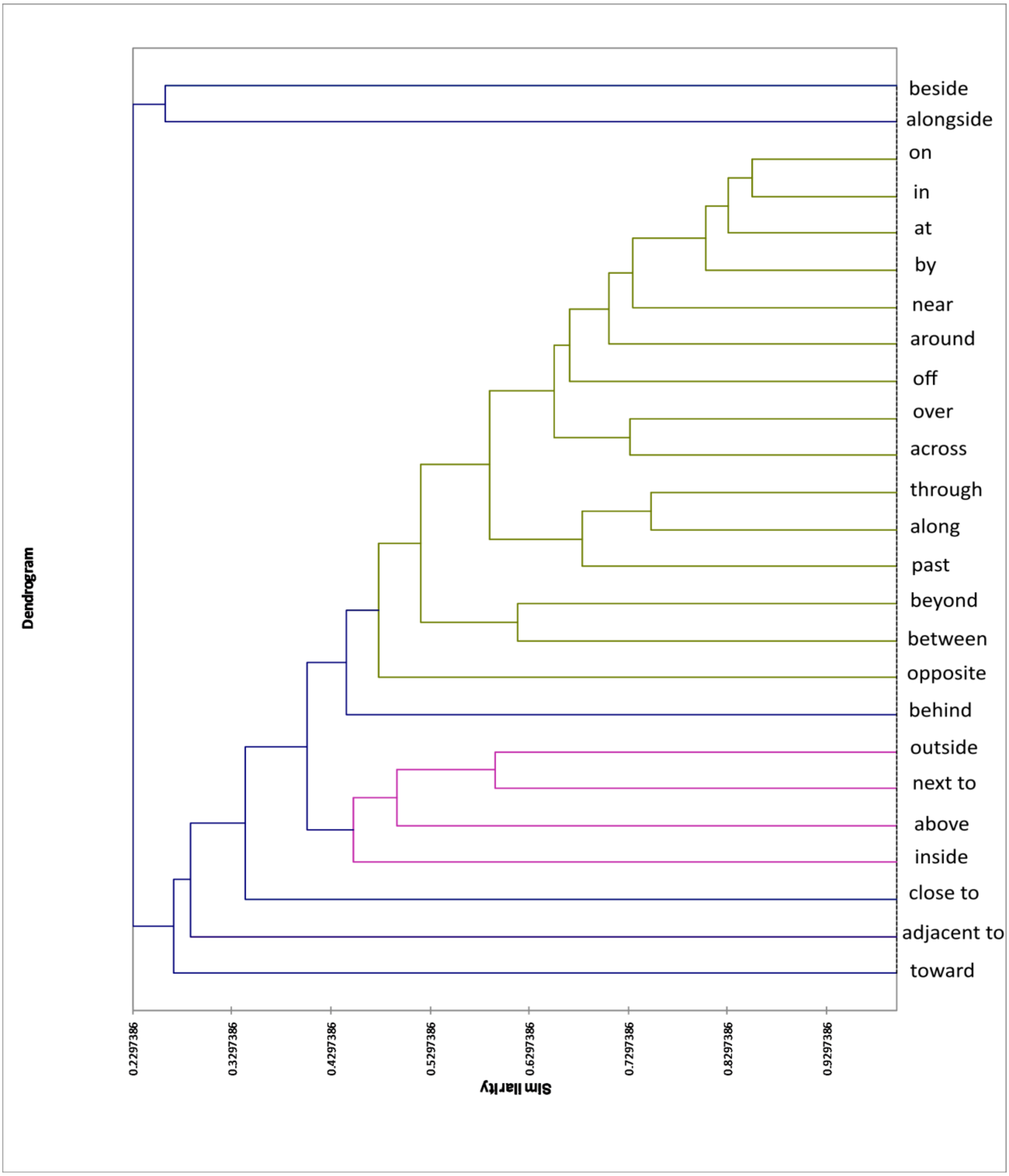


Figure 3.4. Nottingham corpus (method2) dendrogram

3.4.4 Do we see patterns among the highly scoring words in the bag?

We extracted the highest ranked (by tf-idf) words in the bag of word matrix for each spatial preposition and performed part of speech analysis on the top 30 words, classifying the words into 9 of the most frequently occurring parts of speech, accounting for 99% of the words (only 8 words did not fall into these 9 categories, across all spatial prepositions). Figure 3.5 shows the proportions of each part of speech for each spatial preposition in their alphabetic order.

Nouns and prepositions were the most frequently occurring classes, covering 60% of the top 30 words across all spatial prepositions. There is a distinct negative correlation (-0.67 Pearson product moment coefficient) between the frequency of nouns and prepositions across the 25 spatial prepositions.

Table 3.3 ranks the spatial prepositions in order of the frequency of nouns and prepositions, with a group of proximity and adjacency related prepositions (*adjacent, beside, next to, near*) having the highest proportion of noun frequency, and the lowest proportion of preposition frequency. In contrast, the more general prepositions referred to in Section 3.4.1 have lower noun frequencies and higher preposition frequencies, with *in, on* and *at* all appearing near the top of the preposition frequency list.

It is clear from these results that there are differences in the patterns of language used by different prepositions, and this analysis suggests some particular variations by level of specificity of spatial prepositions, and by particular classes of spatial preposition meaning (e.g., topology, proximity, collocation).

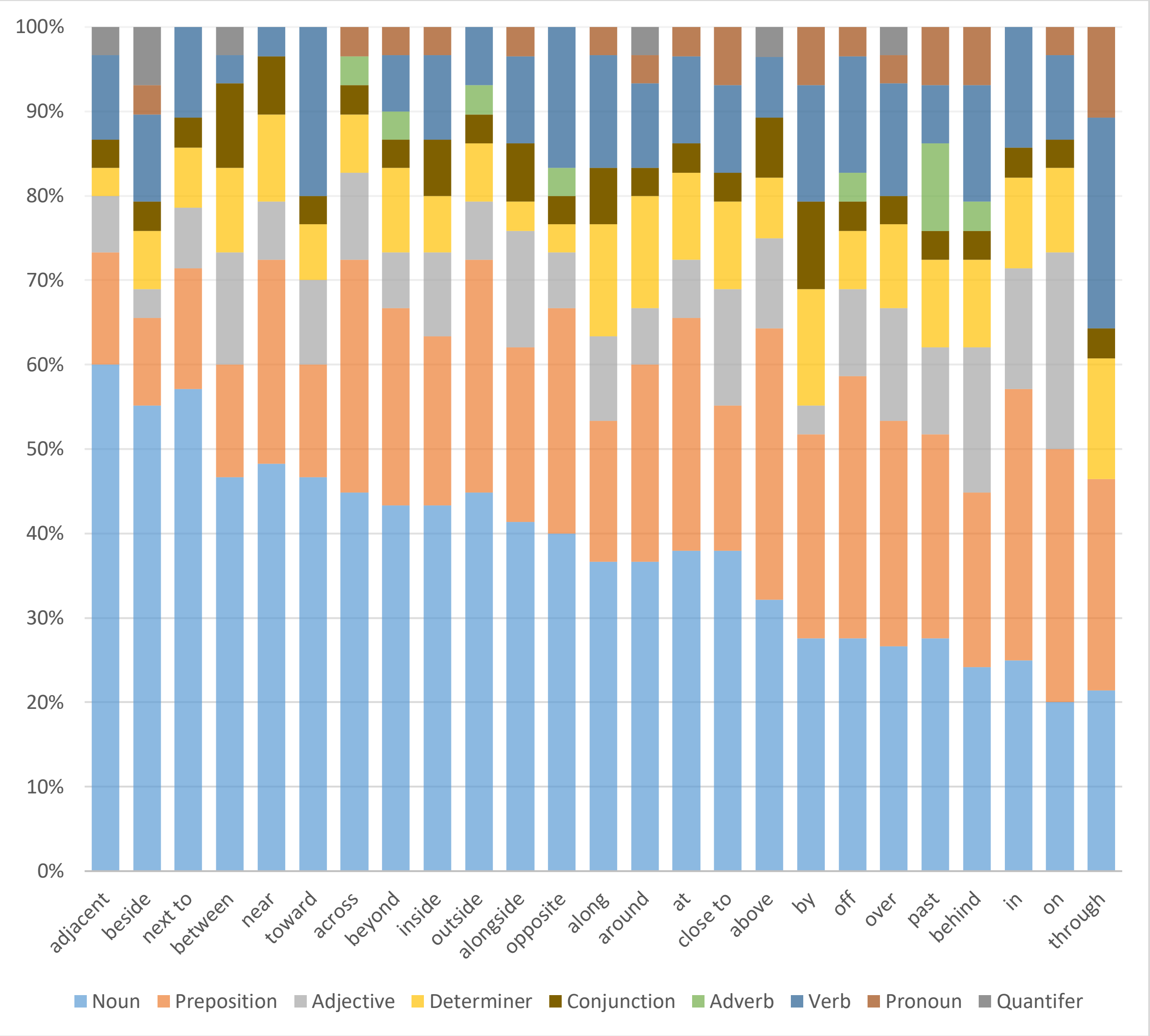


Figure 3.5. Frequency of occurrence of parts of speech among Top 30 words in the bag

3.5 Conclusions and future work

This research suggests that text mining methods show some promise for the identification of semantic similarity and richer relationships among spatial prepositions and are able to identify differences in the way that spatial prepositions are used. Specifically, we identify variations between the spatial prepositions that we consider to be more general (e.g., *at, in, on*) in the sense of being spatial prepositions that could be substituted with other more precise spatial prepositions, and those that have a much

Table 3.3. Rank of spatial prepositions by frequency of parts of speech in top 30 words in the bag

Spatial preposition rank by noun frequency	Spatial preposition rank by preposition frequency
adjacent	*above*
beside	*in*
next to	*off*
between	*on*
near	*across*
toward	*at*
across	*opposite*
beyond	*outside*
inside	*over*
outside	*around*
alongside	*beyond*
opposite	*by*
along	*near*
around	*past*
at	*through*
close to	*alongside*
above	*behind*
by	*inside*
off	*along*
over	*close to*
past	*adjacent*
behind	*between*
in	*next to*
on	*toward*
through	*beside*

narrower meaning. The former, more general spatial prepositions exhibit a higher correlation with other spatial prepositions than the more specific spatial prepositions.

We demonstrate that clustering methods can be used on text data to identify groups of words that have associated meanings, and we show that spatial prepositions vary in the parts of speech that they commonly co-occur with, with proximity spatial prepositions much more commonly co-occurring with nouns than spatial prepositions like (e.g., *at, in* and *on*), which more commonly co-occur with prepositions, potentially due to a need to clarify the spatial preposition in a given context.

Our analysis compares bag of words and word embeddings models on different corpora to see which most closely reflect human cognition. Among the methods tested, the BoW approach with the Nottingham corpus was most highly correlated with the human subjects assessment, but GloVe embedding using vectors extracted from generic data were only slightly worse, and both were much more highly correlated than the BoW method with the BNC. Given that the use of a purely geospatial corpus showed significant improvement for the BoW method over a generic corpus, in future work, we propose to create embeddings from geospatial text, in the hope that this will result in further improvement in the results. This work is a first step towards a broader goal of creating a semantic network of spatial prepositions showing not just the degree of similarity, but also the nature of the relationship between spatial prepositions (e.g., hypernymy, hyponymy, synonymity). It also provides a glimpse of the ambiguous and context-sensitive nature of spatial prepositions, an aspect that must be accommodated in any semantic network.

In the next chapter, we use the semantically similar proximity/adjacency prepositions we obtained from this chapter and Chapter 2 and study the role of context using three different relata on the selection of geospatial prepositions.

Chapter 4 - Geospatial preposition acceptance thresholds and the role of context

ABSTRACT

In natural language location descriptions, people tend to describe object locations relative to other objects (*the house near the river*). As we discussed in previous chapters, geospatial prepositions are a key element of these relative descriptions, and the distances associated with proximity, adjacency and topological prepositions are thought to depend on the context of a specific scene. In this chapter, we use the semantically similar geospatial prepositions obtained from Chapter 2 to extract spatial descriptions from the Google search engine for three sites (as the reference location). We count the frequency with which named locations around them are described (relative to the reference location) using nine geospatial prepositions. Our goal is to compare the acceptance thresholds (distances at which different prepositions are acceptable) for the prepositions, and to study their variations in different contexts using cumulative graphs and scatter plots. Our results show that some proximity/adjacency geospatial prepositions such as *close to* and *near* are used for larger distances than other adjacency geospatial prepositions like *next to, adjacent to* or *beside.* We also found that the characteristics of the reference object (specifically the image schema) influences the selection of geospatial prepositions such as *near* or *in* for a given description.

4.1 Introduction

In natural language location descriptions, people tend to describe their location or that of a point of interest (POI), using relative spatial descriptions. As Kennington (2012) explains relative spatial descriptions describe the locations of objects relative to each other. For example, *garden beside the park* describes the location of the *garden* relative to the *park*. Location descriptions mainly contain three essential elements (Talmy 1983): the locatum (the object for which the location is being described); the relatum (used as a reference location for describing a locatum) and the spatial relation term (specifies the relation in space between the locatum and relatum).

Relative spatial descriptions are important in human communication. Most of the time, people tend to describe location using spatial relation terms instead of formal addresses, as the latter may not be known. They often use known place names as a reference (relatum) to describe their location (for example, *I'm inside the Westfield Albany Mall*). This description locates the person *inside* (the spatial relation term) *Westfield Albany Mall* (the reference object or relatum). Location descriptions can be of critical importance during disaster events, in which they may be used to describe the location of stranded people or dangerous conditions. Similarly, in emergency situations they are often a more usual way for people to describe the location at which assistance is required (Wu and Cui 2018; Hu and Wang 2020) (for instance *there is a fire in the house on Victoria street, next to Mitre 10*). The development of methods to interpret and generate natural language relative location descriptions is thus useful for a number of important applications.

Most of the previous works on georeferencing relative spatial descriptions focused on toponym recognition and disambiguation (Leidner 2008; Lieberman and Samet 2012; Karimzadeh 2016; Kamalloo and Rafiei 2018; Kew et al. 2019) and did not take into account the role of spatial relation terms that modify the spatial description. For example, in a description such as *behind the Shell building*, considering only the toponym is not useful and *behind* should be considered in order to interpret the description accurately. To do this, it is necessary to understand the meaning of spatial relation terms, and the areas in which a given spatial relation term may validly be used to describe location (for example, how near does a locatum have to be to a relatum for *near* to be an appropriate spatial relation term for that locatum-relatum pair). To address this question, a number of models have been developed for specific spatial prepositions, known as acceptance models, applicability models or spatial templates (Moratz and Tenbrink 2006; Hall et al. 2011; Chang et al. 2014; Hall et al. 2015; Skoumas et al. 2016; Yu and Siskind 2017; Du et al. 2017; Platonov and Schubert 2018; Chen et al. 2018; Collell et al. 2018). These models may describe metric properties such as distance or orientation in quantification tasks or the possible area of an object in qualification tasks, and are often probabilistic or predictive, describing areas in which a given preposition is highly suitable, compared to others where it may be borderline. These models may be combined with the known location of the relatum to determine the areas of likely location of a locatum described in a relative location description, to provide automated georeferencing.

In this chapter, we address two gaps in the previous research. Firstly, previous work has mostly focussed on the task of developing acceptance models for individual prepositions. Here, we compare the models for different prepositions in order to study their semantic similarities and differences.

Secondly, the importance of contextual factors on location interpretation has been emphasised in a number of previous works (Herskovits 1985; Johnson 1987; Morrow and Clark 1988; Mark and Frank 1996; Yao and Thill 2005; Gronau et al. 2008; Platanov and Schubert 2018). As Johnson mentioned "Given a center and a periphery we will experience the NEAR-FAR schema as stretching out along our perceptual or conceptual perspective. What is considered near will depend upon the context, but, once that is established, a SCALE is defined for determining relative nearness to the center" (Johnson 1987, p. 125). Herskovits (1985) also claimed that contextual factors influence the selection of prepositions and their interpretation in locative descriptions. She pointed to variation in the locatum/relatum selection based on prepositions. However, context has thus far only been included in acceptance models in limited ways. For example, Chang et al. (2014) and Yu and Siskind (2017) predicted the location of objects in an indoor environment using projective and topological spatial relations (including prepositions), but did not consider context, while works such as Collell et al. (2018) and Malinowski and Fritz (2014) considered contextual factors such as embeddings, relatum type and size to predict the location of objects and identify particular objects in images respectively. We address this gap by comparing differences in acceptance models across three different contexts (three sites) for several qualitative geospatial prepositions (*adjacent to, at, beside, close to, in, on, near, next to* and *outside*) with data scraped from the web through Google searches. We use the frequency with which a specific preposition is used to describe the location of a specific named place with respect to another named place, to produce acceptance profiles that model the distances (between relatum and locatum) for which a given preposition is used. We use acceptance profiles to model distance between locatum and relatum only, rather than acceptance models (also known as spatial templates or applicability models) (Moratz and Tenbrink 2006; Hall et al. 2011; Chang et al. 2014; Hall et al. 2015; Skoumas et al. 2016; Yu and Siskind 2017; Du et al. 2017; Platonov and Schubert 2018; Chen et al. 2018; Collell et al. 2018) which model the two dimensional space around a relatum, because the prepositions we address describe proximity, adjacency, collocation and containment, rather than direction/orientation. While we acknowledge that the use of proximity and adjacency relations may not be exactly concentric around a relatum (Fu et al. 2005), indicating that there may be some directional effects even for proximity and adjacency prepositions, in this chapter our focus is on the distance referred to with these prepositions, so we use acceptance profiles and thresholds (the distance at which the preposition is no longer acceptable). We visualise the distances at which each of the nine prepositions we considered are used in graph form, comparing the prepositions, and considering their semantic similarity, and we study the impact of context by comparing the use of the nine prepositions across three well-known landmarks in London, UK (Trafalgar Square, Buckingham Palace, and Hyde Park).

We address two specific research questions:

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

RQ2: Is the role of context important in the use of geospatial prepositions?

In Section 4.2, we discuss previous work on acceptance models, the similarity of geospatial prepositions and the contextual factors that influence the use of geospatial prepositions. Section 4.3 defines the data extraction method; Section 4.4 presents the results and Section 4.5 provides a discussion and findings. In Section 4.6, we conclude the chapter by summarising the main points and giving some directions for future work.

4.2 Previous work

Spatial relation terms are the core component of locative descriptions. Herskovits (1985) defined locative descriptions as those containing a preposition, the object that belongs to it (locatum) and other elements of a prepositional phrase such as verbs or adjectives. For example, *"a house beside the river"* is a locative description that has three main elements: the locatum, or located object (*house*), a geospatial preposition (*beside*) and the relatum, or reference object (*river*). As Talmy (1983) and Zwarts (1997; 2005) discussed, spatial prepositions specify the connection between the locatum and relatum, including their geometric configuration and orientation. Thus, spatial acceptance models that define the areas (or in our case distances) in which a given preposition may be applied are important for interpretation of relative spatial location descriptions.

4.2.1 Previous work on acceptance models

Spatial acceptance models have been investigated for several goals, including location prediction (Chang et al. 2014; Collell et al. 2018), selection of an appropriate preposition for a description (Du et al. 2017; Platonov and Schubert 2018) and georeferencing (Hall et al. 2011; Chen et al. 2018).

Chang et al. (2014) and Yu and Siskind (2017) used acceptance models to draw a spatial scene in 3D using textual descriptions and to find objects in videos in an indoor environment respectively. They studied projective and topological spatial relation terms as well as motion prepositions and adverbs (Yu and Siskind 2017) with rule-based approaches but did not consider contextual factors or outdoor geographic scenes in their work. In a geographic context, Hall and Jones (2008) extracted descriptions from the *Geograph*[40] website to quantify the distances between the locatum and relatum for cardinal direction spatial relation terms. In another study, Hall et al. (2011;2015) reviewed projective and proximity spatial relations to assign captions to photos and generate photo captions. They defined density fields based on human annotations of spatial relations and produced density models for each spatial relation (highlighted the area/distance that is more likely to specify a spatial relation). This work is similar to our current approach, but the main differences are: the source of data extraction (they used Geograph image captions while we use a wider scope of source material, scraping from the web more generally); the spatial relations we addressed and our focus on comparison of spatial prepositions, and on the role of context in the determination of distances that are acceptable for a given preposition.

In addition to Yu and Siskind (2017), Malinowski and Fritz (2014) and Lan et al. (2012) used deep learning and machine learning models (pooling, CNN and latent ranking SVM) to retrieve specific objects in image configurations, relying on spatial acceptance models. Their goal is different, but their approach (identifying spatial thresholds) is the same as each other. However, again they did not compare different spatial prepositions (and the sets they addressed are not identical but overlap with our set of spatial prepositions) (Lan et al. 2012; Malinowski and Fritz (2014)) or have not used any contextual factors for the identification of acceptance thresholds (Yu and Siskind 2017).

4.2.2 Effects of contextual factors on spatial preposition selection and interpretation

While previous work on acceptance models has developed standard models that apply across a range of different spatial situations or have developed machine learning models that incorporate only limited contextual factors, it is clear that the individual context in a given spatial scene has an influence on the use of spatial prepositions. Herskovits (1985) identified geometric configuration, use types, and salience, relevance, tolerance

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

and typicality as important in determining whether a preposition would apply in a given situation or not. Tyler and Evans (2003) counted context as an important factor for some spatial prepositions such as *over* and stated that other elements of the spatial description such as the locatum and relatum are key to understanding preposition meaning. In a similar approach, Kemmerer (2005) mentions that the best choice of spatial preposition is the one that fits in that specific scene.

Stock and Hall (2017) and Stock and Yousaf (2018) reviewed the impacts of context on location descriptions and we discussed them earlier in Section 2.2.1. In a more similar work on context, Collell et al. (2018) used acceptance profiles to predict the location of objects in photos using some contextual factors such as embeddings and size of locatum. However, their focus is on spatial relations in the form of verbs (implicit) and they did not address prepositions.

In the current chapter, with data collected from the web via Google searches, we investigate the impact of context on the use of several qualitative geospatial prepositions (*adjacent to, at, besides, close to, in, on, near, next to, and outside*). The approach we used in this chapter is different from previous work in several ways. Firstly, we used Google to extract the descriptions associated with specific locata, relata and prepositions, and used frequency of mentions of a locatum-preposition-relatum triple as a proxy for the degree to which a given preposition is acceptable at the distance between the locatum and relatum. This differs from previous work that has addressed more specific types of descriptions, or data from human subjects experiments. Secondly, we compare prepositions, rather the considering them individually. Thirdly, we consider how the acceptance profiles for geospatial prepositions vary across different contexts. This aspect of context has only been addressed through the use of a limited set of features in machine learning models previously.

4.3 Data extraction method

In this section, we discuss the data extraction process. We used a web scraping technique to extract descriptions that contain three elements: *locatum, geospatial* *preposition*, and *relatum* using Google searches. We used actual place names for the locatum and relatum and excluded generic, unnamed geographic features such as a river, street. This enabled us to identify the coordinates of the relatum and locatum used with a specific preposition and calculate the distance between them for further analysis. We use the frequency of references to a locatum by a particular geospatial preposition-relatum combination as a proxy for the applicability of that geospatial preposition. For example, a search for *Green Park next to Buckingham Palace* returned a count of 83 mentions (which we refer to as frequency). We consider that this frequency of use indicates that the *next to* preposition is acceptable for the Green Park-Buckingham Palace locatum-relatum pair.

We selected three relata, being popular tourist attractions in the London area: *Trafalgar Square, Hyde Park*, and *Buckingham Palace* (the actual building and its grounds). In selecting these three landmarks, we aimed for a variety of scales and feature types. Table 4.1 indicates the area and perimeter calculated using QGIS[41] with geometries from OpenStreetMap (OSM[42]) of each site.

Table 4.1. Area and perimeter of the three relata used in our experiment (m^2)

Site	Area	Perimeter
Buckingham Palace	18040	954
Hyde Park	1388013	5629
Trafalgar Square	7741	391

Figure 4.1 shows the locations and sizes of the three relata in the London area. After investigations with a number of other landmarks around London (including *Victoria Embankment Garden, St Martins-in-the-Field Church, National Gallery, Cleopatra's Needle, and Nelson's Column),* it became clear that many landmarks, while popular,

[41] https://qgis.org/en/site/

[42] https://www.openstreetmap.org/

were not sufficiently frequently described as the relatum in spatial relation descriptions found on the web by Google for our analysis. For example, many small landmarks (such as monuments), are rarely used as relata, due to the tendency for relata to be larger, more stable objects (Talmy 1983), and many landmarks are not sufficiently popular to generate frequent mentions on the web.

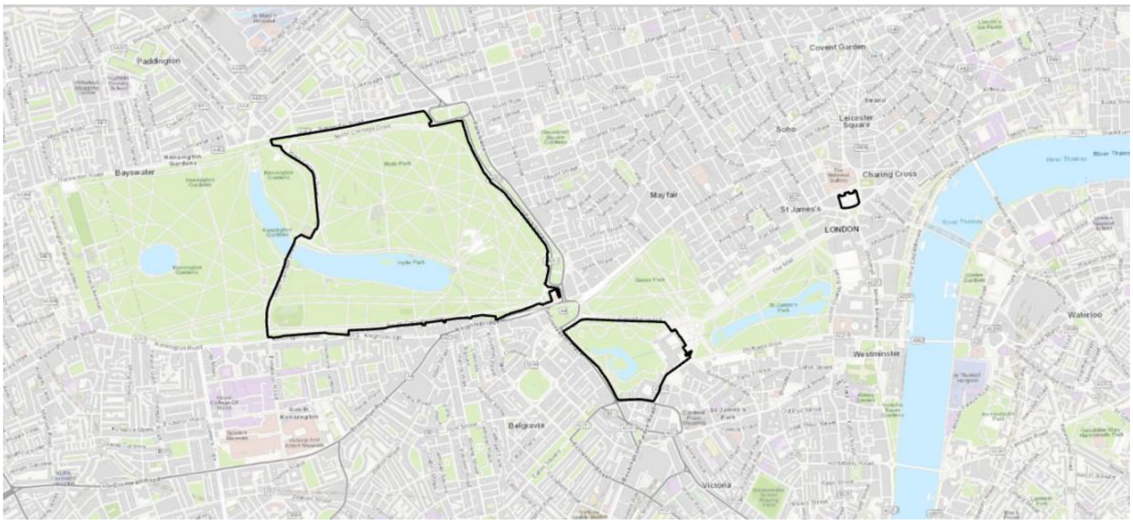


Figure 4.1. Hyde Park (left), Buckingham Palace (middle) and Trafalgar Square (right) on map

We collected data for nine prepositions. Six of the prepositions described adjacency and proximity relations *(adjacent to, next to, close to, beside, near, outside)* and three described topological relations defining contact, collocation or containment *(in, on, at).* We selected these prepositions because our focus was on the variation in the **distance** for which different prepositions are used, rather than orientation, direction or other aspects, and we consider these prepositions to best capture distance variations. Also, topological relations such as *in, on* and *at* have been identified in the literature as being influenced by contextual effects (Mark 1989; Mark and Frank 1996). Furthermore, we are interested in the degree to which the semantics of these topological prepositions 'stretch' the notion of literal containment or collocation to incorporate distances that are some distance from the relatum (Malinowski and Fritz 2014).

Our methodology for extracting the locata that are used with each of geospatial prepositions and each of our relata and their frequency of use consists of 10 steps that we explain in the remainder of this section and display visually in Figure 4.2.

Step 1: Extract all places in the vicinity of each relatum from OpenStreetMap

We used the OpenStreetMap export service to extract all places in the vicinity of the three relata. Given that they are in a similar area, this was done once for all the relata, using a bounding box as shown in Figure 4.3[43].

Step 2: Identify all point and polygon features that are within a specified distance of the centroid of each relatum

From the set returned in Step 1, we identified those features that had centroids within a specified distance (see below) of the centroids of each relatum. We only extracted point and polygon geometries as line objects are segmented into line strings, making analysis more difficult. Also, we consider that prepositions are used differently with linear objects than those that are areal or point-based (e.g., *the house beside the river* indicates proximity, while *the road beside the river* suggests both proximity and alignment). The specified distance for our search was 2km in the case of Trafalgar Square and Buckingham Palace, and 3km in the case of Hyde Park. These distances were selected in order to retrieve a manageable number of locata but with the aim of achieving the extents of acceptable use of the prepositions. Our results indicate that the range of locata analysed for each relata was sufficient in most cases, given that the frequencies reduce to a very low level for the outermost locata that we analysed, but a few of the prepositions required data to extend beyond these distances (e.g., outside), and thus we were not able to establish acceptance thresholds for these prepositions (see Section 4.4.1).

[43] https://www.openstreetmap.org/copyright

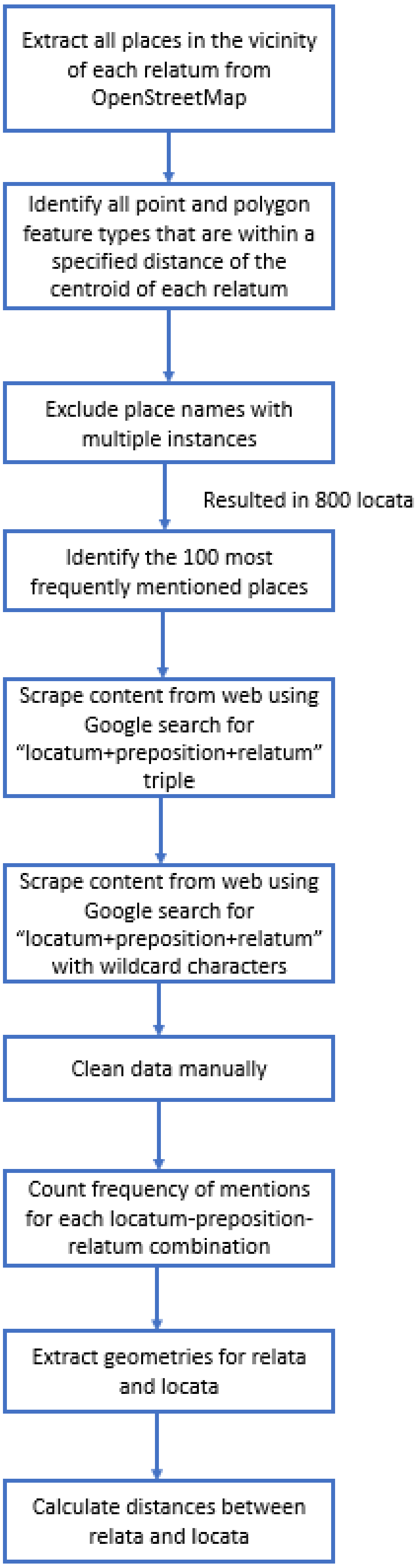
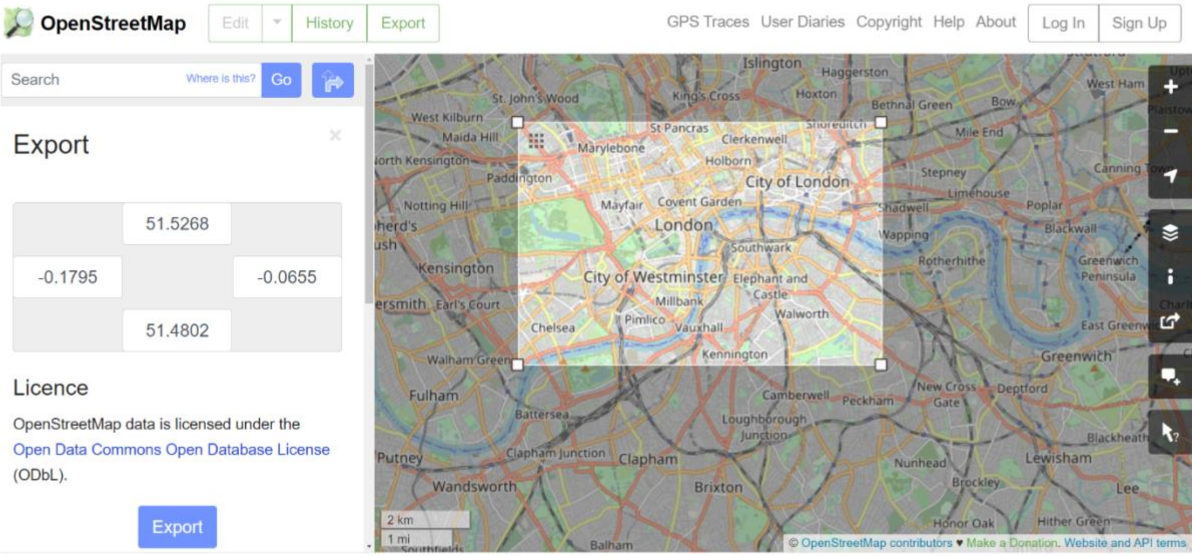

Extract all places in the vicinity of each relatum from OpenStreetMap
Identify all point and polygon feature types that are within a specified distance of the centroid of each relatum
Exclude place names with multiple instances
Resulted in 800 locata
Identify the 100 most frequently mentioned places
Scrape content from web using Google search for "locatum+preposition+relatum" triple
Scrape content from web using Google search for "locatum+preposition+relatum" with wildcard characters
Clean data manually
Count frequency of mentions for each locatum-preposition-relatum combination
Extract geometries for relata and locata
Calculate distances between relata and locata

Figure 4.2. Methodology of data extraction

We selected a larger distance for Hyde Park than for the other two relata because Hyde Park is a lot bigger than the other two sites (as shown in Figure 4.1) and within 2km distance, we could only extract places inside the park and not outside its boundary.



Export

51.5268

-0.1795

-0.0655

51.4802

Licence

OpenStreetMap data is licensed under the Open Data Commons Open Database License (ODbL).

Export

Figure 4.3. Using the export service on the OpenStreetMap to extract place names around each of three relata

Step 3: Exclude place names with multiple instances

After extracting all the place names within 2km of each relatum (3km for Hyde Park), we manually checked and excluded place names that had multiple instances (for example, “McDonald’s” has multiple branches across the London area) in order to avoid ambiguity regarding the coordinate location of the locata referenced by the mentions that we extracted from the web. The manual process involved searching for the place name combined with the word London in Google Maps, and any place names that returned more than one result were excluded.

Step 4: Identify the 100 most frequently mentioned places

After Step 3, we had around 800 locata for each relatum. We next used Google search counts to identify the 100 most frequently mentioned places for each relatum using an automated process. For example, we searched for *The Royal Festival Hall, London*, and recorded the number of times it appeared on Google (the count that appears immediately below the menu options that are below the search input box). After collecting the counts for all candidate locata, we selected the 100 places that had the highest counts. However, the count figure that Google returns is an estimate only, and in particular, often the counts are higher than the actual number of pages available (Matsuo et al. 2007). For example, in some cases the returned results were 2,800,000 on the first page, but when we click on second page, the numbers are decreased because Google omits some results which are repetitive or similar to the ones already returned. Because our main focus was the comparison of numbers and not working with them, we used the first returned counts for each locatum.

Step 5: Scrape content from web using Google search for locatum+preposition+relatum triple

We next generated triples combining each of the 100 locata for a given relatum, each of the prepositions and the relatum itself, surrounded by quotation marks in order to ensure that only the explicit query was returned by Google, rather than other variations (for example, *"National Gallery near Trafalgar Square"*) and used the Python- Beautiful Soup library (Richardson 2007)- to run a query for each triple and scrape the descriptions, including the triple, the URL of the page on which it appeared, and the excerpt from the page that usually appears during the search. By default, Google omits some entries based on similarity, and we adopted this option in Python as well, because we examined this option and most of the returned results were repetitive. During this process, we were asked to enter captcha several times, and used a captcha service to automate this process.

Step 6: Scrape content from web using Google search for locatum+preposition+relatum with wildcard characters

The queries that were run in Step 5 were surrounded by quotation marks to ensure that only relevant searches were returned. However, verbs, adjectives, or other parts of speech are frequently included in location descriptions. For example, *National gallery* is located *next to Trafalgar Square*. We therefore included a version of each triple with wildcard character before preposition "*locatum + * preposition + relatum"* to accommodate this possibility.

Step 7: Manually clean data

We manually reviewed the results returned from the previous steps in order to remove repetitive search results, results with non-spatial use of the selected prepositions (this occurred especially in the descriptions using the wildcard character), and results that refer to a locatum or relatum in another country. Table 4.2 shows the criteria that we followed to exclude the descriptions. Locata are shown with underline, relata are with dashed lines and prepositions are in bold.

After removing the cases mentioned in Table 4.2, we reviewed the descriptions which, while having none of the problems mentioned in this table, were ambiguous, or seemed surprising in some way, mostly because of their topological geospatial prepositions (*in, on,* at) and the high distances between locata and relata. For example, in the following description: "*Istithmar's property portfolio in London, which includes two business parks and West End office buildings the Adelphi and Grand Buildings* ***on*** *Trafalgar Square*", because of the presence of a conjunction (and), the locatum was vague (only the Grand Buildings, or both the Adelphi and the Grand Buildings). However, because the distance between the Adelphi and Trafalgar Square is 422.63m, while the distance between the Grand Buildings and Trafalgar Square is 36.56m, we did not consider the Adelphi as a locatum in this case.

Another ambiguous case was the description "*Great working at the Royal Opera House* ***in*** *Trafalgar Square, providing site safety for the build and derig. We hope everyone had a great time who went!*" We reviewed this case further because of the *in* preposition and the 660m distance between the Royal Opera House and Trafalgar Square. During the data collection, we collected the web page URLs too, and observed on the webpage that there were hashtags such as #trafalgarsquare #royaloperahouse #screening #cinema #outdoorcinema, and discovered that this description refers to a screening *in* Trafalgar Square organised under the auspices of Royal Opera House.

A third group of cases that we manually investigated were descriptions from old newspapers. The locata belong to this group were places for which the locations had changed since the description was published. For instance, the following description is from an 1849 newspaper: "*On Monday a General Assembly of the Academicians was held at the Royal Academy of Arts, **in** Trafalgar Square, when Mr William Dyce and Mr William Calder …* ", but historical documents indicate that the Royal Academy of Arts was relocated in 1868.

Another issue arose in the following 1947 description that describes a vision of an Imperial House that did not exist at the time: "*He wanted to see great Imperial House **in** Trafalgar Square or some where where the British people could have a permanent exhibition of what was going on- ...",* while a new building with that name has been constructed in a different place.

These and other similar descriptions were excluded from our analysis.

Step 8: Count frequency of mentions for each locatum-preposition-relatum combination

After excluding the cases mentioned in Step 7, we counted the frequency of descriptions for each locatum-preposition-relatum combination, adding together those returned through searches with and without the wildcard character.

Step 9: Extract geometries for relata and locata

Because the place names (locata) were unique names in the London area, we were able to search for them through the OpenStreetMap (Nominatim) API and retrieve their geometries. For each locatum and relatum, we retrieved, along with the category, type, centroid coordinates and if a polygon, the coordinates of the boundary.

Table 4.2. Exclusion rules and their examples

Rule for the exclusion	Example (locatum, **preposition,** relatum)	Explanation
Non- spatial preposition	“We're currently thinking perhaps Kings Cross to Charing Cross; a visit to the Benjamin Franklin House then a look **at** Trafalgar Square, walk down the Mall to ...10 answers · Top answer: On the surface looks pretty good, but your first day might be ambitious. The British Museum ...”	**at** is not spatial preposition
Real locatum comes after the specified locatum	“Trafalgar Square, Westminster, WC2N 5DS 3 minutes’ walk from Her Majesty's Theatre. The Fourth Plinth **in** Trafalgar Square remained empty after its ...”	Real locatum is Fourth Plinth and not Her Majesty’s Theatre
Invalid description (Repetitive sentences)	“...Restaurants near National Portrait Gallery · Restaurants **near** Trafalgar Square · Restaurants near ...”	
Locatum and relatum belong to different sentences	“... Trafalgar Voices will be performing in the spectacular Freemasons Hall on 19 December. No concerts **at** St Martin-in-the-Fields this year but #themusicplayson.”	
Real locatum comes before the specified locatum	“The nearest car parks to the Playhouse Theatre are situated **at** Trafalgar Square and Chinatown. These car parks are about 10 to 15 minutes walk away.”	Real locatum is car park not the Playhouse Theatre
Preposition is not the specified preposition (e.g. **in between**)	“Haymarket House is definitely it!! Nestled **in** between Trafalgar Square and Leicester Square you have the delightful little Haymarket Wine House.”	
Complete description doesn't exist	View deals for The Grand **at** Trafalgar Square, including fully refundable rates with free cancellation. Guests praise the locale. The Strand is minutes away. Rating: 8.6/10 · 956 reviews · Price range: from NZ$207	“Her Majesty's Theatre” is missing
Description is in another language but is the exact match with our searched description	“ng mga skyscraper ng Lungsod at ang makitid na kalye ng Soho, matikas na Piccadilly at ang marilag na Tower Bridge, Big Ben **at** Trafalgar Square “	**at** means **and** in Filipino
The locatum name is same as another place/object	“28/10/2010 — Souvenirs of Big Ben are sold in a shop **near** Trafalgar Square on October 28, 2010 in London, England. Get premium, high resolution news ...”	instead of the real Big Ben, this refers to souvenirs of Big Ben
Photo caption, mostly for the case of **in**, meaning standing in the locatum and taking a photo of the relatum.	“King Charles Statue and Big Ben **in** Trafalgar Square at Night with Light Trails in London. London Night View include Big Ben. Lion Statue, seen from Trafalgar ...”	

Step 10: Calculate distances between relata and locata

We then calculated the closest distance between each locatum and relatum using boundary geometries (for polygons). The reason is that, for a large site such as *Hyde Park,* use of the centroid may result in artificially large distances (e.g., for a locatum that is outside, but close to the park boundary). So, if the locatum was inside the relatum, the distance is zero and if it is located outside the relatum the distance was the closest distance between the locatum and relatum boundaries.

Table 4.3 shows the number of locata for each relatum (*Buckingham Palace, Hyde Park* and *Trafalgar Square),* the number of prepositions for which we extracted descriptions for each relatum (out of a total of 9 prepositions) and the total number of descriptions for each site in our final data set, after the above steps had been carried out.

Table 4.3. Description frequencies in the dataset

Site	Descriptions (locata)	Prepositions	Total number of descriptions
Buckingham Palace	78	9	1970
Hyde Park	75	8	1523
Trafalgar Square	76	8	1746

The description extracted from Google, came from different sources. About 30% came from image sharing websites such as *Flickr*[44], *Pinterest*[45] and *Shutterstock*[46], 20% from news websites such as *BBC* and *Telegraph,* 15% came from knowledge databases such

[44] https://www.flickr.com/

[45] https://www.pinterest.com/

[46] https://www.shutterstock.com/

as *Wikipedia*[47] *and Google books*[48]*,* 10% from public social media pages such as *Instagram*[49] and *Facebook*[50]*,* 5% from historical websites such as *British-history*[51] and *historical England*[52] and 20% from other sources.

In section 4.4, we analysed the data using visualisations such as cumulative graphs and scatter plots to investigate the acceptance thresholds for prepositions and study the impact of context on the usage of geospatial prepositions.

4.4 Results

We provide some information about the three relata used for the experiments, their features, and the kinds of locata that were used with them in Table 4.4. Then in Sections 4.4.1 and 4.4.2, we analyse our data in order to address each of the research questions.

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

The term *distance* here refers to the closest distance between the boundary of the relatum and locatum. So, if the locatum is outside the relatum, the distance would be the shortest distance between their boundaries. However, if they overlap or the locatum is inside the relatum, the distance has been set to zero. Figure 4.4 shows the cumulative frequency graphs (Jelinek 1962) for each preposition for each of the three sites.

Table 4.4. Relatum features

[47] https://www.wikipedia.org/

[48] https://books.google.com/

[49] https://www.instagram.com/

[50] https://www.facebook.com/

[51] https://www.british-history.ac.uk/

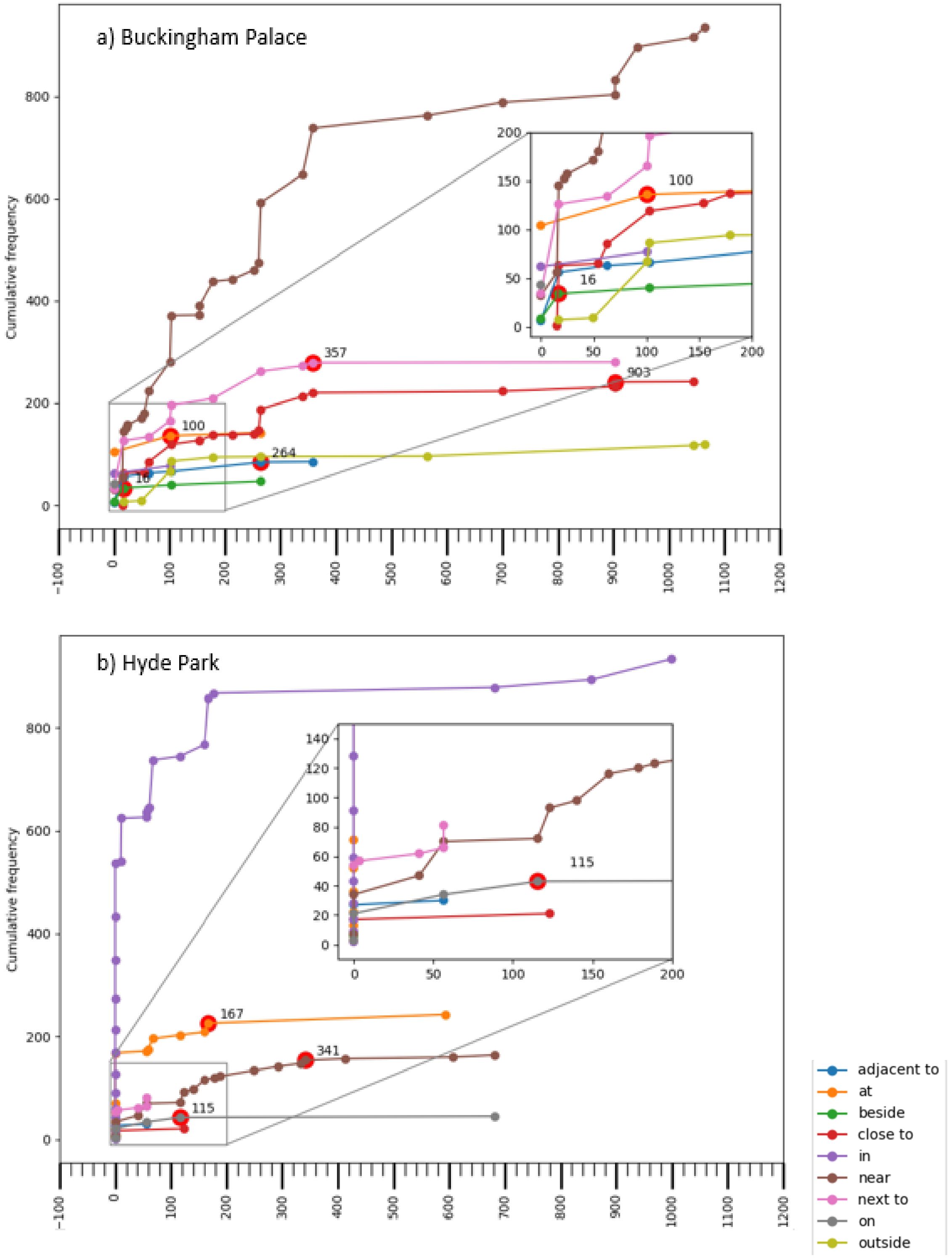
[52] https://historicengland.org.uk/

Site	Features
Buckingham Palace	*Buckingham Palace* is a palace in the City of Westminster. Most of the locata associated with the palace are used with proximity/adjacency prepositions such as *next to, near, close to* and *adjacent to.* This is the only site that the preposition *outside* is used for. In some descriptions, only the palace building is considered but, in some others, the whole grounds and palace are considered together. The most popular locata around Buckingham Palace are *Green Park, Royal Mews,* and *Wellington Barracks* (a military barracks).
Hyde Park	*Hyde Park* is a park located in central London. There are many small locata inside the park, including memorials, statues, fountain, gallery, and swimming club, and there are also some outside it that are mostly referred to with the geospatial preposition *near.*
Trafalgar Square	*Trafalgar Square* is a square in Westminster City in London. It is one of the most popular places in London, due to its location close to many tourist sites, including *the National Gallery, St Martin-in-the-Fields,* and *the River Thames.* Most of the extracted descriptions that used *Trafalgar Square* as a relatum, describe locata close to it, but there are also several statues and memorials that are in the Square itself.

The points on the graphs represent locata and are positioned on the x axis using the distance between the boundary of the relatum and locatum. The vertical axis shows cumulative frequencies, being the total of all mentions of a given locatum plus all locata at closer distances. A slope indicates an increase in frequency (the frequency of mentions in descriptions extracted from Google), while the closer the line is to horizontal, the fewer mentions for the locatum at the right end of the line (although note that the line is never horizontal, as this would only occur if there were no new mentions for the locatum at the rightmost end of the line, but we only add points if there is at least one mention). For example, the line between 176.71 and 682.1 for the *in* preposition for Hyde Park is close to horizontal, as only 11 new mentions were added for the locatum at 682.1 metres. We use cumulative frequency graphs because they provide a clearer picture of the behaviour of each preposition in comparison to raw frequency graphs in which individual locata can obscure the visualization. It is important to note that the point at which each curve flattens is the point at which there are very few new mentions, so we consider this to be the *acceptance threshold* for each preposition. We define the acceptance threshold quantitatively by starting from the last point on each preposition line (from right to left) and moving to the left until the slope of the line connecting two point exceeds 5° (relative to the horizontal) – we chose 5° because we tested other numbers and 5 seemed to be more accurate visually. We define the point to the right of the first line that is less than 5 degrees as the acceptance threshold. If no line for that preposition has a slope of <5°, we consider that our data has not yet reached the point at which the line flattens, and we do not have sufficient data to identify the acceptance threshold. If no line has a slope of >5°, then we calculate the average slope across all lines for the preposition and define the threshold as the point for which the slope between it and the next point (from the right) is greater than that average. Figure 4.5 shows the flowchart for this algorithm.

On the graphs in Figure 4.4 acceptance thresholds are marked with large red dots, and this figure indicates the acceptance thresholds for each preposition for each relatum (Figure 4.4 (a-c)). In addition to the graphs for each relatum Figure 4.4 (a-c), we present a cumulative frequency graph Figure 4.4(d), in which, we aggregate the data for all the relata, and adjust frequency values to account for varying site popularity.

Some relata are more popular than others (i.e., attract more mentions in social media). So, if more popular sites are not adjusted for popularity, their values will have a disproportionately large influence on the shape of the graph that combines the results of all relata, because they would have more mentions for all prepositions. We therefore scale down these more popular sites so that all relata have an equal total, adjusted number of mentions, this being the same as the total minimum total across all relata (in this case it is for Hyde Park, which has 1523 mentions). The reason for doing this rather than normalising (adjusting to values between 0 and 1) is that it still gives some indication of the scale of mentions (Equation 4.1).


a) Buckingham Palace
Cumulative frequency
0
200
400
600
800
−100
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
357
903
100
264
16
200
150
100
50
0
0
50
100
150
200
100
16
b) Hyde Park
Cumulative frequency
0
200
400
600
800
−100
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
167
341
115
140
120
100
80
60
40
20
0
0
50
100
150
200
115
adjacent to
at
beside
close to
in
near
next to
on
outside

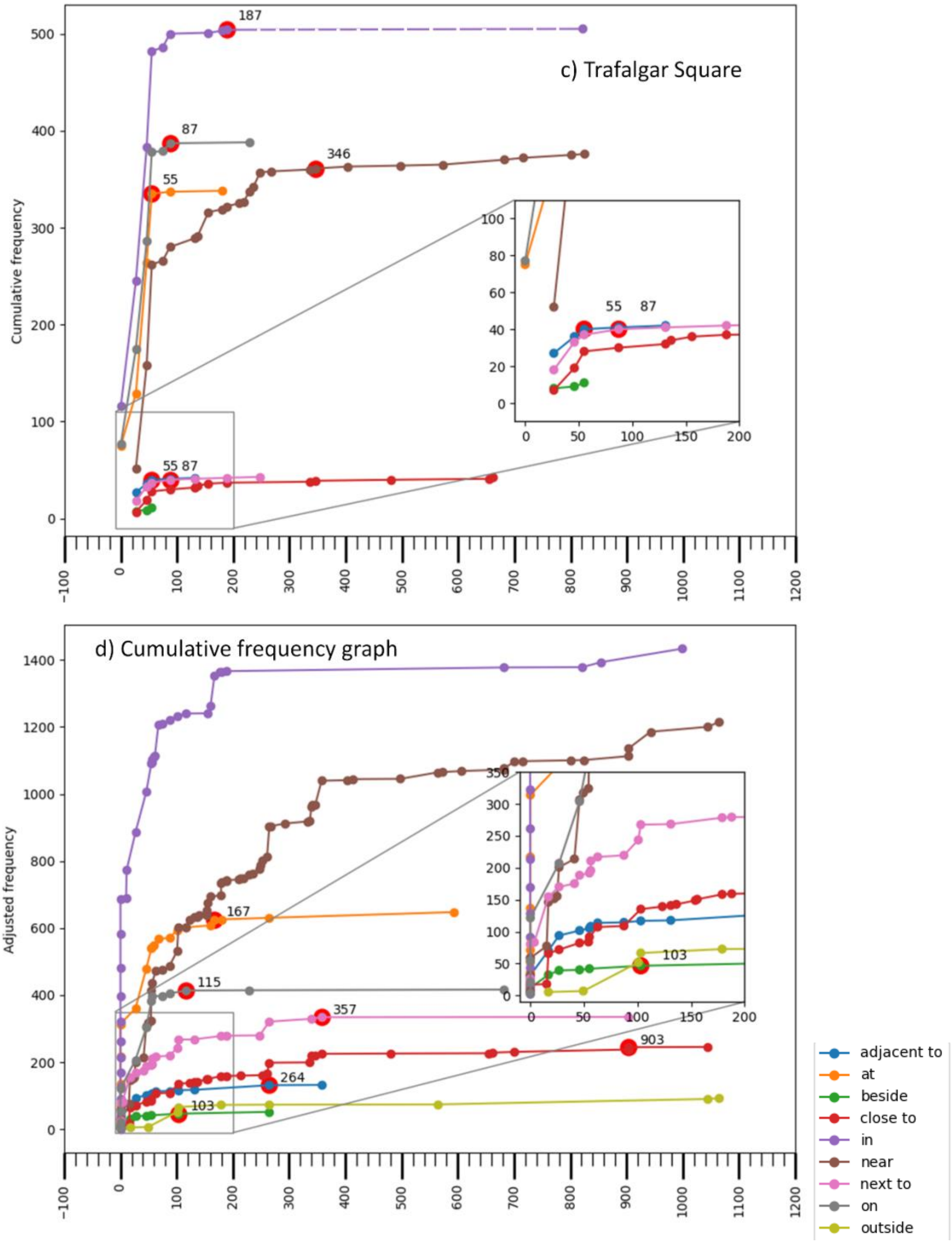


Figure 4.4. Cumulative frequency graphs of all prepositions for (a) Buckingham Palace, (b) Hyde Park and (c) Trafalgar Square(d) cumulative frequency graph with adjusted frequency

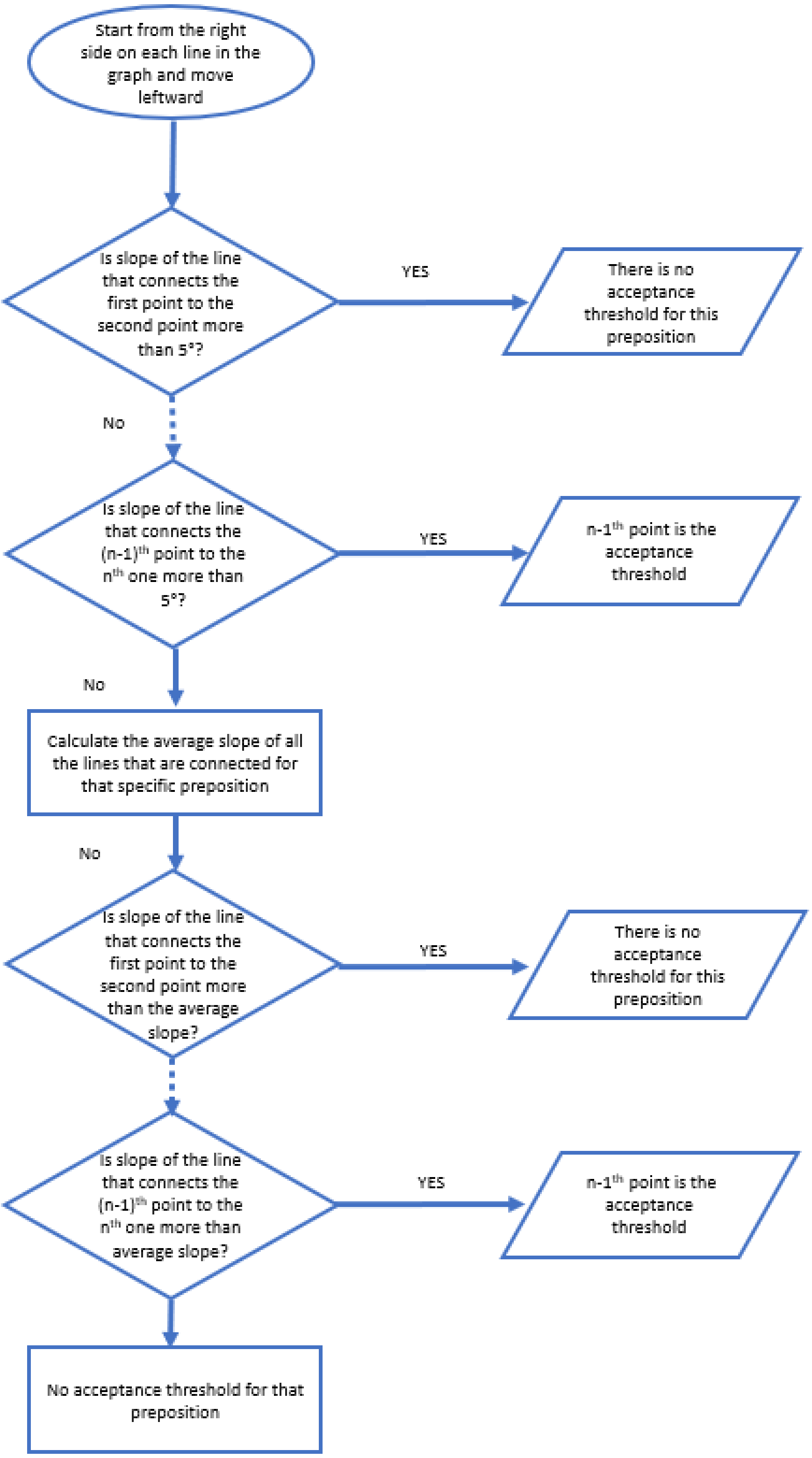

Start from the right side on each line in the graph and move leftward
Is slope of the line that connects the first point to the second point more than 5°?
YES
There is no acceptance threshold for this preposition
No
Is slope of the line that connects the (n-1)th point to the nth one more than 5°?
YES
n-1th point is the acceptance threshold
No
Calculate the average slope of all the lines that are connected for that specific preposition
No
Is slope of the line that connects the first point to the second point more than the average slope?
YES
There is no acceptance threshold for this preposition
Is slope of the line that connects the (n-1)th point to the nth one more than average slope?
YES
n-1th point is the acceptance threshold
No acceptance threshold for that preposition

Figure 4.5. Acceptance thresholds algorithm to identify acceptance threshold (red dots on Figure 4.4 a-d)

$$adj\ freq\ (r_i, l_i) = \frac{freq(r_i, l_i)}{\sum_{i=1}^{n} freq(r_i, l_{i..n})} * \min\ (\sum_{i=1}^{n} freq(r_{i..n}, l_{i..n})) \qquad \text{Equation 4.1}$$

$freq(r_i, l_i)$ indicates the frequency of a given preposition for the locatum$_i$ and relatum$_i$, $\sum_{i=1}^{n} freq(r_i, l_{i..n})$ indicates the sum of the frequencies of the prepositions across all the locata for the relatum$_i$ and $\min\ (\sum_{i=1}^{n} freq(r_{i..n}, l_{i..n}))$ indicates the sum of the frequencies of the prepositions across all the locata for the relatum that has the minimum sum of the preposition frequencies (in this case *Hyde Park).* The total number of mentions for Trafalgar Square and Buckingham Palace were 1746 and 1970 respectively, so the frequencies for these two relata were adjusted down accordingly.

We identify the acceptance thresholds for each relatum in Table 4.5, as well as the mean acceptance threshold (across whichever sites sufficient data was available to calculate a threshold), as well as the acceptance threshold for the aggregated data. This data is visualised in Figure 4.6.

Table 4.5. Acceptance thresholds of each preposition for each site, mean and aggregate

Site/ Preposition	Buckingham Palace threshold	Hyde Park threshold	Trafalgar Square threshold	Mean threshold	Aggregate threshold
adjacent to	265		55	160	264
at	101	167	55	108	167
beside	17			17	103
close to	903			903	903
in			188	188	
near		341	347	344	
next to	358		87	223	357
on		116	87	101	115

The data shows some clear patterns across the prepositions. The *near* preposition has one of the highest acceptance thresholds (distance at which the preposition is no longer acceptable), with a distance of 341m for Hyde Park and 347m for Trafalgar Square. Furthermore, we see *near* being used infrequently for much greater distances than the threshold (700-800m for Hyde Park and Trafalgar Square).

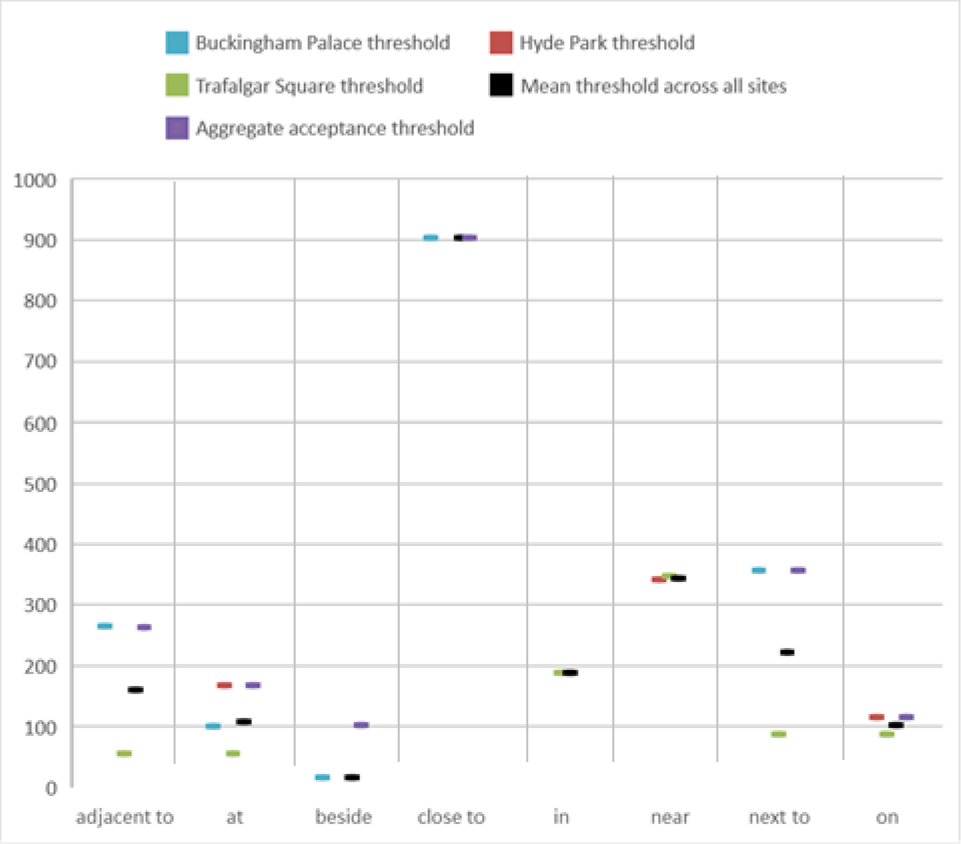


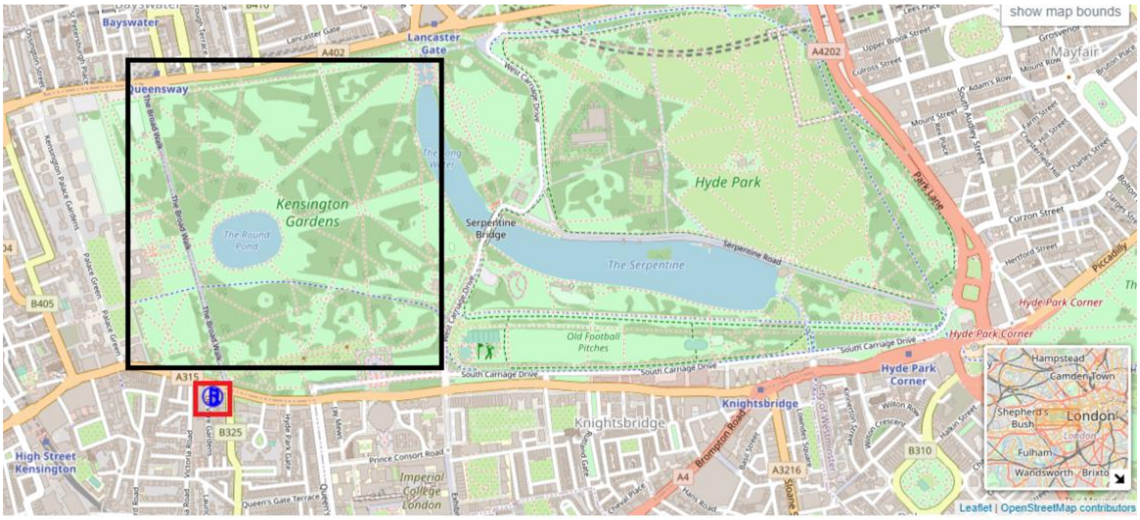
Figure 4.6. Acceptance thresholds of each preposition for each site, mean and aggregate

The acceptance threshold for Buckingham Palace (and for the aggregated data) is much higher (>1100m), as the graph does not level off for this site, indicating that the acceptance threshold is beyond our last data point. Like *near*, the *close to* preposition has among the highest acceptance thresholds, with (903m) for Buckingham Palace and for the aggregated data, and insufficient data to establish a threshold for the other two relata (*close to* is a relatively infrequently used preposition, so we have few mentions in the data).

Next to has the highest mean acceptance threshold (223m) and threshold for the aggregate data (357m) of the adjacency prepositions. However, the range in thresholds between 358m for Buckingham Palace and 87m for Trafalgar Square has some substantial overlap with the range for *adjacent to* (265m to 55m, with mean 160m and threshold for aggregate data of 264m). *Beside* has a much smaller threshold, being 103m for the aggregate data, and 17m for Buckingham Palace (*beside* did not appear in our data for Hyde Park, and only infrequently (with distances of 50m or less) for Trafalgar Square. Based on this limited evidence, we postulate that *beside* is typically limited to much closer locations than *next to* and *adjacent to*, both of which are used for locations within a closer range than the proximity prepositions *near* and *close to*. More data is needed to confirm this. *Outside* has been only used for the Buckingham Palace site, but we couldn't identify any acceptance threshold for it because the slope between its rightmost two points exceeds 5° and thus, we consider that we have not collected sufficient data to determine the acceptance threshold.

Moving to the containment and collocation prepositions *in, at* and *on*; surprisingly, the acceptance thresholds for *in* appear to be large, being beyond our last data point for the aggregate data and for Hyde Park, and to a lesser extent, for Trafalgar Square (187m), even though given that our distances are measured boundary to boundary, we might expect distances of zero (the locatum inside the relatum). For Buckingham Palace, only two locata were used with the *in* preposition. The first one is within its boundary (*The Royal Mews*) and the second is 100m away (*Victoria Memorial*). While the latter is located on the site of Buckingham Palace, it was not within the boundary we extracted from OSM. The Hyde Park data is affected by the location of another garden called Kensington Gardens (see Figure 4.7). People sometimes refer to this garden as Hyde Park. For example, the description *Princess Diana Memorial Playground in Hyde Park* appears, but the playground is on the east side of Kensington Gardens, 900m from the closest boundary of Hyde Park. Kensington Garden was part of Hyde Park until 1728, but given the length of time since their separation and the fact that our sources are more recent than that, we consider it unlikely that this history has influenced of the usage of these two place names.

In the case of Trafalgar Square, the most distant locata within the acceptance threshold are *Her Majesty's Theatre* (187m) and the *Nigerian High Commission* (180m). The reason of using *in* and Trafalgar Square for these two locations is that Trafalgar Square is a well-known landmark in the area. We postulate that in natural language location descriptions, the geographic boundaries of well-known landmarks may be ‘stretched’, but more work is needed to validate this.



© OpenStreetMap contributors

Figure 4.7. Expression Baglioni hotel (red) on Hyde Park on OSM map

We also see an unusual outlier for Trafalgar Square: The Methodist Central Hall, which is marked with the dashed line on the graph because we suspect that it is an error. The source description came from *The Westminster Reporter (The Westminster City Council Magazine)* and reads “*16/09/2015 — The largest air raid shelter in England was at the Methodist Central Hall* ***in*** *Trafalgar Square which could hold 2,000 people each night*.” (September 2015, Issue 120, page 21). However, Figure 4.8 shows that Methodist Central Hall is a substantial distance from Trafalgar Square (821m). There were other cases in our data that showed similarly questionable descriptions, but that could be explained by changes in location of the locatum over time (particularly where the source was a historical document- see Section 3 step 7), but no such evidence could be found for this description. It is clear from other documentation[53] that there was an air raid shelter of this size at the Methodist Central Hall, also referred to as the largest in

[53] https://dmbi.online/index.php?do=app.entry&id=2968

England, but no evidence that the Hall, or the air raid shelter was in Trafalgar Square, other than this single description.

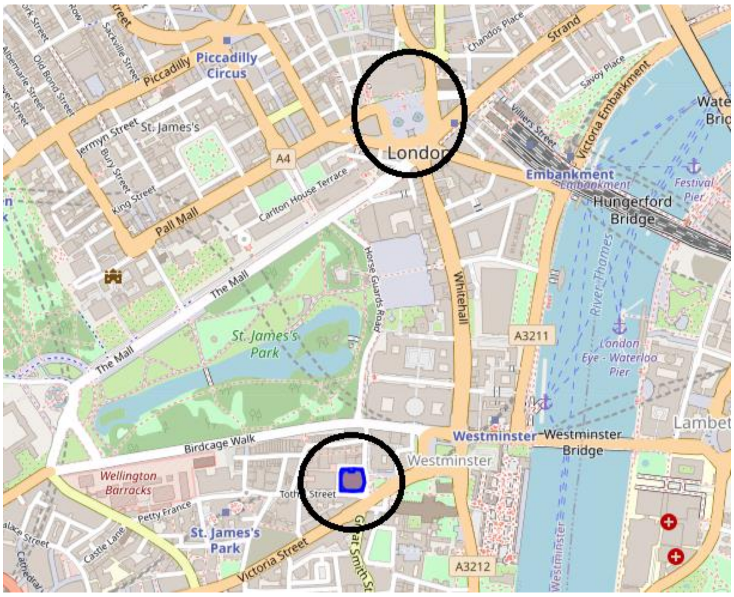
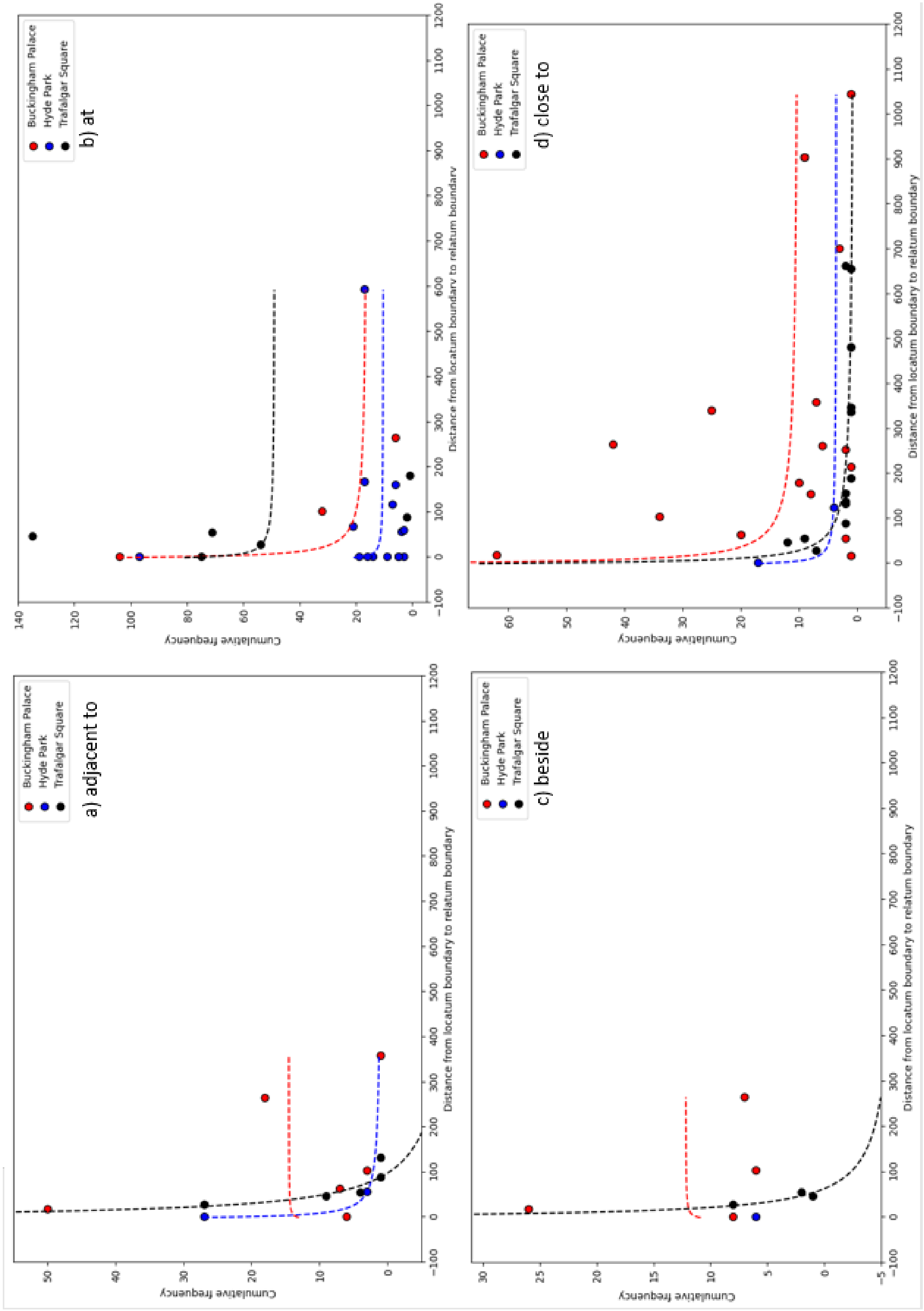


Figure 4.8. The distance between Methodist Central Hall as the locatum (bottom) and Trafalgar Square (up) as a relatum that used preposition 'in'

The *at* preposition has acceptance thresholds of 167m, 100m and 55m for Hyde Park, Buckingham Palace and Trafalgar Square respectively, with a mean of 108m and 167m for the aggregated data, putting it within a similar range to the *in* preposition, but with the lowest acceptance threshold of the three topological prepositions, being for Trafalgar Square (but note that the lowest acceptance threshold across the data is for *beside*, for Buckingham Palace). The *on* preposition has thresholds in a similar range (87m-116m) with a threshold for the aggregate of 115m and a mean of 101m, both of these latter figures being the lowest for the three containment and collocation prepositions.

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

To answer the second research question, we present Figure 4.9 (a-h), which compares the three relata for each of the prepositions using scatter plots, and regression lines for each relatum. We use a reciprocal, linear regression equation to plot the regression line


a) adjacent to
b) at
c) beside
d) close to
● Buckingham Palace
● Hyde Park
● Trafalgar Square
Cumulative frequency
Distance from locatum boundary to relatum boundary
−100
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
0
10
20
30
40
50
60
80
100
120
140
−5
5
15
25

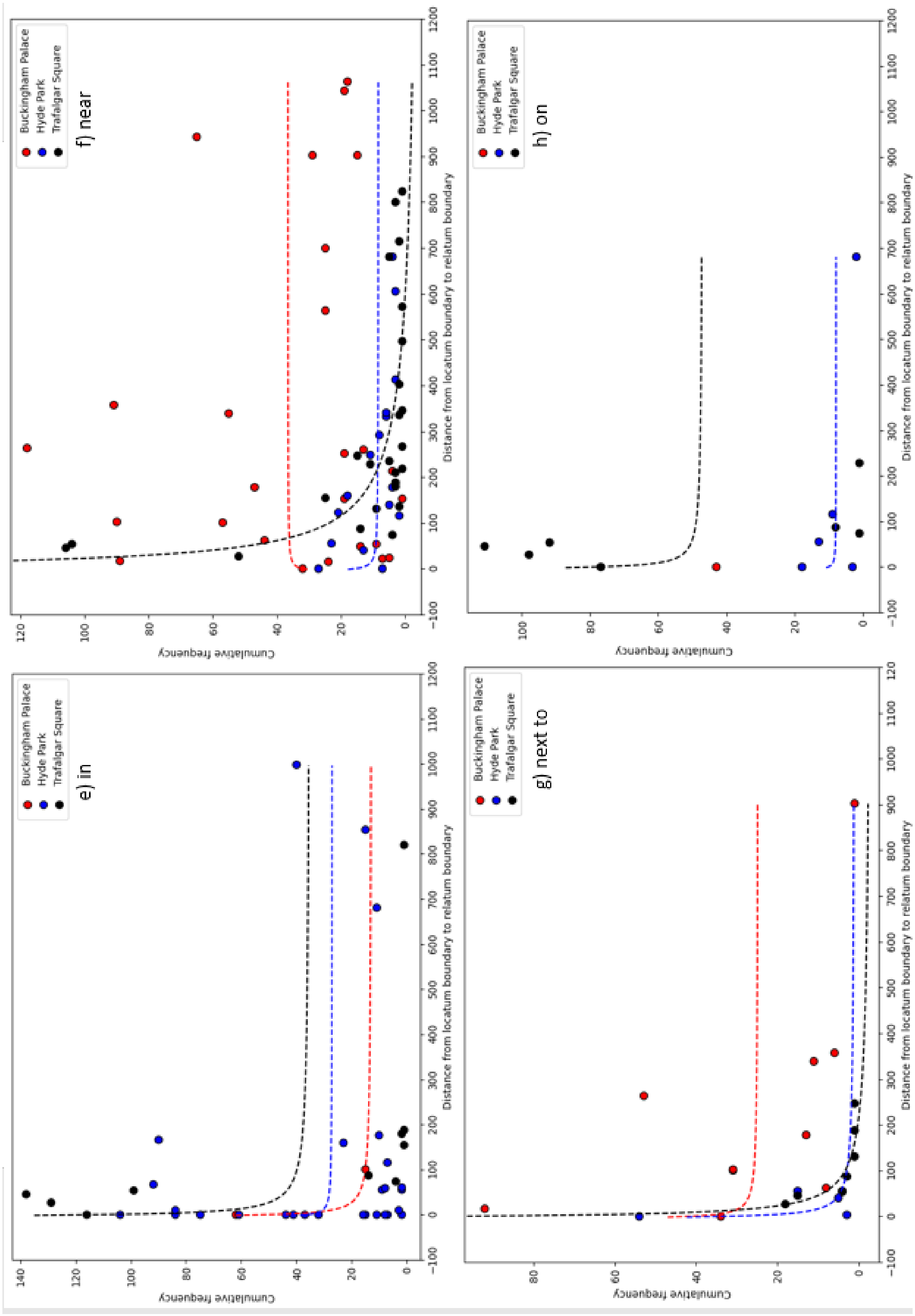


Figure 4.9. Scatter plot and acceptance profiles for the frequency of each preposition across all three sites (a-d)

($y = \frac{1}{x}$), and we refer to these regression lines as acceptance profiles, as they show the profile of the distances at which the preposition is used, including the distances at which it is highly acceptable, as well as those at which is becomes less so. This provides us with more information than the acceptance threshold. Note that we do not show a plot of the *outside* preposition, as our data extraction only identified descriptions that used *outside* with the Buckingham Palace relatum, and in RQ2, our focus is on comparison of the context.

Several of the prepositions show clear similarity across all three sites, including *next to* and *close to*. The curves across the three relata for *next to* are very similar, the main difference being in frequency of mentions, an issue that is discussed further below. The ranges of the data points vary for the three sites, with Buckingham Palace having low frequency mentions for more distant locata, while in contrast, Hyde Park uses *next to* with locata that are relatively close (up to 56m). Although the mentions do not extend as far, for close distances, the *adjacent to* and *beside* preposition graphs are similar to those of *next to*. Both are used with Buckingham Palace for distance up to 250m, but the most frequent uses across the other two sites are less than 100m.

We see a very similar pattern for *close to*, with Buckingham Palace attracting mentions out to approximately 1km, Trafalgar Square to 660m (albeit very low frequency) and Hyde Park to 123m. We consider it likely that the larger distances associated with Buckingham Palace are influenced by the ambiguity in the specific size/area of the relata: the entire grounds, or only the palace building itself. In this analysis, we used the entire grounds, since access to the grounds and palace is limited, so it is less likely that mentions would refer only to the palace (and indeed our comparative analysis confirmed this, with calculations that used the palace geometry only as the relatum being distorted). Despite the differences among the three sites discussed above, the highest frequency uses of *close to* and *next to* (point at which the steeper sections of the graph level out, being less than 100m in both cases) are consistent across all three sites, suggesting limited impact on context among the most common uses of these prepositions. While the *near* graph is similar to that of *close to*, it does show some variation between the three positions: namely the absence of high frequency use at very low distances for Buckingham Palace and Hyde Park. It appears that while *near* may be used for Trafalgar Square for locations very close to (or at) the Square, this does not apply to the other two relata, in contrast to *close to*, which is applied at very small distances for all three relata. Trafalgar Square is a much more open location, which vaguer boundaries than the other two relata, which may explain the more liberal use of *near* in that case.

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

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

Use of the *on* preposition is much more frequent for Trafalgar Square than for Hyde Park and is only used once for Buckingham Palace in our data set, with distance zero. This is likely the result of image schema, with squares and plazas being more frequently associated with a platform schema than parks or palaces (Mark 1989; Mark and Frank 1996). However, we do not see a similar pattern for the *in* and *at* prepositions, which are commonly used for parks and similar types of objects. Trafalgar Square is frequently used with *in, at* and *on*, suggesting that a range of different image schemata are suitable for this feature type, while parks and palaces are more limited. Across all of the prepositions, we see much greater variation between relata of the outer extremes of acceptability of the prepositions. That is, many of the prepositions studied are used less frequently for quite large distances for some sites more than others; while the most frequent uses are much more uniform across the sites, despite the differences in size, feature type and level of urban construction among the relata. Generally, the proximity (*near, close to*) and adjacency (*next to, adjacent to, beside*) prepositions are more frequently used for greater distances for Buckingham Palace than the other two relata, and for smaller distances particularly Trafalgar Square; while *in, on* and *at* are most frequently used for Trafalgar Square across all distances.

In Section 4.5, we discuss the results found in this section with some comparisons with the previous literature.

4.5 Discussion

4.5.1 Acceptance thresholds (RQ1)

Section 4.4.1 provided an analysis of the acceptance thresholds highlighted for our data, as well as the extremes of range within which each preposition is used. Our findings show that among the proximity/adjacency prepositions, *near* and *close to* have the highest acceptance thresholds (are used for largest distances) *near* > 1100m and *close to* = 903m. *Next to* and *adjacent to* are used for distances between 55-358m for Buckingham Palace and Trafalgar Square and no acceptance threshold has been identified for these two prepositions in the case of Hyde Park. The smallest threshold for proximity/adjacency prepositions is for *beside* (17m for the Buckingham Palace) but there is not enough data on the other two sites to confirm this finding. Carlson and Covey (2005) ran a human subjects experiment to estimate the distance associated with some spatial prepositions such as *next to, beside* and *near* and some projective ones.

Similar to our findings, their research showed that *beside* and *next to* are not associated with relatum size and these prepositions specify smaller distances than *near.* Fisher and

Orf (1991) also reviewed the interpretation of *near* and *close* in a university campus area and found that these two terms are used for objects that are proximal, but not if their distance is very close to the locatum. Through their experiment, they found that people did not chose buildings that were very close to the student centre (relatum), but instead those that were further away as *close/near.*

In addition, as mentioned in Section 4.2, some researchers have used methods to identify density fields which provide a probabilistic indication of the areas within which use of a preposition is acceptable. Due to the vagueness of spatial prepositions, most of these works identified the area that prepositions are accepted. As an example, Hall et al. (2015), used Kriging interpolation to model spatial prepositions such as *near, between, at the corner, at* and *next to.* Similar to our work, their findings show a larger distance for *near* in any direction around the relatum and smaller distances for *at* and *next to.* Also, Skoumas et al. (2016) visualised *near*, *in* and *on* and some cardinal relations such as north and south using heatmaps with a grid-based approach. Their visualisation shows that *near* is not restricted to only locations close to the grid centre (relata) but extended further and in all directions around the relata. However, *at* was limited to a small part of the grid in the centre of it which shows more limited usage of *at* to those areas close to the distance of the relata same as our findings. Also, their visualisation of *next to* confirms our findings and shows that *next to* is more limited to the close vicinity of the relatum without any orientational variations.

Our findings show that among all proximity/adjacency prepositions, *beside* has the shortest aggregated threshold and it has been used for the shorter distances (with aggregated threshold 103m which is across all relata). In addition, *next to* has larger aggregated threshold than *adjacent to,* that suggests it might be used in larger distances that *adjacent to.* we also found that *close to* has the largest aggregated threshold among adjacency/proximity geospatial prepositions. However, more data is needed to confirm this as it has been only identified for one site (Buckingham Palace). Among container and collocation prepositions, *at* and *on* have the shortest aggregated thresholds. We could not find an aggregated threshold for *in* and *near.* This is because for *near* the cumulative frequency for Buckingham Palace kept rising, and for *in,* there is not sufficient data for Buckingham Palace and for Hyde Park, the cumulative frequency has never been flattened. This might be because of the size of Hyde Park and the usage of *in* for further area which is not Hyde Park (Kensington Garden discussed in Section 4.5.2).

4.5.2 Contextual factors (RQ2)

The analysis in Section 4.4.2 highlights a number of observations regarding contextual variations in the use of spatial prepositions. Firstly, we note that *near* is used less frequently for locata very close to the relatum for two of the three relata, the exception being Trafalgar Square. This 'doughnut effect' was not evident in the analysis of Hall et al. (2011). For example, their analysis of the expression *Pond near High Boston* (Hall et al. 2011, pp.17) showed that all of the area in the close vicinity of High Boston (a hamlet) was considered *near.* However, Fisher and Orf (1991) identified the doughnut effect and claimed that this might be due to the similarity in place names or functions. For example, a building known as the student service centre is typically considered to be closer to another building known as the student centre, than other buildings with different functions, even though those other buildings are in fact closer. In our work, we consider this effect is likely to be related to the nature of the relatum feature type, and possibly related to vagueness of boundaries and/or openness of the environment. More research is needed to confirm this.

We also noted in Section 4.4.2 the likely impact of image schema on the use of the containment and collocation prepositions. This has been identified by other researchers, who identified the use of the *on* preposition when a platform schema is used, or *in* for a container schema (Mark 1989; Mark and Frank 1996). However, our results identify a variation in image schemata applied to different feature types. Some of our relata were strictly subject to a single image schema (e.g., Hyde Park with the container schema, indicated by the use of the *in* preposition in preference to *at* or *on*), while others were more promiscuous (notably Trafalgar Square, which uses all three of these prepositions liberally, suggesting platform, container and possibly other image schemata such as link are appropriate). Stock and Yousaf (2018) assigned some geolinguistic factors (image-schema, axial structure, solidity, geometry type and scale) to feature types such as highway or river. For example, they claimed that if two relata have similar geolinguistic factors such as image-schema or axial structure, they might be more similar than the ones that have different geolinguistic factors. However, more research is needed to identify the use of different image schema types by feature type.

We also note that the *outside* preposition is only used with Buckingham Palace in our data (see Figure 4.4(a)). While the *outside* preposition would normally be associated with the container image schema, we see low frequency use of *in* (also associated with the container image schema) with Buckingham Palace compared to the other two relata. However, this may be due to the access limitations previously mentioned reducing the frequency of mentions for Buckingham Palace. In contrast, *in* is the most frequently used containment or collocation preposition used for Hyde Park, but *outside* is not used for this relatum in our data. This suggests that *outside* requires a stronger form of containment than *in*, with Buckingham Palace being a stronger container than Hyde Park. It is unclear to what degree the containment that results in the use of *outside* with Buckingham Palace is due to restricted access, rather than feature type or other aspects of the nature of the relatum, but certainly the grounds of Buckingham Palace are similar in nature to Hyde Park, with the exception of accessibility. Clearly the Palace itself is quite different in nature, but neither palace nor grounds can be accessed by the public without payment and in a controlled manner (e.g., organised tour).

In addition to the influence of generalisable characteristics of different relata on the acceptance profiles and thresholds of prepositions, our data shows that individual contexts can influence the use of prepositions, as in the case of Hyde Park and the likely influence of the neighbouring Kensington Gardens. This suggests that general models of preposition applicability, even if they are able to incorporate a rich range of contextual factors such as feature type, accessibility or image schema, are likely to still be limited in accuracy, as they are unable to capture these individual nuances.

One limitation of the current work is that our calculations are based on current OSM data. However, during the data collection from the Google search engine, we extracted several historical texts. For example, in the description *“On Monday a General Assembly of the Academicians was held at the Royal Academy of Arts, in Trafalgar Square, when Mr William Dyce and Mr William Calder … “* from 1844*, The Royal Academy of Arts* is not “in” *Trafalgar Square* now*,* and was relocated in 1868 to a location about 900m away from Trafalgar Square. So, these expressions are not valid anymore. So, this limitation can be seen in both descriptions and the map. We have tried to work with the most up to date data but there is always some data which is not compatible with the current information and in this case, we have not considered the expressions concerned.

Also illustrated by the Hyde Park/Kensington Garden example, the influence of familiarity on the use of prepositions and the associated selection of locata to describe a location is confirmed by our research (Yao and Thill 2005). People use the reference objects whose names they are more familiar with, and this may influence the acceptance profiles and thresholds of prepositions in specific contexts. Also, there are some places for which location descriptions rely on generic relata such as shops, buildings, or other prominent landmarks, in part due to the absence of known named places. This adds to the complexity of spatial descriptions because most of the time these places are not unique, and analysis of the kind described in this chapter is difficult.

Figure 4.10 shows Hyde Park, with locations of locata that used *in, on,* and *at* for Hyde Park descriptions. The symbol indicates the preposition, and the size of the symbol indicates mention frequency. As can be seen, some of the locata are outside the boundary of Hyde Park close to the boundary of Kensington Gardens (e.g., at the bottom left, *Cheval Thorney Court*, *Baglioni Hotel* and *Royal Albert Hall* numbered as 1,2 and 3 respectively*)* and some are inside Kensington Gardens (e.g., on the left side *Princess Diana Memorial Playground* and *Princess Diana Memorial Garden* numbered 4 and 5 respectively*)*. However, as is described above, Hyde Park is still used as their relatum instead of Kensington Gardens, possibly due to the familiarity or prominence of Hyde Park (Epstein et al. 2007).

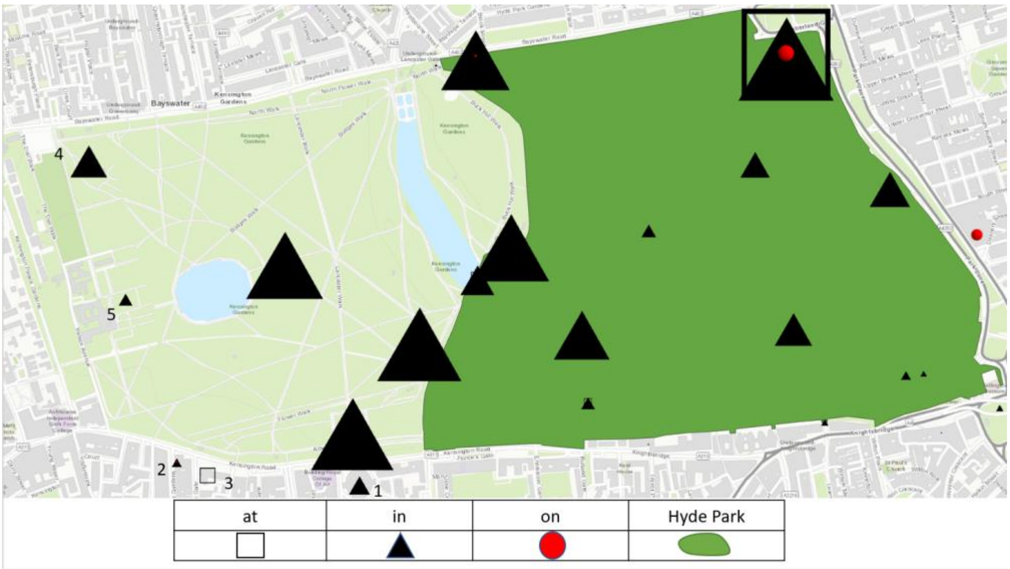


Figure 4.10. Hyde Park and the in, on and at prepositions used with it weighted by their frequencies

We also reviewed the data to find the impact of relatum size on the usage of geospatial prepositions. The only impact of size we found in this data is the frequency of *in, on* and at prepositions in the case of Hyde Park. Due to the size of this relatum, more locata has been found to be located within the boundary or in the close vicinity of this site (Figure 4.10). Thus, there are more descriptions that used these three prepositions than the other two relata (1221 descriptions out of 1523 with three prepositions *in* [943], *on* [45] and *at* [233]).

These results also show that while there are distances that are acceptable for a given preposition across all three of our relata, which we might refer to as 'ranges of agreement', there are also outer extremes of those ranges that are only used in certain circumstances, for only one or two relata, and depending on context. We see some examples of this in Figure 4.10 and Figure 4.11 illustrate the usage of some of the adjacency prepositions with Buckingham Palace as the relatum. Both Hyde Park and Buckingham Palace have data points at much greater distances than Trafalgar Square, usually with low frequency (for example, the *next to* preposition is used at around 900m for Buckingham Palace), but with greater frequency at much lower distances for all three relata. We thus consider it likely that the acceptance areas for geospatial prepositions follow prototype theory (Rosch 1975) in having a range of exemplars, and outliers, which are at least partly determined by context.

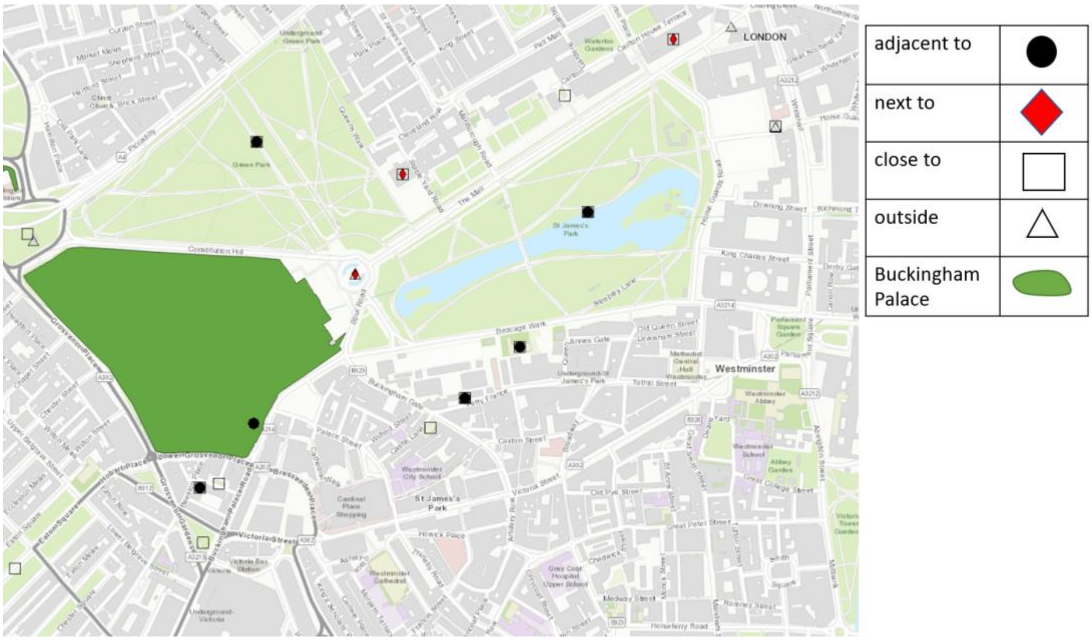


Figure 4.11. Buckingham palace and the adjacent to, next to and close to and outside prepositions used with it (not weighted)

4.5.3 Locata popularity

In addition to the impact of relata on the selection of appropriate geospatial prepositions, we acknowledge the impact of locatum popularity on our dataset. It is likely that people chose the most popular or salient locata when describing a scene, in favour of less noticeable objects. While this may result in bias, since we are looking at the same three relata for all prepositions, the comparison between the relata is still valid, and we see that for a given locatum, a mixture of different prepositions is selected, rather than a single preposition. This means that while our results may be affected by the absence of popular locata at a certain distance from the relatum, and instead a locatum that is further away may be selected, the selection of the preposition will still take distance into account, and still enable valid comparison of preposition selection.

4.6 Conclusion

In this chapter, we investigated the acceptable distances (referred to acceptance profiles and acceptance thresholds in this chapter) for a set of prepositions used in geospatial situations, and the impact of context on those distances. We used a web scraping technique to extract descriptions with three main elements (locatum, preposition and relatum) from Google for three sites in London (*Trafalgar Square, Buckingham Palace,* and *Hyde Park)* and analysed the frequency of mentions of specific locata with each preposition and relatum.

Our experiment led to a number of findings. Firstly, proximity/adjacency geospatial prepositions such as *near* and *close to* are used for larger distances than *adjacent to, next to* and *beside.* Also, *in, on* and *at* have not only been used for locata that are inside the relata, but sometimes these prepositions are used when there is a short distance between locata or relata (depending on the specific context). Another finding is that, based on the visualisations (Figure 4.4 (a-d)) in most cases when the distance increases, the number of mentions decrease. This means that, when we go far away from the relatum, it is more likely that the descriptions that refer to the relata are less frequent. However, this is not always the case. In some cases, frequency of mentions is based on the importance of the locata. As an example, *Boy and Dolphin Fountain in Hyde Park,* has a frequency of 7, even though the distance between its locatum and relatum is zero. In contrast, the description *Serpentine Gallery in Hyde Park,* has a frequency of 92, with a distance between locatum and relatum of 93m (it is located inside Kensington Garden and not Hyde Park). So, in addition to the characteristics of the relatum (as it mentioned in Section 4.5.2), the popularity, familiarity, or size of the locatum has an impact on the mentions and it is more likely that in search engines like Google, we read and find more descriptions about the places that are well-known, and some events might happen around/inside them.

This exploratory research identifies a number of areas for future research. More data collection is needed to identify the acceptance thresholds for the prepositions for which our data did not extend far enough. As an example, the acceptance thresholds we reported in this chapter for *close to* and *in* are only based on one of the sites (Buckingham Palace and Trafalgar Square respectively), as we did not have locata far enough away to identify the acceptance threshold for Buckingham Palace. For *outside* we also have not identified any acceptance threshold as the graph has never reached the point of flattening. We also discussed the doughnut effect for the *near* preposition, and more investigation is required to study this phenomenon in more detail. There is also a need for more future work on the relationship between image-schema and feature types, the flexibility in use of image-schema by a given feature type, and the conditions attached to each use. In addition, it is important to study the impact of geolinguistic factors mentioned in Section 4.5.2 (Stock and Yousaf 2018), on the usage of different geospatial prepositions. Most of the previous works have reviewed the impact of feature types (Manguinhas et al. 2008; Tobin et al. 2010; Alex et al. 2015; Acheson et al. 2017) on the georeferencing task, but there is not much work on the effect of contextual factors on preposition usage and spatial context interpretation.

In the next chapter, we will use the preposition semantic similarity matrix, prepositions' diagram vectors obtained from Chapter 2 as input variables for a machine learning model. Our goal in the next chapter is to predict the distance based on the contextual factors and characteristics of the prepositions, locata and relata.

Chapter 5 - A Machine learning model to predict the distance between geospatial locations using contextual factors

ABSTRACT

People use relative location descriptions, which describe the location of one object relative to another, on a daily basis. Automated understanding of these descriptions is not an easy task, because the interpretation of spatial relation prepositions depends on the context in which they are used. For example, identification of the location referred to in the expression *building beside the Event cinema*, depends on the context of the scene, including aspects such as the location of other buildings, the orientation of the buildings and the location of the roads. In this chapter, we consider two datasets of different scales to predict the distance between locations in a geospatial scene. We use multiple contextual factors and features related to the elements of a location description as they have been shown to be important in Chapter 5 obtained from Chapter 2 and Chapter 3 as inputs into a regression model and predicted the distance between locations in the description. Then we compared the model with three baselines. Our regression model outperforms the best baseline with 25% improvement for both datasets.

5.1 Introduction

People often describe the location of objects using relative geospatial location descriptions (e.g., *the shop beside the cinema, the house behind the bakery*). Automated methods for decoding expressions of this kind are useful for a number of applications, including responding to requests for emergency services if an address is not known (e.g. *we need an ambulance near the public toilets in Metro Park*) and immediate response is important to save lives; identifying the location of disaster impacts reported on social media (e.g. *there is a power line blocking the bridge near the Community Arts Centre*) or identifying the location of species explained in text form in scientific reports (Scott et al. 2021). For this decoding, a machine must be able to understand the main components of a relative location description and interpret the expression based on the relationship between these elements. The main components of relative location descriptions are the locatum (the object whose location is being described), the relatum (the reference object for the locatum) and spatial relation terms (mostly prepositions but could be other parts of speech such as verbs) that connect the two. Relative location descriptions are not easy to interpret due to the vagueness and context sensitivity of spatial relation terms (e.g., prepositions such as *near*). Based on the context, prepositions can have different interpretations. As an example, *the house on the island* and *the house on the road* have different interpretations although the geospatial preposition *on* is used in both descriptions. The *house* in the former expression could be on any part of the island, but the *house* in the latter expression on the road is most likely close to the edge of the road. So, the relatum feature type and meaning change the interpretation of these two expressions.

In this chapter, we predict the distance between the locatum and relatum in geographic space using a machine learning model. We use a number of features consisting of numeric and descriptive values belonging to the locatum, relatum, and the relative location description text to create a regression model and evaluate their effect on the prediction of the distance between locatum and relatum. We focus on 24 qualitative non-directional spatial relation prepositions (we exclude geospatial prepositions that are quantified such as*: 10km to the right of the river*).

We address two research questions:

RQ1: How accurately can we predict distance using machine learning regression methods?
RQ2: How important are specific model features in the success of that prediction?

A large body of previous work addressing georeferencing of location descriptions has focussed on toponym recognition and resolution (Leidner 2008; Lieberman and Samet 2012; Karimzadeh 2016; Kamalloo and Rafiei 2018; Kew et al. 2019) but has not addressed the modification of the georeferenced named places that results from the inclusion of a preposition. For example, *just outside Paris* refers to a place name, but without consideration of the preposition, the georeference will be inaccurate. While a number of previous works have defined so-called spatial templates to generalise and sometimes predict the behaviour of spatial prepositions (Hall et al. 2011; Skoumas et al. 2016), these do not consider the context of the particular situations in which a spatial preposition is used. Other work has developed models that include basic contextual factors to predict location, but these are mostly applied in either indoor or artificial/blocks world environments (Moratz and Tenbrink 2006; Yu and Siskind 2017), or refer to location on images in order to describe or retrieve photos using spatial relations (Hall et al. 2011; Lan et al. 2012; Hall et al. 2015; Collell et al. 2018), and most apply a limited range of contextual factors including object types, urban/rural, embeddings and size (Lan et al. 2012; Hall et al. 2015; Collell et al. 2018).

In this chapter, we predict distance rather than georeferenced location (which is derived from distance and direction) because the prepositions we are working with are largely proximity, collocation and adjacency (non-directional) prepositions, and direction (and thus full georeference) cannot be predicted from these kinds of prepositions. Prediction of distance is useful because it determines the area within which the locatum is located, and for adjacency, collocation, and proximity geospatial prepositions, which do not have specific direction, we cannot narrow down the location further without other additional information. To the best of our knowledge, there is no previous work that uses a wide range of factors affecting spatial language elements for distance prediction between the spatial locations in a description.

In order to develop a model to predict distance for specific relative location descriptions, we use two datasets. The first dataset contains 24 geospatial prepositions (690 geospatial descriptions), and the second dataset contains 5 geospatial prepositions (a subset of 24 geospatial prepositions with 7364 geospatial descriptions). The data sets differ in their scale. We apply SMO regression (Platt 1998; Shevade et al. 2000) to a model that incorporates a range of contextual factors to predict the distance described by the geospatial preposition. As a motivating example, in the expression *there is a fire in a building close to the Sky Tower*, we assume that the location of the relatum (the Sky Tower, a well-known landmark in Auckland, New Zealand) is known, but the location of the building is unknown and there are several buildings around the area. We address the problem of predicting the distance between the *building* and the *Sky Tower,* given a range of information about the features and their properties; the geospatial preposition and the context of the area around the relatum.

The chapter is structured as follows: Section 5.2 reviews some of the previous work on the importance of geospatial prepositions in a text description, the impact of context on the interpretation of a description and some of the machine learning work that has previously used contextual factors to predict object location. In Section 5.3, we discuss the methodology, datasets and attributes used in the model; the regression technique we use to predict the distance between locatum and relatum and the baselines. Section 5.4 presents the results of the evaluation and our analysis of the importance of different attributes in the model. Section 5.5 discusses the results and Section 5.6 concludes the chapter and presents future directions.

5.2 Literature review

Spatial relation terms are one of the main elements of spatial location descriptions (Talmy 1983; 2000; Herskovits 1986; Retz-Schmidt 1988; Mark et al. 1995; Levinson 2003; Tyler and Evans 2003; Langacker 2008), and may be verbs, adverbs, prepositions, or other parts of speech. The geometric relationship between the locatum and relatum in a description is specified by a spatial relation term or terms, although the application of spatial relation terms is governed by cognitive and contextual factors. In this chapter, all the datasets and examples we review contain spatial relation terms in the form of prepositions, referred to in this chapter as spatial prepositions, and we confine our attention to geospatial prepositions.

Spatial prepositions are mainly qualitative (Freksa 1991; Freksa 1992; Frank 1996; Cohn et al. 1997; Yao and Thill 2005; Kunze et al. 2014), meaning that there is no specific quantity assigned to them, and this makes georeferencing challenging. In the example *there is a fire in a building close to the Sky Tower*, the distance between the locations is not specified, so in order to georeference the *building* (or identify which building from a collection of buildings surrounding the Sky Tower is being described), we need to determine the distance indicated by the spatial preposition *close to*. Some previous work has assigned numerical values (or ranges thereof) to prepositions based on the scale of the places (Fu et al. 2005; Delboni et al. 2007; Liu et al. 2009; Hall et al. 2011; Chen et al. 2018). Several works have proposed mathematical, often formal, models to describe the semantics of different spatial relations, including *topological* (Egenhofer and Franzosa 1991), *metric* (Hernández 1991), *directional* (Freksa 1992; Moratz and Tenbrink 2006) or *hybrid* models, which combine topological, metric, and directional spatial relations (Schwering 2007). These works provide theoretical models of the area to which a spatial preposition refers, and do not allow for context. In the case of disjoint proximal relations, such as *near* and *close to*, the topological models provide no guidance on what the actual separation distance might be. Furthermore, they describe the semantics of spatial relations, but are not directly connected to the natural language expression of those spatial relations, and most spatial relations can be expressed by many different lexemes (words or collections of words). Many of these are prepositions, but verbs, adverbs and other phrases may also be used to describe spatial relations.

Bateman et al. (2010) describe projective terms like *left* and *right* as approximate directions because they suggest an approximate, rather than exact, direction. For example, *the road to my right* may not refer to a road that is an exactly 90-degree angle from the direction I am facing but indicates an approximately leftward direction. Thus, projective relations can be considered uncertain in that they indicate degrees of agreement with a particular spatial configuration. A number of works have modelled degree of agreement with specific spatial prepositions, often using probability density fields, referring to them as spatial templates (Hall et al. 2011; Malinowski and Fritz 2014;

Logan and Sadler 1996) or applicability models (Hall et al. 2015). These approaches provide a generalised model of the locations to which a spatial preposition refers, relative to some relatum, usually by amalgamating multiple observation points across different spatial scenes. For example, Hall et al. (2015) gathered data from Geograph to identify the applicability models of the most frequent spatial prepositions (near and cardinal directions), using a human subjects experiment to create density fields. Their main goal was to automatically generate photo captions by applying density fields for a preposition to toponyms. For example, in their density field models, they show that for a cardinal direction such as *north of*, the area north of a toponym is denser (participants in their experiment choose the north part as more applicable than other areas for the north of spatial relation term) than the areas in other directions.

The importance of context in the interpretation of spatial prepositions has been widely acknowledged (Coventry and Garrod 2004; Klien and Lutz 2005; Malinowski and Fritz 2014; Collel et al. 2018). As an example of context-based interpretation, Klien and Lutz (2005) study the *adjacent to* spatial preposition in the expression *floodplain adjacent to a river* for the purposes of defining an ontology. They collect all of the definitions of *adjacent to* from WordNet, these being:

- *nearest in space and connecting without something in between,*
- *touching with common boundary,*
- *near or close without touching or connecting.*

They identify the third definition as the most suitable in the case of a floodplain, for which connection with a river is not essential (the floodplain is simply a place that holds flood water, but there could be other objects in between that do not block water flow). In contrast, adjacent land parcels are required to be touching in order for *adjacent to* to be appropriate. Thus, in addition to the meaning of a spatial preposition, we need to consider the nature of the locatum and relatum in some descriptions to accurately georeferenced an expression, and in these two examples, the area for which the preposition *adjacent to* would be acceptable differs. As another example to clarify the impact of the context on the interpretation, we can mention *A historic house on London's North Bank* versus *A historic house on St James's Pl.* These two descriptions both refer to *Spencer House* in *London.* However, due to the difference between the feature types of *North Bank* (which is a wide linear region) and *St James's Pl* (which is a narrow linear street) the interpretation of these two descriptions is not the same.

As we mentioned in Chapter 2 , there are three different frames of reference (intrinsic, relative, and absolute). Thus, the interpretation of a spatial preposition may vary according to the frame of reference of an observer, and contextual information is needed to determine that frame of reference.

In addition to the *feature type* and *frame of reference*, Stock and Hall (2017) studied some other contextual factors that impact the interpretation of spatial descriptions that have been discussed in Section 2.2.1. Although these works discuss the importance of contextual factors on the interpretation of spatial prepositions and categorise contextual factors in different groups (Stock and Hall 2017), they do not provide either formal or automated models to explicitly describe the ways in which they influence interpretation of spatial prepositions or predict the location of objects or distance between them.

In order to accommodate contextual differences in interpretation, rather than creating generalised models in the form of spatial templates that are generic across all contexts, a number of works have created models that predict location for a specific spatial scene, often using machine learning and incorporating basic contextual factors to aid in prediction (Hall et al. 2015; Skoumas et al. 2016; Collell et al. 2018). Chen et al. (2018) used qualitative spatial relations to model individual relations such as *near, inside, covered, east, north*, then used place graphs to integrate all the descriptions, as well as Kernel Density Estimation (KDE) and regression to create density surfaces and a hexagon tessellation surface. They use granularity of locatum, granularity of relatum, granularity of the discourse and the prominence of the relatum as some contextual factors in their work.

On the other hand, Collell et al. (2018) predicted the location of reference objects (relatum), using the location (coordinates) of the subject (locatum), the embeddings of the expression (subject, relation, object) and the size of the subject using neural network and regression methods. Their focus was on spatial relation verbs that have implicit location meaning (e.g., *the boy riding the horse* indicates that the boy is on the horse). Skoumas et al. (2016) also used distance and orientation between reference point (relatum) and the described point (locatum) to train a Gaussian Mixture Model (GMM) to predict the continuous density distribution of each spatial relation. They used qualitative spatial prepositions such as *in, on, at, near, by* and their focus was to quantify these spatial relations. Bisk et al. (2018) used a different method (neural networks) to predict the location in a block world. The spatial relations they used consisted of projective relations such as *below, right, left, up* and some complicated relations such as *rotate, towers, mirror, degrees.* As contextual factors, they used the embeddings of the instruction given to locate an object. Platonov and Schubert (2018), used a rule-based method to predict the appropriate prepositions in an indoor space. They used 14 relations that consisted of prepositions such as *near, at, in, under* and *in front of, left of.* The contextual factors they incorporated in their work were relatum type, the role, and physical properties of locatum/relatum and other surrounding objects in the room. Lan et al. (2012) and Malinowski and Fritz (2014) followed the same goal to retrieve images that have a specific object in their configuration. Lan et al. (2012) used only three spatial prepositions: *above, below* and *overlaps* and the latent ranking Support Vector Machines (SVM) method in addition to object types as a contextual factor for image retrieval. On the other hand, Malinowski and Fritz (2014) used more contextual factors such as embeddings and visual fragments in addition to object types and spatial pooling, Convolutional Neural Network (CNN) methods with 11 spatial preposition such as *above, across from, behind, below* to retrieve images.

Although these works are similar to what we have done in this chapter, the number of contextual factors they used is limited and do not take into account all the features associated with the locatum and relatum that might influence the prediction of distance. We create a model that incorporate a much richer set of contextual factors, and predict the distance associated with spatial prepositions in particular contexts.

Our work focusses on the role of different elements of a spatial description (relatum, locatum, spatial relation), the characteristics of those elements and the wider context of a spatial scene on distance prediction. The predicted distance can then be used to georeference the region in which a locatum is expected to be located, if the relatum coordinates are known.

5.3 Method

In this Section, we describe the two data sets that we used to train and test our method. We then present the features that were included in our model and explain the regression method that was used. The pre-processing of both datasets is shown in Figure 5.1 For Data Set 2, we mark the stages that has been done by Morris (2020) with a blue box.

5.3.1 Data set 1

The first dataset contains 690 geospatial descriptions collected from two web sources: *Geograph*[54] and *Foursquare*[55].

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

[55] https://foursquare.com/

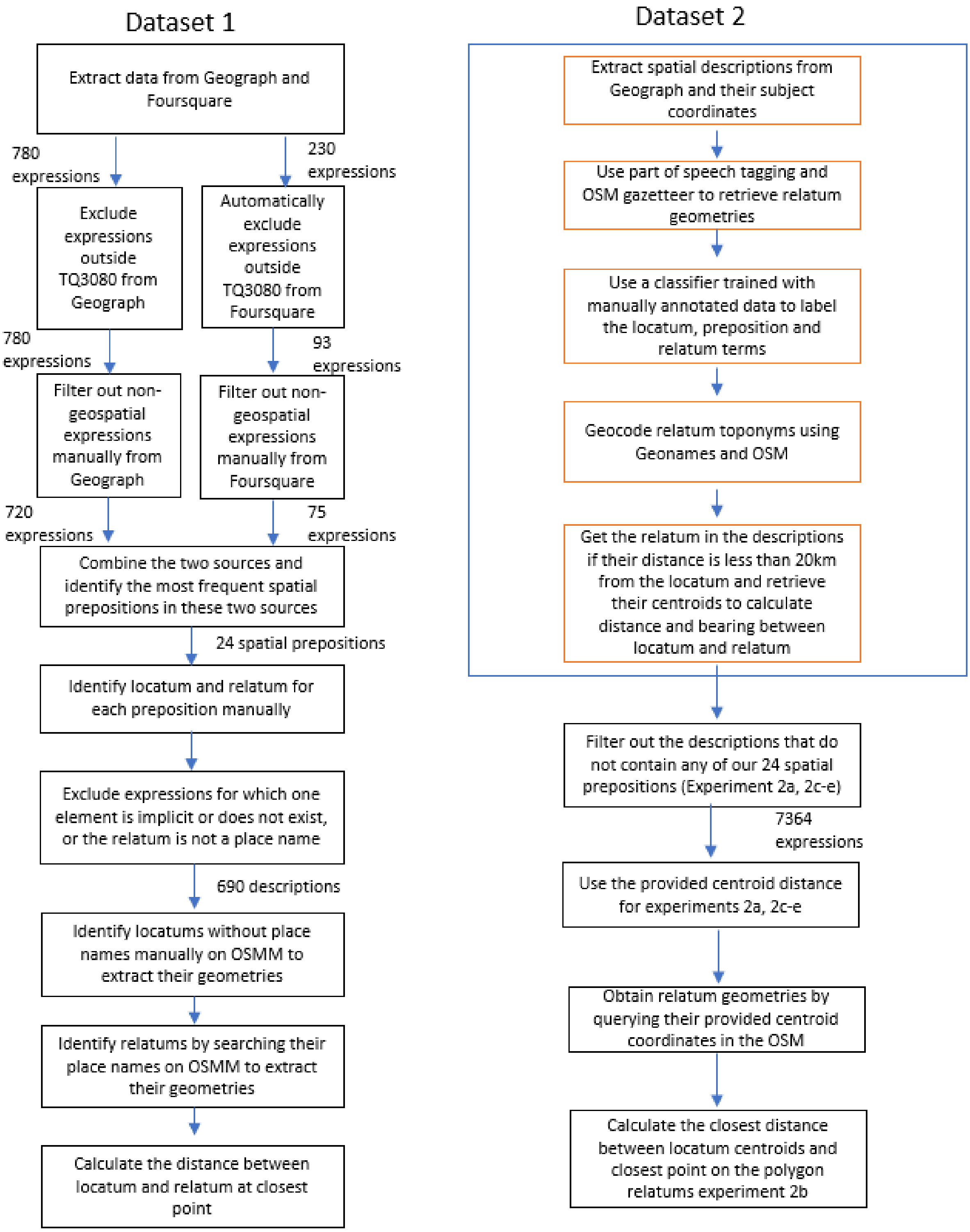

Dataset 1
Extract data from Geograph and Foursquare
780 expressions
230 expressions
Exclude expressions outside TQ3080 from Geograph
Automatically exclude expressions outside TQ3080 from Foursquare
780 expressions
93 expressions
Filter out non-geospatial expressions manually from Geograph
Filter out non-geospatial expressions manually from Foursquare
720 expressions
75 expressions
Combine the two sources and identify the most frequent spatial prepositions in these two sources
24 spatial prepositions
Identify locatum and relatum for each preposition manually
Exclude expressions for which one element is implicit or does not exist, or the relatum is not a place name
690 descriptions
Identify locatums without place names manually on OSMM to extract their geometries
Identify relatums by searching their place names on OSMM to extract their geometries
Calculate the distance between locatum and relatum at closest point
Dataset 2
Extract spatial descriptions from Geograph and their subject coordinates
Use part of speech tagging and OSM gazetteer to retrieve relatum geometries
Use a classifier trained with manually annotated data to label the locatum, preposition and relatum terms
Geocode relatum toponyms using Geonames and OSM
Get the relatum in the descriptions if their distance is less than 20km from the locatum and retrieve their centroids to calculate distance and bearing between locatum and relatum
Filter out the descriptions that do not contain any of our 24 spatial prepositions (Experiment 2a, 2c-e)
7364 expressions
Use the provided centroid distance for experiments 2a, 2c-e
Obtain relatum geometries by querying their provided centroid coordinates in the OSM
Calculate the closest distance between locatum centroids and closest point on the polygon relatums experiment 2b

Figure 5.1. Pre-processing steps for both datasets (the tasks inside the blue box were completed by Morris (2020) prior to this work)

5.3.1.1 Extraction of data in study area

Geograph is a photo sharing site which aims to store photos collected and submitted by members of the public from every square kilometre of Great Britain and Ireland. We downloaded the entire Geograph dataset using bittorrent[56]. By filtering out all the images outside the TQ3080 grid, this resulted in a total of 780 (a grid square in central London using Great Britain's map grid). For each image we extracted the long and short captions and coordinates of the subject of the photograph. Geograph keeps the coordinates of the camera and the subject (locatum), and we used the subject's coordinates to extract full geometries for our locata and relata (see below).

Foursquare is a website that allows people to submit reviews about points of interest (museums, restaurants, cafes, etc) they have visited. We used the Foursquare API to extract venue reviews (known as tips, or texts in Foursquare) in the London area. This resulted in a total of 230 spatial descriptions (similar to long captions in Geograph) in the comments from Foursquare. We automatically excluded the expressions with locations (coordinates) outside TQ3080, resulting in 93 expressions on Foursquare. Then, for both sets, we filtered out non-geospatial descriptions manually (Stock et al. 2013) resulting in 75 expressions from Foursquare and 720 descriptions from Geograph, and automatically extracted the expressions contain any of the prepositions listed by Landau and Jackendoff (1993) and Sithole and Zlatanova (2016) and identified the most frequent geospatial prepositions (by considering the ones that were used in at least 3 descriptions), resulted in 24 prepositions.

5.3.1.2 Identification of preposition, relatum and locatum for each expression

We examined each occurrence of the 24 prepositions in the data set. If the preposition was used geospatially (to describe a geographic location), we manually identified the

[56] http://torrents.geograph.org.uk/

associated relatum and locatum. Descriptions used non-geospatially (e.g., *National Theatre **at** night*) were filtered out. This manual process was validated by manual annotation of a sample of descriptions by one of the supervisors, with average accuracy score 0.84 (0.76 for locatum and 0.88 for preposition and relatum). We excluded:

- the prepositions *to* and *from*, as these are heavily dependent upon the verbs with which they appear, but in this work, we focus only on prepositions; and
- ternary prepositions such as *between*, as our scope in this chapter is limited to binary geospatial prepositions.

We also filtered out expressions whose relatum was not a specific place name such as *the river*, as we use the place name to determine the coordinates of each relatum for later distance calculations. The resulting data set contained 690 geospatial descriptions, their spatial elements (locatum, relatum and geospatial prepositions) and the coordinates of the subject (locatum). Each expression described the location of one subject (the locatum) based on a reference object (the relatum). For instance:

- *Savoy hotel (locatum) near (spatial preposition) Shell-Mex house (relatum)*
- *the walkway (locatum) along (spatial preposition) Savoy buildings (relatum).*

5.3.1.3 Calculation of distance between relatum and locatum for each expression

To train and evaluate our regression model for distance prediction, we next calculated the actual distances between each relatum and locatum in our data set. We consider that the centroid of the locatum and relatum is inadequate for accurate modelling of geospatial prepositions (since for example, the size of the object may influence the use of a preposition) so extracted the actual relatum and locatum geometries (Newstead and Coventry 2000; Kelleher and Costello 2009) from the Ordnance Survey Master Map (OSMM[57]) data set. In addition to calculating more appropriate distance measures than

[57] https://www.ordnancesurvey.co.uk/business-government/products/mastermap-topography

centroid to centroid (see below), this enabled us to later extract other features for our model (see Section 5.3.3).

We extracted the geometry of each locatum by manually identifying its location in OSMM. For the locata that were place names, we extracted the geometries belonging to the place name. For those that were not, we used the coordinates from Geograph (subject coordinates) or Foursquare to identify the closest feature of the correct type on map and extracted their geometries. We extracted the geometry of each relatum, all of which are named places, by searching for the appropriate place in OSMM.

Having extracted the geometries for each relatum and locatum, we calculated the distance between the closest points of the relatum to the locatum using the python Geopy library[58], as the dependent variable for our model.

One of the challenges encountered in using the OSMM data was that many, particularly linear (roads, rivers), geometries are recorded as several segments. For example, *Victoria Embankment* is represented in OSMM as a series of segments that touch each other, but that do not have any mutual identifier to enable all segments for a given place to be extracted. For these relata, we extracted the geometry of nearest segment to the locatum. This was a manual process (identifying all the segments, and selecting the closest), because it required each relatum to be investigated to determine whether the geometry was segmented.

5.3.2 Data set 2

Data Set 1 includes expressions in large scale space (i.e., those that describe specific objects within an urban environment, often buildings, statues). In order to further test our method, we employed a second data set that consists mainly of expressions that

[58] https://geopy.readthedocs.io/en/stable/

describe smaller scale space (involving objects like towns and cities) (see Table 5.1 for comparison of the data sets). Data Set 2 contained 19870 expressions extracted from Geograph from all over the UK and described in Morris (2020). This dataset contains only short captions with a limited set of geospatial prepositions (see Table 5.1). For short captions, a training model was used to tag the locatum, relatum, and preposition terms. The coordinates of the locatum and their categories were obtained from Geograph (locata are either feature types or feature types + place names), and the relata and their categories obtained from OSM.

From the entire data set, we selected 7364 expressions that used the prepositions we had identified as most frequent in our previous data set (*near, next to, close to, in, at*). The remaining prepositions in the data set from Morris (2020) (e.g., *north of, south of*) were excluded from our analysis.

We used the relatum names (place names) from this dataset to query the OpenStreetMap (OSM) Nominatim[59] API, to extract full geometries (polygons/polylines) to calculate distance and for later feature extraction (see Section 5.3.3). For Data Set 2, we calculated the distance between locatum and relatum (our dependent variable) using two methods, which we apply in different experiments (see Section 5.4.1).

- distance from the subject coordinates from Geograph to the relatum centroid (extracted from OSM API, presented by Morris (2020)) for Experiment 1.
- distance between the centroid on locatum to the closest point on the relatum (we queried relata from OSM using their coordinates and got their geometry) using geometries from OSM API (only for relata that have polygon geometries) for Experiment 2a.

[59] https://nominatim.osm.org/ui/search.html

Table 5.1. Comparison of datasets 1 and 2 (distances)

Data Set 1						Data Set 2			
Preposition	Quantity	Max	Min	Average	Stdv	Quantity	Min	Average	Stdv
above	9	111.34	0.00	42.75	39.96	-	-	-	-
across	18	88.56	0.00	16.72	30.32	-	-	-	-
adjacent to	11	104.79	0.00	30.27	30.30	-	-	-	-
along	24	267.54	0.00	30.51	53.95	-	-	-	-
alongside	16	65.08	0.00	18.19	22.65	-	-	-	-
around	8	64.22	0.00	24.75	21.50	-	-	-	-
at	51	84.98	0.00	20.64	21.37	2390	5.55	843.15	2245.83
behind	39	88.20	0.00	24.42	24.39	-	-	-	-
beside	15	55.65	0.00	11.49	18.64	-	-	-	-
beyond	24	436.94	2.99	108.23	105.38	-	-	-	-
by	7	36.17	4.06	17.88	13.30	-	-	-	-
close to	27	149.09	0.00	36.16	44.98	106	10.44	1225.78	2330.04
in	76	63.85	0.00	9.70	17.47	1682	0.97	1118.21	2869.00
inside	3	0.00	0.00	0.00	0.00	-	-	-	-
near	49	221.79	0.00	39.72	50.80	2812	3.81	1430.51	2329.02
next to	18	87.87	0.00	16.75	24.81	374	12.51	1866.59	3897.91
off	12	104.52	13.22	48.49	30.47	-	-	-	-
on	194	131.40	0.00	19.21	27.24	-	-	-	-
opposite	5	69.30	5.27	36.56	28.10	-	-	-	-
outside	18	49.60	0.00	15.72	15.29	-	-	-	-
over	14	160.00	0.00	23.65	55.38	-	-	-	-
past	3	58.10	4.93	27.80	22.33	-	-	-	-
through	11	120.17	0.00	39.03	43.46	-	-	-	-
towards	38	74.05	0.00	20.68	21.87	-	-	-	-
Values across entire set	690	436.94	0	25.37	40.53	7364	0.97	1187.75	2554.16

We extracted the closest points on relata to the locatum centroids only for polygon relatum objects because of the segmentation problem in linear objects. This segmentation caused some problems for distance calculations and resulted in high distances for some prepositions such as *in* and *at*. For example, in a description like *Suzuki dealership in Yarmouth Road,* which has a linear relatum, *Yarmouth Road* is recorded with multiple segments, one of which is 19km from the locatum. Instead, we chose the closest segment.

On the other hand, for polygon objects this problem does not exist and each polygon is recorded as one segment. Use of the centroid-to-centroid distance resulted in high distances in some cases. For example, in the description *the beach at Ryde,* if we use the centroids to calculate the distance, the distance between *the beach* and *Ryde* which is a seaside town is 1406.78 m. However, using the centroid on the locatum and closest point on the relatum returns 0 distance between these two locations because the centroid of the locatum is inside the relatum geometry.

5.3.3 Attribute selection

To feed the regression model with independent variables, we extracted several attributes that we considered might be influential in predicting the distance described by a given prepositional expression. These included preposition related variables, contextual factors, and locatum and relatum related variables. The variables are summarised in Table 5.2, and we discuss some of these in the following text.

5.3.3.1 Locatum and relatum features

5.3.3.1.1 Extracting feature types

We include the 50-dimension, pre-trained GloVe embeddings for the feature types of each relatum and locatum in order to enable patterns among similar types of features to be identified by the model, even if the explicit place name, or term used to describe the feature is different.

Table 5.2: Model Features for Data Sets 1 and 2

Feature	Data set 1	Data set 2
Locatum and Relatum Features		
Area	Area of relatum geometry from OSMM	Area of relatum geometry from OSM
Descriptive group	OSMM Descriptive Group attribute for relatum and locatum using one-hot encoding (see for example Section 5.3.3.1.1 from a total of 20 types.	Not used as the OSMM information was not obtained for this dataset
Feature type GLoVE embeddings	Pretrained average 50-dimension embeddings for relatum and locatum feature type/s (average across feature types returned using method described in Section 5.3.3.1.1).	Pre-trained average 50-dimension embeddings for relatum and locatum using categories and types for both locatum (locata are types rather than specific place names) and relata obtained from OSM (Morris 2020)
Expression GLoVE embeddings	Pretrained average 50-dimension embeddings for the entire expression, after stop-words were removed.	
Liquid/solid (**Lautenschütz et al. 2006**)	See Section 5.3.3.1.2, one-hot encoding of 2 types (liquid, solid).	
Geometry type (**Landau and Jackendoff 1993**)	See Section 5.3.3.1.2, one-hot encoding of 4 types (line, point, polygon, volume).	
Scale (**Lautenschütz et al. 2006**)	See Section 5.3.3.1.2, one-hot encoding of 3 types (District scale feature, Neighbourhood scale feature, and Immediate scale feature).	
Elongation	Not used	Calculated for polygon objects only. See Section 5.3.3.1.3
Preposition features		
Preposition semantic similarity	A semantic similarity matrix acquired from human subject data analysis (Chapter 3) explained in Section 5.3.3.1.2	
Preposition diagram vector	A 55-dimension diagram vector for each preposition using data from the Chapter 3 explained in Section 5.3.3.2.2	
Preposition sense	A 55-dimension vector determined using method described in Section 5.3.3.2.3.	
Other contextual factors		
Density (building) (Gahegan 1995)	Number of buildings (using OSMM descriptive group feature type) within a 20m buffer of the relatum divided by the area of the buffer, Section 5.3.3.3.	Not used
Density (same descriptive group as the relatum)	Number of objects with the same descriptive group as the relatum (OSMM descriptive group feature type) within a 20m buffer of the relatum divided by the area of the buffer, Section 5.3.3.3.	Not used
Density (different descriptive group as the relatum)	Number of objects with different descriptive groups from the relatum (OSMM descriptive group feature type) within a 20m buffer of the relatum divided by the area of the buffer, Section 5.3.3.3.	Not used

Data Set 2 includes feature types for both relata and locata, and many of the locata in Data Set 1 are feature types (e.g., *building close to Trafalgar Square*, in which the locatum is building, which is a feature type), so we extract the GLoVE embeddings for these directly. However, all relata and some locata for Data Set 1 are place names, and in order to extract appropriate embeddings, it is necessary to determine the feature types for these place names. In order to identify feature types for these names, we selected the Geonames feature type typology, which includes types such as *river, park* and *cinema*, because the feature types are diverse, and more likely to reflect the range of types within our data set. To match the place names used in our expressions to the Geonames types, we searched for each place name in two different sources: OpenStreetMap and Wikipedia. While OpenStreetMap returned meaningful feature types for some place names, for others the place names did not describe the feature in sufficient detail (e.g., Cleopatra's Needle category in OSM is recorded as tourism and its type is attraction). Furthermore, the OSM feature types were much more limited in range than the Geonames types. So, we only extracted the types from OSM that matched the Geonames feature types and did not consider the ones that were not. To supplement this data, we extracted all feature types from the first three sentences of Wikipedia for each place name (whether locatum or relatum). This process often resulted in multiple, different kinds of feature types, since Wikipedia sometimes explains not only the actual word, but also close locations/features (e.g., Nelson's Column, London returns column, monument, square and city feature types in the first three sentences). We thus calculated the similarity between all pairs of feature types returned from the above process using WordNet, with Wu and Palmer (1994) similarity measure, which returns a value between 0 and 1 to indicate degree of similarity (Miller 1995; Wu and Palmer 1994). For example, feature types column, monument, square and city were extracted from Wikipedia for Nelson's Column, and we calculated the similarity between each pair. We excluded any feature types that did not have a semantic similarity with any other feature type in the group of >= 0.8. This filtered out feature all the pairs except (column, monument), (column, square) and (monument, square) in the above example. However, in a small number of cases, because of the hierarchical structure of WordNet, some features had high similarity scores but were not necessarily semantically similar (for example, the similarity of all above-mentioned pairs is more than 0.8 but clearly square does not describe the feature type of the Nelson's Column). We therefore manually inspected the set of features to exclude any clearly inappropriate types. So, we ignored square. Square appeared as one of the extracted features because the description for the Nelson's Column was followed by *Trafalgar Square*. Having identified a set of feature types for a given place name (monument and column for this particular example); we then calculated the average embedding (average value for each dimension) to include in our model.

We also add 50-dimension GloVe embeddings for the OS master map descriptive groups mentioned above. OS Master Map (OSMM) uses descriptive groups as an indication of feature type for different locations (e.g., *roadside, landform, tidal water, road or track).* As an example, the *River Thames* has descriptive group *Tidal Water* and *Victoria Embankment* is recorded as *Road or Track.* We extracted the 50-dimensional GloVe embeddings for these descriptive groups and added them as the descriptive group embeddings to the model. If any descriptive group contained more than one word, we calculated the average of all the dimensions using all the words (excluding stop words such as or).

5.3.3.1.2 Additional relatum and locatum features

We extracted a range of additional features that describe the characteristics of the relatum and locatum as they are reported in the literature to influence the interpretation of a description. Lautenschütz et al. (2006) studied the impact of the liquid/solid relatum and locatum on the choice of spatial relations. They found that people choose *across* more when both locatum and relatum were solid; *through* more with liquid-solid locata and relata and rarely chose *next to* in a spatial scene involving only liquid objects. They also identified three categories for scale: *environmental,* *geographic,* and *figural.* The *figural scale* is small and mostly refers to steady indoor objects. *Environmental* and *geographic space* are much bigger and are conceived using maps. In their experiment, the preposition *through* was ranked higher for *environmental* and geographic *space.* In addition, Landau and Jackendoff (1993) discussed the impact of geometry type on the selection of specific prepositions. For example, *along* is used with an elongated relatum.

The relatum and locatum features included in our model were liquid/solid, geometry type (line, point, polygon, volume) and scale (district scale feature, neighbourhood scale feature and immediate scale feature) which are finer models of scales defined by Montello (1993) defined in Stock and Yousaf (2018). District and neighbourhood scale are considered environmental and immediate scale is considered vista based on Montello's categories of scale. Each of these characteristics affect the interpretation of the spatial prepositions. In the following spatial description[60], we can see the impact of geometry type characteristic.

The house on the ***North Bank*** versus *the house on the* ***Embankment Road*** *(geometry type)*

So, for the first description, the house can be *on* any part of North Bank area. However, the house *on* the Embankment Road, is very close to the road or on the edge of the road.

We extracted the locatum and relatum characteristics from the Linguistically Augmented Geospatial Ontology (LAGO) (Stock and Yousaf 2018), an OWL ontology that contains these characteristics for existing geographic feature types using ontologies created by the UK Ordnance Survey (Hart et al. 2004; Mizen et al. 2005). The characteristics included in the LAGO ontology were selected using previous works that

[60] https://www.geograph.org.uk/

identified their importance in the interpretation of spatial prepositions (Landau and Jackendoff 1993; Lautenschütz et al. 2006).

To add these features, we used WordNet to identify the class in the LAGO ontology that was most semantically similar to the feature type for each locatum and relatum (extracted as described in Section 5.3.1.1). We adopted the values for liquid/solid, scale and geometry of the most similar LAGO class, and if more than one class was the most semantically similar (with an equal similarity score, usually 1), we adopted the values for the first returned class. No semantically similar class could be identified in the LAGO ontology for a small number (4) of feature types (e.g., *bridleway*), so the value for these relata and locata were set to 0.

5.3.3.1.3 Elongation

We included elongation in our set of features because we observe in text examples that geospatial prepositions may have different semantics for features that are elongated compared to those that are not (Landau and Jackendoff 1993). For example, *the road goes beside the river* and *the road goes beside the church* describe different spatial configurations between relatum and locatum; while both suggest proximity, the former includes the notion of alignment. While the geometry type feature (see Section 5.3.3.1.2) may capture some aspects of this difference, some objects may have elongated shapes even if their feature type does not suggest this (e.g., a highly elongated building).

We did not calculate elongation for Data Set 1 due to the segmented nature of the OSMM data. As mentioned in Section 5.3.1, long geometry objects are broken into segments with no common identifier, so it is difficult to calculate an elongation value. Furthermore, we calculated elongation only for polygon features in Data Set 2, as elongation cannot be calculated for points and lines (which are segmented in OSM), and we designed different experiments accordingly (see Section 5.4.1).

To calculate elongation (Dey and Santhi 2017), we created a minimum oriented bounding rectangle around each polygon. The formula (Equation 5.1) to calculate the elongation is:

$$Elongation = 1\text{-}((width)/(height)) \qquad \text{Equation 5.1}$$

The width is the shortest edge of the oriented minimum bounding rectangle and height is its longest edge.

5.3.3.2 Preposition features

5.3.3.2.1 Preposition similarity

The geospatial preposition is one of the key determining factors in the interpretation of relative location descriptions and must be incorporated in our regression model. While it is possible to include features for each of the 24 geospatial prepositions that we consider using one-hot encoding (place 1 for the preposition appear in the given spatial description and 0 for the other 23), this approach would fail to capture the highly similar nature of some prepositions (e.g., *near* and *close to*; *next to* and *beside*). We thus use a semantic similarity matrix acquired from human subjects data (Chapter 3). After asking respondents to choose up to three diagrams that matched a given expression using Amazon Mechanical Turk, this chapter created a similarity matrix that included numerical measures of semantic similarity among 24 geospatial prepositions. We thus introduce a feature consisting of a vector of 24 values where each value represents the similarity between the preposition in the expression and each other preposition in the set of 24. Thus, the value in our regression model for the feature for the preposition actually used by the expression is 1, while the values for all prepositions vary between 0 and 1 indicating similarity with the preposition used in the expression. For example, if the expression is *the house next to the church*, the similarity score is 1 for *next to* and since the semantic similarity measure from Chapter 3 is 0.91, this value is given to the feature for the *beside* preposition for this expression, and so on for each preposition.

5.3.3.2.2 Preposition diagram vector

In addition to the semantic similarity measures from Chapter 3 , we included the 55-dimension diagram vector for each preposition using data from the same experiment. In that experiment (also reported in Chapter 3), respondents were asked to select up to three diagrams from a set of 55 spatial configurations that best matched an expression containing a geospatial preposition, and to score how well the diagram fit the expression. Vectors were then calculated containing the average degree of match between each expression used in the human subjects experiment, and each of the 55 diagrams across all respondents. We then calculated an average diagram vector for each preposition by averaging the value for each diagram across all expressions that used the preposition (using the data from Chapter 3). We added these vectors to our model by selecting the average diagram vector for the preposition that appeared in each expression in our data set.

5.3.3.2.3 Preposition senses

It is clear from previous studies (Cresswell 1978; Litkowski and Hargraves 2005) that many spatial prepositions have multiple senses (e.g., *I ran across the field; the house is across the street from me; buildings across the city*). A model that assumes that all uses of a spatial preposition are the same will fail to capture these variations. We therefore incorporated senses represented by diagram vectors as explained below into our model. This was done by clustering the diagram vectors from the human subjects data from Chapter 3 (see Section 5.3.3.2.2) in order to identify similar uses based on the preposition senses. For each preposition, we identified clusters of diagram vectors using agglomerative hierarchical clustering using the complete linkage method and tested different indices such as *silhouette, Hartigan, cindex* and *K1* to get the optimum number of clusters for each preposition, selecting the mode from these different indices. Then we calculated the average diagram vector of each cluster, as a measure of the semantics of that sense. Given that the expressions in our data set are different from those used in Chapter 3 , we then needed to identify which of the sense diagram vectors was most appropriate for each expression in our data sets. For this calculation, we extracted the relatum and locatum feature types of both of our datasets and the data set used in Chapter 3 and compared the similarity between feature types between the data sets using WordNet. We calculated the similarity scores of locatum with locatum and relatum with relatum using the Wu and Palmer (1994) similarity algorithm. The similarity between geospatial prepositions is not available in WordNet as it mostly contains nouns and verbs, so we used the similarity matrix discussed in Section 5.3.3.2.1 to calculate the similarity score between prepositions. Then we calculated the average of these three scores (locatum, relatum, preposition) and identified the most semantically similar expression in the human subjects data for each expression in both of our data sets, and adopted the diagram vector for the sense cluster for that expression. If for instance, expression number 420 in the human subject data from Chapter 3 was the most semantically similar to expression 1 in our data set, and expression 420 belongs to cluster 6 for the preposition *on*, we used the average diagram vector of cluster 6 for *on* as a vector for expression 1 in our data set and added this average diagram vector (55 dimensions) as an extra set of features to our regression model.

5.3.3.3 Contextual factors

In addition to the features that were directly expressed in the relative location descriptions, many of which described rich semantic information about the features and prepositions involved in the description, we incorporated additional contextual features in our model. There is substantial evidence that the context in which an expression is used can affect the interpretation of a spatial preposition (Gahegan 1995; Hall et al. 2015; Platonov and Schubert 2018; Collell et al. 2018), and density is one of these (Gahegan 1995). We therefore include three features that are proxies for different types of density, including not just overall density, but also density of buildings, and density of similar types of features (Table 5.2). For example, the distance indicated by an expression like *the post office is next to the church* may depend on how many other post offices and/or churches are in the area.

Having extracted the features for our model, we ran the regression model using the approach described in the following subsection.

5.3.4 Regression model

We used SMO regression (SMOReg) (Platt 1998; Shevade et al. 2000) to predict the distance between relatum and locatum. Sequential Minimal Optimization (SMO) is a more efficient method of solving the Support Vector Machine (SVM) training optimisation problem than typical QP (quadratic programming) solvers. SMO divides the training challenge into smaller problems that can be solved analytically using heuristics. We tested other regression algorithms including linear but as they did not perform as well, so we only report SMO regression here.

We used 10-fold cross validation to test predictive models by splitting the original data into two parts: a training set for training the model and a test set for testing it. The complexity we set was 400 (Abdiansah and Wardoyo 2015) with the RBF (Radial Basis Function) Kernel (Hearst et al. 1998). The RBF kernel is a function whose value is proportional to the distance between the origin and a given location. We ran the regression with all the attributes, and some subsets of them as explained in Section 5.4.

To evaluate the success of our approach and given that there are few other works that have addressed exactly this problem, we defined three different baselines for the evaluation of the results.

Baseline 1 - Zero Distance (ZD): The first baseline sets all distances to zero. This means that we considered the locatum to be in the same place as the relatum, or the relatum and locatum are touching or overlapping (5.4.1) with the distance described by the geospatial preposition being zero, and simulates the approaches to georeferencing in which only the place name of the relatum is used, with no account being taken of spatial relations terms (e.g. Tobin et al. 2010; Van Laere et al. 2013).

Baseline 2 – Random Distance within Max (RDwM): Baseline 2 predicts the distance between relatum and locatum as a random number generated between zero and the maximum distance for the preposition concerned, across the entire data set (we calculated maximum values for each data set, for each preposition).

Baseline 3 – Average Preposition Distance (APD): Baseline 3 predicts the distance between relatum and locatum as the average distance for each preposition across the data set (we calculated average values for each data set, for each preposition).

5.4 Evaluation

5.4.1 Experimental set up

We ran the following experiments to evaluate the performance of our approach:

Experiment 1 uses Data Set 1, consisting of 690 geospatial descriptions from Geograph and Foursquare, using 24 geospatial prepositions. Table 5.1 shows the number of expressions for each preposition, and Table 2 lists the features that were included in the model for Data Set 1.

Experiment 2a uses the entire Data Set 2, consisting of 7364 expressions with prepositions *at, close to, in, near* and *next to*. These expressions contain relata with polygon, line, and point geometry types. Table 5.1shows the number of expressions for each preposition, and Table 5.2 lists the features that were included in the model for Data Set 2, excluding the elongation features, which was only used for Experiment 2b.

Experiment 2b uses the expressions from Data Set 2 for which the relatum is of polygon geometry type, totalling 1893 expressions. This approach was taken as the elongation feature (see Section 5.3.3.1.3) can only be calculated for polygon relata, so was not included in Experiment 2a. Also, this experiment distances are the closest point on relatum to locatum centroid (the other experiments 2a, 2c-e distances are calculated between locatum and relatum centroids).

We also conducted three experiments (***Experiments 2c, 2d*** and ***2e***) that included only expressions using a single preposition each (*at, in* and *near* with 2390, 1682 and 2812 expressions respectively, selected as they were the most numerous). While our combined regression model that incorporates all spatial prepositions includes a number of features that describe the semantics of the prepositions, and thus we consider that it may be able to differentiate the distinctive behaviour of each preposition, these last three experiments test whether the model can better fit the data for an individual preposition. Distance is calculated between the centroid of the relata and subject coordinates for these experiments.

Experiment 2f also is a showcase of the dataset 2 with smaller range of the distances. This set contains the expressions which the distance between their locatums and relatums is less than 1000m. We examine if our model can perform better on specific range of distance or not.

We evaluated the success of each experiment using the following metrics:

- ***Mean absolute error (MAE)***, being the mean difference between predicted and actual distance between relatum and locatum across all expressions.
- ***Percentage within value (PWV)***, being the percentage of expressions for which the difference between predicted and actual distance between relatum and locatum is <= x metres, where x = 10, 50, 100, 500 and 1000 for Experiment 1, and x = 10, 50, 100, 500, 1000, 2000, 5000, 10000 and 20000 for Experiments 2a-e (due to the scale difference between the two data sets, as described in Sections 5.3.1 and 5.3.2).

For Experiments 1, we predicted the distances between the nearest point on a relatum to the locatum. For experiment 2b, we predicted the distance between the centroid of the locatum and nearest point of the relatum to that centroid. For Experiments 2a and 2c-e, we predicted the distances between the centroids of the locata and relata.

5.4.2 Experimental results

We report the mean absolute errors for all Experiments in Table 5.3. All the MAEs are in metres. In Table 5.4, we present the percentage within value (PWV) metrics for Experiments 1 and 2a-f.

Table 5.3. Minimum absolute errors of the first and second datasets and the baseline (in metres)

Distance predictions	Dataset 1	Dataset 2a- Whole set	Dataset 2b - Polygons	Dataset 2c - at	Dataset 2d- in	Dataset 2e- near	Distances smaller than 1000m- 2f
MAE (baseline 1)	25	1170	1239	843	1118	1437	363
MAE (baseline 2)	52	9141	-	-	-	-	343
MAE (baseline 3)	41	1412	-	-	-	-	214
MAE (Regression predictions)	**19**	**875**	**958**	**670**	**946**	**888**	**176**
Percentage improvement % (PI)	**25%**	**25% B1** **90% B2** **38% B3**	**23%**	**21%**	**15%**	**38%**	**51% B1** **49% B2** **19% B3**

We calculate the percentage improvement (PI) between our model and the best baseline (Baseline 1) using the Equation 5.2 below, with figures presented in the last row of Table 5.3.

$$PI = \frac{MAE\ baseline1 - MAE\ regression\ model}{MAE\ baseline1} * 100 \qquad \text{Equation 5.2}$$

As can be seen in Table 5.3, our model performs better than all three baselines in all experiments. The Experiment 1 MAE for both baselines and our model are smaller than for all the Data Set 2 experiments because of the difference in the scale of the expressions.

Among the specific preposition experiments (2c-e), our model gives better predictions for the *near* preposition than for *in* and *at.* Given that *in* and *at* are typically used to describe containment and collocation relations, which suggest a zero or near-zero distance between the locatum and relatum, it is likely that Baseline 1, which predicts a distance of zero between relatum and locatum, is a good estimate (although note that there are exceptions to this, as the expression *Visitor Centre at Fairburn Ings* (from Data Set 2) has a distance of 1646m between locatum and relatum). In contrast, the distances for *near* are mostly small but non-zero. Our model thus provides a better improvement over Baseline 1 for *near* than for *at* and *in*. We have not calculated baseline 2 and 3 for Experiments 2b-2e as these two baselines were significantly worse than Baseline 1 for Experiments 1 and 2a. Also, there was some large distances between some of the locations in data set 2 (some of them close to 20km) which resulted in the high random number generation and averaging. However, for experiment 2f, we calculated all the three baselines, and the average baseline (baseline 3) was closer to the predictions. This is because we only considered the distances less than 1km for this experiment and average baseline was a better baseline for the distance predictions in this experiment.

As can be seen in Table 5.4, Baseline 1 performs better at the 10m distance for Experiment 1, while our model shows an improvement at greater distances (especially 50m, where our model predicts 93.5% within 50m, compared to 82.2% for the baseline1). It must be noted that Data Set 1 (which Experiment 1 uses) is a small data set, with a wide range of prepositions and very low numbers for some of these (refer to Table 1), making effective training difficult. It is likely that a larger data set would have improved these results. Nevertheless, overall (considering MAE and PWV for distances > 10m) our model is able to give an improvement over all of the baselines, and this suggests that the approach is promising even when a wide range of prepositions (in this case 24) is included.

Table 5.4. Experiments 1 and 2 percentage within values (PWV)

Predictions – baselines / distances	10	50	100	500	1000	2000	5000	10000	20000
Prediction exp 1	43.0	**93.5**	**97.1**	100	100				
Baseline 1 exp 1	**51.3**	82.2	95.1	100	100				
Prediction exp 2a	**2.3**	**11.9**	**23.4**	**73.3**	**87.3**	**93.0**	**95.9**	97.6	100
Baseline 1 exp 2a	0.4	6.3	13.9	51.5	73.0	88.7	95.6	97.6	100
Prediction exp 2b	**2.8**	**11.4**	**23.7**	**73.2**	**87.6**	**92.2**	**95.2**	**97.0**	100
Baseline 1 exp 2b	0.4	6.2	13.1	53.5	75.2	89.5	94.7	96.7	100
Prediction exp 2c	**3.3**	**15.3**	**28.4**	**82.1**	**90.8**	**95.4**	**97.2**	**98.1**	100
Baseline 1 exp 2c	0.5	8.7	18.3	65.6	85.0	94.4	97.2	98.0	100
Prediction exp 2d	**3.2**	**16.6**	**32.9**	**78.0**	**87.9**	**91.7**	**94.5**	**96.9**	100
Baseline 1 exp 2d	0.9	12.3	25.3	67.8	82.5	90.0	94.3	96.5	100
Prediction exp 2e	**1.7**	**7.4**	**14.7**	**63.7**	**84.8**	**92.9**	**96.9**	**98.2**	100
Baseline 1 exp 2e	0.0	0.6	3.0	29.4	58.0	84.2	95.7	97.8	100
Prediction exp 2f	**4.1**	**20.5**	**38.3**	95.5	100				
Baseline 1 exp 2f	0.5	8.5	18.9	70.4	100				
Baseline 2 exp 2f	1.7	10	18.4	73.13	100				
Baseline 3 exp 2f	2.3	11.5	22.5	**96.7**	100				

Table 5.4 shows that for Experiments 2a and 2b-f, our model has higher percentage within value (PWV) measures than all baselines at all distances, with our model predicting 63.7% of expressions within 500m for Experiment 2e (the *near* preposition), compared to only 29.4% using the best baseline.

The regression model can handle the expressions with only three elements, locatum, geospatial preposition and relatum. That is to say, it cannot predict the distance for a geospatial preposition which is not in a set of 24. Also, it is not possible for it to handle the locatum and relatums which their feature type or names is not existed in GloVe embeddings. For example, if there is and expression such as: *The xyz tunnel into the forest,* if the *xyz* is not existed in the GloVe embeddings set, the 50 dimensions for it would be set to 0. Also, *into* is not existed in the set of 24 prepositions, so, the model will return an error.

5.4.3 Feature evaluation

To better understand the model, we analysed the effectiveness of the features for predicting distance between relatum and locatum. As mentioned in Section 5.3.3, we included three categories of features: locatum and relatum features, preposition features and contextual factors in our model. We used Weka's ReliefAttributeEval[61] evaluation function with the Ranker Search Method (Kira and Rendell 1992; Kononenko 1994). Relief attribute evaluation is a statistical attribute selection method that identifies and discriminates the attributes that have high quality in machine learning problems such as regression and classification (Robnik-Šikonja and Kononenko 1997). The features that were most influential in the predictions for the three datasets are presented in Table 5.5. Yellow indicates preposition features (diagram vectors), orange belongs to sense diagram vectors, white indicates the OS descriptive groups for both locatum and relatum, light pink represents the relatum scale, dark pink shows the locatum scale, light green represents the relatum geometry type and dark green shows the locatum geometry type.

[61] https://weka.sourceforge.io/doc.stable/weka/attributeSelection/ReliefFAttributeEval.html

Table 5.5: Most influential ten features for Experiments 1 and 2a (listed in rank order)

(individual vector values are explained in the text)

Experiment 1	Experiment 2a (whole data set)
Diagram vector value 14	Locatum scale – district
Relatum geometry type - polygon	Locatum geometry type - polygon
Relatum scale – immediate	Locatum scale – immediate
Diagram vector value 2	Locatum geometry type - volume
Relatum descriptive group – road or track.	Locatum scale – neighbourhood
Diagram vector value 9	Relatum geometry type - polygon
Diagram vector value 16	Relatum geometry type - point
Locatum descriptive group – building	Relatum scale – immediate
Sense diagram vector value 1	Relatum geometry type - volume
Diagram vector value 37	Glove embedding for Relatum Feature type

This ranking shows the importance of relatum and locatum characteristics and preposition diagram vectors (Section 5.3.3.2). For Experiment 1, relatum characteristics are most important, while Experiment 2b relies more on locatum characteristics, although relatum characteristics also play a role. Of the relatum and locatum characteristics, scale and geometry type are important. This confirms the assertions of Lautenschütz et al. (2006) regarding scale and Clements and Battista (1992) and Coventry (1999) regarding geometry type and supports the need for these characteristics to be included in predictive models. In contrast, solidity/liquidity and elongation do not appear in the top ten (although note that elongation was not included in the model for Data Set 1). Feature types do not appear to have high predictive power, although the one-hot encoding of the relatum as a road or track (most likely correlated with geometry type) and the locatum as a building are important for Data Set 1. The GLoVe embeddings of the feature types only appear in tenth place for Data Set 2, for the relatum, indicating that that specific feature type has less influence on the predictive power of distances associated with geospatial prepositions than more general characteristics such as scale and geometry type.

We also see that a number of the diagram vector (Section 5.3.3.2.2) values are important for Data Set 1. Each value in a diagram vector indicates the degree of agreement with a specific diagram (out of a total of 55), and some particular diagrams are especially important (see Figure 5.2). Values 2, 14 and 16 are sets of polygons with and without observer that indicate proximity and values 9 and 37 are sets of two touching linear objects. It seems that as well as the impact of geometry type and scale, the diagram vector geometries are also important in the prediction of distance. One value (dimension 1) in the preposition sense diagram vector (see Section 5.3.3.2.3) also appears in the top 10 for Data Set 1, indicating that in some cases, the specific sense of a preposition is an important contributor to prediction accuracy, supporting the importance of word senses documented in the literature (Kilgarriff 1997; Kågebäck et al. 2015; Pilehvar and Navigli 2015; Jackson 2019).

5.5 Discussion

5.5.1 Preposition distance patterns

Based on the previous literature (Egenhofer and Franzosa 1991; Clementini et al. 1994; Egenhofer and Shariff 1998; Shariff et al. 1998; Renz et al. 2000; Santos and Moreira 2009), some prepositions, such as *on, in* and *at,* are considered to be topological in nature, and to refer to locata that are inside or (possibly approximately) collocated with the relatum. If the locatum is inside or collocated with the relatum, we would expect to see a distance of zero between locatum and relatum (assuming measurement using boundaries rather than centroids, and a distance of zero if the locatum geometry is inside the relatum geometry, as was done for these experiments). The *in* preposition is associated with the container image schema, suggesting that the relatum is a container that the locatum is inside (Mark 1989; Shintani et al. 2016; Brooks 2018).

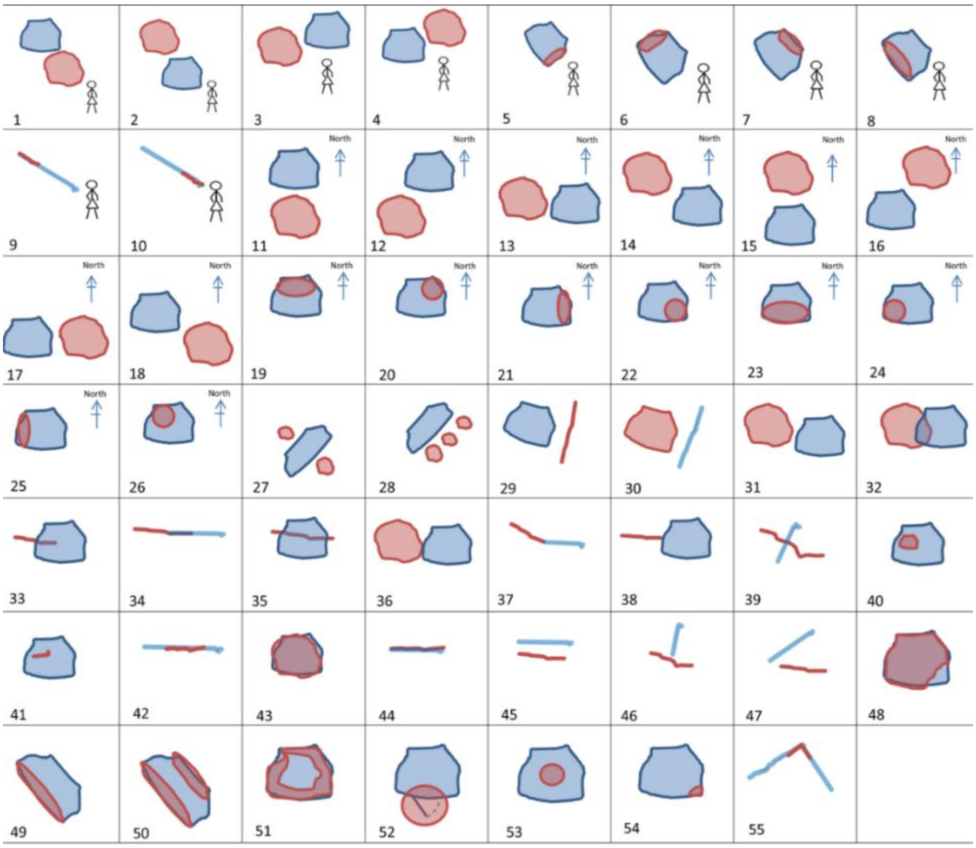


Figure 5.2. Diagrams of the human subject experiment (locatum in red and relatum in blue)

Similarly, the *on* preposition is associated with the platform image schema (Mark 1989), again suggesting that in two-dimensional space (from a survey perspective), the locatum would be inside, or at least collocated with, the relatum. However, our data shows that in some cases these assumptions are incorrect. Many distances associated with expressions using the *on* and *in* prepositions are non-zero, and sometimes, when the descriptions refer to an area that has large scale, these distances are large. Many of the examples used in the literature refer to artificial environments and examples in which object boundaries are much more clearly defined than in the geographic environment. For example, in the spatial description *the ball is on the rug,* the boundaries of the rug are clearly defined, and it is reasonable to expect that this expression would not be used if the ball were not completely on the rug's surface. However, in our real-world, geographic data, geospatial descriptions such as *customers take advantage of the spring sunshine outside this cafe on Trafalgar Square* indicate that these prepositions are used in an approximate way (in this case, the distance between the *cafe* and *Trafalgar Square* is 83m at the closest point). So, this suggests that in a geographic context, individuals may not necessarily regard these containers or platforms to have crisply defined edges. In another example from Data Set 2, *Camping in Pull Woods,* the distance between the camping location and the centroid of *Pull Woods* is 139m, and while it is close to the Woods, it is not inside the area that is considered *Pull Woods* on the map, even though the term *in* is used in natural language.

For some prepositions such as *beyond, near,* and *over,* the maximum distance values are high and the minimum is also low, describing a large distance range. For *beyond* and *over,* there are some expressions that define a location some distance from the relatum. For example, *Shell building beyond London Bridge* shows a large distance (337m) between the building and the bridge. This is one of the challenges in the interpretation of geospatial descriptions and results from the different usages/senses of geospatial prepositions in the geospatial context. In some expressions, the interpretation *beyond* is interpreted as far away (Cooper 1968; Mackenzie 1992; Landau and Jackendoff 1993; Mackenzie 2003; Lindstromberg 2010), while in others, it refers to a much closer locatum (*e.g., Inland Revenue beyond Montreal Place*). This is likely because the semantics of beyond are not primarily distance related, instead referring to something that is on the other side of, or past the relatum, and this could refer to a location that is either close or far away. This is important to note, London Bridge is a large structure that spans the Thames (where the width of the Thames is much larger than that of Montreal Place). So, the characteristics of relata change the interpretation of the preposition.

5.5.2 Data processing challenges

As mentioned, most of the linear objects (e.g., roads, rivers) in OSMM are segmented. A similar approach is seen in many other data sets (e.g., OSM), as it allows attributes to be attached to particular parts of a large, linear object that may vary substantially along its length. However, this data segmentation, together with the lack of an identifier to connect all segments of a given feature forced us to use a manual process in finding the most appropriate segment to the locatum coordinates in Geograph and Foursquare). For the relata, this process was performed manually, by visual inspection to identify the closest segment.

For example, the expression *Temple Church near Victoria Embankment* (where Victoria Embankment is represented in practice by multiple segments that extend over a long distance) (Figure 5.3), shows the segments for *Victoria Embankment (light blue borders)* and a segment for *Temple Church (*yellow border inside the red rectangle*).* As can be seen *Victoria Embankment* has multiple segments and for this one, we chose the closest segment of it (the segment filled with light blue colour) to *Temple Church* that resulted in 194m distance.

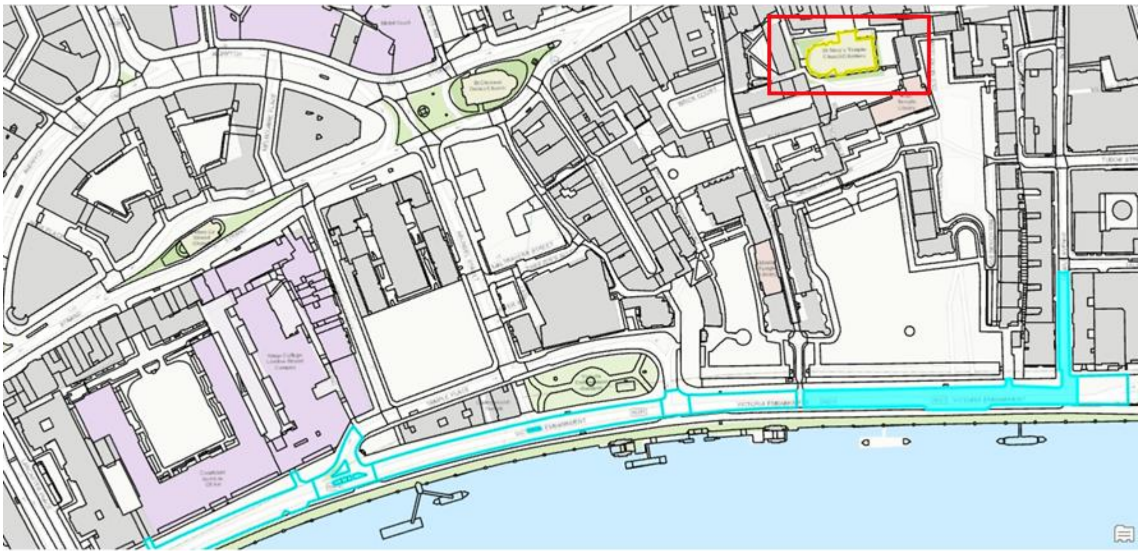


Figure 5.3. Segments of Victoria Embankment and Temple Church in the description Temple Church near Victoria Embankment

5.5.3 Screen shot of the WebApp

Using the regression model described in this paper and identifying the effective factors on this prediction, we have developed a WebApp[62] that uses the factors such as GloVe embeddings of locatum and relatum feature types, preposition diagram vectors and their semantic similarity and the LAGO ontologies to predict the distance between a locatum and relatum. The prototype is a web application displaying a map using the OpenStreetMap[63] data for London. Users can choose one of the 24 geospatial prepositions, locatum types (around 1100 types) and relatums (place names across London). Then the results are shown in a donut shape which shows the area in which the locatum is likely to be located. We used the regression model for dataset 1, as it belongs to the London area and predicted the distance between locatum (unnamed place) and relatum (a place name). After using the regression model to predict the distance for each expression, we calculate the width of the donut by dividing the mean absolute error for dataset 1 (19m) by the average of all distances across all the expressions in dataset 1, (which is 25m, Table 5.1) which is 0.76. So, to calculate the locatum location in each distance prediction, we showed the donut area with the predicted distance plus and minus 0.24 of the predicted distance (Equation 5.3).

$$1 - \frac{MAE\ regression\ model\ Data\ set\ 1}{Average\ of\ all\ distances\ in\ Data\ set1} = 0.24 \qquad \text{Equation 5.3}$$

$$Locatum\ location = predicted\ distance \pm\ (0.24 * predicted\ distance)$$

[62] https://koja.io.ac.nz/

[63] https://www.openstreetmap.org/

Figure 5.4 shows a sample of this WebApp with the Locatum type theatre, preposition *near* and relatum Trafalgar Square. Actual theatres within the donut are marked with orange dots.

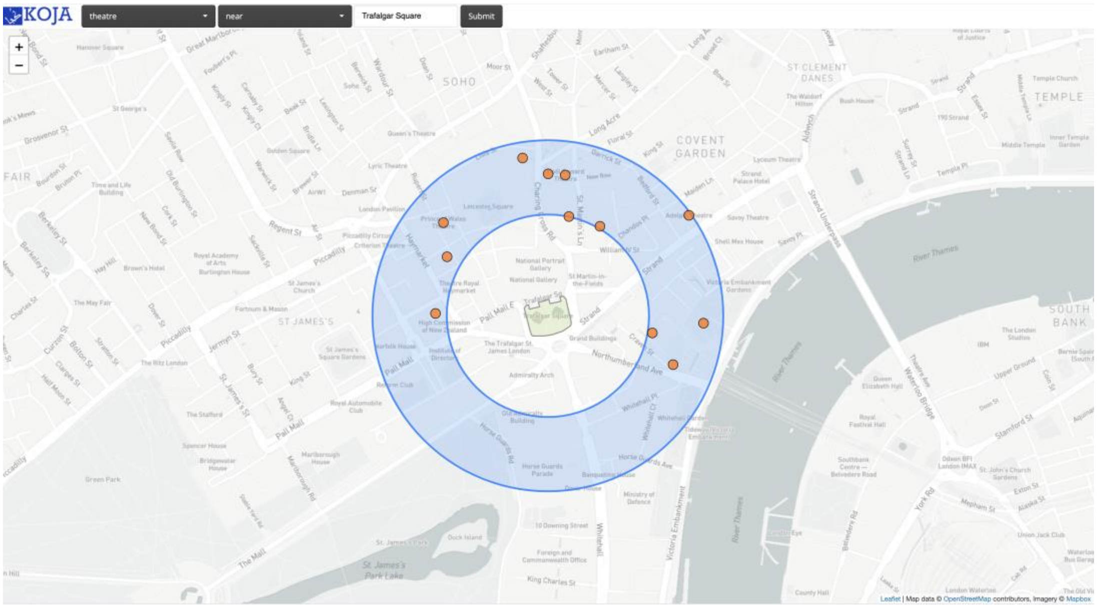


Figure 5.4. The predicted area for theatre type locatum, near preposition and Trafalgar square relatum

5.6 Conclusion

In this chapter, we used machine learning to predict the distance between locations described in relative natural language expressions. We used a regression model and set of input features including the characteristics of the locatum and relatum (feature type via GLoVE embeddings, scale, geometry, liquidity/solidity, elongation); vector models of the geospatial prepositions that describe the relation between the locatum and relatum; and contextual factors (density of objects in the area) with 10-fold cross validation to predict the distance between the locatum and relatum. We studied the importance of different features in our model in predicting distance, finding that scale and geometry type of the locatum and relatum and the vector models that refer to possible diagrammatic representations of the prepositions are among the most influential. We demonstrate the importance of a range of non-geometric features in accurate prediction of distances associated with relative expressions, with applications across a range of domains. Our research answers each research question as follows.

5.6.1 RQ1: How accurately can we predict distance using machine learning regression methods?

We tested our regression model on two different data sets using 6 experiments (Experiment 1 and Experiments 2a-f). For Experiment 1, the model returned a 25% improvement relative to the best baseline. Experiment 2a, which used the data Set 2, showed 25% improvement over the best baseline. Experiments 2b-e tested subsets of data Set 2, and all showed improvement over the best baseline that assumes that locatum and relatum coincide. The highest improvement was for experiment 2e, which included only descriptions that contained the preposition *near.* This showed 38% improvement over the best baseline (baseline 1). Experiment 2f also shows that for the smaller range of distances in data set 2, we can consider baseline 3 and the improvement of this baseline was 19% by our model in this experiment.

5.6.2 RQ2: How important are specific model features in the success of that prediction?

We used a number of features for the first time in our prediction model, compared to previous work that focussed more on the geometric arrangements of the relatum and locatum. Two of these new features (the diagram vectors and preposition senses) were among the most important in distance prediction, indicating the need to represent preposition semantics in predictive models of the distance associated with geospatial descriptions. We also found that the scale and geometry type of the relatum and locatum were important in predicting distance. In contrast, the density features that were included did not appear to have high predictive power in the model.

In future work, we plan to include additional features in our model, and to extend from distance-based prepositions to incorporate other kinds of prepositions, such as projective. For example, we did not consider prepositions that specify orientation such as *left of, right of*, and these prepositions require additional information to be included in the feature sets that covers aspects of orientation relative to some axis. In addition, the senses we consider here are based on a clustering and WordNet similarity algorithm. We plan to evaluate other similarity algorithms and/or instead of comparing the feature types of the locata and relata, compare the similarities in relatum and locatum characteristics (e.g., *image schema, scale, geometry type*) and their embeddings.

In the next chapter, we discuss the accuracy of human annotation of geospatial location descriptions. As we discussed in Chapter 2, 3 and 5, it is used as a ground truth for many machine learning models and we investigate whether this reliance on human annotation is valid.

Chapter 6 - Challenges in creating an annotated set of geospatial natural language descriptions

ABSTRACT

As described in previous chapters, in order to extract and map location information from natural language descriptions, a first step is to identify different language elements within the descriptions. In this chapter, we describe a method and discuss the challenges faced in creating an annotated set of geospatial natural language descriptions using manual tagging, with the purpose of supporting validation and machine learning approaches to annotation and text interpretation. This manual tagging or annotation can be used as a ground truth for further machine learning/deep learning models (e.g., the machine learning model we used in the previous chapter).

6.1 Introduction and literature review

To progress research on the interpretation of geospatial natural language, methods for automated tagging of spatial language are required (Kordjamshidi et al. 2011; Stock et al. 2013). In this chapter, we discuss the challenges that we encountered when trying to create manually tagged annotated data set that addresses the shortcomings of previous data sets, using two experiments. A number of researchers have addressed the problem of annotating geospatial natural language. For example, Stock and Yousaf (2018) annotated a wide range of language elements, including adverb and parts of objects as well as relatum, locatum and spatial relation, mainly by extending POS tags in a rule-based approach. Kordjamshidi et al. (2011) restrict their attention to trajector (locatum), landmark (relatum) and spatial prepositions, although they acknowledge that other parts of speech can be used to express spatial relations. GUM Space specifies a broad range of tags including locatum, relatum, spatial modality (Hois et al. 2009). SpatialML uses mark-up language to tag elements (Mani et al. 2010) including places, coordinate, orientations, form of reference, direction, distance and frame. Work by Zwarts (2005) and Kracht (2008) address spatial prepositions, with a focus on directional prepositions and location. Much of the previous work is either limited to very simple elements

(Kordjamshidi et al. 2011); adopts a complex tag structure (Hois et al. 2009) or assumes a particular syntactic (grammatical) structure (Kordjamshidi et al. 2011; Stock and Yousaf 2018). We propose an annotation scheme that addresses these limitations in that it focuses on semantics rather than syntax. Section 6.2 describes the methodology, results are presented in Section 6.3, and the conclusion and future works are discussed in Section 6.4.

6.2 Methodology

We conduct our exploration of the challenges of creating an annotated data set using two experiments. The first one compares the tagging conducted by pairs of human annotators and discusses discrepancies and issues involved in manual tagging. The second one discusses variations between individual human respondents in matching natural language descriptions to spatial relations, highlighting the lack of consensus.

6.2.1 Experiment 1: creating an annotated data set

The selection of an annotation scheme was based on three criteria: 1. What must be individually identified in order to support effective geocoding of the text? This is difficult to evaluate conclusively, as it depends upon the geocoding approach, and some aspects of spatial language are still not well understood. This criterion influences not only which items we tag, but also which items we identify as separate elements. For example, it is not useful to separate *next to* into two separate tags, because the meaning depends on the combination of the words, and the meaning of *to* in particular is dependent on the presence of *next*. In contrast, adverbs like *right*, or *directly*, have their own meanings which are similar regardless of the preposition they appear with, although the meaning may be influenced by the latter. 2. Can some of the tags or their subcategories be reliably determined automatically? If a particular semantic tag can be reliably identified through an automated approach, then there is little point in annotating in manually. The reliability of an automated approach is a question of degree, but we use the yardstick that if the set of words of interest can be defined by a clear set of specific words, none of which are homonyms, then they might reliably be identified automatically. In practice this is rare, because for example, even though the set of prepositions is a closed word class, since we are interested in semantic tags rather than syntactic, and prepositions normally encode spatial relations, there are examples of spatial relations that are not prepositions *(e.g.in line with)*. 3. What is practical to expect people to reliably annotate? This involves both volume and simplicity. A set of tags that is too complex will be difficult for manual annotators to deal with. The set of tags must be manageable in quantity, and simple enough to understand without specialist knowledge.

In Experiment 1, we develop a generic spatial annotation framework based on the semantic roles of tokens in a sentence. To this end, 1000 sentences were randomly selected from the combined set of three data sources: The Nottingham Corpus of Spatial Language (Stock et al. 2013), The Landcare Research National Soils Database [64] and The Where Am I survey, in which natural language descriptions were elicited from human respondents, as described in (Stock et al. 2015). Table 6.1 identifies, describes, and explains the annotation scheme that was used. Four annotators were given an expanded version of Table 6.1 with a simple explanation of terms and examples. Four research assistants were recruited to assist with the manual annotation and were paid a standard data entry casual rate. Three of them were undergraduate students, with the following areas of study: Software engineering major, linguistics minor; History and chemistry double major; Software engineering major. The fourth person was a professional research assistant specializing in data entry and document transcription, with an undergraduate degree in education. The selection of the annotators was non-random and aimed to achieve a balance of skills in data analysis and language knowledge. All were native English speakers.

The purpose of the work was explained to them in simple terms, and they were given access to the tagging app. Each annotator was then asked to annotate 10 expressions using the tagging app, after which the authors examined the expressions and gave

[64] https://soils.landcareresearch.co.nz/index.php/soil-data/national-soils-data-repository-and-the-national-soils-database

feedback on any issues, before the annotator began annotating in earnest. Each expression was tagged twice by two different annotators.

Table 6.1. Tag labels and descriptions

Title	Explanation
Trajector (locatum)	The object whose location is being described. The important role of the trajector in spatial language has been discussed by a number of researchers and is also known as locatum (Hois et al. 2009) or figure (Talmy 2000).
Landmark (relatum)	The object that is used as a reference point in the description. The landmark also plays an important and well documented role in spatial language and is similar to the relatum and ground identified by other researchers (Talmy 2000).
Spatial relation	The word or words that indicate how two objects are positioned relative to other. The importance of spatial relations has also been well recognised, and they have been widely researched (Coventry and Garrod 2004, Kelleher and Costello 2009, Zwarts 2005). In syntactic terms, spatial relations are most often represented using prepositions, but not always.
Location and movement verb (lmv)	A verb that describes the manner in which one object is positioned relative to the other. The location and movement verb is a subset of the verb syntactic category (Talmy 2000). *The road* ***crosses*** *behind the church.*
Spatial qualifier	A word of set of words that adds more information to the spatial relation and or the location and movement verb. Spatial qualifiers have not been widely recognized as an important carrier of spatial information as yet and may be represented with a range of different parts of speech, including adverbs, adjectives and nouns. *The road goes* ***right*** *beside the church*
Spatial specifier	A word of set of words that describes particular subparts of a feature. E.g., *The* ***north of*** *the country*. Spatial specifiers have also not been widely studied in specific terms, with work instead focusing on general issues of mereology (Hahmann and Gruninger 2011).

6.2.2 Experiment 2: matching of expressions to spatial relations

In the second experiment, we used data collected in earlier work (Stock and Yousaf 2018). In this work, respondents were shown expressions one at a time, and asked to match each expression to one of a series of diagrams that illustrated spatial relations. After viewing the expression and the set of available spatial relation diagrams, each annotator was asked to select values on a Likert scale that included only the positive side of the scale, to indicate his or her opinion about how closely each of the selected spatial relation diagrams matched the expression: *Strongly Agree, Agree, Agree Somewhat*. Only the positive half of the scale was used because users were invited to only select diagrams that they thought reflected the expressions (i.e., if they did not agree, the respective diagram would not be selected). Weights were allocated to each response for a given spatial relation diagram-expression pair, using 1, 0.75 and 0.5 for Strongly agree, Agree and Agree Somewhat respectively. The score for each expression and its geometric configuration was calculated using this formula (Equation 6.1):

$$\mathrm{GCO}Score_{\text{expression, diagram}} = \left(\sum_{k=0}^{n} response_k \ * weight_k\right)/n \qquad \text{Equation 6.1}$$

In which *response k* represents the number of responses with *weight k*, and *n* defines the total number of responses for expression k. Full details of the methodology can be found in (Stock and Yousaf 2018).

6.3 Results

In order to evaluate the reliability of the manual annotation process in Experiment 1, we calculate inter-annotator agreement among the four annotators. Since expressions were randomly allocated to annotator, any combination of pairs of specific annotators may annotate a given expression. Inter-annotator agreement was calculated by comparing the words in a given expression that were given a particular tag by each annotator. Since many of the expressions were complex and contained more than one of some tags, we calculate agreement by proportion of overlap between the words annotated with a particular tag by each user, rather than by a simple true/false agreement. Equation 6.2 expresses this measurement of agreement between annotators for a single expression: For a given tag, ME_k denotes the number of mutual elements (words or multi-word tagged values) that both annotators agree on, and max_k denotes the maximum number of elements that are tagged by either annotator. The total agreement score for the expression is then average of agreement across the populated tags. For example, if user 1 specifies Australia, New Zealand and Canada as landmarks and user 2 specifies Canada and USA as landmarks ME_k for the landmarks would be 1, because just Canada is mutual and the max_k would be three as the maximum number of landmarks by either annotator. The agreement score is calculated for all the tags in an expression, and the average is calculated to determine the agreement across the entire expression.

$$AgreementScore = Average\left(\sum(ME_k/max_k)\right) \qquad \text{Equation 6.2}$$

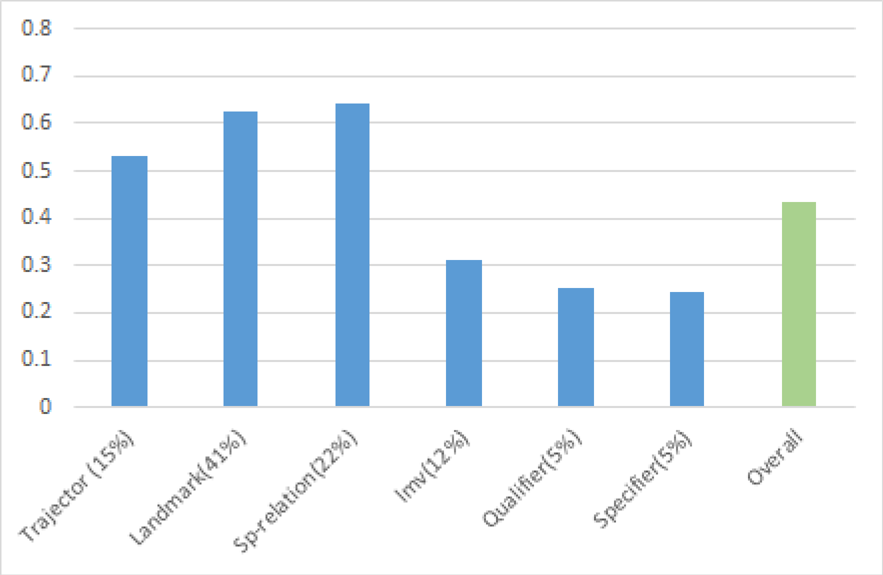


Figure 6.1. Mean inter-annotator agreement by tag type

Figure 6.1 shows the mean inter-annotator agreement for individual tags, as well as overall and also the percentage of tags of each type that were annotated in the 1000 expressions. We used this formula, to have an accurate calculation of each separate tag.

We also explore the role of annotator experience in the manual tagging process and evaluate whether annotator performance improves over time. For each annotator, we calculated inter-annotator agreement for the first, second and third 50 expressions tagged by three annotators through the time to see whether their performance changed by time or not. Only 3 annotators are shown because the remaining did not annotate sufficient expressions. Figure 6.2 (a-c) show the results. We then calculated the inter-annotator agreement of different subsets of annotators, to determine whether some annotators were more successful than others in tagging, either overall of for specific tags. The results (Figure 6.3) show some inconsistency. It is, however, clear that Annotator 2’s contribution is important, with her exclusion resulting in overall deterioration.

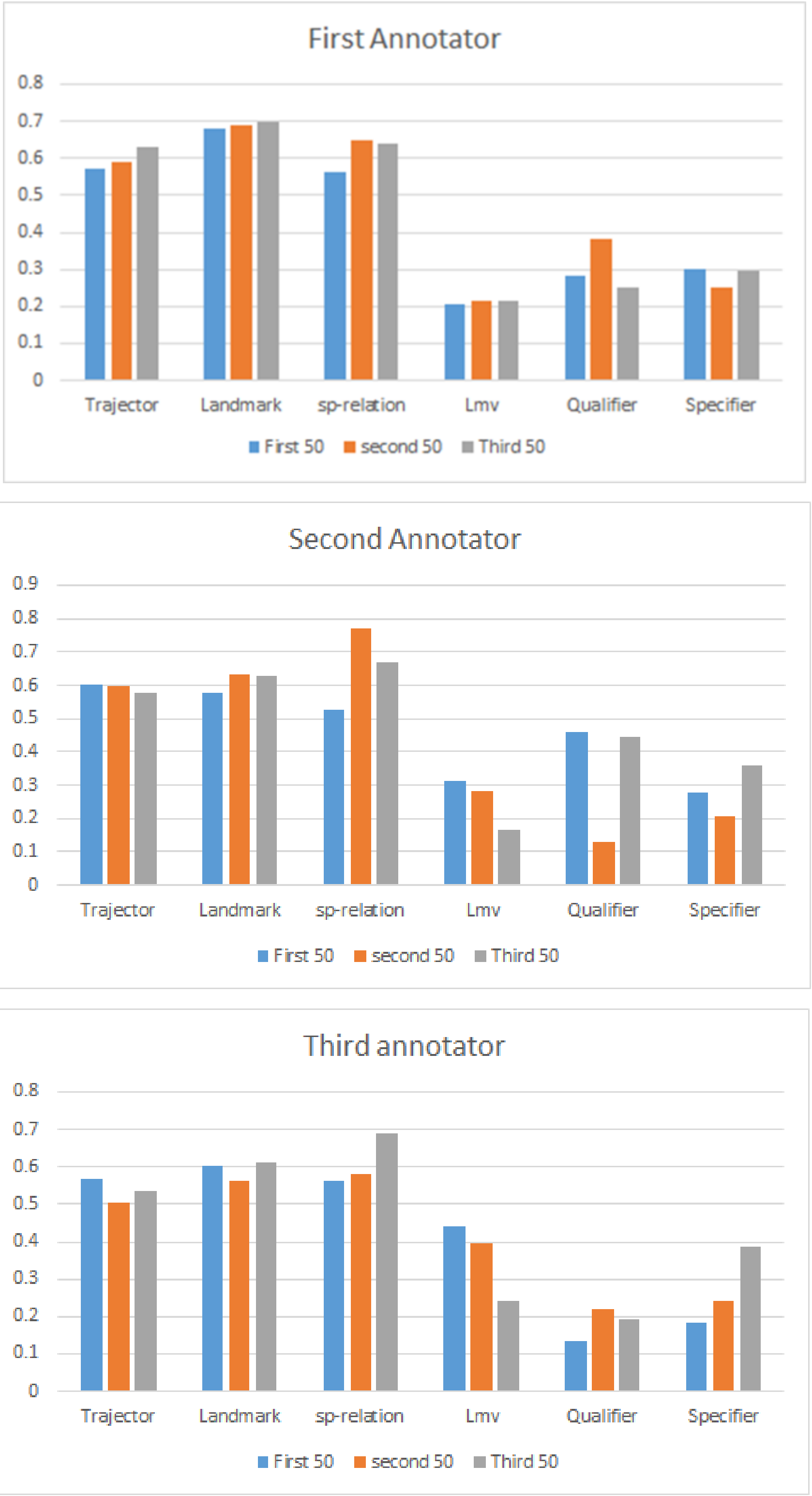


Figure 6.2 (a-c). Annotator performance through the time

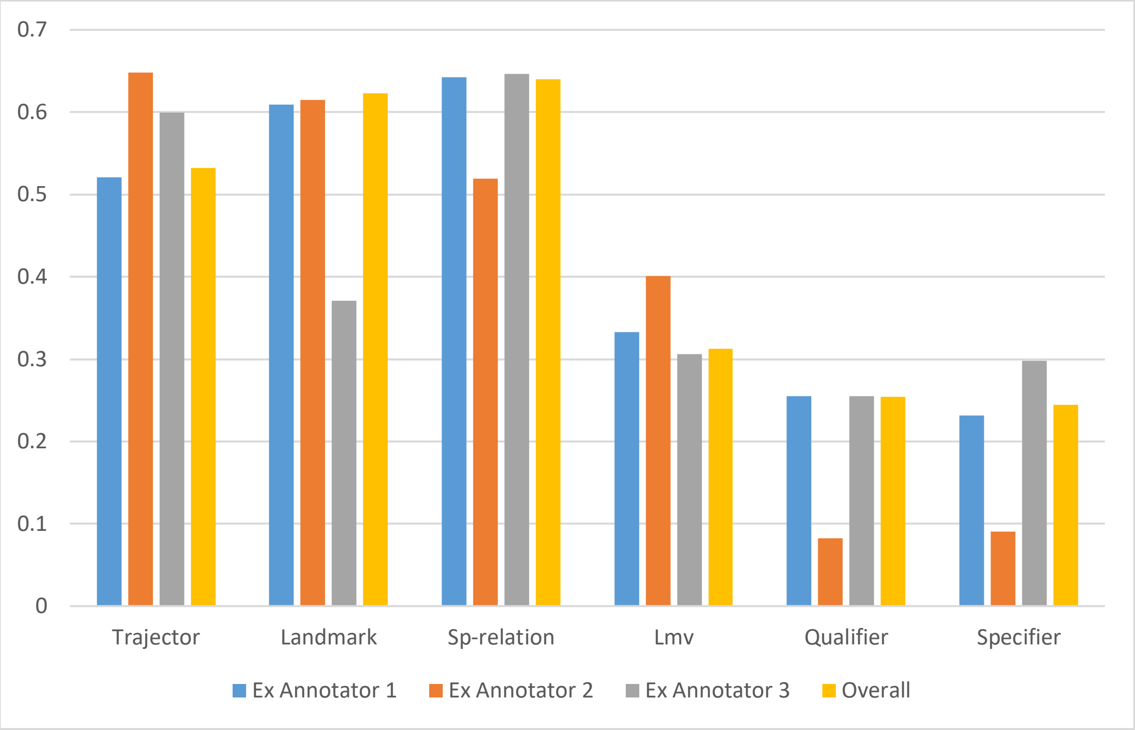


Figure 6.3. Inter-annotator agreement excluding each annotator in turn

Turning to Experiment 2, the results highlight the lack of agreement among individual respondents regarding the spatial relation diagram that best reflects a given expression. The respondents in Experiment 2 were also non experts in geographic information science.

Figure 6.4 (a-b), each show the spread of responses for three example expressions. In contrast to Experiment 1, Experiment 2 used short, simple spatial expressions, and the graph shows the frequency (after weights have been applied as described in Section 2) of selection of each spatial relation for a given expression. Two expressions in Figure 6.4(b) show a number of small peaks, with no clearly dominant relation selected by the respondents. Across the entire data set, a similar pattern was observed, with lack of consensus among respondents in selecting spatial relations to match many expressions.

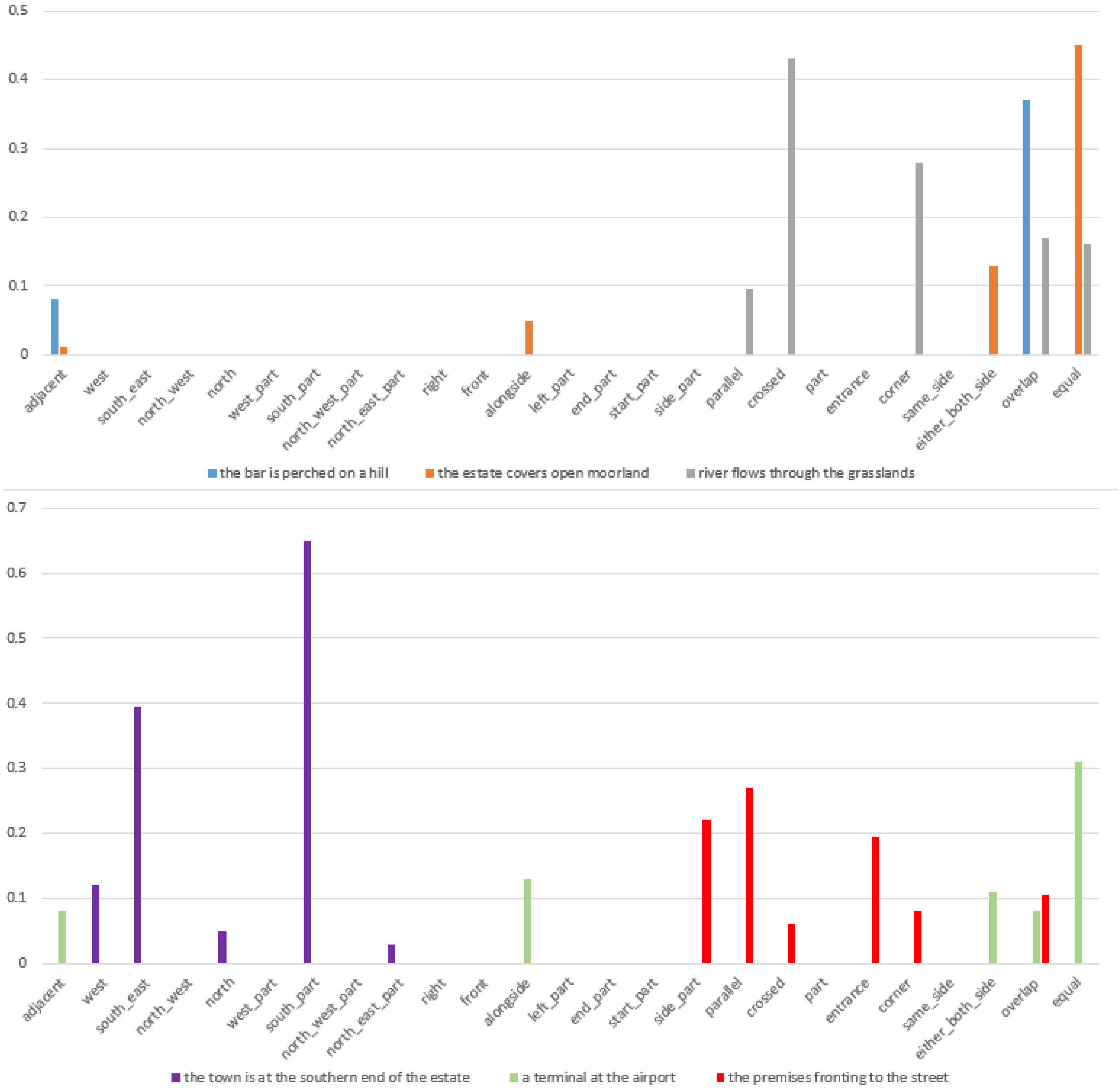


Figure 6.4 a-b. Study 2. GCO score for second three expressions

6.4 Discussion and conclusion

The results clearly show that it is not straightforward to create a manually annotated data set of natural language descriptions with a broad set of language elements that is based on semantics rather than syntax. Obviously, for an annotated data set for use in machine learning and validation, we would like the agreement to be very strong. Considerations of the level of experience of the annotators and the examination of the influence of specific annotators on particular tags did not result in noticeable improvement. The challenges that were encountered can be summarised as follows: Firstly, it is not unusual for the same place name, geographic feature or moving object to be both a trajector and a landmark, and secondly, the landmark/trajector status of a word may be ambiguous. The following example illustrates both of these cases. In the expression *the church stands beside the post office near the bridge*, the structure of the expression could be:

- *trajector+ (lmv) + spatial relation +landmark +spatial relation + landmark*
- *trajector+ (lmv) +spatial relation + (trajector and landmark) +spatial relation + landmark*

In the first case, church is a trajector for both the church landmark and the bridge landmark, and in the second case post office is the trajector for the bridge landmark, as well as the landmark for the church trajector. The annotation scheme used in this chapter allowed each word to be tagged only as a trajector or a landmark, but not both. The creation of a tag that indicates a dual role may be a possible method for addressing this. Resolution of ambiguity is a more difficult problem to solve, and even the most expert and experience annotators may disagree. A final observation from the results is that spatial qualifiers and spatial specifiers had only fair inter-annotator agreement (lower than other tags), and while this may be in part due to confusion about when to use each, when questioned, Annotator 2 was able to accurately explain when the spatial specifier tag was used and claimed to find it easy to understand. Confusion in the tagging process was sometimes caused by considerations of grammar, rather than meaning.

In this chapter, we have described a semantic annotation scheme that is designed to be both useful and practical, and the methodology used to create an annotated data set. We analysed and presented some of the challenges encountered in the process, and the fundamental difficulties resulting from ambiguity and individual discrepancies in the use of spatial language that make it difficult to define a single, reliable annotated data set at a semantic level. In future work, we intend to do more analysis and test different annotation strategies like single tag per annotator, to see if there is any improvement in the results achieved.

Chapter 7 - Conclusion

7.1 Thesis overview

This chapter provides a summary and conclusion of the thesis. Each chapter of the thesis has addressed a specific problem: measuring the semantic similarity between geospatial prepositions; studying the impact of contextual factors on the interpretation of a geospatial prepositions; predicting the distance between locations in a spatial scene and exploring the consistency of human annotations of spatial language. Together, these chapters contribute to the broader goals of advancing the understanding of geospatial prepositions and automating their interpretation.

In the next section, we provide an overview of the research reported in this thesis, and as a reminder, we include an extended version of the table of the objectives, research questions and corresponding manuscripts from the first chapter (Table 7.1), with the research contributions added. We then describe future directions that arise from this research.

7.2 Research overview

As discussed earlier in the introduction, interpretation of geospatial language is essential due to hazards and emergency situations that might happen in everyday life. In an emergency, people do not use formal language that provide exact street addresses to describe their locations. Instead, they use landmarks, known places or prominent geographic features to refer to their locations. They construct relative location descriptions using language elements including the locatum, spatial relation term (such as spatial prepositions, or location or movement verbs), the relatum and other spatial words discussed in previous works (Talmy 1983).

In this thesis, our focus was on spatial relation terms in the form of prepositions. In each chapter, some aspect of the broader goals of understanding and interpretating geospatial prepositions are addressed in order to answer four research questions. These research questions lead on to the future research directions that we describe later in this chapter.

In ***Chapter 2*** , the semantic similarity of 24 geospatial prepositions, and the senses of three groups of prepositions (13 of them) were explored using different methods such as Venn diagrams, extensional maps, t-SNE plots and agglomerative hierarchical clustering. Although there are existing tools such as *WordNet* (Miller 1995) that measure the similarity among different words*,* they are mostly unable to measure the similarity among prepositions and perform poorly at this task. The focus of this work was to provide measures of the semantic similarity of geospatial prepositions as one of the main elements in spatial language descriptions, to find groups of semantically similar prepositions, and to extract the senses for a subset of these, using a human subjects experiment. This study confirmed some of the previous findings on the similarity of some of the geospatial prepositions (Talmy 1983; Herskovits 1985; Tyler and Evans 2003), such as *beside, next to, near* and *adjacent to,* which are considered proximity/adjacency prepositions and the similar senses of geospatial prepositions such as *over, through* and *across* (Cooper 1968; Brugman and Lakoff 1988; Mackenzie 1992; Tyler and Evans 2003; Lakoff 2008; Kreitzer 1997).

The findings on the senses of geospatial prepositions show that some geospatial prepositions share similar senses through different expressions. For instance: *off, past* and *by* share a proximity sense that confirms previous work (Cooper 1968; Landau and Jackendoff 1993; Lindstromberg 2010), but those prepositions also have other senses such as an overlapping sense for *off* and *past* and an enclosure sense for *by.* For the first time our work identified these senses for a wider range of prepositions. We found that the preposition *near* is less likely to be used for line-line locata and relata than *close to.* Also, the analysis of senses shows that the *next to* preposition sense mostly defines proximity rather than immediate adjacency or contact of the kind we see for prepositions such as *adjacent to*.

The study of geospatial preposition senses is important because they determine the correct interpretation of a location description, and understanding these variations is important for successful automated interpretation. The identified senses were validated with an inter-annotator agreement score of 86%.

The results of this chapter (the similarity matrix among geospatial prepositions and diagram vectors that describe the semantics of geospatial prepositions) were used in Chapter 5 as inputs for the machine learning technique. Preposition senses were also incorporated into our machine learning method for predicting distances associated with prepositions.

In contrast to the previous chapter, in which we used human subjects experiments to measure semantic similarity, ***Chapter 3*** measures the semantic similarity among geospatial prepositions using text mining methods. Two methods were used to identify the similarity among geospatial prepositions: Bag of Words (BoW) and GloVe embeddings, and then we evaluated the results against data from a human subjects experiment (Stock and Yousaf 2018) using the Pearson product-moment correlation coefficient. In this chapter, the observations were quite different from the human subjects experiment and the similarity among geospatial prepositions using text mining differed from that determined with the experiment. We used two different corpora to perform these methods. The first corpus was purely geospatial (the NCGL (Stock et al. 2013)), and the measurements of similarity in this corpus showed high correlation with the human subject experiment using the Bag of Words method (47% for all spatial prepositions and 76% with the less frequent spatial prepositions). On the other hand, a more general corpus, the British National Corpus (BNC), showed a lower correlation with the human subjects' experiment (29%). This is most likely due to the fact that many prepositions such as *in, on* and *at* have multiple senses or meanings, many of which are not spatial in the general corpus, causing the semantic representations used (BoW and GloVe embeddings) to be too coarse grained to detect the spatial semantics accurately. Our results also showed that prepositions that have fewer non-spatial uses senses (e.g., *beyond, opposite*) correlate better with the human subjects data in both general and geospatial corpora. We also showed that GloVe embeddings similarity measures correlate well with the NCGL corpus (45% for all spatial prepositions and 77% with the less frequent spatial prepositions).

In this chapter we also analysed the parts of speech of words that co-occur with geospatial prepositions (we identified the highest ranked words by tf-idf across all 1000 tf-idf values and extracted the top 30 words that their tf-idf values ranked high for that geospatial preposition). For some geospatial prepositions such as *beside, adjacent to (proximity/adjacency prepositions),* nouns are the most frequent co-occurred elements, while more general prepositions like *on* and *in* co-occurred more with other prepositions.

In addition to the study of geospatial prepositions to identify their acceptance thresholds, in ***Chapter 4*** we studied the role of context on the use of geospatial prepositions. We used web scraping to extract location descriptions using the Google search engine with three sites as relata in the London area (*Buckingham Palace, Trafalgar Square* and *Hyde Park)*. We counted the frequency of expressions that combined a given named locatum and geospatial preposition with each of the relata and calculated the distances between each locatum-relatum pair. We used frequency of use of a preposition with locata at different distances to indicate the range of distances at which use of a given preposition is acceptable. We then used frequency graphs (cumulative and non-cumulative) to examine the acceptance profiles for each preposition-relatum combination, and calculated acceptance thresholds for each preposition. We used this graphic and numerical data to explore the difference between the distances at which preposition are used across all three relata, as well as the impact of contextual factors on the choice of specific preposition. The results show that the preposition *outside* is used only with *Buckingham Palace*, and we consider this to be connected to use of the container image-schema for the palace, together with restricted access by the public. Also, for the biggest relatum of the three, *Hyde Park,* topological prepositions (*in, on, at*) are used more commonly than the other two relata. The graphs also indicate the similarity between some proximity/adjacency prepositions like *beside,* *next to* and *adjacent to*, confirming the results from Chapter 2 and Chapter 3 on the study of the semantic similarity of geospatial prepositions.

After studying the characteristics of the main spatial elements in a geospatial description (like the semantics and similarity of geospatial prepositions and the role of context on the interpretation of a description) in ***Chapter 5*** , we took a step towards the development of automated methods for interpretation of geospatial location descriptions by building a model to predict the distance between locations in a spatial scene. The contribution in this chapter is a method and model that uses machine learning to predict the distance between the locatum and relatum using features including the similarity matrix of geospatial prepositions and diagram vectors of geospatial prepositions obtained from Chapter 2 ; characteristics of the relatum and locatum factors including scale, liquidity or solidity, geometry type, area and elongation; preposition sense diagram vectors and contextual factors regarding the surrounding area such as object density. The complete list of these features and their explanations are defined in Chapter 5 . These features were used as input variables for a regression model. The results of running this model on two different datasets (Data Set 1: 690 spatial expressions in London TQ3080 and Data Set 2: 7364 spatial expressions across all of the UK) were evaluated against three different baselines and achieved results that were up to 25% (for Data Set 1) and 35% (for a subset of Data Set 2) better than the best baseline. For Data Set 1, the mean absolute error (MAE) of distance between locatum and relatum is 25m for the best baseline (that assumes that locatum and relatum coincide), compared with 19m for our model. For Data Set 2, the best baseline has 1170m MAE, compared to 875m for our model. Our analysis of the role of individual features in the model show that preposition diagram vectors, some of the contextual factors such as scale and geometry types and preposition senses have a high impact on the distance prediction between relatum and locatum. To the best of our knowledge, there is no previous work using the wide range of features used in our study to predict the distance for specific geospatial location descriptions across multiple scales. This study confirmed the essential role of the characteristics of geospatial prepositions and contextual factors on the prediction of distance, leading to more accurate automated interpretation of a location descriptions.

Chapter 6 describes an experiment to review the accuracy and consistency of human subjects in annotating spatial language. In this chapter, we reviewed the accuracy of annotation using a human subjects experiment and calculated the inter-annotator agreement between annotators for various spatial elements (locatum, relatum, spatial relation term, location and movement verb). The results show that, for the spatial elements such as *trajector (locatum), landmark (relatum)* and *spatial relation (mostly prepositions)* the identification of spatial elements is more reliable than for other elements like *location and movement verbs* and *special specifiers.* However, despite a training stage, the results show that human annotators are not very consistent with each other for many spatial language elements, and we thus recommend further exploration of the use of human annotators as ground truth (and other alternatives) in spatial natural language processing research.

Throughout this thesis I have explored the semantics of geospatial prepositions, their senses and similarities and showed how important prepositions are on the accurate interpretation of location descriptions. Also, I showed the important role of contextual factors and their characteristics on this interpretation. In Chapter 5, I showed that I can use all this information such as the senses of geospatial prepositions, their similarity, and the characteristics of context to predict the distance between two locations.

By using the context surrounding a simple geospatial description, I was able to predict the distance associated with two locations (non-place names, place names) with a 19m mean absolute error for a dense area such as London. This is a notable improvement on the current models of distance prediction between two locations. This finding can possibly narrow down the search area in location descriptions and make future georeferencing tasks easier. Table 7.1 presents the research questions, objectives, and contributions of each chapter in more details.

Table 7.1. Research questions, contributions, and results

Research question	Objective	Manuscript	Contribution
RQ1: Which geospatial prepositions (or groups of prepositions) and senses are semantically similar?	•To understand the semantic similarity of geospatial prepositions and their senses using a human subjects study	Manuscript 1 (Chapter 2) •Measures the semantic similarity among 24 geospatial prepositions and describes the senses of three groups of them	•Measured and quantified the degree of semantic similarity among 24 geospatial prepositions •Identified groups of semantically similar geospatial prepositions •Identified the nature of the semantic differences and similarities among the prepositions •Identified the senses of 13 geospatial prepositions in 3 groups, and the relations between those senses •Validated the identified senses with 86% inter-annotator agreement •Identified that the preposition *near* is less used for line-line relations than *close to* •Identified that *next to* is mostly used to show proximity rather than adjacency, while *adjacent to* is used more for adjacency relations
	•To understand the semantic similarity of geospatial prepositions using text mining methods •To evaluate how well text mining methods could be used to determine the semantic similarity of geospatial prepositions	Manuscript 2 (Chapter 3) •Measures the semantic similarity among 25 geospatial prepositions using text mining methods including bag of words and word embeddings and compares the results to ground truth data from a human subjects experiment	•Identified the high correlation between the prepositions used in the Geospatial corpus (NCGL) and the human subject experiment (47%) compared to a general corpus like BNC (29%) •Showed that less frequent geospatial prepositions like *opposite* and *beyond* have higher correlation with human subjects data (76%) •Demonstrated that general embeddings such as GloVe correlate well with geospatial corpora prepositions like NCGL for geospatial prepositions (77%) •Identified the co-occurrence of some geospatial prepositions with other parts of speech (proximity and adjacency prepositions mostly co-occur with nouns)
RQ2: Is the role of context important in the interpretation of location descriptions?	•To review the effect of contextual factors and distance between the locatum and relatum and understand the use of specific prepositions within different contexts	Manuscript 3 (Chapter 4) •Identifies the impact of the contextual factors of the relatum on choosing prepositions in a spatial scene.	•Showed that some prepositions (*near, outside, close to)* may be used acceptably at large distances, while *beside*, *next to, adjacent to, at* and *on* are applied mostly at very small distances •Demonstrated the importance of some specific contextual factors (e.g., image schema, accessibility of relatum) on the usage and distance at which specific geospatial prepositions can be acceptably used •Showed that a preposition like *near* is less commonly used when locatum and relatum are very close for two sites

RQ3: Is the object distance associated with geospatial prepositions across a range of spatial scenes and scales accurately predictable using machine learning methods?	•To predict the distance between the locatum and relatum using contextual information and extract the most important factors that influence this prediction	Manuscript 4 (Chapter 5) •Predicts the distance between the locatum and relatum in a given expression using contextual factors, word embeddings of geographic features, the semantic similarity of geospatial prepositions (Chapter 3) and environmental factors	•Developed a method to successfully predict the distances associated with geospatial prepositions in different contexts with 25% and 35% percent improvement over the best baseline for the first and a subset of second datasets respectively •Included contextual and preposition sense data in predictive models of geospatial preposition interpretation for the first time, including: ○ Spatial prepositions similarity matrix ○ Spatial prepositions sense diagram vectors ○ Contextual factors like scale, geometry type and liquidity solidity of the locatum and relatum, and object density •Demonstrated the importance of preposition diagram vectors, their senses and relatum and locatum geometry type and scale to accurately predict the distance associated with geospatial prepositions
RQ4: Is human annotation a reliable form of annotation for the further analysis on location descriptions?	•To understand the reliability of human subjects in annotating geospatial language	Manuscript 5 (Chapter 6) Creates an annotated dataset using six different spatial language elements, as marked by specific labels that use an annotation scheme) and evaluates the degree to which annotators can consistently identify the elements	•Evaluated the role and importance of human annotators on the annotation of spatial elements •Demonstrated that human annotators were able to annotate locatum, relatum and spatial relations more reliably than location and movement verbs, spatial qualifiers and spatial specifiers.

7.3 Future research directions

The research described in this thesis suggests a number of possible avenues for future research, as described in this section.

In this work we considered 24 geospatial prepositions, focussing on topological, proximity and adjacency prepositions. In future work, this set of prepositions could be expanded to include directional prepositions. This would require additional features to be added into the predictive model described in Chapter 5 , in order to capture aspects of direction and orientation. Furthermore, spatial relations that are described with non-prepositional language elements (location and movement verbs) are common in location descriptions (e.g., push, locate, ride, follow, stand, sit), but previous research has only addressed these in limited ways. An application of some of the work in this thesis to non-prepositional spatial relation terms provides substantial potential to further the goal of automation of natural language processing of location descriptions.

Another possible direction is using three dimensional diagrams in the human subjects experiment (Chapter 2) to better understand the nature of some of the prepositions such as *above* that have vertical component. The diagrams we used in this thesis are represented in a 2-dimensional space that focusses on the position of the locatum relative to the relatum. However, the world has three dimensions and some of the senses belong to a preposition might be captured using 3-D visualisations.

As an addition to our work, a semantic network could be created to show the nature of the similarity such as (hyponym, hypernym and synonymity) between geospatial prepositions. For example, can we say that *near* is more general than *close to?* or can we say every *close to* is *near,* but not every *near* is *close to?* In this work we extracted the senses of 13 geospatial prepositions, and a more detailed network at the sense level would further elucidate the understanding of prepositions and provide a useful tool for automated georeferencing (similar to the approach demonstrated in Chapter 5).

We also consider the inclusion of additional features into our predictive model would be useful. For example, density field maps could be useful to add direction for directional spatial relations (*north of, south of)*. Density field maps created by previous works in a grid space represent visualisations of geospatial preposition acceptance thresholds. A variable (feature) that can be extracted from these maps is the area that they cover in a grid space as well as their direction, and these can be added as features to a machine learning model for the georeferencing task.

Extension of our predictive model (Chapter 5) to obtain more spatial expressions from different sources and run neural network models in addition to regression methods to predict the distance is another useful future step. However, collecting more spatial data is not an easy task and requires lots of pre-processing and manual annotation to create a suitable training set. We did not used a neural model (Huang and Carley 2017) in our research because we considered both datasets in Chapter 5 did not reach the amount of data that is needed for deep learning tasks, particularly given the large number of variables included in the model. This issue might however be re-visited with regard to the use of transfer learning methods such as the BERT-based transformer models.

We would also like to extend this work by developing more complex predictive models that people might use in specific situations such as emergencies, or that incorporate other dialects of English such as American English speakers, New Zealand English speakers, or some countries that English is not their first language, but they can speak English.

Finally, we would like to expand the work in Chapter 4 to evaluate the impact of contextual factors on the use of geospatial prepositions by considering a wider range of relata with more variations in characteristics (particularly different feature types). Although most of the previous works have considered the impact of a single context on the usage of prepositions, this work studied three sites, but more sites are needed to generalise the findings more broadly.

7.4 Concluding summary

This thesis has presented five chapters that advance understanding of geospatial prepositions and that demonstrate the importance of a range of factors in accurate interpretation of natural language location descriptions. The semantic similarity and senses of geospatial prepositions were examined in Chapter 2 and Chapter 5 . Distance prediction using semantic similarity between geospatial prepositions and a wide range of contextual and other features were discussed in Chapter 5 and the importance of contextual factors was demonstrated in Chapter 4 and Chapter 5 . Chapter 6 evaluated the consistency and reliability of human annotators in annotating the spatial elements that are used as inputs for experiments on spatial language and distance prediction in this research field.

As has been stated in the introduction, there is a need to automate the processing of geospatial location descriptions for situations that require fast response (e.g., emergencies, disaster response etc). This thesis presented a series of steps that led to increased understanding of the spatial language elements involved in location descriptions (especially geospatial prepositions) and the development of methods for predicting distances associated with location descriptions, using a range of available contextual information.

References

Abdiansah, A., & Wardoyo, R. (2015). Time complexity analysis of support vector machines (SVM) in LibSVM. *International Journal Computer and Application*, 128(3), 28–34.

Acheson, E., De Sabbata, S., & Purves, R. S. (2017). A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. *Computers, Environment and Urban Systems*, 64, 309-320.

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S. H., & Liu, B. (2011, April). Predicting flu trends using twitter data. In *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 702-707). IEEE.

Aflaki, N., Jones, C., & Stock, K. (2019, September 18). Mining the Semantic Similarity of Spatial Relations from Text. *Geocomputation 2019 - Adventures in GeoComputation, Queenstown, NZ*.

Aflaki, N., Russell, S., & Stock, K. (2018). Challenges in Creating an Annotated Set of Geospatial Natural Language Descriptions (Short Paper). *10th International Conference on Geographic Information Science (GIScience 2018).*

Agirre E, Alfonseca E, Hall K, Kravalova J, Paşca M, Soroa A (2009) A study on similarity and relatedness using distributional and WordNet-based approaches. In: *Proceedings of Human language technologies: the 2009 annual conference of the North American chapter of the association for computational linguistics. ACL*, pp 19–27

Alex, B., Byrne, K., Grover, C., & Tobin, R. (2015). Adapting the Edinburgh geoparser for historical georeferencing. *International Journal of Humanities and Arts Computing*, 9(1), 15-35.

Al-Olimat, H. S., Shalin, V. L., Thirunarayan, K., & Sain, J. P. (2019, November). Towards geocoding spatial expressions (vision paper). In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 75-78).

Andrea Rodriguez, M., & Egenhofer, M. J. (2004). Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure. *International Journal of Geographical Information Science*, *18*(3), 229-256.

Attia, M., S. Maharjan, Y. Samih, L. Kallmeyer, and T. Solorio (2016). CogALex-V Shared Task: GHHH - Detecting Semantic Relations via Word Embeddings. *In Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex-V)* Pp. 86–91.

Au, L. (2010). Semantic network methods to disambiguate natural language meaning. U.S. Patent No. 7,711,672. Washington, DC: U.S. Patent and Trademark Office.

Ballatore, A., Bertolotto, M., & Wilson, D. C. (2014). An evaluative baseline for geo-semantic relatedness and similarity. *GeoInformatica*, 18(4), 747-767.

Bateman, J. A., Hois, J., Ross, R., & Tenbrink, T. (2010). A linguistic ontology of space for natural language processing. *Artificial Intelligence*, 174(14), 1027-1071.

Beaman, R., & Conn, B. (2003). Automated geoparsing and georeferencing of Malesian collection locality data. *Telopea*, 10(1), 43-52.

Bennett, D. C. (1972). Some observations concerning the locative-directional distinction. *Semiotica*, 5(1), 58-88.

Bennett, D. C. (1975). Spatial and temporal uses of English prepositions. *An Essay in Stratificational Semantics,* London, England: Longman Group.

Bisk, Y., Shih, K. J., Choi, Y., & Marcu, D. (2018, April). Learning interpretable spatial operations in a rich 3d blocks world. In *Thirty-Second AAAI Conference on Artificial Intelligence.*

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing*, 14(3), 191-205.

Bitters, B. (2009). Spatial relationship networks: Network theory applied to GIS data. *Cartography and Geographic Information Science*, 36(1), 81-93.

Bittner, T., Donnelly, M., & Winter, S. (2005). Ontology and semantic interoperability. In *Large-scale 3D Data Integration (pp. 139-160). CRC Press.*

Blaylock, N., Swain, B., & Allen, J. (2009). Mining geospatial path data from natural language descriptions. Proc. *ACM SIGSPATIAL GIS International Workshop on Querying and Mining Uncertain Spatio-Temporal Data*.

Brooks, S. (2018). Effects of Image Schema-Based Instruction on the Acquisition of Prepositions. *Studies in Applied Linguistics*, 29.

Brugman, C., & Lakoff, G. (1988). Cognitive topology and lexical networks. In *Lexical Ambiguity Resolution* (pp. 477-508). Morgan Kaufmann.

Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1), 13-47.

Cannesson, E., & Saint-Dizier, P. (2002, July). Defining and representing preposition senses: A preliminary analysis. In *Proceedings of the ACL-02 workshop on Word Sense Disambiguation: Recent Successes and Future Directions* (pp. 25-31).

Cardoso, A. B., Martins, B., & Estima, J. (2019, September). Using Recurrent Neural Networks for Toponym Resolution in Text. In *EPIA Conference on Artificial Intelligence* (pp. 769-780). Springer, Cham.

Carlson, L. A., & Covey, E. S. (2005). How far is near? Inferring distance from spatial descriptions. *Language and Cognitive Processes*, 20(5), 617-631.

Chaffin, R., & Herrmann, D. J. (1984). The similarity and diversity of semantic relations. *Memory & Cognition*, 12(2), 134-141.

Chang, A., Savva, M., & Manning, C. D. (2014, October). Learning spatial knowledge for text to 3D scene generation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2028-2038).

Chatterjee, A. (2008). The neural organization of spatial thought and language. In *Seminars in Speech and Language* (Vol. 29, No. 03, pp. 226-238). © Thieme Medical Publishers.

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

Chen, T., Hui, E. C., Wu, J., Lang, W., & Li, X. (2019). Identifying urban spatial structure and urban vibrancy in highly dense cities using georeferenced social media data. *Habitat International*, 89, 102005.

Chodorow, M., Gamon, M., & Tetreault, J. (2010). The utility of article and preposition error correction systems for English language learners: Feedback and assessment. *Language Testing*, 27(3), 419-436.

Clementini, E., Sharma, J., & Egenhofer, M. J. (1994). Modelling topological spatial relations: Strategies for query processing. *Computers and Graphics*, 18(6), 815-822.

Clements, D. H., & Battista, M. T. (1992). Geometry and spatial reasoning. *Handbook of Research on Mathematics Teaching and Learning*, 420-464.

Cohn, A. G., Bennett, B., Gooday, J., & Gotts, N. M. (1997). Qualitative spatial representation and reasoning with the region connection calculus. *GeoInformatica*, 1(3), 275–316.

Collell, G., Van Gool, L., & Moens, M.-F. (2018). Acquiring common sense spatial knowledge through implicit spatial templates. *Thirty-second AAAI Conference on Artificial Intelligence*.

Cooper, G. S. (1968). *A Semantic Analysts of English Locative Prepositions* (Report No.1587). Cambridge, MA: Bolt Beranek and Newman.

Coventry, K. R. (1999). Function, geometry and spatial prepositions: Three experiments. *Spatial Cognition and Computation*, 1(2), 145-154.

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

Coventry, K. R., Carmichael, R., & Garrod, S. C. (1994). Spatial prepositions, object-specific function, and task requirements. *Journal of Semantics*, 11(4), 289–309.

Coventry, K. R., Tenbrink, T., & Bateman, J. (2009). *Spatial Language and Dialogue* (Vol. 3). OUP Oxford.

Cresswell, M. J. (1978). Prepositions and points of view. *Linguistics and Philosophy*, 2(1), 1-41.

Cuyckens, H. A. (1991). The semantics of spatial prepositions in Dutch: A cognitive-linguistic exercise (Doctoral dissertation, Universitaire Instelling Antwerpen (Belgium)).

Dahlmeier, D., Ng, H. T., & Schultz, T. (2009, August). Joint learning of preposition senses and semantic roles of prepositional phrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 450-458).

Danziger, E. (2010). Deixis, gesture, and cognition in spatial frame of reference typology. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 34(1), 167-185.

De Felice, R., & Pulman, S. (2008, August). A classifier-based approach to preposition and determiner error correction in L2 English. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)* (pp. 169-176).

Delboni, T. M., Borges, K. A., Laender, A. H., & Davis Jr, C. A. (2007). Semantic expansion of geographic web queries based on natural language positioning expressions. *Transactions in GIS*, 11(3), 377–397.

Dey, N., & Santhi, V. (Eds.). (2017). *Intelligent techniques in signal processing for multimedia security*. Springer International Publishing.

Dirven, René (1993) . Dividing up physical and mental space into conceptual categories by means of English prepositions. In *The semantics of Prepositions* (pp. 73-98). De Gruyter Mouton.

Doherty, P., Guo, Q., Liu, Y., Wieczorek, J., & Doke, J. (2011). Georeferencing incidents from locality descriptions and its applications: a case study from Yosemite National Park search and rescue. *Transactions in GIS*, 15(6), 775-793.

Du, S., Wang, X., Feng, C. C., & Zhang, X. (2017). Classifying natural-language spatial relation terms with random forest algorithm. *International Journal of Geographical Information Science*, 31(3), 542-568.

Egenhofer, M. J., & Franzosa, R. D. (1991). Point-set topological spatial relations. *International Journal of Geographical Information System*, 5(2), 161–174.

Egenhofer, M. J., & Shariff, A. R. B. (1998). Metric details for natural-language spatial relations. *ACM Transactions on Information Systems* (TOIS), 16(4), 295–321.

Emmorey, K., Kosslyn, S. M., & Bellugi, U. (1993). Visual imagery and visual-spatial language: Enhanced imagery abilities in deaf and hearing ASL signers. *Cognition*, 46(2), 139–181.

Epstein, R. A., Higgins, J. S., Jablonski, K., & Feiler, A. M. (2007). Visual scene processing in familiar and unfamiliar environments. *Journal of Neurophysiology*, 97(5), 3670-3683.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, No. 34, pp. 226-231).

Fellbaum, C. (1998). A semantic network of english: the mother of all WordNets. In *EuroWordNet: A Multilingual Database with Lexical Semantic Networks* (pp. 137-148). Springer, Dordrecht.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. Studies in linguistic analysis.

Fisher, P. F., & 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.

Frank, A. U. (1996). Qualitative spatial reasoning: Cardinal directions as an example. *International Journal of Geographical Information Science*, 10(3), 269–290.

Freeman, J. (1975). The modelling of spatial relations. *Computer Graphics and Image processing,* 4(2), 156-171.

Freksa, C. (1991). Qualitative Spatial Reasoning. In *Cognitive and Linguistic Aspects of Geographic Space* (pp. 361-372). eds. D. Mark, A. Frank. Springer, Kluwer, Dordrecht.

Freksa, C. (1992). Using Orientation Information for Qualitative Spatial Reasoning. In *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space* (pp. 162-178). eds. A. Frank, I. Campari, Springer Verlag.

Fu, G., Jones, C. B., & Abdelmoty, A. I. (2005, October). Ontology-based spatial query expansion in information retrieval. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 1466-1482). Springer, Berlin, Heidelberg.

Fu, R., Guo, J., Qin, B., Che, W., Wang, H., & Liu, T. (2014, June). Learning semantic hierarchies via word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1199-1209).

Gahegan, M. (1995, September). Proximity operators for qualitative spatial reasoning. In *International Conference on Spatial Information Theory* (pp. 31-44). Springer, Berlin, Heidelberg.

Gärdenfors, P. (2004). *Conceptual spaces: The Geometry of Thought*. MIT press.

Geeraerts, D., & Cuyckens, H. (Eds.). (2007). Introducing cognitive linguistics. In *The Oxford Handbook of Cognitive Linguistics*.

Goldstone R, Son J (2005) Similarity. In: Holyoak K, Morrison R (eds) *Cambridge Handbook of Thinking and Reasoning*. Cambridge University Press, New York, pp 13–36

Goodman, N. (1972). Seven Strictures on Similarity. In *Problems and Projects*. Bobs-Merril.

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

Hahmann, T., & Gruninger, M. (2011, March). A naïve theory of dimension for qualitative spatial relations. In *2011 AAAI Spring Symposium Series*.

Hahn, U., & Chater, N. (1997). Concepts and similarity. *Knowledge, Concepts and Categories*, 43-92.

Hall, M. M., & Jones, C. B. (2008, November). Quantifying spatial prepositions: an experimental study. In *Proceedings of the 16th ACM SIGSPATIAL, International Conference on Advances in Geographic Information Systems* (pp. 1-4).

Hall, M. M., & Jones, C. B. (2021). Generating geographical location descriptions with spatial templates: A salient toponym driven approach. *International Journal of Geographical Information Science*, 1–32.

Hall, M. M., Jones, C. B., & Smart, P. (2015, October). Spatial natural language generation for location description in photo captions. In *International Conference on Spatial Information Theory* (pp. 196-223). Springer, Cham.

Hall, M. M., Smart, P. D., & Jones, C. B. (2011). Interpreting spatial language in image captions. *Cognitive Processing*, 12(1), 67-94.

Hart, G., Temple, S., & Mizen, H. (2004). Tales of the river bank, first thoughts in the development of a topographic ontology. In *7th Conference on Geographic Information Science (AGILE 2004), Heraklion, Greece* (pp. 169-178).

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. series c (Applied Statistics),* 28(1), 100-108.

Hayward, W. G., & Tarr, M. J. (1995). Spatial language and spatial representation. *Cognition*, 55(1), 39–84.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications*, 13(4), 18–28.

Hernández, D. (1991). Relative representation of spatial knowledge: The 2-D case. In *Cognitive and Linguistic Aspects of Geographic Space* (pp. 373-385). Springer, Dordrecht.

Herskovits, A. (1980, June). On the spatial uses of prepositions. In *18th Annual Meeting of the Association for Computational Linguistics* (pp. 1-5).

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

Herskovits, A. (1986). *Language and spatial cognition: An interdisciplinary study of the prepositions in English*. London: Cambridge University Press.

Hill, A.W., Guralnick, R., Flemons, P., Beaman, R., Wieczorek, J., Ranipeta, A., Chavan, V. and Remsen, D. (2009). Location, location, location: utilizing pipelines and services to more effectively georeference the world's biodiversity data. *BMC Bioinformatics*, 10(14), 1-9.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1988). *Applied Statistics for the Behavioral Sciences*, chapter 13. Houghton Mifflin.

Hirtle, W. H. (1985). Linguistics and the dimensions of language: An overview of Guillaume’s theory. *Lingua*, 67(1), 65–83.

Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E. G., & Milios, E. (2006). Information retrieval by semantic similarity. *International Journal on Semantic Web and Information Systems* (IJSWIS), 2(3), 55-73.

Hois, J., & Kutz, O. (2008, September). Natural language meets spatial calculi. In *International Conference on Spatial Cognition* (pp. 266-282). Springer, Berlin, Heidelberg.

Hois, J., Tenbrink, T., Ross, R., & Bateman, J. (2009). *The Generalized Upper Model spatial extension: a linguistically-motivated ontology for the semantics of spatial language.* SFB(Vol. 3). TR8 internal report, Collaborative Research Center for Spatial Cognition, University of Bremen, Germany.

Hu, Y., & Wang, J. (2020). How do people describe locations during a natural disaster: an analysis of tweets from Hurricane Harvey. *arXiv preprint* arXiv:2009.12914.

Huang, B., & Carley, K. M. (2017, July). On predicting geolocation of tweets using convolutional neural networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation* (pp. 281-291). Springer, Cham.

Jackson Jr, P. C. (2019). I do believe in word senses. *Proceedings ACS*, 321, 340.

Janowicz K, Raubal M (2007) Affordance-based similarity measurement for entity types. In *International Conference on Spatial Information Theory* (pp. 133-151). Springer, Berlin, Heidelberg.

Janowicz, K., Raubal, M., & Kuhn, W. (2011). The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science*, 2011(2), 29-57.

Jelinek, A. J. (1962). Use of the cumulative graph in temporal ordering. *American Antiquity*, 28(2), 241-243.

Johnson, M. (1987). *The body in the mind: The bodily basis of meaning, imagination, and reason*. Chicago, IL: University of Chicago Press.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3), 241-254.

Jones, C. B., Purves, R. S., Clough, P. D., & Joho, H. (2008). Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science*, 22(10), 1045–1065.

Joseph, K., Tan, C. H., & Carley, K. M. (2012, September). Beyond" local"," categories" and" friends" clustering foursquare users with latent" topics". In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (pp. 919-926).

Kågebäck, M., Johansson, F., Johansson, R., & Dubhashi, D. (2015, June). Neural context embeddings for automatic discovery of word senses. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing* (pp. 25-32).

Kamalloo, E., & Rafiei, D. (2018, April). A coherent unsupervised model for toponym resolution. In *Proceedings of the 2018 World Wide Web Conference* (pp. 1287-1296).

Karimzadeh, M. (2016, October). Performance evaluation measures for toponym resolution. In *Proceedings of the 10th Workshop on Geographic Information retrieval* (pp. 1-2).

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

Kelm, P., Murdock, V., Schmiedeke, S., Schockaert, S., Serdyukov, P., & Van Laere, O. (2013). Georeferencing in social networks. In *Social Media Retrieval* (pp. 115-141). Springer, London.

Kemmerer, D. (2005). The spatial and temporal meanings of English prepositions can be independently impaired. *Neuropsychologia*, 43(5), 797–806.

Kemmerer, D. (2006). The semantics of space: integrating linguistic typology and cognitive neuroscience. *Neuropsychologia*, 44, 1607-1621.

Kennington, C. (2012). Markov Logic Networks for Spatial Language in Reference Resolution. *ESSLLI 2012 Student Session*, 54.

Kew, T., Shaitarova, A., Meraner, I., Goldzycher, J., Clematide, S., & Volk, M. (2019, September). Geotagging a Diachronic Corpus of Alpine Texts: Comparing Distinct Approaches to Toponym Recognition. In *Proceedings of the Workshop on Language Technology for Digital Historical Archives* (pp. 11-18).

Khan, A., Vasardani, M., and Winter, S. (2013). Extracting spatial information from place descriptions. In: S. Scheider, et al., eds. *Proceedings of the First ACM SIGSPATIAL International Workshop on Computational Models of Place*, 5 November, Orlando, FL. New York, NY: ACM, 62–69.

Kilgarriff, A. (1997). I don’t believe in word senses. *Computers and the Humanities*, 31(2), 91-113.

Kim, J., Vasardani, M., & Winter, S. (2016). From descriptions to depictions: A dynamic sketch map drawing strategy. *Spatial Cognition & Computation*, 16(1), 29–53.

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Machine Learning Proceedings* 1992 (pp. 249-256). Morgan Kaufmann.

Klien, E., & Lutz, M. (2005, September). The role of spatial relations in automating the semantic annotation of geodata. In *International Conference on Spatial Information Theory* (pp. 133-148). Springer, Berlin, Heidelberg.

Klippel, A., Xu, S., Li, R., & Yang, J. (2011, July). Spatial event language across domains. In *Workshop on Computational Models for Spatial Language Interpretation and Generation, CoSLI-2*.

Kononenko, I. (1994, April). Estimating attributes: Analysis and extensions of RELIEF. In *European Conference on Machine Learning* (pp. 171-182). Springer, Berlin, Heidelberg.

Kordjamshidi, P., Van Otterlo, M., & Moens, M.-F. (2011). Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing* (TSLP), 8(3), 1–36.

Kracht, M. (2008). The fine structure of spatial expressions. *Syntax and Semantics of Spatial P*, 120, 35.

Kreitzer, A. (1997). Multiple levels of schematization: A study in the conceptualization of space. *Cognitive Linguistics*, 8(4), 291-326.

Kunze, L., Doreswamy, K. K., & Hawes, N. (2014, May). Using qualitative spatial relations for indirect object search. In *2014 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 163-168). IEEE.

Lakoff, G. (2008). *Women, fire, and dangerous things: What categories reveal about the mind.* University of Chicago Press, Chicago.

Lakoff, G., & Johnson, M. (1980). Conceptual metaphor in everyday language. *The Journal of Philosophy*, 77(8), 453–486.

Lakoff, G., & Johnson, M. (1980). The metaphorical structure of the human conceptual system. *Cognitive Science*, 4(2), 195–208.

Lan, T., Yang, W., Wang, Y., & Mori, G. (2012, October). Image retrieval with structured object queries using latent ranking svm. In *European Conference on Computer Vision* (pp. 129-142). Springer, Berlin, Heidelberg.

Landau, B., & Jackendoff, R. (1993). “What” and “where” in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2), 217–238.

Landau, B., & Jackendoff, R. (1993). Whence and whither in spatial language and spatial cognition? *Behavioral and Brain Sciences*, 16(2), 255–265.

Landwehr, P. M., & Carley, K. M. (2014). Social media in disaster relief. In *Data Mining and Knowledge Discovery for Big Data* (pp. 225-257). Springer, Berlin, Heidelberg.

Langacker, R. (1987). *Foundations of Cognitive Grammar. Volume 1: Theoretical Prerequisites*. Stanford University Press, Stanford, California.

Langacker, R. (2008). *Cognitive grammar: A basic introduction*. Oxford: Oxford University Press.

Lautenschütz, A. K., Davies, C., Raubal, M., Schwering, A., & Pederson, E. (2006, September). The influence of scale, context and spatial preposition in linguistic topology. In *International Conference on Spatial Cognition* (pp. 439-452). Springer, Berlin, Heidelberg.

Leech, G. N. (1970). *Towards a Semantic Description of English*. Indiana Studies in the History and Theory of Linguistics. Indiana University Press, Bloomington, Indiana.

Lehmann, C. (1983). Latin preverbs and cases. In *Latin Linguistics and Linguistic Theory*, 145-161.

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

Levinson, S. C.(2003). *Space in language and cognition: Explorations in cognitive diversity* (Issue 5). Cambridge University Press.

Levinson, S. C., & Meira, S. (2003). “Natural concepts” in the spatial topological domain—adpositional meanings in crosslinguistic perspective: an exercise in semantic typology. *Language,* 79(3), 485-516.

Lieberman, M. D., & Samet, H. (2012, August). Adaptive context features for toponym resolution in streaming news. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 731-740).

Lim, K. H., Karunasekera, S., Harwood, A., & Falzon, L. (2017, December). Spatial-based topic modelling using wikidata knowledge base. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 4786-4788). IEEE.

Lindstromberg, S. (2010). *English prepositions explained*. John Benjamins Publishing.

Litkowski, K. C., & Hargraves, O. (2005, April). The preposition project. In *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications* (pp. 171-179).

Liu, Y., Guo, Q. H., Wieczorek, J., & Goodchild, M. F. (2009). Positioning localities based on spatial assertions. *International Journal of Geographical Information Science*, 23(11), 1471–1501.

Logan, G. D., & Sadler, D, (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and Space* (pp. 493-529). Cambridge, MA: MIT Press.

Long, X., Jin, L., & Joshi, J. (2012, September). Exploring trajectory-driven local geographic topics in foursquare. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (pp. 927-934).

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research,* 9(Nov), 2579-2605.

Mackenzie, J. L. (1992). English spatial prepositions in Functional Grammar. *Working Papers in Functional Grammar 46.*

Mackenzie, J. L. (2003). One sense for beyond? Not beyond us. *Belgian Journal of English Language and Literatures*, 1(New Series), 7-16.

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

Manguinhas, H., Martins, B., & Borbinha, J. (2008, November). A geo-temporal web gazetteer integrating data from multiple sources. In *2008 Third International Conference on Digital Information Management* (pp. 146-153). IEEE.

Mani, I., Doran, C., Harris, D., Hitzeman, J., Quimby, R., Richer, J., Wellner, B., Mardis, S. and Clancy, S. (2010). SpatialML: annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, 44(3), 263-280.

Marchi Fagundes, C. K., Stock, K., & Delazari, L. S. (2021). A cross-linguistic study of spatial location descriptions in New Zealand English and Brazilian Portuguese natural language. *Transactions in GIS*.

Mark, D. M. (1989, November). Cognitive image-schemata for geographic information: Relations to user views and GIS interfaces. In *GIS/LIS* (Vol. 89, No. 2, pp. 551-560).

Mark, D. M., & Egenhofer, M. J. (1994). Modeling spatial relations between lines and regions: combining formal mathematical models and human subjects testing. *Cartography and Geographic Information Systems*, 21(4), 195-212.

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

Mark, D. M., Comas, D., Egenhofer, M. J., Freundschuh, S. M., Gould, M. D., & Nunes, J. (1995, September). Evaluating and refining computational models of spatial relations through cross-linguistic human-subjects testing. In *International Conference on Spatial Information Theory* (pp. 553-568). Springer, Berlin, Heidelberg.

Mark, D. M., & Egenhofer, M. J. (1994, September). Calibrating the meanings of spatial predicates from natural language: Line-region relations. In *Proceedings, Spatial Data Handling 1994* (Vol. 1, pp. 538-553).

Mark, D. M., & Egenhofer, M. J. (1994). Modeling spatial relations between lines and regions: combining formal mathematical models and human subjects testing. *Cartography and Geographic Information Systems*, *21*(4), 195-212.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44(1), 1-23.

Matsuo, Y., Tomobe, H., & Nishimura, T. (2007, July). Robust estimation of google counts for social network extraction. In *AAAI* (Vol. 7, pp. 1395-1401).

Mikolov, T., Yih, W. T., & Zweig, G. (2013, June). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746-751).

Miller G, Charles W (1991) Contextual correlates of semantic similarity*. Language and Cognitive Processes*, 6(1), 1-28.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11), 39–41.

Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and perception*. Belknap Press.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4), 235-244.

Mizen, H., Dolbear, C., & Hart, G. (2005, November). Ontology ontogeny: Understanding how an ontology is created and developed. In *International Conference on GeoSpatial Sematics* (pp. 15-29). Springer, Berlin, Heidelberg.

Montello, D. R. (1993, September). Scale and multiple psychologies of space. In *European Conference on Spatial Information Theory* (pp. 312-321). Springer, Berlin, Heidelberg.

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

Morris, C. (2020). Modelling the Use of Spatial Language. Undergraduate final dissertation (supervised by Chris Jones), Cardiff University.

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

Munnich, E., & Landau, B. (2003). The effects of spatial language on spatial representation: Setting some boundaries. *Language in Mind: Advances in the Study of Language and Thought*, 113–155.

Newstead, S. E., & Coventry, K. R. (2000). The role of context and functionality in the interpretation of quantifiers. *European Journal of Cognitive Psychology*, 12(2), 243-259.

Novel, M., Grütter, R., Boley, H., & Bernstein, A. (2020). Nearness as context-dependent expression: An integrative review of modeling, measurement and contextual properties. *Spatial Cognition & Computation*, 20(3), 161–233.

Olivier, P., & Gapp, K.-P. (1998). *Representation and processing of spatial expressions*. Psychology Press.

Pedersen T, Patwardhan S, Michelizzi J (2004). WordNet:: Similarity-Measuring the Relatedness of Concepts. In *AAAI* (Vol. 4, pp. 25-29).

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532-1543).

Pilehvar, M. T., & Navigli, R. (2015). From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence*, 228, 95–128.

Platonov, G., & Schubert, L. (2018, June). Computational models for spatial prepositions. In *Proceedings of the First International Workshop on Spatial Language Understanding* (pp. 21-30).

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines, in *Advances in Kernel Method: Support Vector Learning*, Scholkopf, Burges, and Smola, Eds. Cambridge, MA: MIT Press. pp. 185–208.

Purves, R.S., Clough, P., Jones, C.B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A.K., Vaid, S. and Yang, B. (2007). The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science*, 21(7), 717-745.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive English grammar*. London and New York: Longman.

Radke, M., Das, P., Stock, K., & Jones, C. B. (2019). Detecting the Geospatialness of Prepositions from Natural Language Text (Short Paper). In *14th International Conference on Spatial Information Theory (COSIT 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Renz, J., Rauh, R., & Knauff, M. (2000). Towards cognitive adequacy of topological spatial relations. In *Spatial Cognition II* (pp. 184-197). Springer, Berlin, Heidelberg.

Retz-Schmidt, G. (1988). Various views on spatial prepositions. *AI Magazine*, 9(2), 95–95.

Rice, S. (1993). Far afield in lexical fields: The English prepositions. In *ESCOL* (Vol. 92, pp. 206-17).

Richardson, D. C., Spivey, M. J., Edelman, S., & Naples, A. J. (2001). ' Language is spatial': Experimental evidence for image schemas of concrete and abstract verbs. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 23, No. 23).

Richardson, L. (2007). Beautiful soup documentation. Dosegljivo: https://www. crummy. com/software/BeautifulSoup/bs4/doc/.[Dostopano: 7. 7. 2018].

Richter, K.-F., Schmid, F., & Laube, P. (2012). Semantic trajectory compression: Representing urban movement in a nutshell. *Journal of Spatial Information Science*, 4, 3–30.

Robnik-Šikonja, M., & Kononenko, I. (1997, July). An adaptation of Relief for attribute estimation in regression. In *Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97)* (Vol. 5, pp. 296-304).

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

Rose-Redwood, R., Alderman, D., & Azaryahu, M. (2010). Geographies of toponymic inscription: New directions in critical place-name studies. *Progress in Human Geography*, 34(4), 453–470.

Rubenstein H, Goodenough J (1965) Contextual correlates of synonymy. *Communications of the ACM*, 8(10), 627-633.

Rupp, C. J., Rayson, P., Baron, A., Donaldson, C., Gregory, I., Hardie, A., & Murrieta-Flores, P. (2013, October). Customising geoparsing and georeferencing for historical texts. In *2013 IEEE International Conference on Big Data* (pp. 59-62). IEEE.

Saint-Dizier, P. (2006). Introduction to the syntax and semantics of prepositions. In *Syntax and Semantics of Prepositions* (pp. 1-25). Springer, Dordrecht.

Sánchez, D., Batet, M., Isern, D., & Valls, A. (2012). Ontology-based semantic similarity: A new feature-based approach. *Expert Systems with Applications*, 39(9), 7718-7728.

Santos, M. Y., & Moreira, A. (2009, July). Conceptual neighborhood graphs for topological spatial relations. In *Proc. of the World Congress on Engineering* (Vol. 1, pp. 12-18).

Schnoebelen, T., & Kuperman, V. (2010). Using Amazon mechanical turk for linguistic research. *Psihologija*, 43(4), 441-464.

Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval* (Vol. 39, pp. 234-265). Cambridge: Cambridge University Press.

Schwering, A. (2007, September). Evaluation of a semantic similarity measure for natural language spatial relations. In *International Conference on Spatial Information Theory* (pp. 116-132). Springer, Berlin, Heidelberg.

Scott, J., Stock, K., Morgan, F., Whitehead, B. and Medyckyj-Scott, D. (2021). Automated georeferencing of Antarctic species. Full paper to be presented at GIScience 2021, September 27-30.

Shariff, A. R. B., Egenhofer, M. J., & Mark, D. M. (1998). Natural-language spatial relations between linear and areal objects: The topology and metric of English-language terms. *International Journal of Geographical Information Science*, 12(3), 215–245.

Shevade, S. K., Keerthi, S. S., Bhattacharyya, C., & Murthy, K. R. K. (2000). Improvements to the SMO algorithm for SVM regression. *IEEE Transactions on Neural Networks*, 11(5), 1188-1193.

Shintani, M., Mori, K., & Ohmori, T. (2016). Image schema-based instruction in English grammar. *Focus on the Learner*, 285–296.

Sithole, G., & Zlatanova, S. (2016). Position, location, place and area: An indoor perspective. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci*, 3(4), 89-96.

Skoumas, G., Pfoser, D., Kyrillidis, A., & Sellis, T. (2016). Location estimation using crowdsourced spatial relations. *ACM Transactions on Spatial Algorithms and Systems (TSAS),* 2(2), 1-23.

Steels, L., & Loetzsch, M. (2006). Perspective alignment in spatial language. *ArXiv Preprint Cs/0605012*.

Stock, K. (2008). Determining semantic similarity of behaviour using natural semantic metalanguage to match user objectives to available web services. *Transactions in GIS*, 12(6), 733-755.

Stock, K. (2014). A geometric configuration ontology to support spatial querying. In: 17th AGILE Conference on Geographic Information Science, 3-6 June, Castellon, Spain.

Stock, K. M., & Delazari, L. S. (2011). Where am I?/Onde Estou? Automated Interpretation of Human Language Descriptions of Current Location.

Stock, K., & Hall, M. (2017, September). The role of context in the interpretation of natural language location descriptions. In *International Conference on Spatial Information Theory* (pp. 245-254). Springer, Cham.

Stock, K., & 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.

Stock, K., Leibovici, D., Delazari, L., & Santos, R. (2015). Discovering order in chaos: Using a heuristic ontology to derive spatio-temporal sequences for cadastral data. *Spatial Cognition & Computation*, 15(2), 115-141.

Stock, K., Pasley, R. C., Gardner, Z., Brindley, P., Morley, J., & Cialone, C. (2013, September). Creating a corpus of geospatial natural language. In *International Conference on Spatial Information Theory* (pp. 279-298). Springer, Cham.

Stock, K., Jones, C. B., Russell, S., Radke, M., Das, P., & Aflaki, N. (2021). Detecting geospatial location descriptions in natural language text. *International Journal of Geographical Information Science*, 1-38.

Takenobu, T., Tomofumi, K., & Suguru, S. (2005). Meaning of Japanese spatial nouns. In Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications (pp. 93-100).

Talmy, L. (1975, September). Figure and ground in complex sentences. In *Annual Meeting of the Berkeley Linguistics Society* (Vol. 1, pp. 419-430).

Talmy, L. (1978). The relation of grammar to cognition--a synopsis. *American Journal of Computational Linguistics*, 16-26.

Talmy, L. (1983). How language structures space. *In Spatial Orientation* (pp. 225-282). Springer, Boston, MA.

Talmy, L. (2000). *Toward a cognitive semantics: Volume I: Concept structuring systems* (pp. 1-565). Cambridge, MA: MIT press.

Taylor, H. A., & Tversky, B. (1996). Perspective in spatial descriptions. *Journal of Memory and Language,* 35(3), 371-391.

Tenbrink, T. (2008). *Space, time, and the use of language*: An investigation of relationships (Vol. 36). Walter de Gruyter.

Tezuka, T., & Tanaka, K. (2005, September). Landmark extraction: A web mining approach. In *International Conference on Spatial Information Theory* (pp. 379-396). Springer, Berlin, Heidelberg.

Tobin, R., Grover, C., Byrne, K., Reid, J., & Walsh, J. (2010, February). Evaluation of georeferencing. In *Proceedings of the 6th Workshop on Geographic Information Retrieval* (pp. 1-8).

Tratz, S., & Hovy, D. (2009, June). Disambiguation of preposition sense using linguistically motivated features. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium* (pp. 96-100).

Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, *32*(3), 379-416.

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

Tyler, A., & Evans, V. (2001). Reconsidering prepositional polysemy networks: The case of over. *Language*, 724-765

Usery, E. L. (2020). A conceptual framework and fuzzy set implementation for geographic features. In *Geographic Objects with Indeterminate Boundaries* (pp. 71–85). CRC Press.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

van Erp, M., Hensel, R., Ceolin, D., & Van der Meij, M. (2015). Georeferencing animal specimen datasets. *Transactions in GIS*, 19(4), 563-581.

Van Laere, O., Schockaert, S., & Dhoedt, B. (2013). Georeferencing Flickr resources based on textual meta-data. *Information Sciences*, 238, 52-74.

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

Vorwerg, C., & Rickheit, G. (1998). Typicality effects in the categorization of spatial relations. In *Spatial Cognition* (pp. 203-222). Springer, Berlin, Heidelberg.

Wang B, Fei T, Kang Y, Li M, Du Q, Han M, et al. (2020) Understanding the spatial dimension of natural language by measuring the spatial semantic similarity of words through a scalable geospatial context window. *PLoS ONE* 15(7): e0236347

Wu, D., & Cui, Y. (2018). Disaster early warning and damage assessment analysis using social media data and geo-location information. *Decision Support Systems*, 111, 48-59.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. In*: Proc. 32nd Annual Meeting of the Association for Computational Linguistics (ACL),* Las Cruces, NM, USA, pp. 133–138.

Yao, X., & Thill, J. (2005). How Far Is Too Far?–A Statistical Approach to Context-contingent Proximity Modeling. *Transactions in GIS*, 9(2), 157–178.

Ying, J. J.-C., Lee, W.-C., Weng, T.-C., & Tseng, V. S. (2011). Semantic trajectory mining for location prediction. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 34-43).

Yu, H., & Siskind, J. M. (2017). Sentence directed video object codiscovery. *International Journal of Computer Vision*, 124(3), 312-334.

Zelinsky-Wibbelt, C. (1990). The semantic representation of spatial configurations: A conceptual motivation for generation in machine translation. In *COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*.

Zelinsky-Wibbelt, C. (Ed.). (1993). *The semantics of prepositions: From mental processing to natural language processing (Vol. 3)*. Berlin: Walter de Gruyter.

Zenasni, S., Kergosien, E., Roche, M., & Teisseire, M. (2015, October). Discovering types of spatial relations with a text mining approach. In *International Symposium on Methodologies for Intelligent Systems* (pp. 442-451). Springer, Cham.

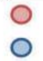
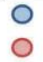
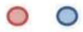
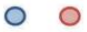
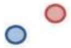
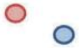
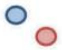
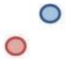
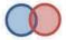
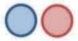
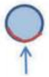
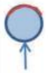
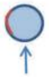
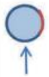
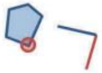
Zwarts, J. (1997). Vectors as Relative Positions: A Compositional Semantics of Modified PPs1. *Journal of Semantics*, 14(1), 57–86.

Zwarts, J., & Winter, Y. (2000). Vector space semantics: A model-theoretic analysis of locative prepositions. *Journal of Logic, Language and Information*, 9(2), 169-211.

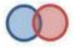
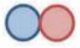
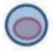
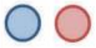
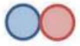
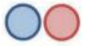
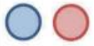
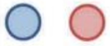
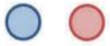
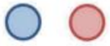
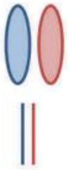
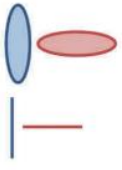
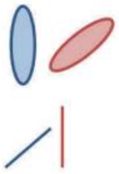
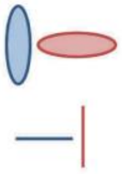
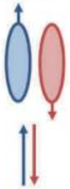
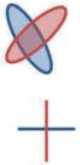
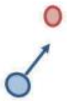
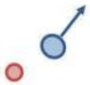
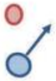
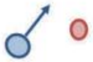
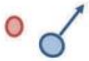
Zwarts, J. (2005). Prepositional aspect and the algebra of paths. *Linguistics and Philosophy*, 28(6), 739-779.

Zwarts, J. (2017). Spatial semantics: Modeling the meaning of prepositions. *Language and Linguistics Compass*, 11(5), e12241.

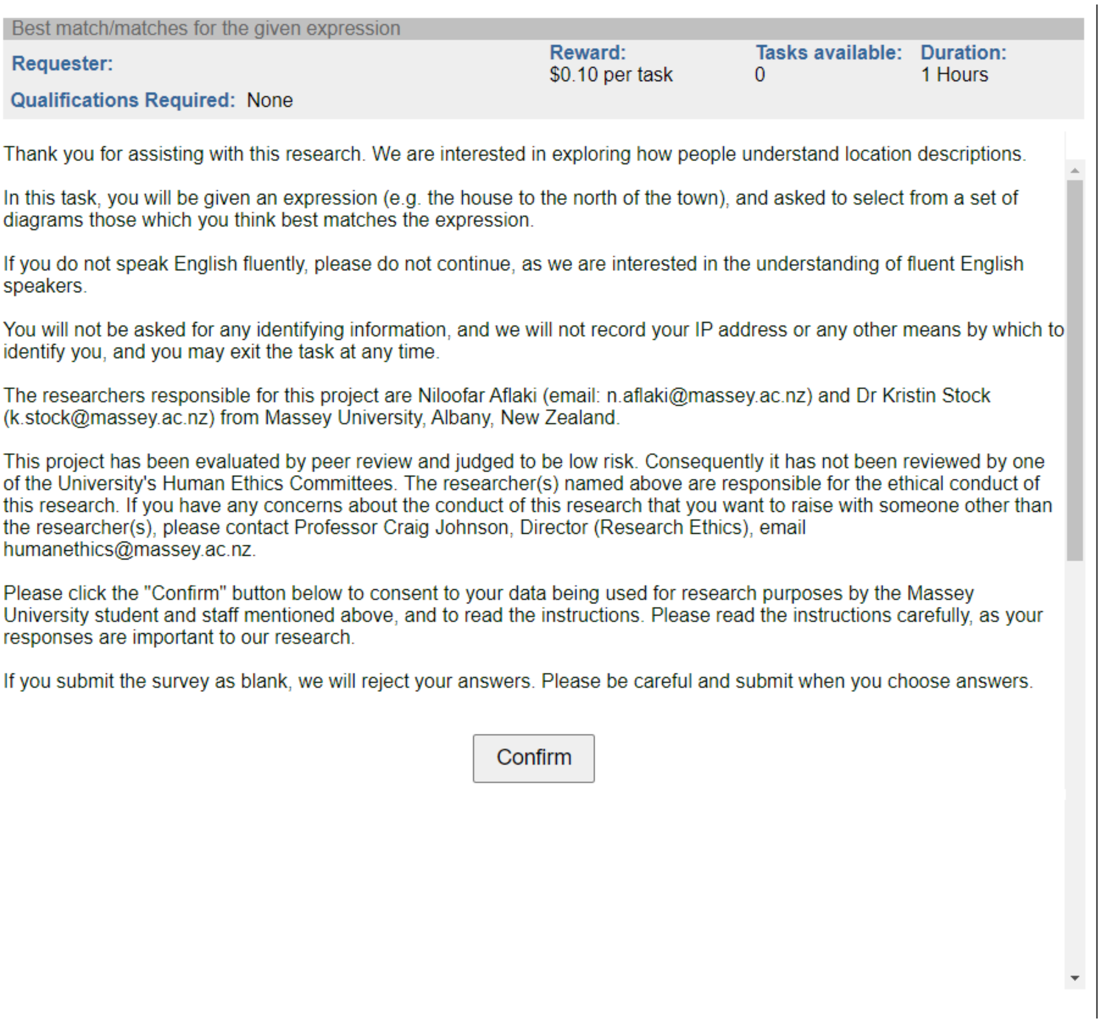
Appendix A - Geometric configurations Stock (2014)

Parameter		Values								
DIRECTION (dr): What is the cardinal direction from one object to the other?	**Label**	north(a,b)	south(a,b)	west(a,b)	east(a,b)	northEast(a,b)	northWest(a,b)	southEast(a,b)	southWest(a,b)	
	Illustration									
	Query (WHERE clause)	MinY[5](ST_Envelope(b)) >= MaxY(ST_Envelope(a)) AND MinX(ST_Envelope(b)) >= MinX(ST_Envelope(a)) AND MaxX(ST_Envelope(b)) <= MaxX(ST_Envelope(a))	MaxY (ST_Envelope(b)) <= MinY(ST_Envelope(a)) AND MinX(ST_Envelope(b)) >= MinX(ST_Envelope(a)) AND MaxX(ST_Envelope(b)) <= MaxX(ST_Envelope(a))	MaxX (ST_Envelope(b)) <= MinX(ST_Envelope(a)) AND MinY(ST_Envelope(b)) >= MinY(ST_Envelope(a)) AND MaxY(ST_Envelope(b)) <= MaxY(ST_Envelope(a))	MinX (ST_Envelope(b)) >= MaxX(ST_Envelope(a)) AND MinY(ST_Envelope(b)) >= MinY(ST_Envelope(a)) AND MaxY(ST_Envelope(b)) <= MaxY(ST_Envelope(a))	MinX (ST_Envelope(b)) >= MaxX(ST_Envelope(a)) AND MinY(ST_Envelope(b)) >= MaxY(ST_Envelope(a))	MaxX (ST_Envelope(b)) <= MinX(ST_Envelope(a)) AND MinY(ST_Envelope(b)) >= MaxY(ST_Envelope(a))	MinX (ST_Envelope(b)) >= MaxX(ST_Envelope(a)) AND MaxY(ST_Envelope(b)) <= MinY(ST_Envelope(a))	MaxX (ST_Envelope(b)) <= MinX(ST_Envelope(a)) AND MaxY(ST_Envelope(b)) <= MinY(ST_Envelope(a))	
	Axioms									
ADJACENCY (a): Are objects adjacent to each other?	**Label**	adjacent(a,b)								
	Illustration									
	Query (WHERE clause)	(ST_DWithin(a,b,σ)) AND ((ST_Touches(a,b) =1) OR (ST_Disjoint(a,b) =1))								
	Axioms	a.adjacent(a,b) ⇒ ds.near								
COLLOCATION (cl): Are objects in the same place?	**Label**	within collocated(a,b)	exactly collocated(a,b)	substantially collocated(a,b)	approximately collocated(a,b)					
	Illustration									
	Query (WHERE clause)	ST_Contains(a,b) = 1	ST_Equals(a,b) = 1	(ST_Overlaps(a,b) = 1) AND (ST_Area(ST_Difference(a,b) > ST_Area(a)/2[6])	ST_DWithin(a,b,σ)					
	Axioms	cl.within collocated (a,b) ≡ t.contain(a,b)	cl.exactly collocated (a,b) ≡ t.equal(a,b)	cl.substantially collocated (a,b) ≡ t.overlap(a,b) cl.exactly collocated (a,b) ⇒ cl.substantially collocated (a,b)	cl.substantially collocated (a,b) cl.exactly collocated (a,b) cl.substantially collocated (a,b) ⇒ cl.approximately collocated (a,b) cl.exactly collocated (a,b) ⇒ cl.approximately collocated (a,b) cl.within collocated (a,b) ⇒ cl.approximately collocated (a,b)					
OBJECT PARTHOOD (op): Which part of the object is of interest?	**Label**	part(a,b)	whole(a,b)	rest(a,b,c)	front(a,b,D)	back(a,b,D)	left side(a,b,D)	right side(a,b,D)	middle (a,b)	corner (a,b, σ)
	Illustration									
	Query (WHERE clause)	ST_Contains(a,b) = 1	ST_Equals(a,b) = 1	ST_Equals(ST_Union(a,b),c) = 1	FrontGeometry[7](a, D) = b	BackGeometry (a, D) = b	LeftSideGeometry (a, D) = b	RightSideGeometry (a, D) = b	ST_Centroid(a) = b	PolygonAngle[8](ST_Intersection(a,b)) < σ AND PolygonAngle(ST_Intersection (a,b)) > 0 (ST_Touches(a,b) = 1) AND ST_Intersection(ST_Boundary(a), ST_Boundary(b) IS NOT NULL AND (ST_Azimuth(a) <> ST_Azimuth(b))

	Axioms	op.part(a,b) ≡ t.contain(a,b)	op.whole(a,b) ≡ t.equal(a,b)	op.rest(a,b,c) ⇒ op.part(a,c) Λ op.part(a,b)
				op.rest(a,b,c) ⇒ ¬t.overlap (a,b)
				op.rest(a,b,c) ⇒ t.touch (a,b)
					op.front(a,b,D) ⇒ ¬ op.back (a,b,D) Λ ¬ op.left side (a,b,D) Λ ¬ op.right side(a,b,D)	op.back(a,b,D) ⇒ ¬ op.front(a,b,D) Λ ¬ op.left side (a,b,D) Λ ¬ op.right side(a,b,D)	op.left side(a,b,D) ⇒ ¬ op.back (a,b,D) Λ ¬ op.front (a,b,D) Λ ¬ op.right side(a,b,D)	op.right side(a,b,D) ⇒ ¬ op.back (a,b,D) Λ ¬ op.front (a,b,D) Λ ¬ op.left side(a,b,D)	op.middle (a,b,D) ⇒ ¬ op.back (a,b,D) Λ ¬ op.front (a,b,D) Λ ¬ op.left side(a,b,D) Λ ¬ op.right side(a,b,D)	op.corner (a,b,D) ⇒ ¬ op.junction (a,b,D)

Parameter		Values								
TOPOLOGY (t): Are the objects connected and how?	**Label**	overlap(a,b)	touch(a,b)	contain(a,b)	disjoint(a,b)	equal(a,b)				
	Illustration									
	Query	ST_Overlaps(a,b) = 1	ST_Touches(a,b) =1	ST_Contains(a,b) = 1	ST_Disjoint(a,b) = 1	ST_Equals(a,b) = 1				
	Axioms									
DISTANCE (ds): How close are the objects to each other?	**Label**	distance 0 all points(a,b)	distance 0 any point(a,b)	very near(a,b)	near(a,b)	neither near nor far(a,b)	far(a,b)	x spatial units apart (a,b,x)	x temporal units apart by travel at y velocity[1](a,b,x,y)	
	Illustration									
	Query (WHERE clause)	ST_Equals(a,b) = 1	ST_Touches(a,b) =1	ST_DWithin(a,b,σ)	ST_DWithin(a,b,2σ)	(ST_Distance(a,b) > 2σ) AND (ST_Distance(a,b) < 4σ)	NOT ST_DWithin(a,b,4σ)	ST_Distance(a,b) = x ST_Distance(ST_Centroid(a), ST_Centroid(b)) = x	ST_Distance(a,b) = xy ST_Distance(ST_Centroid(a), ST_Centroid(b)) = xy	
	Axioms	ds.zeroAllPoints(a,b) ≡ t.equal(a,b)	ds.zeroAnyPoint(a,b) ≡ t.touch(a,b)	ds.veryNear(a,b) ≡ (t.disjoint(a,b) ∨ t.touch(a,b))	ds.near ≡ (t.disjoint(a,b) ∨ t.touch(a,b))	ds.neitherNearNorFar(a,b) ≡ t.disjoint(a,b)	ds.far(a,b) ≡ t.disjoint(a,b)	ds.spatialUnitsApart(a,b) ≡ t.disjoint(a,b)	ds.temporalUnitsApart(a,b) ≡ t.disjoint(a,b)	
LINEAR ORIENTATION(lo): How are linear objects oriented relative to each other?	**Label**	parallel(a,b)	perpendicular(a,b)	diagonal(a,b)	orthogonal(a,b)	antiparallel(a,b)	crossed(a,b)			
	Illustration									
	Query (WHERE clause)	MaxAzimuth[2](ST_Boundary(SSRectangle[3](ST _ConvexHull(a)))) – MaxAzimuth[2] (ST_Boundary(SSRectangle[3] (ST_ConvexHull(b)))) = 0 ST_Azimuth(ST_StartPoint(a), ST_EndPoint(a)) - ST_Azimuth(ST_StartPoint(b), ST_EndPoint(b)) = 0	MaxAzimuth[2] (ST_Boundary(SSRectangle[3] (ST_ConvexHull(a)))) – MaxAzimuth[2] (ST_Boundary(SSRectangle[3] (ST_ConvexHull(b)))) IN (π/2, 3π/2) ST_Azimuth(ST_StartPoint(a), ST_EndPoint(a)) - (ST_Azimuth(ST_StartPoint(b), ST_EndPoint(b)) = IN (π/2, 3π/2)	MaxAzimuth[2] (ST_Boundary(SSRectangle[3] (ST_ConvexHull(a)))) – MaxAzimuth[2] (ST_Boundary(SSRectangle[3] (ST_ConvexHull(b)))) IN (π/4, 3π/4, 5π/4, 7π/4) ST_Azimuth(ST_StartPoint(a), ST_EndPoint(a)) - (ST_Azimuth(ST_StartPoint(b), ST_EndPoint(b)) = IN (π/4, 3π/4, 5π/4, 7π/4)	MaxAzimuth[2] (ST_Boundary(SSRectangle[3] (ST_ConvexHull(a)))) – MaxAzimuth[2] (ST_Boundary(SSRectangle[3] (ST_ConvexHull(b)))) IN (0, π/2, π , 3π/2) ST_Azimuth(ST_StartPoint(a), ST_EndPoint(a)) - (ST_Azimuth(ST_StartPoint(b), ST_EndPoint(b)) = IN (0, π/2, π , 3π/2)	MaxAzimuth[2] (ST_Boundary(SSRectangle[3] (ST_ConvexHull(a)))) – MaxAzimuth[2] (ST_Boundary(SSRectangle[3] (ST_ConvexHull(b)))) = π ST_Azimuth(ST_StartPoint(a), ST_EndPoint(a)) - (ST_Azimuth(ST_StartPoint(b), ST_EndPoint(b)) = π	MaxAzimuth[2] (ST_Boundary(SSRectangle[3] (ST_ConvexHull(a)))) – MaxAzimuth[2] (ST_Boundary(SSRectangle[3] (ST_ConvexHull(b)))) IN (π/2, 3π/2) AND ST_Overlaps(a,b) ST_Azimuth(ST_StartPoint(a), ST_EndPoint(a)) - (ST_Azimuth(ST_StartPoint(b), ST_EndPoint(b)) = IN (π/2, 3π/2) AND ST_Overlaps(a,b)			
	Axioms	lo.parallel(a,b)⇒lo.orthogonal (a,b)	lo.perpendicularl(a,b)⇒ lo.orthogonal(a,b)			lo.antiparallel(a,b)⇒ lo.orthogonal(a,b)	lo.crosses(a,b)⇒t.overlap(a,b)			
HORIZONTAL PROJECTIVE ORIENTATION(hpo): How are objects oriented to each other relative to a projected axis?	**Label**	in front of(a,θ[4],b)	behind(a,θ,b)	left(a,θ,b)	right(a,θ,b)	alongside(a,θ,b)				
	Illustration									
	Query	ST_Angle(ST_Azimuth(a,b), θ) < π/2	ST_Angle(ST_Azimuth(a,b), θ) > π/2	(ST_Azimuth(a,b) < θ) AND (ST_Azimuth(a,b) > θ±2 π)	(ST_Azimuth(a,b) > θ) AND (ST_Azimuth(a,b) < θ±2 π)	ST_Angle(ST_Azimuth(a,b), θ) IN (π, 3π/2)				
	Axioms									

Appendix B - Full instructions for experiment



Requester:	**Reward:** $0.10 per task	**Tasks available:** 0	**Duration:** 1 Hours
Qualifications Required: None			

Thank you for assisting with this research. We are interested in exploring how people understand location descriptions.

In this task, you will be given an expression (e.g. the house to the north of the town), and asked to select from a set of diagrams those which you think best matches the expression.

If you do not speak English fluently, please do not continue, as we are interested in the understanding of fluent English speakers.

You will not be asked for any identifying information, and we will not record your IP address or any other means by which to identify you, and you may exit the task at any time.

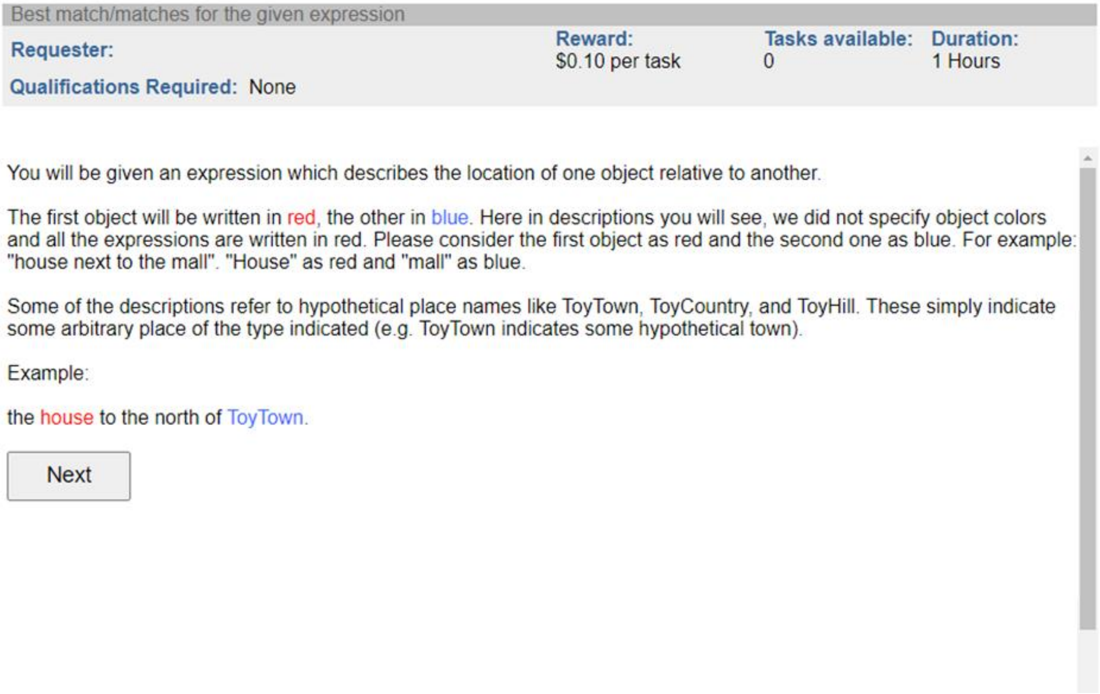
The researchers responsible for this project are Niloofar Aflaki (email: n.aflaki@massey.ac.nz) and Dr Kristin Stock (k.stock@massey.ac.nz) from Massey University, Albany, New Zealand.

This project has been evaluated by peer review and judged to be low risk. Consequently it has not been reviewed by one of the University's Human Ethics Committees. The researcher(s) named above are responsible for the ethical conduct of this research. If you have any concerns about the conduct of this research that you want to raise with someone other than the researcher(s), please contact Professor Craig Johnson, Director (Research Ethics), email humanethics@massey.ac.nz.

Please click the "Confirm" button below to consent to your data being used for research purposes by the Massey University student and staff mentioned above, and to read the instructions. Please read the instructions carefully, as your responses are important to our research.

If you submit the survey as blank, we will reject your answers. Please be careful and submit when you choose answers.

Page 1



Best match/matches for the given expression

Requester:	Reward: $0.10 per task	Tasks available: 0	Duration: 1 Hours
Qualifications Required: None			

You will be given an expression which describes the location of one object relative to another.

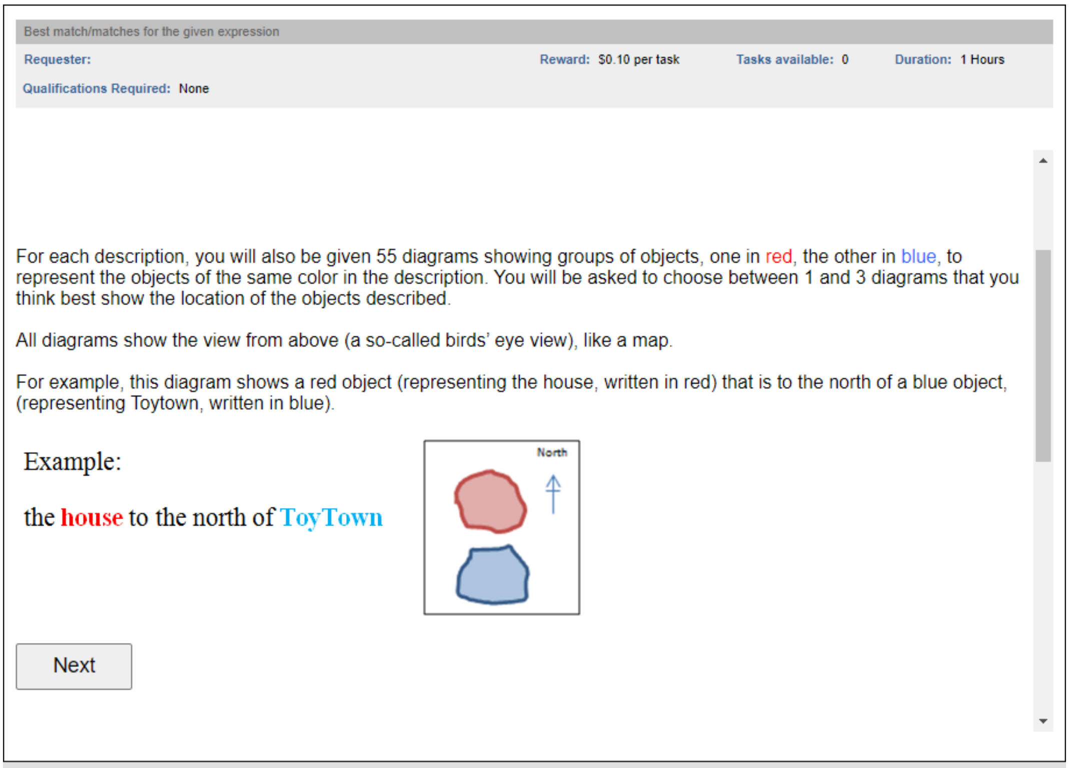
The first object will be written in red, the other in blue. Here in descriptions you will see, we did not specify object colors and all the expressions are written in red. Please consider the first object as red and the second one as blue. For example: "house next to the mall". "House" as red and "mall" as blue.

Some of the descriptions refer to hypothetical place names like ToyTown, ToyCountry, and ToyHill. These simply indicate some arbitrary place of the type indicated (e.g. ToyTown indicates some hypothetical town).

Example:

the house to the north of ToyTown.

Page 2



Best match/matches for the given expression

Requester: Reward: $0.10 per task Tasks available: 0 Duration: 1 Hours

Qualifications Required: None

For each description, you will also be given 55 diagrams showing groups of objects, one in red, the other in blue, to represent the objects of the same color in the description. You will be asked to choose between 1 and 3 diagrams that you think best show the location of the objects described.

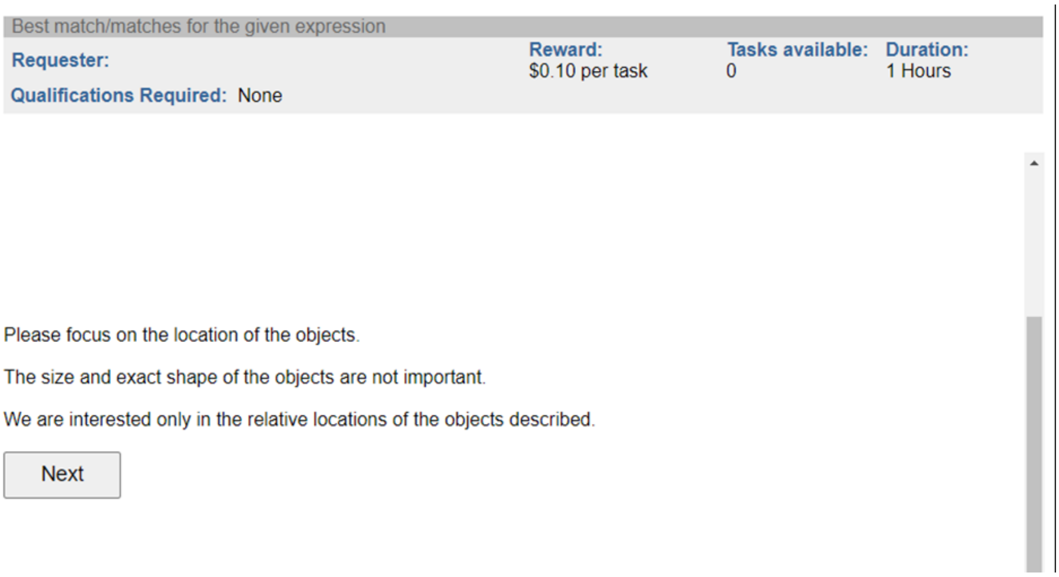
All diagrams show the view from above (a so-called birds' eye view), like a map.

For example, this diagram shows a red object (representing the house, written in red) that is to the north of a blue object, (representing Toytown, written in blue).

Example:

the house to the north of ToyTown

Page 3



Best match/matches for the given expression

Requester: Reward: $0.10 per task Tasks available: 0 Duration: 1 Hours

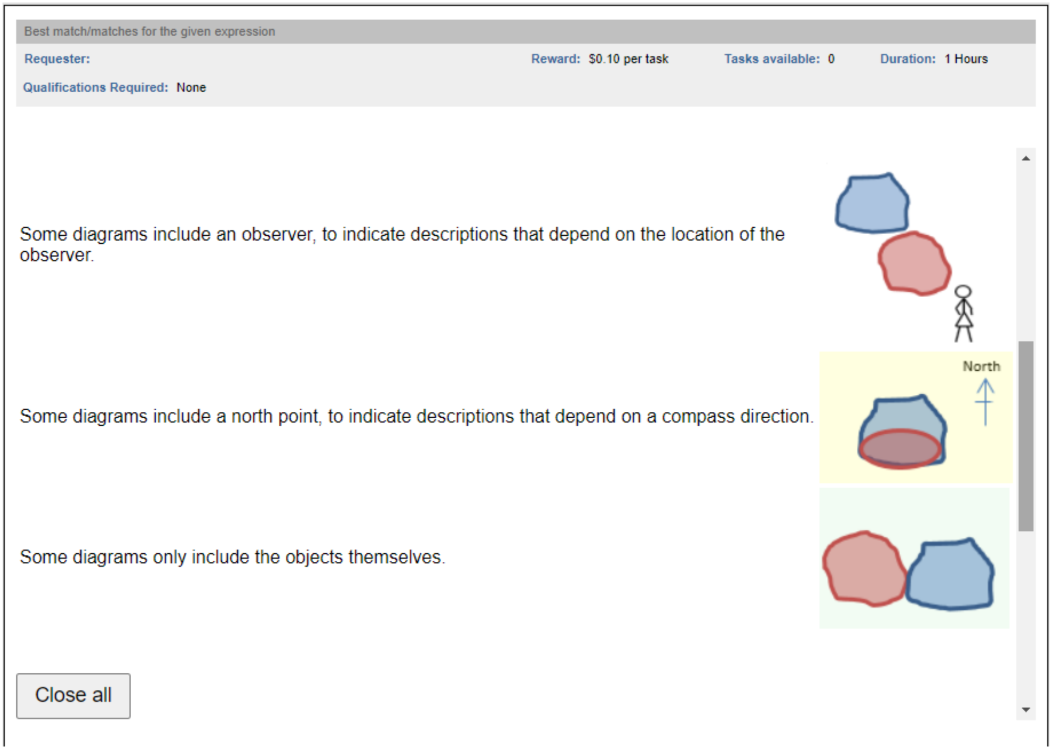
Qualifications Required: None

Please focus on the location of the objects.

The size and exact shape of the objects are not important.

We are interested only in the relative locations of the objects described.

Page 4



Best match/matches for the given expression

Requester: Reward: $0.10 per task Tasks available: 0 Duration: 1 Hours

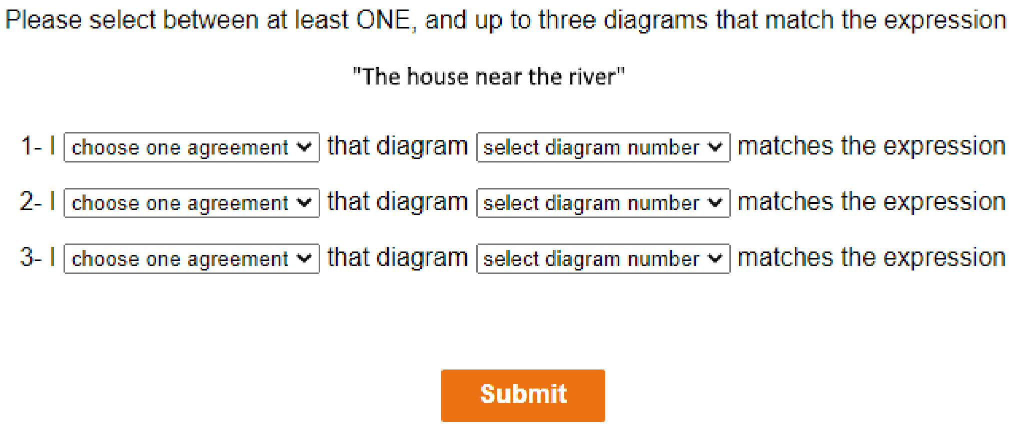
Qualifications Required: None

Some diagrams include an observer, to indicate descriptions that depend on the location of the observer.

Some diagrams include a north point, to indicate descriptions that depend on a compass direction.

Some diagrams only include the objects themselves.

Page 5



Please select between at least ONE, and up to three diagrams that match the expression

"The house near the river"

1- I [choose one agreement ▾] that diagram [select diagram number ▾] matches the expression

2- I [choose one agreement ▾] that diagram [select diagram number ▾] matches the expression

3- I [choose one agreement ▾] that diagram [select diagram number ▾] matches the expression

Final page (Experiment)

Appendix C - Similarity matrix of geospatial prepositions

	above	across	adjacent to	along	alongside	around	at	behind	beside	beyond	by	close to	in	inside	near	next to	off	on	opposite	outside	over	past	through	towards
above	1	0.371	0.495	0.531	0.449	0.389	0.627	0.852	0.432	0.774	0.566	0.53	0.399	0.396	0.571	0.491	0.542	0.718	0.729	0.651	0.54	0.457	0.309	0.274
across	0.371	1	0.295	0.449	0.335	0.3	0.419	0.36	0.257	0.369	0.407	0.344	0.253	0.261	0.292	0.258	0.436	0.442	0.275	0.308	0.843	0.676	0.95	0.322
adjacent to	0.495	0.295	1	0.549	0.596	0.304	0.447	0.607	0.87	0.649	0.784	0.906	0.199	0.199	0.948	0.933	0.627	0.592	0.785	0.867	0.334	0.512	0.241	0.419
along	0.531	0.449	0.549	1	0.922	0.364	0.434	0.569	0.584	0.637	0.763	0.679	0.245	0.255	0.54	0.488	0.684	0.616	0.572	0.534	0.551	0.606	0.388	0.327
alongside	0.449	0.335	0.596	0.922	1	0.277	0.351	0.583	0.732	0.697	0.796	0.786	0.175	0.188	0.614	0.58	0.729	0.535	0.647	0.597	0.409	0.674	0.26	0.315
around	0.389	0.3	0.304	0.364	0.277	1	0.465	0.318	0.254	0.306	0.502	0.281	0.442	0.439	0.332	0.251	0.312	0.421	0.288	0.35	0.272	0.27	0.193	0.142
at	0.627	0.419	0.447	0.434	0.351	0.465	1	0.53	0.354	0.488	0.654	0.465	0.892	0.881	0.523	0.402	0.493	0.804	0.406	0.509	0.522	0.43	0.379	0.342
behind	0.852	0.36	0.607	0.569	0.583	0.318	0.53	1	0.609	0.897	0.686	0.69	0.26	0.275	0.703	0.594	0.68	0.642	0.823	0.81	0.433	0.646	0.289	0.386
beside	0.432	0.257	0.87	0.584	0.732	0.254	0.354	0.609	1	0.711	0.849	0.928	0.153	0.152	0.902	0.893	0.768	0.573	0.833	0.826	0.28	0.631	0.171	0.334
beyond	0.774	0.369	0.649	0.637	0.697	0.306	0.488	0.897	0.711	1	0.763	0.797	0.224	0.236	0.735	0.624	0.837	0.632	0.814	0.819	0.451	0.787	0.301	0.538
by	0.566	0.407	0.784	0.763	0.796	0.502	0.654	0.686	0.849	0.763	1	0.872	0.472	0.479	0.834	0.749	0.848	0.766	0.757	0.802	0.429	0.721	0.325	0.43
close to	0.53	0.344	0.906	0.679	0.786	0.281	0.465	0.69	0.928	0.797	0.872	1	0.219	0.223	0.926	0.889	0.817	0.634	0.83	0.876	0.387	0.691	0.268	0.497
in	0.399	0.253	0.199	0.245	0.175	0.442	0.892	0.26	0.153	0.224	0.472	0.219	1	0.987	0.274	0.182	0.268	0.632	0.176	0.253	0.333	0.204	0.231	0.142
inside	0.396	0.261	0.199	0.255	0.188	0.439	0.881	0.275	0.152	0.236	0.479	0.223	0.987	1	0.276	0.183	0.274	0.633	0.179	0.255	0.334	0.217	0.248	0.155
near	0.571	0.292	0.948	0.54	0.614	0.332	0.523	0.703	0.902	0.735	0.834	0.926	0.274	0.276	1	0.95	0.694	0.685	0.868	0.947	0.339	0.575	0.232	0.42
next to	0.491	0.258	0.933	0.488	0.58	0.251	0.402	0.594	0.893	0.624	0.749	0.889	0.182	0.183	0.95	1	0.568	0.599	0.865	0.872	0.289	0.444	0.198	0.281
off	0.542	0.436	0.627	0.684	0.729	0.312	0.493	0.68	0.768	0.837	0.848	0.817	0.268	0.274	0.694	0.568	1	0.663	0.659	0.692	0.491	0.819	0.355	0.554
on	0.718	0.442	0.592	0.616	0.535	0.421	0.804	0.642	0.573	0.632	0.766	0.634	0.632	0.633	0.685	0.599	0.663	1	0.665	0.658	0.613	0.518	0.393	0.316
opposite	0.729	0.275	0.785	0.572	0.647	0.288	0.406	0.823	0.833	0.814	0.757	0.83	0.176	0.179	0.868	0.865	0.659	0.665	1	0.89	0.336	0.552	0.194	0.281
outside	0.651	0.308	0.867	0.534	0.597	0.35	0.509	0.81	0.826	0.819	0.802	0.876	0.253	0.255	0.947	0.872	0.692	0.658	0.89	1	0.345	0.626	0.243	0.43
over	0.54	0.843	0.334	0.551	0.409	0.272	0.522	0.433	0.28	0.451	0.429	0.387	0.333	0.334	0.339	0.289	0.491	0.613	0.336	0.345	1	0.612	0.771	0.335
past	0.457	0.676	0.512	0.606	0.674	0.27	0.43	0.646	0.631	0.787	0.721	0.691	0.204	0.217	0.575	0.444	0.819	0.518	0.552	0.626	0.612	1	0.632	0.561
through	0.309	0.95	0.241	0.388	0.26	0.193	0.379	0.289	0.171	0.301	0.325	0.268	0.231	0.248	0.232	0.198	0.355	0.393	0.194	0.243	0.771	0.632	1	0.361
towards	0.274	0.322	0.419	0.327	0.315	0.142	0.342	0.386	0.334	0.538	0.43	0.497	0.142	0.155	0.42	0.281	0.554	0.316	0.281	0.43	0.335	0.561	0.361	1

Appendix D - Summary of sense extraction

Preposition	Candidate Senses from Venn Diagram	Sense
Across	Core of preposition: -overlaps (35)	Sense 1: Objects that are overlapping
	Cluster 1: -overlaps (32, 33)	Does not justify separate sense as only geometry type differs Sense 1.
	Cluster 2: -across some other object, indicated by nearness diagrams (29, 30)	Sense 2: Objects that are across some other object from
	Cluster 3: -covering (43, 48, 51)	Sense 3: Objects that are covering (multiple)
Through	Core of preposition: -overlaps (35, 39, 33)	Sense 1: Objects that are overlapping
	Cluster 1: -polygon geometries	Does not justify separate sense as only geometry type differs from Sense 1.
	Cluster 2: -linear geometries	Does not justify separate sense as only geometry type differs from Sense 1.
Over	Core of preposition: -overlaps (39)	Sense1: objects are overlapping/crossing Sense 3: overlap + alignment
	Cluster 1: -mainly dominated by overlap using linear and polygon objects	Does not justify separate sense as only geometry type differs from Sense 1
	Cluster 2: -emphasis on verticality, often polygon/point like objects that sit in a vertically dominant position, so more like one object on top of (or nearly on top of) another	Sense 2: One object is above another object
	Cluster 3:	Does not justify separate sense as only geometry type differs from Sense 1

	-pairs of linear objects (whether aligned or not aligned)	
Adjacent to	Core of preposition: -touches (all senses) (36) -semantically similar proximity also important	Sense 1: objects are touching or nearly touching
	Cluster 1: -overlaps (32)	Sense 2: there is some overlap in the objects (with vague boundaries) – probably not actually an extra sense, just a stretching of the main sense
	Cluster 2: -linear features, proximity and touching	Does not justify separate sense as only geometry type differs from Sense 1
	Cluster 3: -multiple objects-sides of (50)	Does not justify separate sense as only the frequency differs from Sense 1
Beside	Core of preposition: -touching relation (all three clusters) -proximity also important	Sense 1: objects are touching or close to each other
	Cluster 1 -closeness and touching	Sense 1: objects are touching or close to each other
	Cluster 2: -close and touching -line and polygon	Does not justify separate sense as only geometry type differs from Sense 1
	Cluster 3 -close and parallel -line types	Sense 2: objects are close, linear, and parallel Line types alone doesn't justify separate sense, but parallelism does
Close to	Core of preposition: -no three-way core -proximity	Sense 1: objects are close to each other

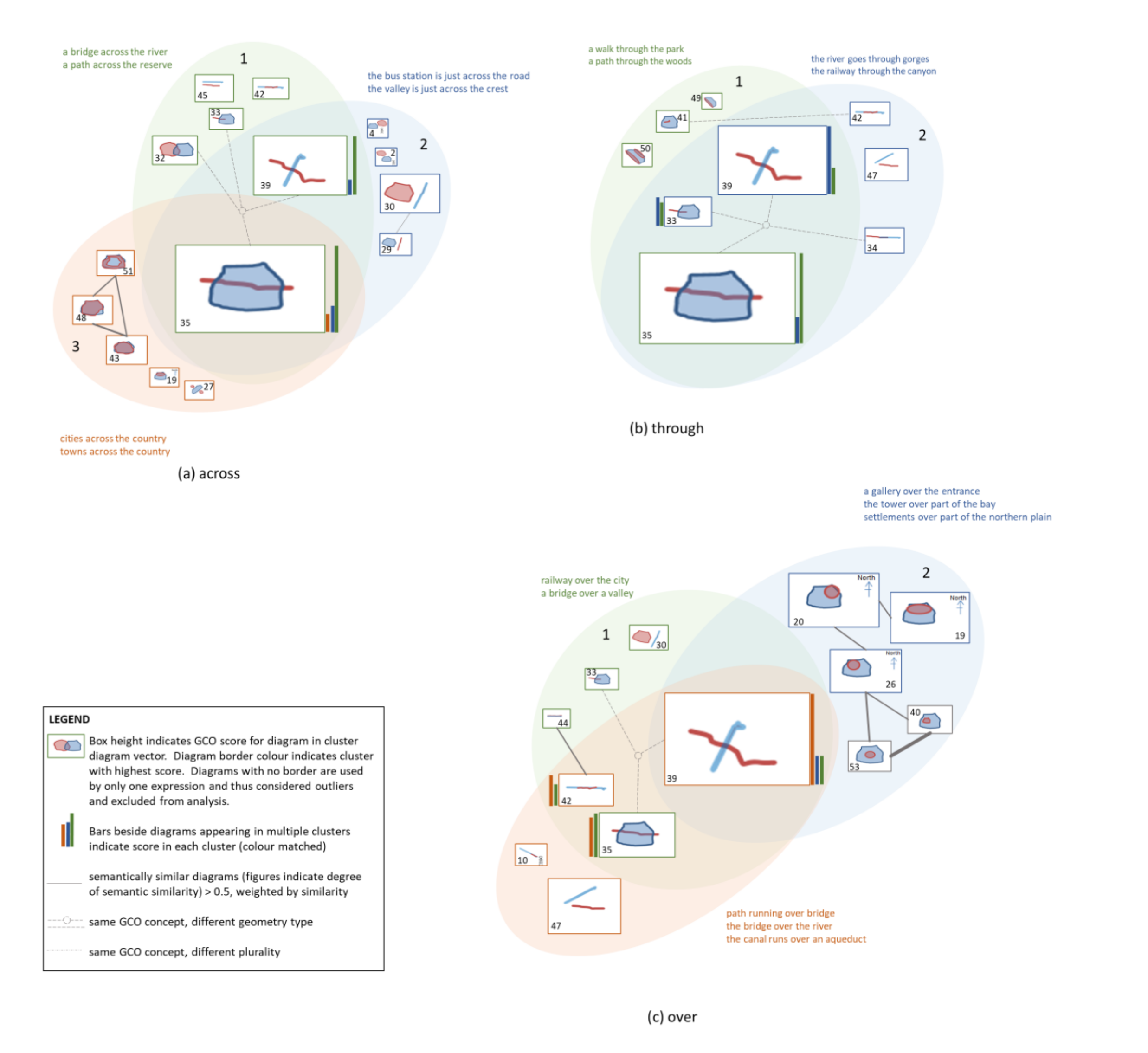
	-touching less important	
	Cluster 1: -polygons, close to each other, but mostly not touching	Sense 1: objects are close to each other
	Cluster 2: -linear, parallel most important, but other orientations also permitted	Linear parallelism not so strong as for beside (other orientations score more highly), so does not justify separate sense
	Cluster 3: -no separate sense	Same as sense 1
Near	Core of preposition: -no three-way core -proximity -touching even less important than for close to	Sense 1: near (proximity) Only one sense
	Cluster 1: -proximity -3 and 53 are for expressions that involve parts (eastern part, centre part) -28 is disjoint (similar to proximity)	Sense 1: near (proximity)
	Cluster 2 (only 2 expressions): -proximity, for line-polygon pairs	Does not justify separate sense as only geometry type differs from Sense 1
	Cluster 3 (only 2 expressions): -touches and proximity for pairs of polygons	Does not justify separate sense, for some reason diagrams showing a person are selected
Next to	Core of preposition:	Sense 1: proximity or touching between objects

	-proximity most important -touching close second	
	Cluster 1: -touching most important -proximity second -polygon pairs	Sense 1: proximity or touching between objects
	Cluster 2: -proximity -49 sometimes used to indicate linear object (a nature reserve located *next to* the coast), sometimes to describe part of an object (the side of the garden *next to* main street) -some degree of overlap -linear and parallel	Sense 2: linear parallelism Diagram 49 sometimes used to indicate linear object
Outside	Core of preposition: -proximity most important, touching much lower	Sense 1: proximity of objects
	Cluster 1: -proximity, pairs of polygons, multiple objects	Does not justify separate sense as objects are either multiple or have different geometries
	Cluster 2: -proximity -overlaps (a trail outside of the park) -line and polygon	Sense 2: objects have partial overlap
Off	Core of preposition:	Sense 1: proximity of objects

	-no diagrams in all four clusters -proximity most dominant	
	Cluster 1: -proximity, or touching	Sense 1: proximity of objects
	Cluster 2: -two linear objects, all sorts of relations, but more involving touching/overlapping	Sense 2: Overlapping of linear objects
	Cluster 3: -linear objects overlapping or touching	Does not justify separate sense as only geometry type differs Sense 2
	Cluster 4: -proximity -includes multiple object types	Does not justify separate sense as objects are either multiple or have different geometry types
Past	Core of preposition: -proximity, but no diagram is in all three clusters	Sense 1: proximity (includes by the side of)
	Cluster 1: -proximity, or touching (less important)	Sense 1: proximity (includes by the side of)
	Cluster 2: -with verb, bounded by, flanked by or -by the side of (probably two senses)	Sense 2: enclosure (with appropriate verb)
	Cluster 3: -proximity, multiple objects	Does not justify separate sense as objects are multiple (sense 1)
By	Core of preposition: -proximity, but no diagram is in all three clusters	Sense 1: proximity (includes by the side of)
	Cluster 1: -proximity, or touching (less important)	Sense 1: proximity (includes by the side of)
	Cluster 2: -with verb, bounded by, flanked by or	Sense 2: enclosure (with appropriate verb)

	-by the side of (probably two senses)	
	Cluster 3: -proximity, multiple objects	Does not justify separate sense as objects are multiple (sense 1)
Beyond	Core of preposition: -no overlap	Sense 1 (only sense): proximity, with locatum on the other side of relatum from implied observer position
	Cluster 1: -proximity, but implies observer position (diagram 2)	Does not justify separate sense as objects are close with a presence of an observer
	Cluster 2: -touching lines, again implies observer position (diagram 9)	Does not justify separate sense as linear objects are close with a presence of an observer

Appendix E - Venn diagrams for *across, through* and *over*


a bridge across the river
a path across the reserve
1
45
42
33
32
39
the bus station is just across the road
the valley is just across the crest
2
4
2
30
29
51
48
3
43
35
19
27
cities across the country
towns across the country
a walk through the park
a path through the woods
1
49
41
50
39
33
35
the river goes through gorges
the railway through the canyon
2
42
47
34
a gallery over the entrance
the tower over part of the bay
settlements over part of the northern plain
railway over the city
a bridge over a valley
1
30
33
44
42
10
35
47
39
2
North
20
North
19
North
26
40
53
path running over bridge
the bridge over the river
the canal runs over an aqueduct

(a) across

(b) through

(c) over

LEGEND

Box height indicates GCO score for diagram in cluster diagram vector. Diagram border colour indicates cluster with highest score. Diagrams with no border are used by only one expression and thus considered outliers and excluded from analysis.

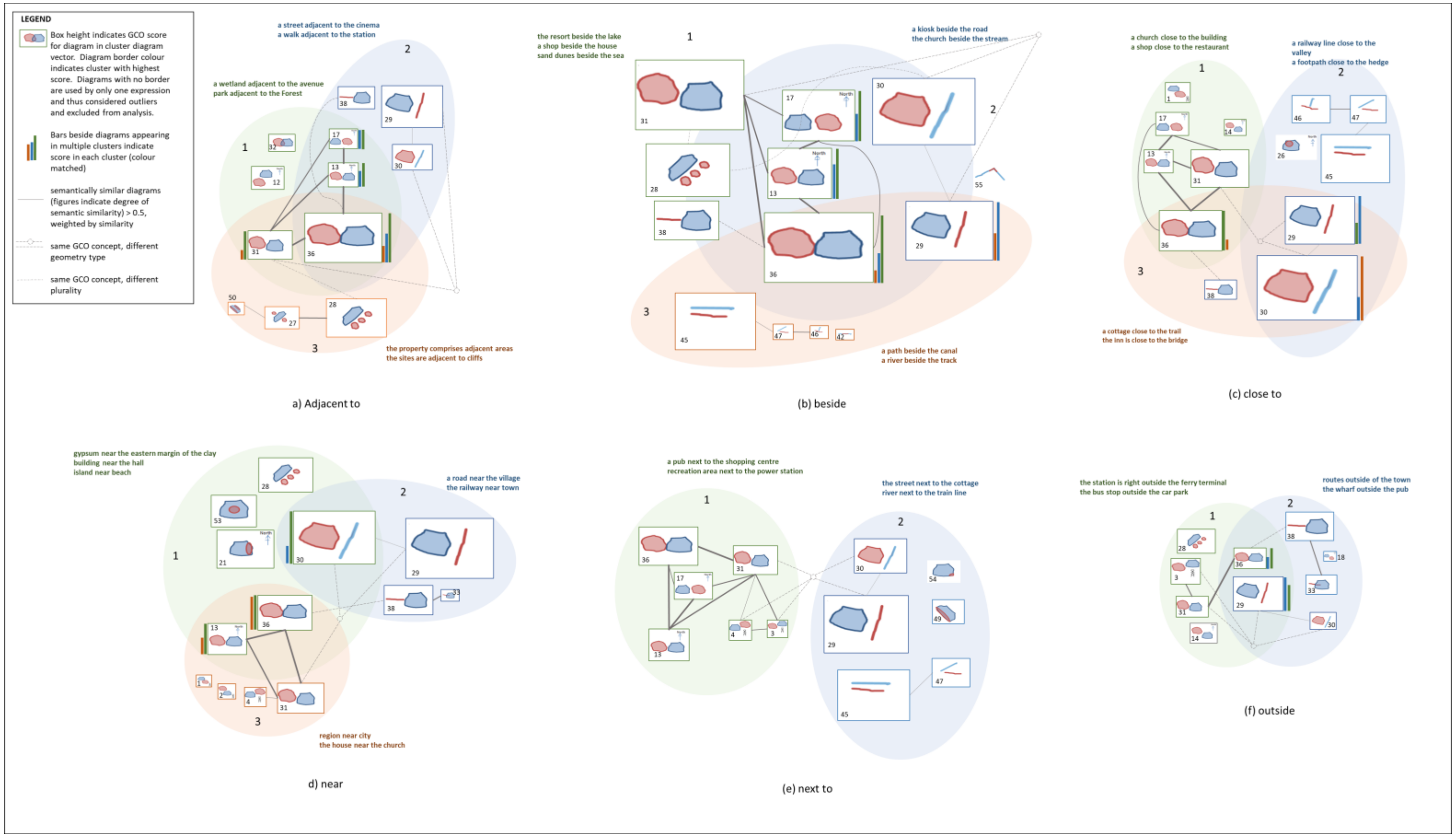
Bars beside diagrams appearing in multiple clusters indicate score in each cluster (colour matched)

semantically similar diagrams (figures indicate degree of semantic similarity) > 0.5, weighted by similarity

same GCO concept, different geometry type

same GCO concept, different plurality

Appendix F - Venn diagrams for adjacency/proximity prepositions


LEGEND
Box height indicates GCO score for diagram in cluster diagram vector. Diagram border colour indicates cluster with highest score. Diagrams with no border are used by only one expression and thus considered outliers and excluded from analysis.
Bars beside diagrams appearing in multiple clusters indicate score in each cluster (colour matched)
semantically similar diagrams (figures indicate degree of semantic similarity) > 0.5, weighted by similarity
same GCO concept, different geometry type
same GCO concept, different plurality
a street adjacent to the cinema
a walk adjacent to the station
a wetland adjacent to the avenue
park adjacent to the Forest
the property comprises adjacent areas
the sites are adjacent to cliffs
the resort beside the lake
a shop beside the house
sand dunes beside the sea
a kiosk beside the road
the church beside the stream
a path beside the canal
a river beside the track
a church close to the building
a shop close to the restaurant
a railway line close to the valley
a footpath close to the hedge
a cottage close to the trail
the inn is close to the bridge
gypsum near the eastern margin of the clay
building near the hall
island near beach
a road near the village
the railway near town
region near city
the house near the church
a pub next to the shopping centre
recreation area next to the power station
the street next to the cottage
river next to the train line
the station is right outside the ferry terminal
the bus stop outside the car park
routes outside of the town
the wharf outside the pub

a) Adjacent to

(b) beside

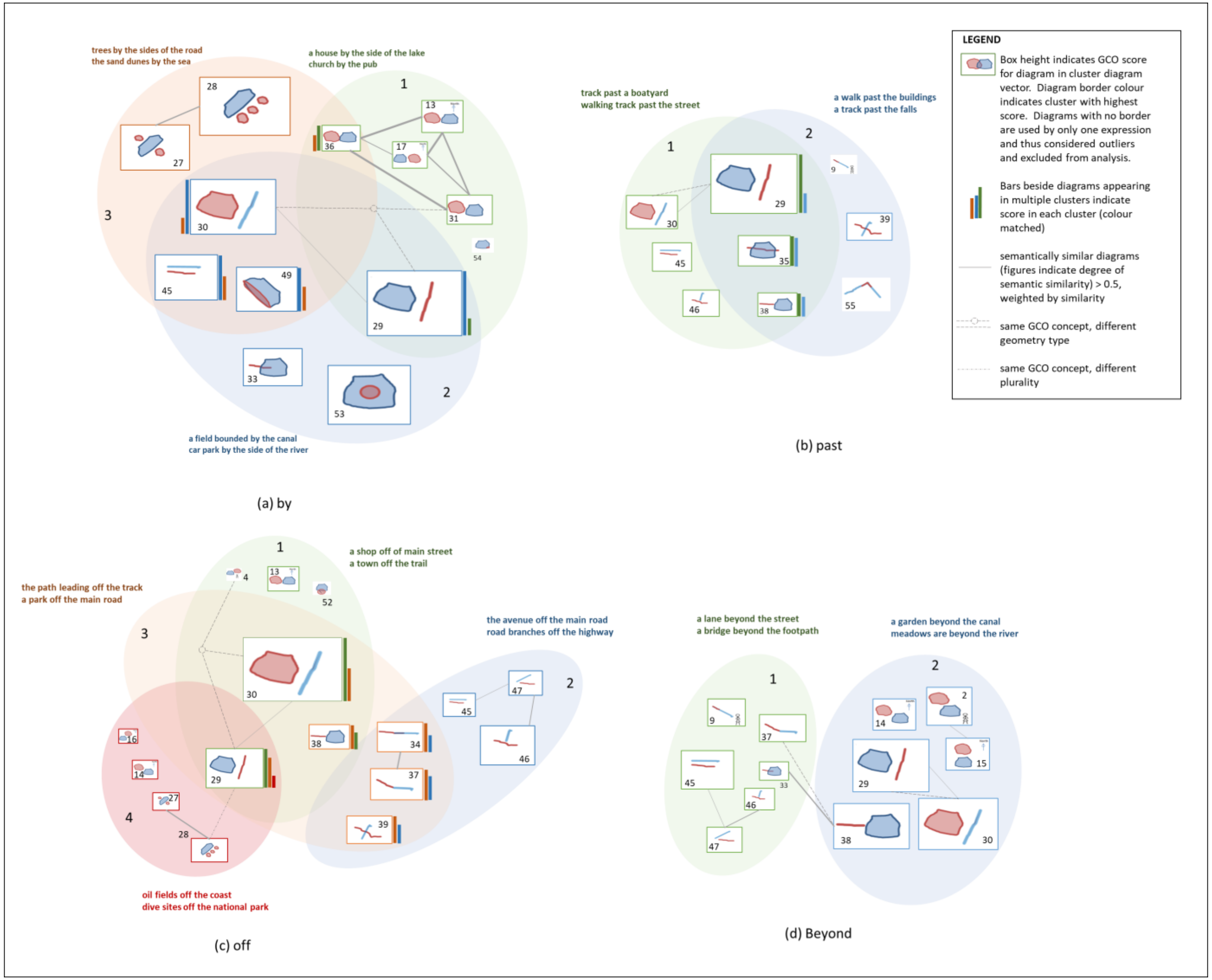
(c) close to

d) near

(e) next to

(f) outside

Appendix G - Venn diagrams for *by, past, off, beyond*


trees by the sides of the road
the sand dunes by the sea
a house by the side of the lake
church by the pub
a field bounded by the canal
car park by the side of the river
1
2
3
28
27
36
13
17
31
54
30
45
49
29
33
53
track past a boatyard
walking track past the street
a walk past the buildings
a track past the falls
1
2
30
29
9
39
45
35
46
38
55
LEGEND
Box height indicates GCO score for diagram in cluster diagram vector. Diagram border colour indicates cluster with highest score. Diagrams with no border are used by only one expression and thus considered outliers and excluded from analysis.
Bars beside diagrams appearing in multiple clusters indicate score in each cluster (colour matched)
semantically similar diagrams (figures indicate degree of semantic similarity) > 0.5, weighted by similarity
same GCO concept, different geometry type
same GCO concept, different plurality
the path leading off the track
a park off the main road
a shop off of main street
a town off the trail
the avenue off the main road
road branches off the highway
oil fields off the coast
dive sites off the national park
1
2
3
4
4
13
52
30
47
45
46
16
14
27
28
29
38
34
37
39
a lane beyond the street
a bridge beyond the footpath
a garden beyond the canal
meadows are beyond the river
1
2
9
37
45
33
46
47
14
2
29
15
38
30

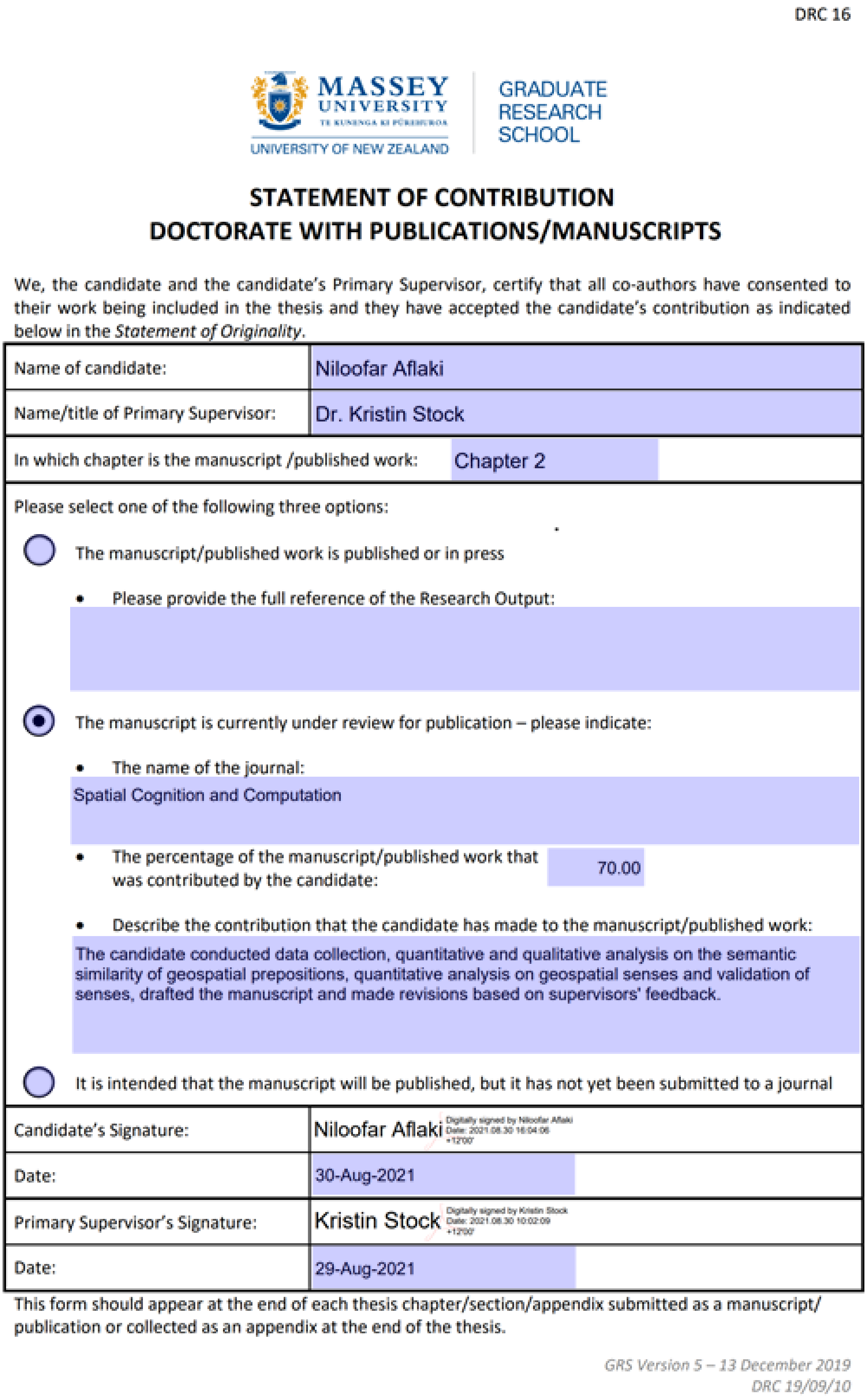
(a) by

(b) past

(c) off

(d) Beyond

DRC 16 form (Chapter 2)



DRC 16

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Niloofar Aflaki
Name/title of Primary Supervisor:	Dr. Kristin Stock
In which chapter is the manuscript /published work:	Chapter 2

Please select one of the following three options:

○ The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

● The manuscript is currently under review for publication – please indicate:

- The name of the journal:

Spatial Cognition and Computation

- The percentage of the manuscript/published work that was contributed by the candidate: 70.00

- Describe the contribution that the candidate has made to the manuscript/published work:

The candidate conducted data collection, quantitative and qualitative analysis on the semantic similarity of geospatial prepositions, quantitative analysis on geospatial senses and validation of senses, drafted the manuscript and made revisions based on supervisors' feedback.

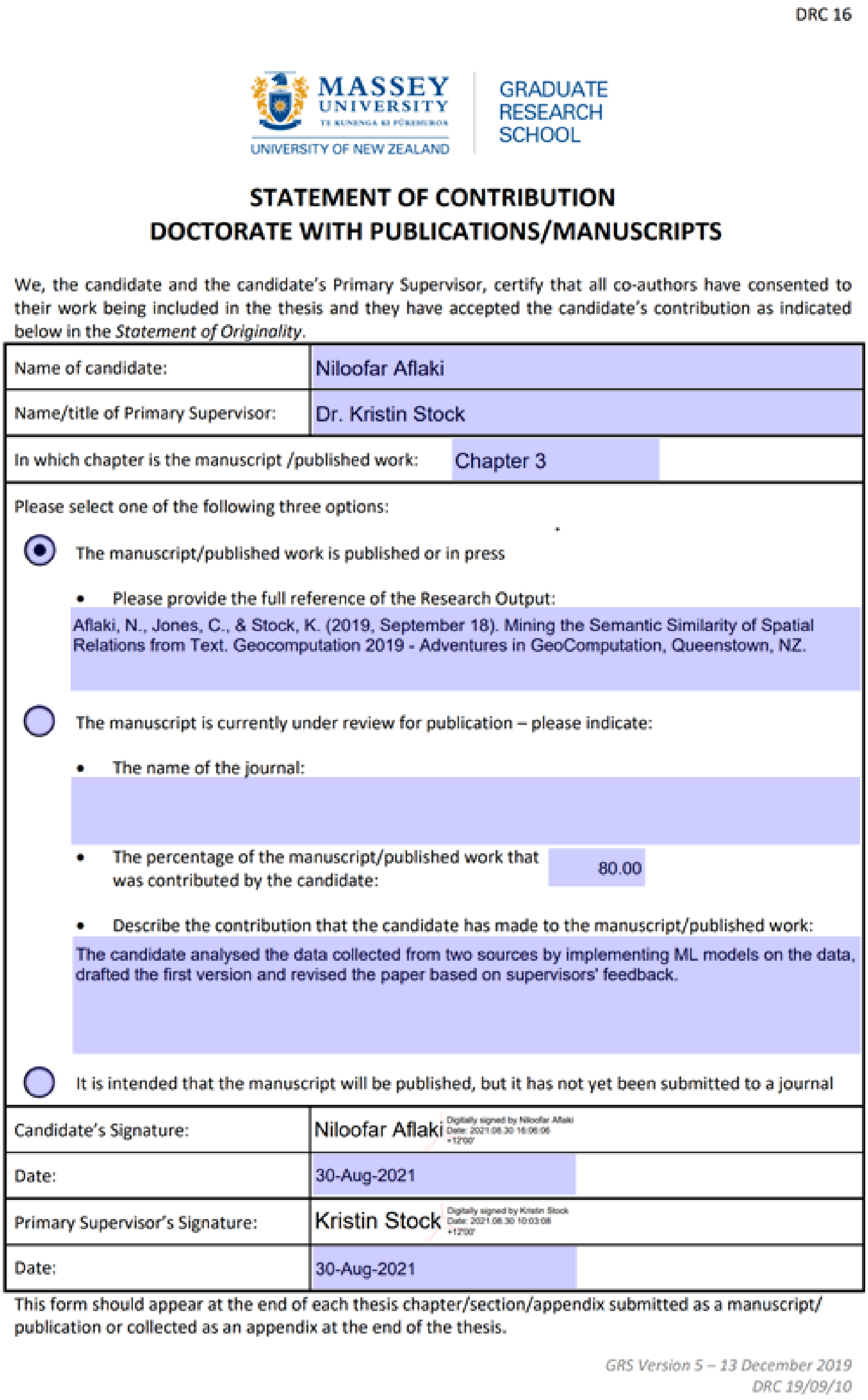
○ It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Niloofar Aflaki Digitally signed by Niloofar Aflaki Date: 2021.08.30 16:04:06 +12'00'
Date:	30-Aug-2021
Primary Supervisor's Signature:	Kristin Stock Digitally signed by Kristin Stock Date: 2021.08.30 10:02:09 +12'00'
Date:	29-Aug-2021

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

GRS Version 5 – 13 December 2019
DRC 19/09/10

DRC 16 form (Chapter 3)



DRC 16

MASSEY UNIVERSITY TE KUNENGA KI PŪREHUROA UNIVERSITY OF NEW ZEALAND | GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Niloofar Aflaki
Name/title of Primary Supervisor:	Dr. Kristin Stock
In which chapter is the manuscript /published work:	Chapter 3

Please select one of the following three options:

● The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Aflaki, N., Jones, C., & Stock, K. (2019, September 18). Mining the Semantic Similarity of Spatial Relations from Text. Geocomputation 2019 - Adventures in GeoComputation, Queenstown, NZ.

○ The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate: 80.00

- Describe the contribution that the candidate has made to the manuscript/published work:

The candidate analysed the data collected from two sources by implementing ML models on the data, drafted the first version and revised the paper based on supervisors' feedback.

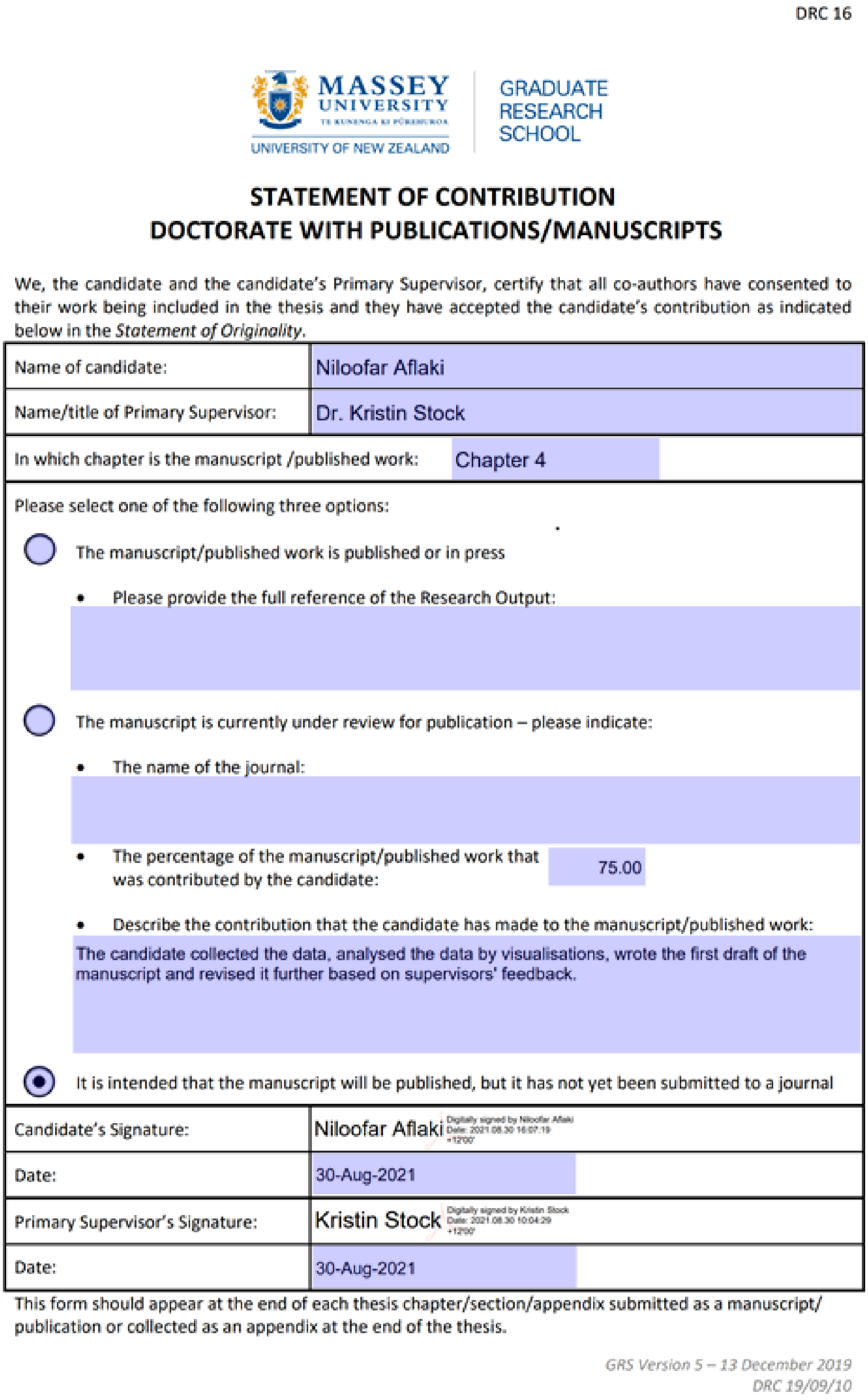
○ It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Niloofar Aflaki Digitally signed by Niloofar Aflaki Date: 2021.08.30 16:06:06 +12'00'
Date:	30-Aug-2021
Primary Supervisor's Signature:	Kristin Stock Digitally signed by Kristin Stock Date: 2021.08.30 10:03:08 +12'00'
Date:	30-Aug-2021

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

GRS Version 5 – 13 December 2019
DRC 19/09/10

DRC 16 form (Chapter 4)



DRC 16

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Niloofar Aflaki
Name/title of Primary Supervisor:	Dr. Kristin Stock
In which chapter is the manuscript /published work:	Chapter 4

Please select one of the following three options:

○ The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

○ The manuscript is currently under review for publication – please indicate:

- The name of the journal:
- The percentage of the manuscript/published work that was contributed by the candidate: 75.00
- Describe the contribution that the candidate has made to the manuscript/published work:

The candidate collected the data, analysed the data by visualisations, wrote the first draft of the manuscript and revised it further based on supervisors' feedback.

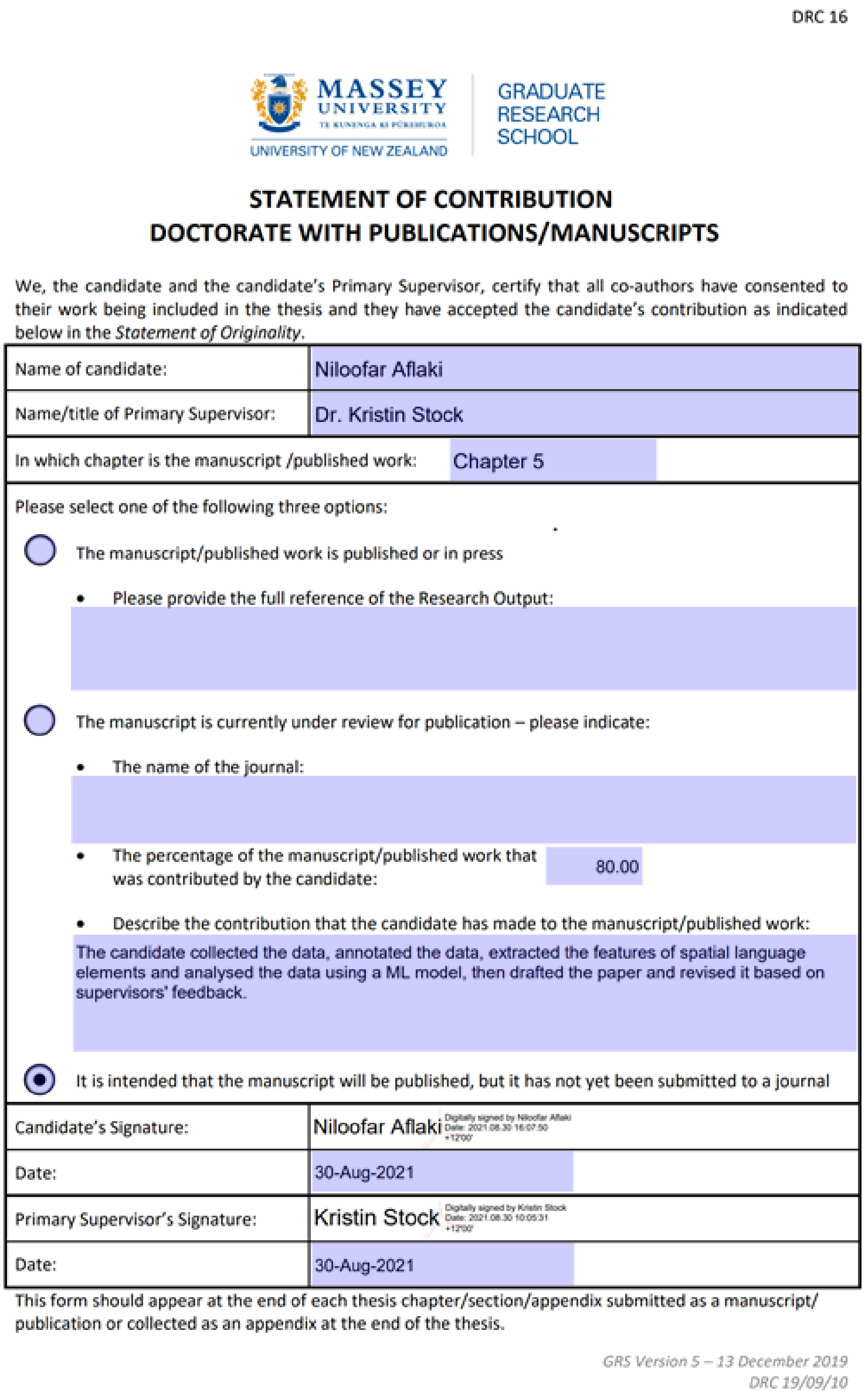
● It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Niloofar Aflaki Digitally signed by Niloofar Aflaki Date: 2021.08.30 16:07:19 +12'00'
Date:	30-Aug-2021
Primary Supervisor's Signature:	Kristin Stock Digitally signed by Kristin Stock Date: 2021.08.30 10:04:29 +12'00'
Date:	30-Aug-2021

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

GRS Version 5 – 13 December 2019
DRC 19/09/10

DRC 16 form (Chapter 5)



DRC 16

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Niloofar Aflaki
Name/title of Primary Supervisor:	Dr. Kristin Stock
In which chapter is the manuscript /published work:	Chapter 5

Please select one of the following three options:

○ The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

○ The manuscript is currently under review for publication – please indicate:

- The name of the journal:
- The percentage of the manuscript/published work that was contributed by the candidate: 80.00
- Describe the contribution that the candidate has made to the manuscript/published work:

The candidate collected the data, annotated the data, extracted the features of spatial language elements and analysed the data using a ML model, then drafted the paper and revised it based on supervisors' feedback.

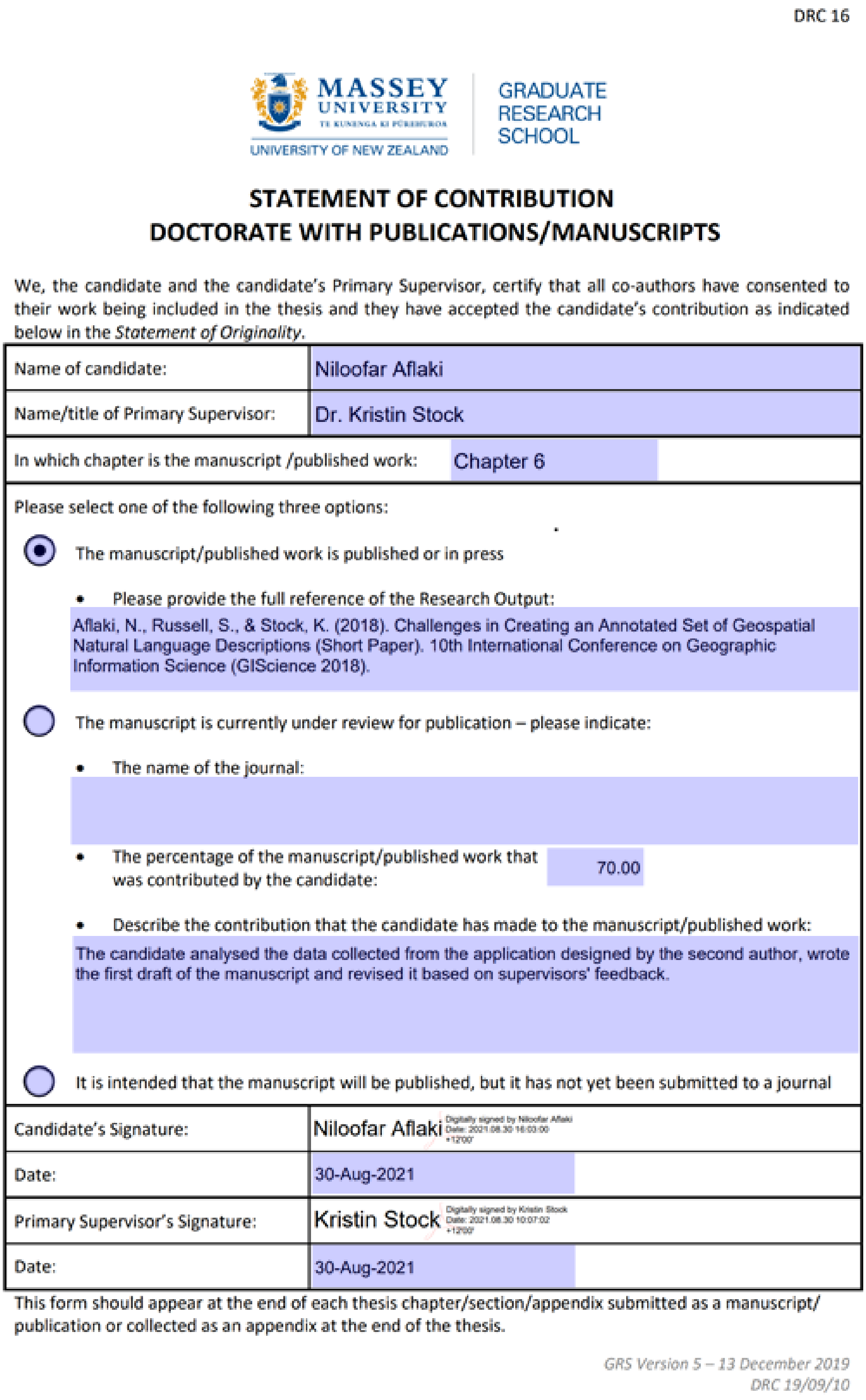
● It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Niloofar Aflaki Digitally signed by Niloofar Aflaki Date: 2021.08.30 16:07:50 +12'00'
Date:	30-Aug-2021
Primary Supervisor's Signature:	Kristin Stock Digitally signed by Kristin Stock Date: 2021.08.30 10:05:31 +12'00'
Date:	30-Aug-2021

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

GRS Version 5 – 13 December 2019
DRC 19/09/10

DRC 16 form (Chapter 6)



DRC 16

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Niloofar Aflaki
Name/title of Primary Supervisor:	Dr. Kristin Stock
In which chapter is the manuscript /published work:	Chapter 6

Please select one of the following three options:

● The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Aflaki, N., Russell, S., & Stock, K. (2018). Challenges in Creating an Annotated Set of Geospatial Natural Language Descriptions (Short Paper). 10th International Conference on Geographic Information Science (GIScience 2018).

○ The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate: 70.00

- Describe the contribution that the candidate has made to the manuscript/published work:

The candidate analysed the data collected from the application designed by the second author, wrote the first draft of the manuscript and revised it based on supervisors' feedback.

○ It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Niloofar Aflaki Digitally signed by Niloofar Aflaki Date: 2021.08.30 16:03:00 +12'00'
Date:	30-Aug-2021
Primary Supervisor's Signature:	Kristin Stock Digitally signed by Kristin Stock Date: 2021.08.30 10:07:02 +12'00'
Date:	30-Aug-2021

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

GRS Version 5 – 13 December 2019
DRC 19/09/10