UNIVERSITY OF LEEDS

Modelling the Semantic Variability of Spatial Prepositions in Referring Expressions

Adam Louis Richard-Bollans

University of Leeds

School of Computing

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

May, 2021

"Everything is vague to a degree you do not realise till you have tried to make it precise."

– Bertrand Russell, Lecture Series on The Philosophy of Logical Atomism (1918-19)

This comment was made when highlighting the common occurrence of statements for which, though one may be sure that they are true, providing a precise definition is a perilous task. This is a commonly recurring sentiment when tackling the semantics of spatial language, and natural language in general.

Intellectual Property Statement

The candidate confirms that the work submitted is his own, except where work has been included which has formed part of jointly authored publications. The contributions of the candidate and coauthors to this jointly authored work is explicitly indicated below. The candidate also confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter 4 elaborates on work published in:

- Adam Richard-Bollans, Brandon Bennett, and Anthony G. Cohn. Automatic generation of typicality measures for spatial language in grounded settings. In *Proceedings of 24th European Conference on Artificial Intelligence*, 2020. doi: https://doi.org/10.3233/FAIA200341

Chapter 5 elaborates on work published in:

- Adam Richard-Bollans, Lucía Gómez Álvarez, and Anthony G. Cohn. Modelling the polysemy of spatial prepositions in referring expressions. In *Proceedings of 17th International Conference on Principles of Knowledge Representation and Reasoning*, 2020. doi: https://doi.org/10.24963/kr.2020/72

Chapter 6 elaborates on work published in:

- Adam Richard-Bollans, Lucía Gómez Álvarez, and Anthony Cohn. Categorisation, typicality & object-specific features in spatial referring expressions. In *Proceedings of the Third International Workshop on Spatial Language Understanding*, pages 39–49. Association for Computational Linguistics, 2020. doi: http://doi.org/10.18653/v1/2020.splu-1.5

Both Chapter 2 & Chapter 3 include material from the above as well as from:

- Adam Richard-Bollans. Towards a cognitive model of the semantics of spatial prepositions. In *ESSLLI Student Session Proceedings*. Springer, 2018
- Adam Richard-Bollans, Lucía Gómez Álvarez, Brandon Bennett, and Anthony G. Cohn. Investigating the dimensions of spatial language. In *Proceedings of Speaking of Location 2019: Communicating about Space*. CEUR Workshop Proceedings, 2019

The candidate confirms that he is the first author of all these jointly-authored publications, that the work contained within these publications is directly attributable to him and the contributions of the co-authors has been in terms of general advice and assistance.

Acknowledgements

I would like to thank those people who have stimulated my scientific interest throughout my life — primarily my brother, Jack, and friend, John Mooney.

Thank you to all the people in Leeds who have made being here so enjoyable and have provided so much support along the way. In particular, thank you to Lucía Gómez Álvarez who first introduced me to the world of Computing; your insights, feedback and shared enthusiasm for understanding the semantics of natural language have been invaluable.

Thank you to my supervisors, Professor Anthony G. Cohn and Dr. Brandon Bennett, for your support and the interesting discussions we've had over the past four years about spatial language.

Finally, thank you to all the people who participated in the studies we conducted and thank you to the Engineering and Physical Sciences Research Council for the funding that made this research possible.

Abstract

Spatial prepositions in the English language can be used to denote a vast array of configurations which may greatly diverge from any canonical meaning and this semantic variability poses challenges for many systems where commands or queries are given in natural language. There have been many accounts from Linguistics and Cognitive Science highlighting the various phenomena which contribute to this semantic variability — primarily, spatial prepositions appear to encode functional as well as spatial information and to also exhibit polysemy. Both these issues represent significant challenges for grounded natural language systems and have not yet been accounted for in semantic models of spatial language.

To begin exploring the semantic variability of spatial prepositions, I will compare various cognitive accounts which incorporate the functional notions of *support* and *location control* and I will provide methods for constructing a semantic model based on Prototype Theory. In order to incorporate polysemy into this model, I will contribute methods for identifying polysemes based on Herskovits' notion of 'ideal meanings' as well as a modification of the 'principled polysemy' framework of Tyler and Evans. I will also introduce a notion of 'polyseme hierarchy' which will allow these polysemes to be incorporated in the semantic model.

By including functional relationships as well as polysemy into a semantic model we are able to provide a measure of typicality which is useful in interpreting referring expressions. However, this model does not yet account for 'object-specific features' which seem to influence categorisation decisions involving spatial prepositions. In the final chapter I will provide insight into the nature of categorisation and typicality for spatial prepositions and highlight the importance of object-specific features. Though a concrete solution to including these features will not be provided, I will provide suggestions of how these features may be included in our semantic model.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Spatial prepositions in the English language can be used to denote a vast array of configurations which may greatly diverge from any canonical meaning. These terms have evolved to be broad and flexible in their meaning and pose challenges for many systems where commands or queries are given in natural language. Following [11], we call this flexibility of meaning 'semantic variability' as opposed to 'vagueness', 'ambiguity' etc.. as we intend to investigate in general the varied ways that the semantics of a term may differ; and moreover, 'vagueness' and 'ambiguity' are often used in inconsistent ways across various fields [12].

There have been many accounts from Linguistics and Cognitive Science highlighting the various phenomena which contribute to the semantic variability of spatial prepositions. Firstly, spatial prepositions appear to not be purely spatial and seem to also encode functional information. For example, the functional notion of *support* appears to strongly influence the use of the preposition 'on'. Furthermore, functional properties of objects, e.g. affordances and roles, also have a strong influence. For example, the preposition 'in' is strongly associated with the role of being a *container*. Secondly, spatial prepositions appear to encode multiple distinct but related senses i.e. spatial prepositions exhibit polysemy. For example, in usual settings, the way a book is on a table is different to the way a clock is on a wall.[1]

Both these issues represent significant challenges for grounded systems required to interpret and use spatial language in a similar way to humans. Firstly, systems must account for contextual features from the geometric and functional domains as well as object-specific knowledge. Secondly, systems must understand the varied senses that a spatial preposition can encode and be able to reason about when a particular sense is being intended or is appropriate to use.

The primary motivation for this work is to explore semantic issues of spatial language in order to provide methods for tackling these issues, which as yet have not been incorporated in computational models. We hope that the methods developed here will be employed in semantic models which deal with spatial language in grounded settings, for example in human-robot interaction. Through exploring semantics in situated dialogue we also aim to provide analysis which furthers the theoretical work on spatial

[1] Whether this type of distinction actual constitutes polysemy is a matter of debate and will be discussed later.

language and cognition as well as cognitive models of concepts more generally.

In general in this thesis emphasis is given to semantics in grounded settings as opposed to simple textual occurrences. There has been some attention to the problem of interpreting spatial language in text, motivating the SpaceEval task [13], however a distinction should be made between uncontextualised textual usage and contextualised grounded usage. This issue will be discussed further in Section 2.3.4.

The particular challenge motivating this work is how to handle *referring expressions* — noun phrases which serve to identify entities e.g. 'the book under the table' — that situated agents may encounter and produce in indoor environments. Humans often prefer brief ambiguous descriptions over lengthy unambiguous descriptions [14], and locative expressions often fulfil this desire for brevity. For example, rather than referring to objects based on elaborate visual attributes like 'the yellow cup with two pink dots on it', humans often refer to objects using simple locative expressions, say 'the cup next to the stapler'. We also see many examples of these expressions in the SemEval-2014 corpus [15] and the HuRIC corpus [16], both of which consider natural language commands given to robots, as well as the GRED3D corpus [17] which contains referring expressions for block worlds.

Modelling the semantics of referring expressions in grounded settings provides a particular challenge where existing corpus-based methods from the field of Natural Language Processing are not immediately applicable. For example, a popular corpus-based approach is to represent words in text as real-valued vectors using a word embedding, e.g. GloVe [18]. Such methods learn the semantics of terms primarily through measuring co-occurrences with other words and provide semantic representations which can be useful in a variety of tasks, such as question answering [19]. In the context of modelling spatial prepositions, these methods may be used to capture the conventions associated with prepositions and decide when a particular preposition is appropriate in a sentence in some text corpus, e.g. [20]. However, when interpreting grounded referring expressions the important semantic features relate to the physical relationships of objects in a scene and such features cannot be extracted from word patterns that may be derived from corpus-based methods.

1.2 Scope

This work is generally intended to explore the semantics of spatial language in such a way that will be useful in interpreting referring expressions. However, this thesis is not concerned with modelling pragmatic issues[2] such as reasoning about the salience of objects in a scene or about the possible choices available to a speaker to refer to an object.

It is apparent that contextual factors relating to scale [22, 23] and domain [24] influence the usage of spatial prepositions. In this thesis we consider the usage of spatial prepositions in single rooms containing objects on or around a tabletop.

The spatial prepositions analysed in this thesis are those considered to have a functional component as well as those prepositions that seem to act as their geometric counterpart. For the 'functional' prepositions, object affordances and functional relationships, such as *support* and *location control*, appear to be salient [8, 10] compared to the geometric counterparts where geometric features and relative positions of objects appear to be more salient. In English, we consider the functional prepositions to be: 'in', 'on', 'over' and 'under'; and their respective geometric counterparts to be: 'inside', 'on top of', 'above' and 'below'. We also consider 'against' to be a functional preposition [25] which does not have a clear geometric counterpart (though there are possible candidates e.g. 'next to', 'near', 'by' or 'at'). The motivation for the split into functional and geometric prepositions will be discussed further in Section 2.2.3. The prepositions analysed in this thesis are therefore 'in', 'inside', 'on', 'on top of', 'over', 'above', 'under', 'below' and 'against'.

1.3 Gaps in Existing Models

In general, approaches to modelling the semantics of spatial prepositions do not capture the semantic variability that they appear to exhibit. In particular, semantic models do not allow for spatial prepositions to represent distinct senses i.e. existing semantic models do not incorporate polysemy. This is partly due to polysemy being a particularly difficult semantic phenomenon to capture but also a result of spatial prepositions

[2]The field of pragmatics is generally understood as the field of linguistics concerned with understanding speakers' *intended* meaning, however, it is not always clear where the line should be drawn between semantics and pragmatics [21].

being modelled by simple geometric relationships even though functional aspects are recognised as being salient. By limiting representations of spatial prepositions to one or two salient geometric dimensions it is difficult to envisage how multiple senses may be encoded.

In semantic models it is also assumed that the underlying semantics are the same when making category and typicality judgements. However, various accounts of cognition and semantic representations have highlighted that, for some concepts, different factors may influence category and typicality judgements [26, 27]. In particular, some features may be more salient in categorisation tasks while other features are more salient when assessing typicality. The possibility of such a distinction has not been explored in the context of spatial language and may have important ramifications for the processing of referring expressions.

1.4 Contributions

To begin exploring the semantic variability of spatial prepositions, we will provide methods for extracting the functional notions of *support* and *location control* from 3D virtual scenes. In its current form this feature extraction process relies on the rich information that can be extracted from such virtual scenes, however we will also discuss the potential to apply this to real world scenarios.

In order to incorporate these features into a semantic model and provide semantic models based on a diverse set of features, we will compare various cognitive accounts and provide methods for constructing a semantic model of spatial prepositions based on Prototype Theory [28]. This approach will initially model spatial prepositions as a single sense; however, the provided methods will allow us to explore how to model the polysemy that spatial prepositions appear to exhibit.

To model this polysemy, we will firstly provide methods for identifying polysemes based on Herskovits' notion of 'ideal meanings' [29] as well as a modification of the 'principled polysemy' framework of Tyler and Evans [30]. In order to incorporate these polysemes into the semantic model, we will also introduce a notion of 'polyseme hierarchy' and methods to quantify this notion.

By incorporating functional relationships as well as polysemy into a semantic model we are able to provide a measure of typicality which is useful in interpreting referring expressions. However, this model does not yet account for object-specific features —

related to object properties and affordances — which appear to influence spatial preposition usage. So far existing studies relate object-specific features to categorisation tasks [5, 31, 32] but not to utterance interpretation where the notion of typicality is often more salient. In order to include these features in semantic models it is important to understand the influence these features have when interpreting and generating utterances.

In the final chapter we will provide insight into the nature of categorisation and typicality for spatial prepositions and highlight the importance of object-specific features. Though a concrete solution to including these features will not be provided, we will provide some suggestions of how these features may be included in our semantic model.

1.5 Intended Audience

This thesis is primarily of interest to those interested in the semantics of spatial language in grounded settings. However, in Chapter 5 we tackle the phenomenon of polysemy which is not limited to spatial prepositions and the proposed methods may be applicable to a wider class of lexical items. Moreover, the discussion of Chapter 6 will highlight that assessing the semantics of a term can be task-dependent and that this may influence pragmatic decisions in referring expressions. Such influences may be of interest to researchers developing pragmatic accounts of referring expressions.

1.6 Thesis Outline

In Chapter 2 we begin by providing some linguistic background which highlights the semantic complexity exhibited by spatial prepositions and identifies various aspects that ought to be accounted for in semantic models. We will then consider the issue of modelling semantics in the context of referring expressions, outlining the cognitive accounts of concepts that will be compared and features that may be included in the concept representations. We will also highlight that there is a large body of work relating to mapping spatial language to some semantic representation and that, though this work is related to modelling spatial language for referring expressions, there are some important distinctions. Next we will provide some theoretical background on the distinction between categorisation and typicality. We will then provide an overview of

attempts to model spatial language in grounded settings and outline various gaps in existing approaches. Finally, we will consider existing datasets which may be used to generate and test semantic models of spatial language for referring expressions. This discussion will motivate the new studies we have conducted, which are described in detail in Chapter 3 and formally archived in the Leeds Research Data Repository.[3]

In Chapter 4 we explore general issues of representing spatial prepositions for handling referring expressions, in particular when making typicality judgements. Various underlying conceptual representations will be compared and the issue of generating appropriate parameters for these models from data will be addressed. The main outcomes of this chapter are the Baseline Prototype Model of spatial prepositions based on Prototype Theory. We will also explore the utility of including functional features in the model. In providing suitable methods for generating the model parameters from data, we allow for similar models to be constructed for concepts where the semantics may not be easily defined; and this will also allow us to model the semantics of distinct polysemes in the following chapter.

In Chapter 5 we will explore how to model the polysemy that spatial prepositions appear to exhibit and refine the Baseline Prototype Model by accounting for polysemy. We will provide novel methods for distinguishing separate polysemes, modelling the semantics of these polysemes and incorporating these into a Polysemy Model which outperforms the Baseline Prototype Model. Once the performance of the Polysemy Model has been assessed we will then explore refinements of the model and simple methods for reducing the reliance on intuition to build the model. Finally, we will analyse the properties and behaviour of the generated polysemy models, providing some insight into the improvement in performance over the Baseline Prototype Model, as well as justification for the given methods.

In Chapter 6 we will analyse whether the influence of object-specific features is limited to categorisation and discuss the implications for pragmatic strategies and semantic models. The main hypothesis of Chapter 6 is that object-specific features are more salient in categorisation, while geometric and physical relationships between ob-

[3] http://archive.researchdata.leeds.ac.uk/view/collections/Spatial_Prepositions_and_Situated_Dialogue.html (Some extra analysis provided in this thesis as well as more up to date code can be found here: https://github.com/alrichardbollans/spatial-preposition-annotation-tool-unity3d/tree/master/Analysis/extra%20thesis%20results).

jects are more salient in typicality judgements. Based on the collected data we cannot verify the hypothesis and will conclude that object-specific features appear to be salient in both category and typicality judgements, further evidencing the need to include these types of features in semantic models. We will then propose how such features may be incorporated into semantic models.

Finally, in Chapter 7 we will summarise what has been achieved in the thesis as well as the limitations of the generated semantic models. We will also discuss possibilities for future work.

Chapter 2

Background

2.1 Terminology

Before providing a detailed overview of linguistic and computational models of spatial prepositions we will first address some important terminology.

Regarding the names of the objects being discussed we use *figure* (also known as: target, trajector, referent) to denote the entity whose location is important e.g. 'the **bike** next to the house' and *ground* (also known as: reference, landmark, relatum) to denote the entity used as a reference point in order to locate the figure e.g. 'the bike next to the **house**'. We call potential figure-ground pairs *configurations*.

It is possible that ambiguity may arise in the use of 'ground' where it may be confused with the ground/floor. In this thesis however, 'ground' will never be used to mean 'floor' unless explicitly stated.

Spatial prepositions are often categorised as either 'topological' or 'projective' terms. Topological terms refer to static topological relations, involving notions of containment and proximity, and locate the figure in some neighbouring region of the ground. The 'topological' terms considered in this thesis are 'in', 'inside', 'on', 'on top of' and 'against'.

In contrast, projective terms often convey information about the direction that an object is located in relative to another e.g. 'the light *above* the desk', 'the ball *in front of* the car'. The 'projective' terms considered in this thesis are 'above', 'over', 'below' and 'under'.

Projective spatial prepositions depend on a particular frame of reference being adopted for their interpretation. A common system of frames is given by Levinson [33]:

- Intrinsic: This frame locates figure objects with respect to intrinsic properties of the ground. For example, 'the bike in front of the person' locates the bike with respect to the intrinsic front of the person
- Relative: Locates objects with respect to the viewpoint of an agent. For example, 'the bike is on the right of the person' may locate the bike from the perspective of an onlooker
- Absolute: Locates objects with respect to fixed properties of the environment, such as cardinal direction or the direction of gravity. For example, 'the bike is north of the person'

It is possible to use 'above', 'over', 'below' and 'under' with any of these reference

frames. However, in many environments all objects and agents are normally oriented and therefore the intrinsic, relative and absolute frames are equivalent for these prepositions. In some cases, where, say, the ground is upside-down, the relative and absolute frames are still equivalent for these prepositions and seem to take precedence over the intrinsic frame. As a result, much work on reference frames is conducted with projective prepositions such as 'in front of', 'behind', 'right of' etc... where there is greater ambiguity as to which frame is being used. For this reason, issues regarding reference frames will not be considered in detail in this thesis.

2.2 Semantic Variability

At first one may believe that the semantics of some of the discussed terms are relatively simple and easy to model. For example, 'in' signifies inclusion or containment; can it really be so difficult to interpret or generate appropriate expressions involving 'in'? In this section we explore this issue in some detail, highlighting the challenges of representing spatial language in grounded settings.

Considering the examples of 'on' in Figure 2.1 we can quickly see the variability these terms may exhibit. Example a. is the canonical usage of 'on' where the cup is supported by the table from below. The figure doesn't need to always be above the ground however, as seen in the other examples and in the case of e. the figure is actually below the ground. Example d. shows that spatial prepositions can also be used to express relations between an object and one of its parts — the handle is a part of the door.

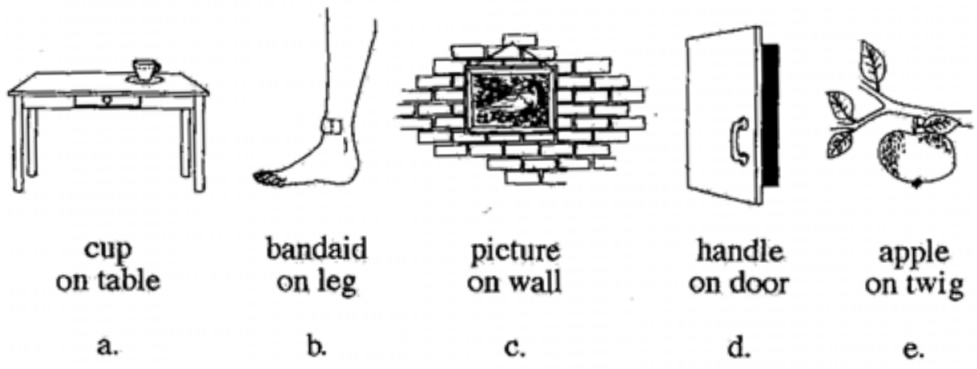


Figure 2.1: Examples of 'on' from [6].

In the given examples, the figure is always in contact with and supported by the ground. However, 'on' may also be used when these features are not present. For example, Miller and Johnson-Laird [34] suggest that 'on' may indicate that the figure is in contact with some functional region of the ground rather than the ground itself. A good example of this is provided in Figure 2.2, where one may describe the atlas as being 'on' the table even though it is not in contact with it. Also, from experimental studies, in the phrases 'the bookcase on the wall' from [35] and 'the balloon on the ceiling' from the study of [5], 'on' is used to describe a contact relationship where no support is apparent.

Figure 2.2: 'The atlas on the desk' from [7].

2.2.1 Simple Relations

Spatial prepositions may exhibit a large degree of variability, nonetheless it is apparent that some spatial prepositions encode basic general notions such as 'in' expressing containment and 'on' expressing contact or support. In the simple case of 'in', having a single defining feature, one may define 'in' as follows:

Definition 2.2.1 *X is in Y to the extent that X is located in the interior of Y:*

$$in(X, Y) \iff Located(X, Interior(Y))^4$$

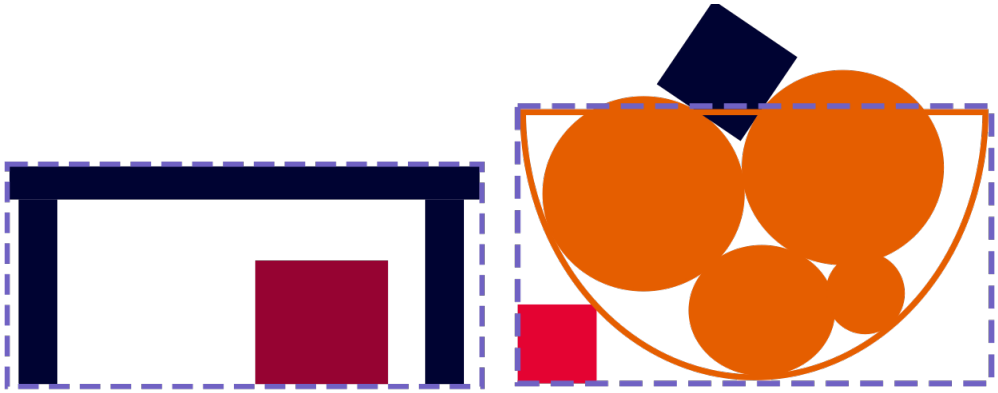
Then following this definition, in the context of processing referring expressions, configurations in a scene may be compared for how well they fit this definition.

Herskovits [7] refers to the assumption that spatial prepositions simply encode basic relations between objects as the *simple-relations* model of spatial prepositions. Herskovits provides many examples of the inadequacy of this assumption in her work and

[4] Taken from [7] which is adapted from [36].

here we will briefly elaborate on this by providing further examples related to 'in' particularly in the context of creating semantic models for processing referring expressions.

Following Definition 2.2.1, suppose that $Interior(Y)$ can be modelled as the convex hull of Y, then X may be fully contained in the convex hull of Y and therefore $in(X, Y)$, but not what we would think of as 'in'. For example, in Figure 2.3(a) — is the box 'in' the table? We may want to overcome this by stipulating that Y must be a type of *container*, which is reasonable for many situations [5]. However, for computational reasons features may be somewhat crudely approximated — as is common, we model the 'interior' of an object by using axis-aligned bounding boxes. As a result, in Figure 2.3(b), using this definition the red cube is more 'in' the bowl than the black cube. We see examples of this in our data, discussed in Section 4.6.1.



(a) (b)

Figure 2.3: Containment issues.

Another solution may be to refine the notion of 'interior' and allow for different types of containment based on the type of interior, as discussed in [37, 38]. For example in Figure 2.3(b), using the terminology of [37], the red cube is *geometrically inside*[5] the bowl, while the black cube is partially contained in the *containable inside*[6] of the bowl. As being geometrically inside is containment in a weak sense, compared to the stronger sense of being in the containable inside, we may be able to deduce the black cube is a

[5]In [37] 'geometrically inside' denotes being contained in the *convex hull*, however in this example here we use the bounding box instead.

[6]In [37] the 'containable inside' of an object is essentially a region that could be enclosed with an appropriate lid e.g. in a bowl or cup it would be the region which usually contains food/drink.

better fit of the description of 'the cube in the bowl'. In order to make this deduction, one would need to construct an appropriate hierarchy or reasoning mechanism for different types of containment and their influence in different contexts.

However, extracting these relations from scenes is extremely difficult even when the scenes are 3D virtual models. Also, simply refining the notion of containment does not overcome the problem presented by instances of 'in' where there is no apparent containment. Such an example, provided by Garrod et al. [8], is given in Figure 2.4. This is clearly a contrived example but illustrates the point well — 'the pear in the bowl' may be used to describe the pear in (a) but this would be strange for the pear in (b), even though the pear in (b) is more geometrically contained in the bowl than in (a). This appears to be a result of the functional relationships that some spatial prepositions encode and it would appear that such relationships should therefore be accounted for in our semantic models, which is discussed further in Section 2.2.3.

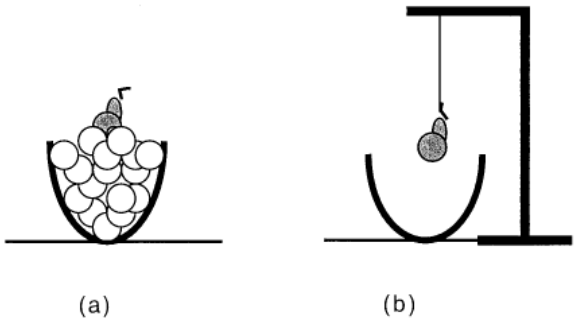


Figure 2.4: Example given in [8]. In which case is the pear more 'in' the bowl?

It should hopefully be clear that the initial intuitions one may have about the semantics of these terms may be over-simplified, and that simple models do not align with how the terms are used. In the next section we explore how this semantic complexity may arise.

2.2.2 Ideal Meanings

As we have seen, it is apparent that some spatial prepositions encode basic notions but that understanding these basic notions is not enough to represent the semantics of spatial prepositions, particularly in grounded settings. Nevertheless, these notions are

conceptually primitive or pre-linguistic, for example there is some evidence to suggest that young infants have some understanding of the concepts of contact and support [39], and it is plausible that our understanding of the meanings of spatial prepositions is initially built on these conceptual primitives.

For example, suppose that support is a basic concept held by infants and that a support relation between two objects is recognised following the image-schematic representation of [9], see Figure 2.5. When learning the meaning of 'on' it may initially be associated with this simple notion of support and this would seem to represent the canonical usage of 'on'. 'on' may then become more varied as we develop our understanding of support and also as we attach other concepts, such as contact, to the term 'on'.

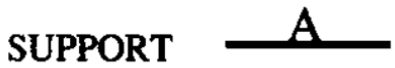


Figure 2.5: Image-schema for *support* [9].

The pioneering work of Herskovits [29] explored the role of these basic notions, or 'ideal meanings' in the nomenclature of Herskovits, in the semantics of spatial prepositions. Ideal meanings are to be understood as geometric abstractions which represent something similar to a prototypical notion of a concept. For example, the ideal meaning of the preposition 'in' is 'inclusion of a geometric construct within another geometric construct'. This is roughly captured by the *containment* image schema in Figure 2.6.

Figure 2.6: Image-schema for *containment* [9].

Spatial prepositions manage to meaningfully convey a variety of differing yet connected meanings, and Herskovits proposes that these meanings are connected by and derived from the ideal meaning via 'sense' and 'tolerance' shifts. Before explaining the process of sense and tolerance shifts, we will briefly explain how these geometric abstractions may be related to real world scenarios.

Schematization

When confronted with a configuration in the real world a process of abstraction may be necessary in order to relate the information-rich configuration with a simple geometric abstraction. This process is referred to as 'schematization' [40, 41].

For example, 'in' is the conventional preposition used to describe how an object relates to a field, e.g. 'the scarecrow in the field', and it sounds odd to say 'the scarecrow on the field'. This is in spite of the physical relationships in this scenario resembling 'support' (Figure 2.5) rather than 'containment' (Figure 2.6). The process of schematization provides one explanation of this — we imagine a 2D scene where the field is a flat plane containing the scarecrow. Similarly we say 'in the car' but 'on the train', possibly explained by conceptualising the car as a container while the train is imagined as a flat platform.

Modelling the process of schematization and finding the factors that influence this process is an interesting research challenge which would require a sophisticated integration of spatial cognition and commonsense. In this thesis we will not directly tackle how the process of schematization should be modelled, though some of the methods developed may help to overcome some of the issues raised by the complexities of schematization.

Sense and Tolerance Shifts

With ideal meanings as a starting point, Herskovits proposes that the full myriad of preposition use is then achieved via 'sense shifts' and 'tolerance shifts'.

Sense shifts appear in a discontinuous manner where the relations expressed by the ideal meaning are substituted for conceptually similar relations, and new senses of the terms are generated. Herskovits provides the instructive example of 'the muscles in his leg' where the relation being expressed by 'in' is no longer containment but *parthood*. It is also plausible, however, that this usage could be explained via a particular schematization rather than a sense shift — if we conceptualise the leg as its outline which contains the muscles.

Tolerance shifts occur in a continuous manner and allow for usages of a preposition when its ideal meaning is only approximately represented. For example, in Figure 2.6 we may still consider 'A' to be 'in' the circle if it is moved slightly outwards so that not

all of 'A' is contained within the circle. The *sorites* vagueness[7] that spatial prepositions exhibit can be viewed as a result of such tolerance shifts.

Revisiting the example of Figure 2.2 where an atlas in on a desk, this may be considered as an example of a tolerance shift if we suppose that the atlas is still in contact with desk although in a diminished fashion i.e. in indirect contact. However, it is also plausible that this represents some sort of sense shift in which contact is no longer salient. The extent to which such instances are genuinely distinct senses will be explored further in Section 2.2.4.

A Worked Example

Consider the example given in Figure 2.7, similar to Garrod's pear in the bowl (Figure 2.4 a.). Though there is no geometric containment of the red ball in the bowl, it may be common to refer to the ball as 'the red ball in the bowl'. This example can however be related to the ideal meaning of 'in' via appropriate schematization and tolerance shifts. Firstly, we may abstract the scene by viewing the collection of balls as a single entity — denoted by the blue dashed line. This single mass is now partially contained in the bowl and may be obtained via a tolerance shift applied to the ideal meaning of 'in'. Similarly, one may abstract the scene by imagining the bowl as a larger region of functional influence which extends above the bowl [31].

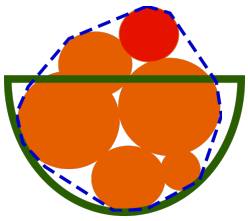


Figure 2.7: 'the red ball in the bowl': Object aggregates are contained.

Alternatively, instead of explaining non-containment instances of 'in' via abstractions and manipulations, one may argue that 'in' in fact encodes a functional relationship. In developing an account of spatial prepositions which argued that the underlying semantics are both geometric and functional, Garrod and Sanford [42] introduced the notion of *location control* in order to explain instances such as Figure 2.7. Location

[7] Sorites vagueness refers to the ambiguity which arises when the applicability of a predicate depends on specific parameters whose thresholds for applicability are undetermined.

control is generally understood as the degree to which the ground constrains the location of the figure and a movement of the ground causing a movement of the figure is an indication that location control is apparent.

It is plausible that, from a computational perspective, it is simpler to represent functional features, such as location control, directly rather than modelling schematizations and sense shifts. In the following section we will explore in more detail the various functional influences on spatial preposition usage.

2.2.3 Functional Influences

Initial attempts to understand and model spatial language naturally focused heavily on geometry. However, as has been recognised in the past couple of decades, spatial constraints are not enough to fully characterise spatial prepositions [8, 25, 29, 31, 43, 44]. The use of spatial prepositions appears to be determined by geometric, functional *and* conventional considerations, as evidenced in [8, 31, 32].

Talmy [25] introduced and highlighted the importance of 'force-dynamics' in language and cognition, considering the force interactions of objects as a primitive notion that pervades language through metaphor. This work inspired future researchers to pay more attention to the force interactions present; most notably in the investigations of [8, 44] which considered the interactions of geometry and functionality in spatial semantics, in particular highlighting that the functional control of the ground over the figure strongly influences preposition usage.

Garrod et al. [8] give the well-cited example, considered above, that a pear may be considered as 'in' a bowl when it is not even partially contained by the convex hull of the bowl — if it is sat on top of a pile of other pears in the bowl. We also see examples of this in our collected data. It has also been shown that the way objects are labelled and conceptualised affects preposition use. Coventry et al. [31] found that when given exactly the same scene of an object on a plate/dish, humans will describe the configuration as 'in' when the 'plate' is labelled as a dish, and 'on' when labelled as a plate. Coventry et al. suggest that this is due to the affordances associated with the concepts 'plate' and 'dish'.

Functional Relationships

There is a significant body of work involving numerous experimental studies exploring the non-geometric aspects of spatial prepositions. Central to many of these studies has been the idea that objects may interact in a functional way that is not simply geometric in nature. Of particular salience for the prepositions considered in this thesis are the functional relationships:

- Location control [42]
- Support [8]
- Covering/Protection [10, 30]

As previously described, location control is the ability for one object to constrain the movement of another. Location control arising through some form of enclosure of one object inside another, what Garrod et al. [8] refers to as 'fcontainment', is seen to be salient for the preposition 'in'.

The notion of *support* may be considered as a particular type of location control which is constrained to the vertical direction — X supports Y if X resists the acceleration of Y due to gravity. Support is most often associated with the preposition 'on', as we have already seen in many of the examples.

Finally, the prepositions 'over' and 'under' appear in some instances to encode a functional relationship of *covering* or *protection*. This sense of covering does not simply reflect a geometric relationship but is also concerned with properties and affordances of the figure and ground objects in a given context. A good example of this is provided in [10], see Figure 2.8.

Participants in general gave higher ratings to 'over' and 'under' in scenes where objects fulfilled a protecting function. For example, in (a) the man is protected from the rain by the umbrella, but he is not in (b); therefore (a) is a better instance of 'the umbrella over the man'/'the man under the umbrella' than (b).

Though functional relationships appear to have a significant influence on the usage of some spatial prepositions, geometric relationships are still clearly influential and often even more so. The degree to which functional aspects influence the semantics differs for each preposition — some prepositions, e.g. 'in' and 'on', are more functionally biased while others, e.g. 'above' and 'left of', are more geometrically biased. Though some

(a) An umbrella protecting from the rain (b) An umbrella not protecting from the rain

Figure 2.8: Examples of functional interaction from [10].

prepositions are functionally biased, clearly they are still affected by spatial constraints and also geometrically biased prepositions are affected by functional aspects [45].

Even though the distinctions may not be clear-cut, I think it is informative to split some prepositions into two classes — functional and geometric, as in Section 1.2. The functional prepositions have a strong functional component as well as a geometric reading which is associated with a corresponding geometric preposition. For example, the preposition 'in' appears to have a functional component of location control and a geometric component of geometric containment which is associated with the preposition 'inside'.

This is not to say that the geometric prepositions are purely geometric. For example, in [10], though the effect was weaker than for 'over' and 'under', the functional interactions were also found to influence the prepositions 'above' and 'below'. We will also see similar effects for 'on top of' and 'inside' in Section 4.4.3.

Moreover, as noted by an anonymous reviewer of [3], context and phrasing may coerce a more geometric reading of a functional preposition and vice versa. For example, 'the box is *partially* under the table'.

Object-Specific Features

As opposed to functional relationships concerned with the physical interactions between objects, it is apparent that various object properties and affordances influence the

usage of spatial prepositions [31, 32]. For example, the animacy of the figure object may influence a decision to use 'in' or 'on' [32].[8] Following [31], we call these kinds of features 'object-specific' features.

In each of the experimental studies that have considered the influence of object-specific features [5, 31, 32], the measured effects of these features has been regarding their role in judgements when participants must label a configuration with a preposition i.e. in *categorisation* tasks; and it is not clear to what extent object-specific influence decisions when interpreting spatial language or assessing *typicality*. The notions of categorisation and typicality will be discussed further in Section 2.3.5 and the role of object-specific features will be analysed in more depth in Chapter 6 where it is hypothesised that these types of features provide a source of disagreement between category and typicality judgements. Here we will provide some examples of salient object-specific features for each of the functional prepositions.

In As 'in' expresses a notion of *containment*, the ability of the ground to contain the figure is often salient whether or not the ground does in fact contain the figure in a geometric sense. Therefore, whether the ground is a type of container appears to be salient for 'in' [5, 31, 32] and this may be considered a salient object-specific feature.

Over/Under The 'covering' sense of 'over' [30] appears to be closely related to the functions of the figure and ground [46]. For example, a covering object like a lid may exhibit this sense of 'over' when covering a container. Therefore, whether or not the figure is a covering object or the ground is a type of container may be salient object-specific features.

There is also a non-covering sense where a specific functional interaction exists between part of the figure and ground. For example, a tap may be 'over' a sink if only the spout of the tap is above the sink. Similarly, an object may be 'under' a lamp when the object is not under the lamp in a geometric sense but the light from the lamp shines on the object. These specific functional interactions rely on particular properties of the figure or ground and so we consider them to be object-specific features.

Relating to the functional interactions of the figure and ground, an intermediary object between the figure and ground may serve to block any functional interaction, as

[8] People were found to prefer 'in' when describing an inanimate figure (a coin) and 'on' when describing an animate figure (a firefly).

studied in [31], and diminish the effect of any object-specific features which are present.[9]

Against 'against' is commonly used to denote contact between two objects and, as argued in [29], is more applicable in situations where the ground object is fixed and the figure is mobile. For example, one may describe a chair as being 'against a wall' but it would be odd to describe a wall as being 'against a chair'.

On 'on' is ubiquitous in the English language and is applied to many situations where usually at least one of the following hold: the figure is supported by the ground, the figure is above the ground or the figure is in contact with the ground. As a result, it is not clear that there are particular properties of figure or ground objects at table-top scales which create strong preferences for 'on'.

As discussed above, the preposition 'in' is often preferred when the ground object is a container. 'on' is therefore used less frequently in these scenarios [32], even though the physical relationships between the objects often fulfil the requirements for 'on'. As a result, whether or not the ground is a container appears to be a salient object-specific feature for 'on'.

Finally, 'on' may be used to denote attachment of the figure to the ground. It is therefore plausible that, similarly to 'against', 'on' is more applicable in situations where the ground object is fixed relative to the figure.

One may argue that some object-specific features appear to be salient for a given preposition simply because they often co-occur with salient geometric features. For example, objects which are containers naturally have containing parts and therefore objects in typical scenes often relate to containers via containment. However, in the study of [31] when shown exactly the same scene of an object on a plate/dish, participants describe the configuration as 'in' when the 'plate' is labelled as a dish, and 'on' when it is labelled as a plate.

Single Domain Thesis

As we have seen, extra-geometric information appears to influence the usage and understanding of spatial prepositions. At first glance, this appears to present a tension for accounts which rely on rooting the representation of spatial prepositions in a geometric

[9] Note however that no significant effect of this type of blocking was found.

domain, or any single domain as in [47]. This tension, however, seems to simply be a matter of at what level extra-geometric information is processed e.g. we may take 'in' to represent geometric containment and that location control appears to become salient due to some abstract transformations (as discussed in Section 2.2.2), or on the other hand 'in' may represent a degree of both containment and location control.

To give an example, let us revisit the study of [10], where an umbrella is providing protection from the rain.

(a) An umbrella protecting from the rain (b) An umbrella protecting from the rain (re-oriented)

Figure 2.9: The umbrella over the man.

Suppose that the image in Figure 2.9(a) provides a good instance of 'the umbrella over the man'. If we allow spatial prepositions to encode multiple domains, this instance may be simply explained by saying that even though the umbrella is not geometrically above and over the man, this provides a good instance of 'over' as 'over' encodes a functional relationship of protection which is apparent in the image. On the other hand, following a single domain thesis, one may say that 'over' encodes a geometric notion of covering from above and that the presence of the rain triggers a reorientation of the scene, as in Figure 2.9(b), where the image is rotated to reflect the salience of the

direction of the rain.[10] In this rotated scene, the umbrella is now geometrically covering the man and therefore 'over' is applicable.

Though the processes may be distinct, in both accounts the functional influence of the rain is salient and must be accounted for. From a computational perspective, to take the single domain approach would require modelling these discussed abstractions and transformations. However, it is not clear what guides these processes; take for example the use of 'on' for trains and 'in' for cars discussed in Section 2.2.2. Suppose that these conventions arise because in the process of schematization cars are imagined as containers while trains are imagined as flat platforms — therefore transforming the situations into appropriate geometric representations of the ideal meaning of each preposition. It is not clear why such abstractions arise and how one would construct a model to reliably perform this in a similar way to humans.

A more practical approach is to allow spatial prepositions to be represented in multiple domains and add extra-geometric information as features in semantic models. For example, in the context of table top environments we may add the functional feature of *location control* as well as geometric features capturing *containment*.

2.2.4 Polysemy

A significant portion of this thesis is dedicated to modelling the *polysemy* that spatial prepositions appear to exhibit. This phenomenon is well known in linguistics and it is pervasive in natural language [48]. The ability for terms to represent distinct but related meanings is unexplored in the work on grounded semantics and referring expressions, where even homonymy is rarely considered, as noted in [49]. The evidence from both philosophy of language and linguistics is that many terms display some degree of polysemy [50, 51], and following the theoretical literature one would expect that existing semantic models could benefit from accounting for this phenomenon.

The definition of polysemy is the subject of much debate in cognitive linguistics [52], and moreover the notion of polysemy overlaps with vagueness and ambiguity which may result in a varied theoretical treatment [12]. The purpose of this section is not to provide a definition of polysemy, but simply to introduce the notion and how it manifests in spatial language. In this section we will give some background on the nature of polysemy and some examples of the kind of polysemy that will be modelled.

[10] Usually an implicit salient direction is provided by gravity.

What is Polysemy?

As opposed to homonymy[11] where a term may express semantically distinct senses, a term is considered to exhibit polysemy if it denotes multiple *related* senses and we call these distinct senses *polysemes*. For example, the term 'wood' has the following senses (taken from WordNet [53]):

> Sense 1: the hard fibrous substance under the bark of trees
>
> Sense 2: the trees and other plants in a large densely wooded area

There appear to be two main contrasting accounts of how polysemous uses of a term arise — the under-specification account and over-specification account [48].

In the under-specification account, the meaning of a term is some abstract representation which is applied in context and the polysemous variation arises from mapping the representation to the context. For example, the 'ideal meanings' account of Herskovits [29] (Section 2.2.2) may be viewed in this way — the varied usages of spatial prepositions arise via schematization and sense and tolerance shifts applied to ideal meanings.

There is some disagreement regarding the extent to which semantic variation arising from under-specification may be considered polysemy, e.g. [30] propose that for a sense to be a truly distinct polyseme there must be instances of the sense where its meaning cannot be derived from the context (along with knowledge of the other senses).

In the over-specification account, the meaning of a term is composed of a collection of distinct senses. For example, in the radial category approach of Lakoff [54] the meaning of a term is a collection of categories which are organised around a basic or ideal sense. Such accounts are however criticised as they seem to require humans to store an unnecessarily large collection of varied senses for each polysemous term [48].

Following the example above of 'wood', an under-specified account may assert that 'wood' has a central meaning like 'relating to trees' from which a more specific meaning is generated in context. When interpreting the phrase 'going to the woods' it is easy to understand that 'woods' in this context refers to a collection of trees whereas in 'making it out of wood', 'wood' refers to the material derived from trees. In contrast, in an over-

[11] Homonymy denotes the capacity of a sign to convey two or more unrelated meanings e.g. 'bank' may refer to financial institution or river bank.

specified account reasoning isn't required in order to generate these interpretations as these two distinct meanings are stored and simply retrieved in context.

To highlight some of the difficulties of processing polysemous terms, let us consider the challenge of interpreting the meaning of the term 'bank'. A common approach to dealing with the semantic variability of terms, in particular in the field of Word Sense Disambiguation, is to take an over-specified account and begin by constructing an inventory of senses that a term may exhibit [55].

One must then make various choices about the granularity of the sense distinctions in the inventory. For example, an inventory may distinguish the following senses for 'bank':

> Sense 1: a financial institution that accepts deposits and channels the money into lending activities
>
> Sense 2: a building in which the business of banking [is] transacted
>
> Sense 3: sloping land (especially the slope beside a body of water)

each of which appear in the sense inventory WordNet [53]. A coarser sense inventory may not distinguish the polysemous Senses (1) and (2).

The challenge for a computational word sense disambiguation model is then to label instances of 'bank' with the correct sense(s). For example, consider the following statements:

a. 'I watched the swans swim by as I sat on the bank'.

b. 'I'm going to make a deposit with the bank'.

c. 'I'll wait for you outside the bank after depositing my cheque'.

The instance of 'bank' in (a.) is a relatively unambiguous case of Sense 3 and relatively simple models can be generated which can determine this by considering the semantic closeness of the senses of 'bank' to other terms appearing in the statement, e.g. Sense 3 $\sim$ 'river' $\sim$ 'aquatic habitat' $\sim$ 'swans'.

A similar method can be used to determine that 'bank' in (b.) is not a case of Sense 3. This is clearly an instance of Sense 1, the financial institution, but it is unclear to what extent it is an instance of Sense 2, the physical building. In (c.), however this may

be considered a case of both Sense 1 and Sense 2, which we are able to determine due to the implications of 'waiting outside'.

This approach to word sense disambiguation requires a sense inventory to be drawn up which meaningfully distinguishes different senses and provides the semantics of each sense. However, it is unclear that such an inventory can be generated for spatial prepositions in grounded contexts.

As opposed to homonymous senses, it is often the case that polysemous senses co-occur and deciding between polysemous senses can require more developed commonsense reasoning, as seen in the above example. As the senses of polysemous terms are so closely intertwined, the theoretical and computational treatment of polysemy presents a difficult challenge for semantic models.

Spatial Language and Polysemy

The polysemy of spatial prepositions is well recognised in the literature [29, 56] which includes both detailed analysis of the semantic variation of spatial prepositions, e.g. [30], and attempts to provide a formal treatment of them, such as [57, 58]. However, polysemy is rarely, if ever, accounted for in computational models for situated dialogue.

In Section 2.2 we have already provided some detail regarding the semantic complexity and variability of spatial prepositions. In this section we will outline how this previously discussed semantic complexity relates to the notion of polysemy and also demarcate the kind of polysemy that will be tackled in the rest of this thesis.

The Polysemy of 'on' As we have discussed in Section 2.2.1, the meanings of spatial prepositions may initially appear rather simple in situated environments, e.g. 'in' $\iff$ *containment*, however there are many instances of spatial prepositions where they don't appear to adhere to these ideal notions. This type of semantic variability underlies the type of polysemy we attempt to model in this thesis and here we will provide a brief overview of this kind of variability in the case of 'on'.

To begin, let's consider a simple definition of the preposition 'on' from [36]:

Definition 2.2.2

$on(X, Y) \iff$ *A surface of X is contiguous with a surface of Y, and Y supports X*

As seen in Section 2.2.1 however, 'on' may be appropriately used to describe a relationship between objects, X, Y, where X and Y are not in contact and where Y does not support X.

For example, X may be 'on' Y if it is (Sense 1) resting on top of it e.g. 'a book on a table' (Sense 2) attached to the side of it e.g. 'a clock on a wall' (Sense 3) simply in contact with it e.g. 'a balloon on the ceiling'.

To what extent are Sense 1, 2 and 3 distinct polysemous senses? One approach to answering this question is provided by the 'principled polysemy' approach of Tyler and Evans [30] — a sense is considered to be distinct from other senses if the following criteria are met:

1. The sense includes a non-spatial component which distinguishes it from other senses and/or where the spatial configuration is meaningfully different from other senses

2. There are instances of the sense where its meaning cannot simply be derived from the context along with knowledge of the other senses

Whether or not these senses satisfy the criteria is not immediately clear. For example, if we suppose that the meaning of 'on' is highly under-specified and that its main, primary sense simply requires that X is supported by Y, then Senses 1 and 2 would not satisfy Criterion 1 and would be instances of this 'support' sense. Sense 3 would appear to satisfy Criterion 1 as the configuration meaningfully differs from the 'support' sense i.e. there is no support apparent. Moreover, from the knowledge of this 'support' sense it is not clear that Sense 3 — a 'contact without support' sense — could be derived and therefore Sense 3 may be genuinely distinct from the other senses.

However, let's suppose 'on' has a more constrained primary meaning, similar to Definition 2.2.2, which specifies that X is in contact with Y, X is supported by Y and X is above Y. In this case, the spatial configurations of Sense 1, 2 and 3 do appear to meaningfully differ as there are features, e.g. *support* or the extent to which X is *above* Y, which are apparent in some of the senses but not in others and therefore each fulfil Criterion 1. It seems however that these senses do not fulfil Criterion 2 as each can be derived either from the primary sense or by simply relaxing the constraints of the primary sense.

Regardless of whether these senses constitute distinct polysemes in any particular theoretical framework, the semantic variability that arises from these senses will be important for semantic models to capture if they are to reliably use and interpret spatial language and it is these distinctions which are tackled in this thesis.

Clearly the distinctions being considered here are particularly fine-grained and are not concerned with the wider usages of spatial prepositions which may provide better examples of polysemy. For example, the phrase 'John is on TV' has little concrete spatial sense as, presumably, we are talking about a projection of an image representing John which is made by the TV. However, once abstracted via an appropriate schematization the spatial sense becomes clear — we imagine the image produced of John is a distinct entity which is contiguous with the TV. Furthermore, there also appear to be senses which are not so clearly derived from the spatial senses. For example, 'on' may be used to relate an entity with some state e.g. 'To be on alert' [59].

2.3 Modelling Semantics in Referring Expressions

A significant body of work has been dedicated to creating computational models for the tasks of Referring Expression Generation and Comprehension (REG/C), see [60] for a detailed overview. However, most of this work avoids expressions involving *vague* language i.e. where the extension (set of things that could be referred to) of lexical items are ambiguous. When vagueness is explored in REG, it is usually with respect to *gradable* properties whose parameters are clearly defined, e.g. height [61]. We explore the issue of reference using spatial language, where the semantics are not so clear and therefore a more thorough challenge is presented for semantic representations.

2.3.1 Pragmatic Accounts

In order to understand in more detail the requirements of a semantic model in the context of referring expressions, here we give a brief overview of recent approaches to modelling the pragmatics involved in referring expressions.

The 'Rational Speech Act' (RSA) model of [62, 63] has been a popular approach in recent years for attempting to model the pragmatics of referring expressions. Goodman and Frank [63] provide empirical support that Bayesian inference is a reasonable model for how a listener can recover a speaker's intended meaning in simple scenarios and

assumes that speakers act in a rational, helpful manner which maximises the informativeness of their utterance while avoiding costly (redundant) utterances. In the context of referring expressions [62], the probability that a speaker is referring to a particular object, r_s, given a particular referring expression, w, and context, C, is given as:

$$P(r_s|w,C) = \frac{P(w|r_s,C)P(r_s)}{\sum_{r' \in C} P(w|r',C)P(r')} \tag{2.1}$$

$P(r_s)$ refers to the probability that an object would be referred to independent of the actual utterance i.e. a measure of its *salience*. $P(w|r_s,C)$ refers to the probability that the speaker would utter w to refer to the object, calculated as:

$$P(w|r_s,C) = \frac{|w|^{-1}}{\sum_{w' \in W} |w'|^{-1}} \tag{2.2}$$

where $|w|$ is the number of objects in the context, C, that w could apply to (the extension of w) and W is the set of words that apply to the speaker's intended referent.

The calculation of $P(w|r_s,C)$ assumes that each word has a definite extension, $|w|$, in the context. However, such a crisp extension is not always available when w exhibits vagueness or ambiguity.

van Deemter [60] suggests that vague terms may be accounted for in such an approach by instead calculating $P(w|r_s,C)$ as a function of $P(r_s|w,C)$:

$$P(w|r_s,C) = \frac{P(r_s|w,C) \times P(w|C)}{P(r_s|C)} \tag{2.3}$$

where $P(r_s|w,C)$ is to be determined by the salience of r_s and how *typical* r_s is for w.

Mast et al. [64] propose a similar pragmatic strategy which aims to maximise both the *acceptability* and *discriminatory power* of a description. Acceptability is defined as $P(D|x)$: the probability of accepting D as a description when given object x; while discriminatory power is defined as $P(x|D)$: the probability of choosing object x when given description D. Acceptability is calculated by considering how well the description D fits the object x and this is achieved by calculating the semantic similarity of x to prototypes for concepts appearing in D.

Another probabilistic account is provided in early work by Horacek [65] where the author explores the various types of uncertainty that may arise in generating and interpreting referring expressions. One kind of uncertainty considered is the uncertainty around 'conceptual agreement' i.e. to what extent a concept is applicable to some

instance. This type of uncertainty is modelled as a probability in their given model. Similarly, Degen et al. [66] show that various phenomena can be explained in the RSA model when a continuous rather than binary semantics is used. Also, in the context of modelling spatial language for referring expressions, Spranger and Pauw [67] argue for a 'lenient' semantics which considers similarity of entities to concepts rather than strict concept membership.

There are of course referring expressions involving vague terms where the referent is definite and unambiguous. For example, imagine a setting with two men who are, say, 160cm tall and a third man who is 180cm tall. The utterance 'the tall man' refers to this third man in an unambiguous way even though the term 'tall' exhibits sorites vagueness and the speaker and listener may disagree which height constitutes 'tall'. To arrive at the correct referent, one does not need to have any measure of how well 'tall' fits any of the men, only that the third man is a better instance of 'tall' than the others.

As in [61], such referring expressions do not need to be treated in a probabilistic fashion. However, this relies on the vague terms being 'gradable' and there being a clear method of determining when one instance is a better instance of the term than some other instance e.g. the third man is more more 'tall' than the other two men.

In general, as one may expect, it appears that the semantics of vague terms ought to be modelled as a matter of *degree* in order to be incorporated in pragmatic accounts of REG/C. This also fits with the findings of Logan and Sadler [68], that humans provide natural gradations of the applicability of spatial prepositions in grounded settings.

2.3.2 Concept Representations

In Chapters 4 and 5 we explore how to model typicality judgements for spatial prepositions in referring expressions and take into account a variety of the phenomena discussed in Section 2.2.

A useful semantic model in this context is one that agrees with human judgements and we therefore believe that semantic models for this task should ideally reflect the cognitive processes humans use to understand and represent these terms. One way that we intend to achieve this is to base the semantic model on a cognitively plausible conceptual representation. Here we will outline various cognitive accounts of concept representation, which will be compared in Chapter 4.

Prototype Models Based on Rosch's Prototype Theory [28], prototype models assess typicality of an instance by measuring its semantic distance to the *prototype*, where the prototype is a central member of the category. In geometric representations where concepts are represented in a continuous n-dimensional feature space, the prototype is usually taken to be the geometric centre of the category [69]. In feature models where concepts are represented by sets of binary properties, this takes the form of family resemblance [70] where prototypical members of a category are those members with the most properties in common with other members of the category.

Exemplar Models In exemplar models concepts are represented by a set of exemplars — typical instances of a concept. Typicality in these models is then calculated by considering the similarity of an instance to the given exemplars [71, 72].

Conceptual Spaces A more recent approach that has been considered as a unification of both the prototype and exemplar view is that of Conceptual Spaces [73]. As with prototype models, typicality in Conceptual Spaces is often represented by the distance to a prototypical point or region in the space. This prototypical point or region is often taken to be the centre of the area represented by the concept [74, 75].

Rule-Based Models The previous conceptual models rely on a notion of semantic *similarity* either to a prototypical instance or collection of instances where some metric is used to model semantic distance. In contrast, one may take a more classical approach to the semantics of these terms and generate rules which capture their meaning.

Rule-based accounts of concepts usually provide necessary and sufficient conditions for when an entity may be considered an example of the concept, similar to a dictionary entry [76]. However, for some concepts, as with spatial prepositions, entities vary in the *degree* to which they are a member of a concept rather than simply being a member or not a member. One may nevertheless encode rules which account for this and assign a measure of how well some entity fits a concept, for example the models of [77, 78] in the context of spatial language.

Of course, one may create rules which replicate the above similarity-based models and so rule-based models are in this sense a generalisation of the above models. The important distinction here is how these models are generated.

Rule-based models often rely on expert intuition, as opposed to similarity models

which lend themselves more readily to being generated empirically. This intuitive approach can prove successful where the semantics of terms involve a small number of well-understood features. This is reflected in general in rule-based accounts of spatial language which tend to be limited to modelling a single geometric sense of a preposition.

Neural Representations Neural networks are biologically inspired and so at a lower level may be more cognitively valid than the abstract representations provided above. Though neural networks have proven extremely popular for many AI applications, this approach has not received much attention for modelling the semantics of spatial language in situated contexts [79, 80]. This is possibly due to a lack of suitable datasets for training [16, 81, 82].

Though machine learning and big data techniques for such commonsense tasks can be attractive, as I've discussed previously in [83], an over-reliance on trained methods can be problematic when dealing with the intricacies of natural language. It may be the case that with enough training data a neural network model creates an internal representation that is closely aligned with a satisfactory cognitive model. However, such models are likely to be highly context sensitive and subject to dataset bias [84–86], uninterpretable by humans and difficult to update on-the-fly. Part of the intention of this thesis is to better understand the nature of spatial language; and due to the black-box nature of neural networks this is an unattractive approach.

2.3.3 Features

In order for the conceptual representations we generate to sufficiently capture the semantics of the given terms we ideally aim to incorporate any features that may be considered salient. To this end, we will give a brief overview here of features that appear in existing computational models, outlining geometric and functional relations that are used to model the the prepositions considered in this thesis.

Geometric Features

Unsurprisingly, geometric features have been well covered in the field. The principal and most commonly occurring geometric features are:

- Contact [78]

- Distance [77, 78, 87–91]
- Overlap with projection from objects [77, 78, 88]
- Height difference [77, 78]
- Object alignment [77, 87, 89–91]
- Containment [77, 78, 88, 89]

Various subtle differences may exist between these features in semantic models e.g. distance between objects may be calculated between object bounds or centres of mass. Similarly, these features may be made more or less general; for example in [64] object alignment is measured by two separate features: centre point angular deviation and bounding box angular deviation. Also, simplifications are often made for computational reasons; e.g. calculations are often made using bounding boxes of objects.

Functional Features

As discussed in Section 2.2.3, spatial prepositions appear to encode functional notions. This aspect of spatial language, however, has not been much explored in computational models. The functional notions of *support* and *location control* are often cited as crucial for an understanding of the prepositions 'on' and 'in'; however there is very little with regards to how these features should be modelled. Regarding *support*, Kalita and Badler [92] do provide a crude interpretation but it is not clear how this would be implemented in practice. With regards to *location control*, there is some work which focuses on overlap with region of influence [90, 91, 93, 94] which could be considered as something like a proxy for location control, but other than this, the feature does not appear in existing work.

Qualitative Representations

So far in discussing features for semantic models, the feature representations have generally been quantitative in nature. However, a particularly popular approach for understanding and reasoning about space has been to model qualitative relationships. For example in the Region Connection Calculus (RCC) [95], the notion of connection, $C(x, y)$: 'x is connected to y', is modelled and axioms for connection are provided.

Though the expressiveness of such topological logics is limited [96], this notion of connection along with that of *convexity*[12] may be used to distinguish various types of containment.

The main advantage of qualitative representations is the ability to perform symbolic spatial reasoning and this approach has many potential practical applications including computer vision, geographical information systems and understanding spatial content in natural language [97]. In Section 2.3.4 we will see some examples of qualitative representations being used to interpret spatial language in text.

However, spatial prepositions appear to exhibit sorites vagueness and humans provide natural gradations of the applicability of spatial prepositions in grounded settings [68]. We therefore desire a semantic model which captures this variability and, as we have seen in Section 2.3.1, in processing referring expressions we require a semantic model which provides a measure of how *well* a configuration fits a particular spatial preposition.

As discussed in Section 2.2.1, it may be possible to achieve this to some degree while using qualitative representations e.g. by modelling different *types* of containment or support. However, such an approach would be unable to distinguish configurations which share the same type of containment, say, but where the containment differs by a matter of degree.[13] Moreover, extracting distinct qualitative notions from scenes, e.g. 'contained in the convex hull' vs. 'contained in the containable inside', is particularly difficult compared to measuring the degree to which two bounding boxes overlap.

2.3.4 Interpreting Spatial Language

Modelling the semantics of spatial language is important in a variety of domains and tasks. However, approaches to modelling spatial language in one domain are not always applicable to the domain explored in this thesis. In particular, much work has been done regarding mapping spatial language to some semantic representation where generally the ground truth is limited. We will provide a brief overview of these tasks here and outline that though these tasks are related to modelling spatial language for referring expressions, there are some important distinctions.

[12] A region, X, is convex if for any two points in X, the straight line joining them is also in X.

[13] It may be possible to employ a two-level approach where configurations are assessed qualitatively first and then quantitatively, and it is plausible that this is a human-like approach to the problem.

Clearly an important challenge for the field of Artificial Intelligence is to model and process the semantics of text where no ground truth is given; and there is much work on the topic, see for example the range of works attempting to tackle the Winograd Schema Challenge [98]. The precise meanings of a spatial preposition in these instances is often unclear and the spatial preposition may only be used to invoke some general notion or even be little more than a syntactic placeholder. For example, in the Stanford SNLI corpus [99] we see:

> Text: A black race car starts up in front of a crowd of people.
>
> Hypothesis: A man is driving down a lonely road.

One does not need to have a precise understanding of 'in front of' in order to determine the validity of the hypothesis in this case.

Bateman et al. [43] recognised a tendency to over-committed interpretations of spatial terms in the field. While accepting that spatial terms can be interpreted precisely to allow for inferences to be made, Bateman et al. highlight that this is only possible with appropriate contextualisation. They argue for a 'two-level semantics', first assigning a pre-contextualised, linguistic, general meaning to spatial terms (using a spatial extension of the GUM ontology [100]) before assigning a precise meaning appropriate to the context.

It is generally recognised in work related to mapping text to a semantic representation, e.g. [101, 102], that the representation must be sufficiently under-specified in order to accommodate the variability of natural language terms. In the context of spatial language, qualitative spatial representations seem at first glance to provide such an under-specified method of representation for representing and reasoning with spatial language in text. There is much work dedicated to formal qualitative representations and their relation to natural language, e.g. [103–105].

To consider a particular example, in an attempt to provide a logical framework for handling polysemy, Rodrigues et al. [57] give an in depth study of the semantics of 'in' and explore the polysemy that it exhibits. In their framework possible interpretations of 'in' are formally defined based on abstract concepts and qualitative spatial relations. Each interpretation is formed of a range of components, for example one interpretation may be that the figure is contained in a container medium where the figure is a solid object and the figure is partly or fully geometrically contained in the ground. An

algorithm is then presented which maps input sentences to a set of plausible interpretations. This work highlights how object roles and types may affect preposition usage and also the variety of senses that 'in' may represent. However, as is the case with many such text-based tasks, due to the lack of ground truth it is not clear exactly when the algorithm is correct and there is a tendency to generate over-committed interpretations of the language. Herskovits [29] provides the example of 'the nail in the box' which clearly displays the ability for a phrase with no physical context to have an ambiguous geometric representation — the nail may be 'in' the box following the usual role of nails being in things *or* the usual role of boxes in containing things. Moreover, for the current purposes it is not clear how the framework could be exploited to aid in referring expression tasks.

Similarly, there are various tasks and implementations regarding processing spatial language in descriptions of images, for which there are numerous datasets, e.g. [86, 106]. In such tasks, a ground truth is provided in the image though there are still challenges for verifying semantic interpretations of the language.

For example Kordjamshidi et al. [107] outline such a task which comprises extracting spatial information from text. The descriptions to be interpreted are descriptions of images given in the IAPR TC-12 image Benchmark [106]. As opposed to earlier task specifications given for the SemEval series [108, 109], in this task more fine-grained interpretations are necessary which map topological spatial prepositions to RCC8 [95] and directional prepositions to 'left, right, above, below, back and front'.

The ground truth provided by the images can provide partial verification of any mapping. However, as we have seen in Section 2.2 spatial prepositions exhibit a high degree of semantic variability and there may be many plausible interpretations of the language. For example, so far in this thesis we have seen examples of 'in' which, in 2D, are instances of the RCC8 relations **PO** (black square and bowl in Figure 2.3), **TPP** (orange balls and bowl in Figure 2.3) and **DC** (pear and bowl in Figure 2.4). Similarly, Bennett and Cialone [104] have highlighted the various ways spatial terms may be interpreted in RCC8.

In general when interpreting the spatial information encoded in an expression containing spatial language, there are many plausible interpretations within a given semantic representation. Such semantic representations seem to provide collections of what a spatial preposition *could* mean, and this may be useful in certain tasks but it

is not clear how these may be exploited in processing referring expressions.

In contrast, the ground truth of a referring expression is the intended referent rather than the ambiguous semantic content of the language. Interpreting referring expressions requires interpreting spatial information in order to arrive at a definite answer — the intended referent of the speaker.

2.3.5 Categorisation and Typicality

So far when discussing the semantics of spatial prepositions the notions of categorisation and typicality have been conflated. However, it may be the case that these notions ought to be modelled separately for referring expressions. Here we will provide some background on this issue, which will be explored in the context of spatial prepositions in Chapter 6.

We suppose that a category decision is when an entity is labelled with a category or concept and, though priming and context may certainly be factors, the judgement is not made in direct comparison with another entity. For example, a categorisation judgement occurs when an agent is asked whether 'the apple is *in* the bowl'. In order to reply, the agent judges the membership of the instance in the relevant category and the wider context plays a relatively minor role.

Typicality usually refers to the extent to which an entity is a good example of a concept — how similar it is to some ideal conceptual representation. In our studies we ground the notion of typicality in comparison and preference — an entity, x, is more typical of a category than entity y if, when x is compared with y, people in general pick x as a better category member. Note that here we use typicality to refer to how *well* an entity fits a concept, rather than simply frequency of occurrence.

For example, imagine a table-top scene containing an orangey-red ball, o, and a red ball, r. Suppose an agent utters to a listener 'the red ball'. If they use this utterance to refer to o, they would be flouting the Gricean Maxim of Manner [110], as by committing to o being red they are also committing to r being red and therefore making an ambiguous description. We would therefore usually assume, or make the *conversational implicature*, that they are referring to r. What is important here is that r is closer to an ideal and generally agreed on notion of 'red'.

By assessing typicality in this way the notion of typicality is distinguished from graded category membership. The typicality data that we have collected does not arise

from graded membership judgements where study participants are asked to assign a value of how well the concept fits the category, e.g. in [111, 112], but instead from tasks in which participants select the best fitting instance from a given description.

This also differs from much work on typicality, e.g. [112], where the typicality of subconcepts are compared with respect to some concept — for example in determining if a robin is a more typical bird than a penguin. In contrast, in the current thesis we are considering instance-level typicality where the typicality of situated entities are compared for a given concept.

Following various criticisms of definitional representations of concepts in human cognition, e.g. [113], Rosch's Prototype Theory [28] provided an account based on family resemblance which does not presuppose that concepts have necessary and sufficient conditions for making category judgements. By relying on a degree of resemblance to some prototypical notion of a concept, category membership in this account may be treated as a matter of degree. With such an account it becomes appealing to conflate the notions of categorisation and typicality — the more an entity resembles a prototype the more likely it is to be labelled with the category *and* the more typical it is. There are however various accounts of concept analysis which suggest that category and typicality judgements are fundamentally different.

Smith et al. [26] consider the influences of category decisions and propose a model to account for experimental findings. Central to the model is the 'characteristic feature assumption', which supposes that features vary in the extent to which they define a concept. Smith et al. suppose that there is a distinction in types of features — 'defining' features which strongly influence category judgements and 'characteristic' features which strongly influence typicality judgements — and give the example of 'robin' to illustrate this. For the concept 'robin', 'have wings' is an important defining feature relating to the categorisation of an entity as 'a robin', while 'perches in trees' is a characteristic feature which relates to how typical an entity is of 'a robin'.

Rips [27] argued that categorisation of some entity is more than simply a judgement of how similar the entity is to some typical representation of the category. Rips provides support for this in an experiment where participants are asked to imagine an object of a given size and are asked which of two concepts, A, B say, the object is more similar to and which the object is most likely to be. For the given concept pairs, e.g. 'pizza' and 'quarter', the hypothetical object may be considered more similar to one of the

concepts yet more likely to be the other. In the case of the pizza and quarter, a round object with a three inch diameter is regarded as more similar to a quarter as pizzas are rarely so small, but more likely to be a pizza as the size of a quarter is generally fixed and is less than 3 inches.

This issue is explored further by Osherson and Smith [114] where it is again argued that concept membership and typicality are distinct phenomena. Based on a study in [115], Osherson and Smith argue that the notions differ using the seemingly extreme example of the concept 'red', even though it may be hard to imagine distinct defining and characteristic features for a concept with such simple semantics. However, the noted difference in judgements arises as people recognise particular wavelengths of light as being unambiguously red yet less typical than prototypical red. This difference in judgements preserves a monotonic relationship between category and typicality judgements i.e. if an entity, x, is a better category instance than y, then it is not less typical than y.

We believe that such monontonicity results offer a trivial case which may be explained without any fundamental modification of the underlying semantics or how they are processed. In the simple case of the colour red, we may represent the semantics in both categorisation and typicality judgements by considering the *dominant wavelength* — an instance is more or less typical and a better or worse category instance based on the similarity of the dominant wavelength to prototypical red and the distinction in category and typicality judgements is explained via a threshold which is applied in category judgements.

However, in the case of spatial language, the semantics are more complex and are influenced by a variety of geometric, functional and object-specific features. We believe that as a result, the relationship between typicality and categorisation may in fact be non-monotonic — there may be instances of a spatial preposition, i_1, i_2, such that i_1 is a better category member but less typical than i_2.

With regards to existing models of spatial language and models of referring expressions more generally, it is generally assumed that the underlying semantics of categorisation and typicality are essentially the same. However, any distinction in category and typicality judgements may have important ramifications for how referring should be processed. This will be discussed further in Chapter 6.

2.4 Related Models

There is a vast body of work concerning the semantics of spatial language and how they should be modelled. In this section we will provide an overview of attempts to model spatial language in grounded settings.

One approach to modelling the semantics of spatial prepositions has been to generate rules which capture their meaning, for example [77, 78]. One advantage of rule-based models is the ability to precisely explore and incorporate a particular aspect of spatial language. For example, Platonov and Schubert [78] provide a rule-based computational model of spatial prepositions which encodes various senses of the terms and also aims to account for synecdoche[14] by tagging and iterating over 'salient parts' of objects. As an example, the canonical sense of the preposition 'on' is measured by the extent to which the figure is above and touching the ground and the model also checks if this sense of 'on' applies better to any of the 'interactive parts' of the ground.

These models, however, largely rely on expert intuition to generate rules and, as a result, such approaches often lead to over-simplified representations which are susceptible to the pitfalls associated with the simple-relations model, as discussed in Section 2.2.1. For example, in [78] 'in' is simply measured using geometric containment.

The early work of Abella and Kender [77] aimed to provide a computational model of spatial prepositions which accounts for (sorites) vagueness. The underlying representation, similar to that of a Conceptual Space, represents the semantics of spatial prepositions in a multidimensional geometric feature space where 'ideal regions' are defined for each preposition by way of constraints on the space. Due to the inherent vagueness of spatial prepositions Abella and Kender argue for 'fuzzification' to be incorporated in the model, which is achieved using *fuzzy sets* [116] — the 'ideal region' for each preposition is a crisp set and a fuzzy set is generated by measuring distance from the ideal region.

Using this approach, they define the prepositions 'near', 'far', 'inside', 'above', 'below', 'aligned' and 'next' based on physical properties such as object area, centre of mass and elongation. For example, the ideal region of the preposition 'inside' is defined

[14] A synecdoche is a phrase in which a part is used to refer to the whole, or vice versa. For example, in the context of spatial language, one may say 'the car is under the bridge' to communicate that the car is geometrically under the platform part of the bridge rather than the bridge as a whole (including its supporting pillars).

such that the bounding box of the figure is fully contained in the bounding box of the ground and any such configuration is considered an unambiguous member of the concept 'inside'. This appears to be a plausible account of how sorites vagueness may be modelled for the ideal senses of these terms, and is similar to the account of [47]. However, on their own such definitions do not capture the variability expressed by these terms.

The idea that spatial prepositions are associated with regions for which the prepositions unambiguously apply and that deviations from the acceptable region cause the terms to be less applicable is central to many models, e.g. [117, 118], and was largely popularised in [68]. Logan and Sadler [68], focusing on projective prepositions in simple 2D grid scenarios, provided experimental data suggesting that humans fit 'spatial templates' to objects defined by 'good', 'acceptable' and 'bad' regions of acceptability. These regions provide natural gradations of the applicability of spatial prepositions centred around the 'good' acceptability region and a focus of subsequent work [94, 118, 119] has been to quantify the deviation in acceptability.

There have been various differing approaches to modelling this deviation. For example, Kelleher and Costello [91] consider how this deviation in acceptability may be influenced by various contextual factors. They begin with the idea of a 'potential field' — providing a measure of acceptability of a preposition for each point in space — which also incorporates a measure of salience of the ground object and is then modified considering the effects of other potential ground objects in the scene, providing a 'relative potential field'.

Mast et al. [64] model this deviation in acceptability of spatial prepositions using Prototype Theory, where a prototypical point is given in a feature space and the acceptability is measured by the distance from the prototype. Mast et al. take this approach in developing a semantic component of a dialogue system to tackle problems involving referring expressions. The use of a prototype in a feature space rather than spatial template means that the semantics are not constrained to simple geometric features. However, as with the majority of work on computational models of spatial prepositions Mast et al. focus on modelling projective prepositions (in particular, 'left of', 'right of', 'in front of' and 'behind') and the features considered are 'centre point angular deviation', 'bounding box angular deviation' and 'physical distance'. In Chapter 4 we will extend this approach to model the prepositions considered in this thesis as well as

provide an empirical method for generating the model parameters from data.

Much work on modelling spatial language is focused on modelling projective spatial prepositions, e.g. [64, 118–122]. Clearly this is relevant to this thesis as we are attempting to model 'over', 'under', 'above' and 'below'. However, often this work only considers a simple geometric representation of the terms and is focused on the pragmatic and/or grammatical complexities that arise, e.g. [64, 118, 121, 122]. While exploring the problem of representing projective prepositions using spatial calculi, Hois and Kutz [121] highlight the following common pragmatic considerations for interpreting these terms:

1. Position and orientation of speaker. For example, in 'in front of me' one must know the position and orientation of the speaker in order to locate the figure.

2. Reference system. For example, in 'to the right of that table' a decision must be made about which reference frame to adopt. This could be intrinsic if the table has an obvious front, but most likely will be relative in this utterance.

3. Various aspects of domain-specific knowledge. Most commonly 'intrinsic fronts' e.g. in 'the bike in front of the bus' it is necessary to understand which part of the bus is considered its front.

4. Dialogue history. For example, when the speaker makes a clarification such as 'no, further to the right'.

These considerations are often the focus of work on computational accounts of projective terms. For example, Moratz and Tenbrink [118] create a computational model of projective prepositions for scenarios where a robot agent must identify an object based on locative descriptions given by humans. The semantic portion of the model is based on simple geometric features and the topics of analysis are pragmatic in nature e.g. how does the direction of view of the robot influence the choice of frame of reference?

It should be noted, however, that there are instances of quite detailed models for projective terms. For example, Regier and Carlson [94] provide the Attention Vector Sum (AVS) model of some projective prepositions. To determine the acceptability of a preposition in this model for a given configuration, a set of vectors is constructed from various points on the ground object pointing towards the figure object. These vectors are weighted by the 'attention' paid by the viewer to the origin of the vector and the vectors are then summed and compared to a canonical direction. For the preposition

'above', the closer this summation is to the upright vertical the more acceptable the configuration is (with some adjustments made for height).

The models discussed so far have been based on assumptions about the underlying conceptual model, either representing the semantics in the form of rules or as central acceptability regions from which semantic distance can be measured. However, various more recent modelling approaches have relied more on data and training while limiting the conceptual assumptions. Such approaches are appealing as it is a difficult challenge to generate rules or conceptual models which sufficiently capture the varied meanings of spatial prepositions.

For example, Doğan et al. [79] consider the problem of grounding spatial prepositions for human-robot interaction in scenarios where a robot must identify an object on a tabletop given a locative expression. To model the semantics of spatial terms, Doğan et al. train a 'Relation Presence Network' — a multilayer perceptron which takes feature values of a configuration as input and outputs, for each preposition, the probabilities that the spatial preposition is present. Similar work has been carried out for 3D blocks world environments by Yan et al. [123]. However, as we have discussed in Section 2.3.2, these types of representations are problematic and do not provide much insight with regards to a semantic analysis.

A different approach by Fichtl et al. [124] has been to train a model using a Random Forests algorithm [125] to classify configurations with spatial prepositions. This work is focused mainly on the pipeline of computer vision, converting point clouds into useful histogram representations. The spatial prepositions 'on top of' and 'inside' are classified in a binary manner as either present or not and given the vagueness exhibited by spatial prepositions, this is not a useful approach for referring expressions.

To conclude, in general there are various issues which are not covered so far in computational accounts of spatial prepositions. Firstly, functional features such as *support* and *location control* are not represented, though they are often cited as important. This will be explored in Chapter 4.

Secondly, the sorites vagueness exhibited by spatial prepositions is well recognised and is captured by most models. However, the conceptual vagueness, or *polysemy*, exhibited by spatial prepositions has not been addressed, with the possible exception of the preposition 'on' in [78]. The inclusion of polysemy in our semantic model will be explored in Chapter 5.

Finally, in general features in the models are *relational* and features specific to the figure or ground are not taken into account. The influence of such object-specific features will be explored in Chapter 6 in the context of their influence on categorisation and typicality judgements. A concrete solution to including these features will not be provided but we will provide some suggestions of how this may be achieved with our semantic model.

2.5 Datasets

Over a decade ago Barclay and Galton [81] highlighted the lack of, and need for, a comprehensive 'scene corpus' which encompasses all aspects of spatial language. Such a corpus would include a large variety of scenes and situations, e.g. spatial relationships, functional relationships, viewpoints of speakers and listeners, scales etc.. in order to provide a useful tool for training and testing spatial dialogue systems. Though there is a large number of datasets for related challenges in computer vision research,[15] there are very few concerned with spatial language. In this section, we will provide an overview of related datasets and provide motivation for conducting the studies described in Chapter 3.

In order to provide the semantic analysis desired in this thesis there are some requirements for any prospective dataset. Datasets should include a variety of relationships between objects, both functional and geometric, and a variety of objects which have associated object-specific features. Datasets should also include a rich set of features for each configuration, or allow the extraction of such features.

Regarding the construction and testing of semantic models, to generate the concept representations described in Section 2.3.2 we require a collection of preposition instances. To test the applicability of these concept representations for referring expressions, we require a separate collection of varied referring expressions which are grounded in some way e.g. where the referent is labelled.

There exist a variety of broad datasets containing images annotated with descriptions such as ImageCLEF [106], PASCAL [126], HuRIC [16], Visual Genome [127] and SUN09 [128]; and similar datasets restricted to spatial language such as SpatialSense [86]. However, extracting rich features from images is extremely difficult and hinders

[15] See the collected list of databases for computer vision research:http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm Date Accessed: 01/12/2020

any deep semantic analysis.[16] Datasets comprising 3D scenes are therefore desirable.

There are various instances of implementations for real world environments, particularly in robotic applications, for example [118, 131, 132]. However, the training and testing processes for such implementations are generally not useful in the current thesis. For example, there may be no method for reproducing the real world set up or the models are trained on video data which suffers from similar problems to 2D images with respect to feature extraction.

There are exceptions however, for example in [131] where a robotic system is evaluated using the Train Robots dataset [133]. Train Robots provides virtual scenes of a robot arm performing actions on different objects along with natural language descriptions of the actions. Human annotators were provided with pairs of scenes to annotate, an initial scene and a final scene, where the annotators were asked to provide appropriate commands for the robotic arm to produce the final scene from the initial scenes. As the focus is on robotic manipulation and object placement, the spatial terms in the study are being modified by verbs and so it is harder to provide in depth analysis of the semantics. For example, if somebody says 'put the green block on the red block', 'on' will most often be interpreted in its canonical 'on top of' sense and so from such a dataset we may not see the true semantic variability of these terms.

Similarly, there has been much recent interest in the interplay of vision and commonsense in the field of Visual Question Answering (VQA). For example, Gordon et al. [134] developed the IQUAD V1 dataset using the interactive 3D environment AI2-THOR [135]. The IQUAD V1 dataset provides virtual scenes and multiple choice questions for a situated agent to answer. The focus of the dataset is on interaction with the environment, so for example to answer the question 'is there a cup in the microwave?' an agent must navigate the room, find the microwave and see if a cup is in it. As a result, the dataset does contain a large number of instances of spatial prepositions. However, as the focus is on commonsense visual reasoning and not natural language semantics, the questions and ground truth answers are automatically generated and, moreover, the spatial prepositions used are limited to 'in' and 'on'. Automatic generation of questions including spatial relations is challenging, as also noted by Johnson et al. [136] when constructing the CLEVR VQA dataset. Therefore, datasets using automatically

[16] Recent advances in reconstructing 3D scenes from 2D images, e.g. [129, 130], may make this more plausible in future however.

generated spatial expressions are generally limited to unambiguous cases of the spatial terms, as is the case with IQUAD V1.

As a subset of VQA tasks, there has been recent interest in multimodal communication, primarily focused on *gesture*, in the interpretation of situated referring expressions [137–139]. Though this is a strongly related topic, the focus of this thesis is on the semantic content of spatial prepositions and we therefore desire data which does not include extra pragmatic factors, such as gesture.

There are various datasets relating to spatial language in referring expressions which are based on blocks worlds i.e. scenes containing simple geometric objects such as blocks and balls [17, 90, 91, 138, 140]. Such datasets may be easy to generate and allow researchers to test specific pragmatic or semantic issues of spatial language; however, as we would like to explore the influence of object-specific features such datasets are not appropriate.

As a result of this lack of data, similar research has relied on constructing small scale datasets [78, 89]. Using scenes from the Google Sketchup 3D warehouse Golland et al. [89] collect referring expressions and interpretations of referring expressions from human annotators. However, the provided data is limited to referring expressions and simple categorical judgements are not provided. Moreover, the dataset does not appear to contain the prepositions 'in', 'on top of', 'against' or 'over'.

In providing a rule-based model of spatial prepositions, Platonov and Schubert [78] provide a dataset which is similar to our requirements. Human annotators annotated configurations in screenshots of virtual scenes according to two tasks: a 'truth judgement task' and a 'description task'. In the truth judgement task, participants are asked to provide a response from the Likert scale ('Yes', 'Rather yes', 'Uncertain', 'Rather no' and 'No') assessing the degree to which a given spatial relation holds between two objects (e.g. 'Is the chair near the table?'). In the description task participants are given an object, by referring to the object label e.g. 'Where is Pencil 1?', and asked to provide a description of the location of the object. In this way the description task provides a collection of grounded referring expressions.

We have decided not to make use of this dataset as, firstly, the prepositions 'on top of', 'inside' and 'against' are not tested in this data. Secondly, the dataset is comprised of only 48 scenes with a large portion being blocks worlds. Finally, I believe that the question phrasing of the truth judgement task may prime a more geometric reading of

the prepositions. For example, reconsidering the many examples given above where 'in' is used to describe a configuration displaying no geometric containment (e.g. the pear and bowl), it is plausible that when asked 'Is the pear in the bowl?' people in general respond 'no' even though they may refer to the pear as 'the pear in the bowl' or label the (pear, bowl) configuration with 'in'.

This nevertheless represents a valuable groundwork for the data collection approach of this thesis, which we enrich by supporting tasks in 3D environments (instead of screenshots) where participants can navigate the scenes and select objects. We believe that such additional features may be important in providing more flexibility for the exploration of borderline configurations.

Since the beginning of the empirical studies contained in this thesis, these same issues regarding grounding spatial prepositions have been highlighted in the very recent paper of Goyal et al. [82]. Goyal et al. provide the Rel3D dataset in order to overcome some of these issues, which may be beneficial for future work.

To conclude this section, in general we have seen that there isn't a suitable existing dataset for providing an in depth semantic analysis of a range of spatial prepositions. Of course, this isn't to say that there are no existing in depth semantic analyses of spatial prepositions. There have been many experimental studies conducted over the past couple of decades in order to analyse particular aspects of spatial prepositions [8, 10, 32, 118, 141–143]. However, such studies are necessarily limited to test some hypothesis and therefore cannot be used to generate or test broad semantic models.

Where there are potential candidate datasets they are small scale and don't include all the prepositions we investigate in this thesis. Moreover, none of the discussed datasets contain representations of functional relationships which we will explore and include in the dataset provided in this thesis.

Overall, we find that there is a lack of detailed geometric, functional and contextual data which hinders the capacity to properly investigate the semantic complexity of spatial prepositions and provide a semantic model reflecting how they are used to achieve communicative success. As a result, we have created a new data collection environment and experimental studies, described in Chapter 3.

Chapter 3

Data Collection

Due to the lack of appropriate existing datasets, as discussed in Section 2.5, we have conducted three studies to collect data on spatial prepositions which are described in the following sections.

3.1 Preliminary Study

In order to begin exploring the semantics of spatial prepositions in grounded settings, suggest further directions and inform future studies we initially conducted a preliminary study. We give a very brief overview here and the study is discussed in detail in [5] and is archived in the Leeds research data repository.[17]

The framework[18] for the preliminary study is built on the 3D modelling software Blender.[19] Two distinct tasks were created — a Selection Task and a Description Task. In the Selection Task participants are given a preposition on screen and asked to select all figure-ground pairs in the scene which fit the preposition. The Selection Task was designed to efficiently collect large amounts of data regarding the semantics, with minimal pragmatic considerations. In the Description Task objects are highlighted and participants are able to type in a spatial description of the object. The Description Task provides data on how spatial prepositions are used in referring expressions. In both tasks, participants are given a first person view of a scene which they can navigate using the mouse and keyboard.

The use of virtual 3D environments allows for the extraction of a wide range of features that would not be immediately available in real-world or image based studies. A wide range of geometric features were extracted as well as the functional notion of *support*. Further, we extracted some object properties using the relational knowledgebase ConceptNet [144].

3.1.1 Issues and Insights

Some tentative insights were gained from the preliminary study which help to highlight the complexities of modelling spatial language [5]. Firstly, the ground of almost all configurations labelled with 'in' was a type of *container*.[20] This is somewhat unsurprising

[17] https://doi.org/10.5518/620

[18] https://github.com/alrichardbollans/spatial-preposition-annotation-tool-blender

[19] https://www.blender.org/

[20] We call an object a container if there exists an 'IsA' edge between it and 'container' in ConceptNet.

but suggests that object roles and properties as well as physical relationships between objects should be accounted for in semantic models. Secondly, we found that there was significant overlap between 'in' and 'on'. This highlights some of the complexity of modelling this language — borderline scenarios are common and often more than one preposition can be used in a given situation. Finally, *support* appeared salient for 'on' but not 'in'. This suggested that functional features ought to be accounted for but that *support* was not a good proxy for *location_control* in relation to 'in'.

The Selection Task relied on the thoroughness of participants in selecting all admissible configurations for each preposition; however we found that our scenes likely contained too many objects for this to be a reliable outcome for every participant. This hampered our efforts to provide any significant analysis.

When analysing descriptions given by participants in the Description Task many appear genuinely ambiguous. This could be as a result of participants not being fully aware of all the objects in the scene, not having properly read the instructions or that the aim of the task was not made clear enough. From this point we decided to restrict the potential pragmatic influences that may arise when collecting data on referring expressions.

Following the insights of this preliminary study, a new study was conducted which provides the main dataset of this thesis and which is described below.

3.2 Study on the Semantics of Spatial Language (Study 1)

In order to train and test typicality measures of spatial language (see Chapters 4 and 5), we collected data on spatial prepositions, again using 3D virtual environments, which is described in [1]. Collected data and details of the framework can be found in the Leeds research data repository.[21,22] The latest version of the data collection environment and code for analysis can be found on the GitHub repository.[23]

The data collection framework is built on the Unity3D[24] game development software, which provides ample functionality for the kind of tasks we implement. Unity3D

[21] https://doi.org/10.5518/764
[22] https://doi.org/10.5518/825
[23] https://github.com/alrichardbollans/spatial-preposition-annotation-tool-unity3d
[24] https://unity.com/

was used instead of Blender as the Unity3D game engine is more sophisticated and allows the resulting environments to be embedded easily in websites, which facilitated online studies.

3.2.1 Tasks

Two tasks were created for our study — a Preposition Selection Task and a Comparative Task. The former allows for the collection of categorical data with which models can be constructed and the latter provides typicality judgements on which the models can be tested.

Preposition Selection Task

In the Preposition Selection Task participants are shown a figure-ground pair (highlighted and with text description, see Figure 3.1) and asked to select *all* prepositions in the list which fit the configuration. Participants may select 'None of the above' if they deem none of the prepositions to be appropriate.

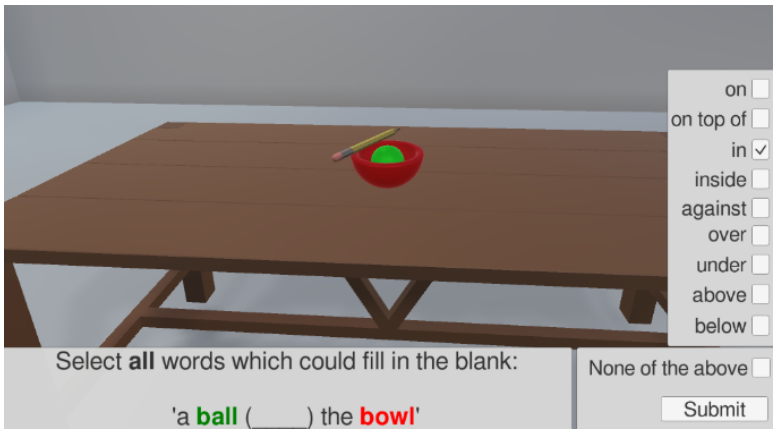


Figure 3.1: Preposition Selection Task

Often concepts are viewed as antagonistic entities; for example work in Conceptual Spaces is often concerned with comparison of categories, e.g. partitioning a feature space [145], and data collection for exemplar models is often presented as a choice *between* categories. We believe however that the vagueness present in spatial language is so severe that it is not clear that a meaningful model distinguishing the categories is possible, as also evidenced in the preliminary study. It is for this reason that in

the Preposition Selection Task participants are asked to select *all* possible prepositions rather than a single best-fitting preposition.

Comparative Task

In the Comparative Task a description is given with a single preposition and ground object where the figure is left ambiguous, see Figure 3.2. Participants are asked to select an object in the scene which *best fits* the description. Again, participants can select none if they deem none of the objects appropriate.

Figure 3.2: Comparative Task

This task is restricted compared to the Description Task described in Section 3.1 in a number of ways in order to limit pragmatic influences and allow a better semantic analysis. In this task rather than providing descriptions to identify a given figure, participants interpret the given locative expression by selecting a figure object. Also, the ground object is clearly marked so there is no ambiguity relating to the selection of the ground and, moreover, the resulting annotation provides an unambiguous configuration which can be compared with other configurations in the scene.

In both tasks, participants are given a first person view of an indoor scene which they can navigate using the mouse and keyboard. To allow for easy selection, objects in the scene are indivisible entities e.g. a table in the scene can be selected but not a particular table leg.

3.2.2 Scenes

For the study 67 separate scenes were created in the Unity3D editor in order to capture a variety of tabletop configurations. Each scene is a small collection of objects which provide test configurations for each task. For the Preposition Selection Task, configurations to test in each scene are predetermined and when a participant is tested on a scene they are tested on each of these configurations. For the Comparative Task, ground objects to test are predetermined and when a participant is tested on a scene they are tested on each ground object with a randomly selected preposition. All salient objects are made to be visible from the initial view of the camera.

3.2.3 Feature Extraction

The use of virtual 3D environments allows for the extraction of a wide range of features that would not be immediately available in real-world or image-based studies. In this section we describe the features extracted from scenes and used in our analysis. Exact details of how each feature is calculated are given in the data archive[21] and diagrams providing more details for each of the features are provided in Appendix A.

In our analysis we have represented in some form each relational feature discussed in Section 2.3.3, which we believe accounts for the majority of features given in computational models of spatial prepositions.

Geometric Features

Geometric features (distance between objects, bounding box overlap etc..) are in general simple to extract. We made use of eight geometric features:

- *shortest_distance*: the smallest distance between figure and ground
- *contact*: the proportion of the figure which is touching the ground
- *above_proportion*: the proportion of the figure which is above the ground
- *below_proportion*: the proportion of the figure which is below the ground
- *containment*: the proportion of the bounding box of the figure which is contained in the bounding box of the ground

- *horizontal_distance*: the horizontal distance between the centre of mass of each object
- *f_covers_g*: this feature takes the area of the figure and ground in the horizontal plane and measures the proportion of the area of the ground which overlaps with the area of the figure (with some adjustments made with respect to vertical separation)
- *g_covers_f*: As above, with figure and ground reversed

Some simplifications have been made in the calculations of these features. For example, we measured *contact* as the proportion of the vertices of the figure mesh which are under a threshold distance[25] to an approximation of the ground.

Functional Features

Building on the preliminary investigation, we explore the relationship between spatial prepositions and the functional features *support* and *location control* and consider how to extend existing semantic models to account for them.

We take *support* to express that the ground impedes motion of the figure due to gravity, while *location control* expresses that a horizontal movement of the ground causes a movement of the figure. As discussed in Section 2.3.3, useful methods of quantifying these notions in a given scene are not apparent. Rather than attempting to formally define these notions, as in [92, 146], we quantified these notions via *simulation* using Unity3D's built-in physics engine.

Support To assess the degree to which an object, G, gives *support* to another object, F; we analyse how F falls when G is removed from the scene by measuring the distance fallen, d, by the centre of mass of F. We would like *support*, S, to be 1 when F is fully supported and 0 when no support is apparent.

A simple way to achieve this is to normalise d by the height, h, of G and then limit S to between 0 and 1:

[25] The threshold distance used is the 'Default Contact Offset' used by Unity3D — when the distance between two objects is under the sum of the Default Contact Offset of the objects then they are considered to be in contact.

$$S = \begin{cases} \frac{d}{h}, & \text{if } d \leq h \\ 1, & \text{otherwise} \end{cases} \tag{3.1}$$

This works well in canonical cases where F is supported on top of the highest surface of G. However, this is not always the case e.g. if F is attached to the side of G. We therefore modify h to obtain a more appropriate normalising factor, h'.

h' is calculated as follows:

- If the bottom of F is above the top of G, then $h' = h$
- Else, if the bottom of F is above the bottom of G, then $h' = F_b - G_b$ where F_b, G_b are the lowest points of F and G respectively
- Otherwise, if the initial centre of mass of F is above the G_b then $h' = F_{\text{com}} - G_b$
- In all other cases $h' = h$

Note that there is still room for improvement e.g. this method may produce a value less than 1 when G fully supports F in the case that F falls onto another object which catches it. However, this method appropriately models many cases and this is supported by later results discussed in Section 4.4.3.

Location Control To assess the degree to which an object, G, gives *location_control* to another object, F; we analyse how F moves when forces are applied to G. We take four separate measurements, applying a force to G in the four cardinal directions, which are averaged. For each measurement, the horizontal movement of the centre of mass of F in the direction of the force is measured, this is then normalised by the movement of the centre of mass of G in the direction of the force. Again this value is limited to between 0 and 1.

Standardising Features

In order for the feature weights calculated in the following chapters to be meaningful and comparable, it is necessary to standardise the feature values. As in [75], we achieve this using the standard statistical method of z-transformation — where a calculated feature value, x, is converted to a standardised form, z, as follows:

$$z = \frac{x - \bar{x}}{\sigma} \tag{3.2}$$

where $\bar{x}$ is the mean of the given feature and σ is the standard deviation. In this thesis, where feature values are discussed or given in plots, the unstandardised values are given for readability.

3.2.4 Study

The study was conducted online and participants from the university were recruited via internal mailing lists along with recruitment of friends and family.[26] Each participant performed first the Preposition Selection Task on 10 randomly selected scenes and then the Comparative Task on 10 randomly selected scenes, which took participants roughly 15 minutes. Some scenes were removed towards the end of the study to make sure each scene was completed at least 3 times for each task. 32 native English speakers participated in the Preposition Selection Task providing 635 annotations, and 29 participated in the Comparative Task providing 1379 annotations.

As the study was hosted online we first asked participants to show basic competence. This was assessed by showing participants two simple scenes with an unambiguous description of an object. Participants are asked to select the object which best fits the description in a similar way to the Comparative Task. If the participant makes an incorrect guess in either scene they are taken back to the start menu.

3.2.5 Annotator Agreement

In order to assess annotator agreement we calculate Cohen's Kappa for each pair of annotators in each task, Table 3.1 provides a summary. Cohen's kappa for a pair of annotators is calculated as $\frac{p_o - p_e}{1 - p_e}$ where p_o is the observed agreement and p_e is the expected agreement. For the Comparative Task p_e is approximated, see the data archive[21] for details.

The observed agreement is significantly higher for the Preposition Selection Task, however chance agreement is higher in this task due to the distribution of responses — for a given preposition, participants were very likely to not select the preposition for a given configuration in our scenes. Expected agreement in the Preposition Selection

[26] University of Leeds Ethics Approval Code: 271016/IM/216. Participants were recruited without incentive.

Task	Shared Annotations	Average Expected Agreement	Average Observed Agreement	Average Cohen's Kappa
Preposition Selection	11880	0.78	0.893	0.718
Comparative	1325	0.566	0.766	0.717

Table 3.1: Summary of annotator agreements in Study 1

Task is therefore higher than in the Comparative Task and when we account for this using Cohen's Kappa, the average agreement of participants is very similar in both tasks.

Often semantic models are trained on and attempt to model the kind of categorisation judgements given in the Preposition Selection Task. Given the similarity of annotator agreement for both tasks, we conclude that it is also reasonable to attempt to construct a model which represents the kind of typicality judgements that are given in the Comparative Task.

3.2.6 Model Evaluation

While the Preposition Selection Task provides categorical data from each participant, the Comparative Task provides qualitative judgements regarding which configurations of objects better fit a description. We suppose that the configuration (figure-ground pair) which best fits a given description should be more typical, for the given preposition, than other potential configurations in the scene. We therefore use these judgements to test models of typicality — a model agrees with a participant if the model assigns a higher typicality score to the configuration selected by the participant than other possible configurations.

As there is some disagreement between annotators (see Table 3.1) it is not possible to make a model which agrees perfectly with participants. We therefore create a metric which represents agreement with participants in general.

Taking the aggregate of participant judgements for a particular preposition-ground pair in a given scene, we can order possible figures in the scene by how often they were chosen. This creates a ranking of configurations within a scene from most to least

typical for a given preposition and ground. We turn the collection of obtained rankings into inequalities, or *constraints*, which the models should satisfy. This provides a metric for testing the models.

As an example, consider an instance from the Comparative Task — a ground, g, and preposition, p, are given and participants select a figure. Suppose that there are three possible figures to select, f_1,f_2 and f_3, which are selected x_1, x_2 and x_3 times respectively. Let $\mathbb{M}$ be a model we are testing and $\mathbb{M}_p(f, g)$ denote the typicality, for preposition p, assigned to the configuration (f, g) by the model $\mathbb{M}$.

Suppose that $x_1 > x_2 > x_3$, then we want $\mathbb{M}_p(f_1, g) > \mathbb{M}_p(f_2, g)$, $\mathbb{M}_p(f_1, g) > \mathbb{M}_p(f_3, g)$ and $\mathbb{M}_p(f_2, g) > \mathbb{M}_p(f_3, g)$. Let's say that $x_1 = 10$, $x_2 = 1$, $x_3 = 0$. As the distinction between (f_1, g) and (f_2, g) is greater than for (f_2, g) and (f_2, g), it is more important that the model satisfies the first constraint than the last constraint. For this reason we assign weights to the constraints which account for their importance.

A constraint is more important if there is clearer evidence for it — if more people have done that specific instance and if the number of participants selecting one figure over another is larger. We assign weights to the constraints by taking the difference in the number of selections e.g. in the first constraint above, we would assign a weight of $x_1 - x_2$.

In this way we generate a set of weighted constraints to be satisfied. The 'overall' score given to the models is then equal to the sum of weights of all satisfied constraints divided by the total weight of all constraints. A higher score then implies better agreement with participants in general.

In the following we separate the scores given for each preposition in order to assess differences across the prepositions. We also give an average score across prepositions which is simply the sum of scores for each preposition divided by the number of prepositions.

3.3 Comparing Category and Typicality Judgements for Spatial Prepositions (Study 2)

In order to investigate the influence of object-specific features on typicality and categorisation judgements (see Chapter 6), we conducted a study which is described below. Collected data, details of the framework and results of the analysis can be found in the

Leeds research data repository.[27] The latest version of the data collection environment and code for analysis can be found in the Github repository.[23]

3.3.1 Tasks

The data collection framework is again built on the Unity3D[28] game development software and two tasks were created for our study — a Categorisation Task and a Typicality Task.

Categorisation Task

The Categorisation Task is a modified version of the Preposition Selection Task which will allow better comparison with typicality judgements, this is discussed in more detail in Chapter 6. In the Categorisation Task participants are shown a figure-ground pair (highlighted and with text description, see Figure 3.3) and asked to select *all* prepositions in the list which fit the configuration. Participants may select 'None of the above' if they deem none of the prepositions to be appropriate.

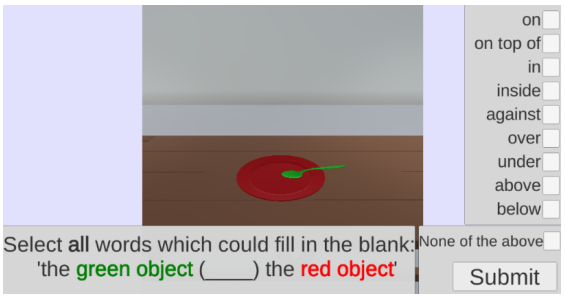


Figure 3.3: Categorisation Task

Typicality Task

In the Typicality Task participants are given a description and shown two configurations, see Figure 3.4. Participants are asked to select the configuration which *best fits* the description. Again, participants can select none if they deem none of the configurations to be appropriate.

[27] https://doi.org/10.5518/873
[28] https://unity.com/

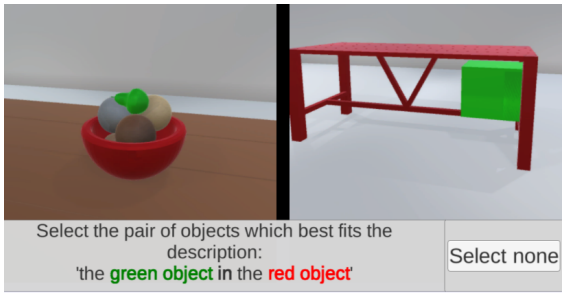


Figure 3.4: Typicality Task

In order to minimise differences in the tasks that may elicit different conceptualisations of objects in the scenes, the phrasing of the descriptions is the same in both the Categorisation Task and Typicality Task e.g. both tasks use the definite determiner 'the' and objects are referred to by their colour.

3.3.2 Scenes

We hypothesise that object-specific features strongly influence category decisions while the geometric ideals associated to the prepositions are more salient in typicality decisions. Scenes are therefore created for each preposition which vary the degree to which these object-specific features are present and also vary how similar the relational aspects of the configurations are to the geometric ideals associated with the given preposition. For example for the preposition 'in', we have a scene where the ground is a container and the figure is not very well contained in it and a scene where the ground is not a type of container but the figure is well contained in it. In this case, if the hypothesis is correct, we would expect a preference for categorisation in the former and a preference for typicality in the latter.

We have created 18 virtual 3D scenes, given in Appendix C, each containing a single highlighted figure-ground pair. Four scenes each were created for 'in', 'on', 'over' and 'under' and these scenes were also shared with their respective geometric counterparts: 'inside', 'on top of', 'above' and 'below'. Two scenes were created for 'against'. In the Typicality Task, participants compare scenes/configurations associated with the preposition given in the description.

3.3.3 Study

The study was conducted online and participants from the university were recruited via internal mailing lists along with recruitment of friends and family.[29] Each participant performed first the Categorisation Task on 6 randomly selected scenes and then the Typicality Task on 15 randomly selected scenes, which took participants roughly 5 minutes. 30 native English speakers participated providing 180 annotations in the Categorisation Task and 447 annotations in the Typicality Task.

Again, as the study was hosted online, we first asked participants to show basic competence. This was assessed by showing participants two simple scenes with an unambiguous description of an object. Participants are asked to select the object which best fits the description in a similar way to the Comparative Task. If the participant makes an incorrect guess in either scene they are taken back to the start menu.

3.3.4 Annotator Agreement

In contrast to the annotator agreements in Section 3.2.5, participants agreed substantially more in categorisation judgements than typicality judgements for which inter-annotator agreement is surprisingly low, see Table 3.2. This study contained a small number of purposely difficult comparisons for the Typicality Task, and it is possible that with a broader set of comparisons agreement would be higher in this task.

Task	Shared Annotations	Average Expected Agreement	Average Observed Agreement	Average Cohen's Kappa
Categorisation	7731	0.738	0.845	0.689
Typicality	982	0.476	0.683	0.413

Table 3.2: Summary of annotator agreements in Study 2

[29] University of Leeds Ethics Approval Code: MEEC 19-030. Participants were recruited without incentive.

Chapter 4

Comparing Cognitive Models

In this chapter we explore general issues of representing spatial prepositions for handling referring expressions, in particular when making typicality judgements. Various underlying conceptual representations, as discussed in Section 2.3.2, will be compared and the issue of generating appropriate parameters for these models from data will be addressed. We will also explore the utility of including the functional features discussed in Section 3.2.3.

The main outcomes of this chapter are the Baseline Prototype Model of spatial prepositions based on Prototype Theory, and methods for determining its parameters. In providing suitable methods for generating the model parameters from data, we allow for similar models to be constructed for concepts where the semantics may not be easily defined; and this will allow us to model the semantics of distinct polysemes in the following chapter. To demonstrate the suitability of this approach, three 'Simple Relation' models relying on expert intuition of the author and two other models which are generated using data from the Preposition Selection Task are set up to provide a comparison. We will conclude that the Baseline Prototype Model provides significant improvement over the other given models and also discuss the improvements given by a novel inclusion of functional features in our model. Finally, limitations of the model will be discussed.

4.1 Semantic Distance and Semantic Similarity

The models in this chapter rely on a notion of semantic distance to measure typicality and following much of the existing literature, e.g. [71], semantic similarity between two points x and y in a feature space is measured as a decaying function of the distance, $d(x, y)$:

$$s(x, y) = e^{-d(x,y)} \tag{4.1}$$

As is common, we take the distance, $d(x, y)$, to be the weighted Euclidean metric:

$$d(x, y) = \sqrt{w_1(x_1 - y_1)^2 + \cdots + w_n(x_n - y_n)^2} \tag{4.2}$$

where w_i is the weight for the i^{th} feature and x_i, y_i are values of the i^{th} feature for points x and y.

With the exception of the Exemplar model, each of the following models are then defined by a prototype and set of feature weights for each preposition:

1. $P = (x_1, ..., x_n)$ the prototype in the feature space
2. $W = (w_1, ..., w_n)$ the weights assigned to each feature

where typicality of a configuration, x, is then calculated as the semantic similarity to the prototype using Equation 4.1:

$$T(x) = s(x, P) = e^{-d(x,P)} \tag{4.3}$$

4.2 Simple Relation Models

In this section we outline some 'Simple Relation' models of spatial prepositions which will be used for comparison. The intention is for these models to replicate rule-based models which are often given in the literature.

A specific advantage of rule-based models is usually the ability to construct complex dependent statements, e.g. we could define in(X, Y) by saying that if Y is a container then either X must be at least partly contained in Y or Y must provide location control to X, and if Y is not a container then X must be at least partly contained in Y. This kind of definition is rare, however, due to the complexities which arise in attempting to handle the many varied situations that may occur.

One example of this comes from [78], where 'on' is defined using various rules which aim to capture different senses of the preposition. For example, where the ground is 'planar' (e.g. a wall), the ground is larger than the figure and the centre of the figure is above half of the ground's height, 'on' is measured as the extent to which the figure is touching the ground. The condition that the centre of the figure be above half the height of the ground appears somewhat arbitrary and highlights an issue with constructing these types of rules. Moreover, the final generated value of 'on' is the maximum value given by any such sense and this doesn't appear to account for the apparent hierarchy of senses that prepositions exhibit, which is discussed further in Chapter 5.

It should be noted that for some prepositions there are instances of quite detailed rule-based or template-based models which are not precisely replicated here. For example, [94] provide the attention vector sum (AVS) model of some projective prepositions, discussed in further detail in Section 2.4. Suitably reconstructing such a model

would be an intensive process and the model we provide below for 'above' would arguably provide similar results in many cases. In general, it is a challenge to accurately replicate other computational models as there are important differences in the features being used and some of the features involve an intensive process of labelling salient working parts of objects which may not easily translate into real environments.

Instead of reconstructing some of the more involved models, we provide more basic simple relation models here which are based on various rule-based accounts given in the literature [36, 77, 78, 92]. We have set up a simple geometric model and an intuitive best guess model which includes functional features.

For readability in the following model definitions, the models are specified by salient features and prototypical feature values for each preposition. The typicality in the models is then calculated using Equation 5.1, where the feature weights are 1 for each of the given salient features and 0 for remaining features. Note that the given prototypical feature values, e.g. 1 for *above_proportion* and 0 for *horizontal_distance* when specifying 'above' and 'over' in the Simple Model, are values prior to standarisation, but these values are standardised, as discussed in Section 3.2.3, in the actual models.

4.2.1 Simple Model

The Simple Model is based on what can be found in many computational models of spatial prepositions. 'in' and 'inside' are measured by *containment*; 'on' and 'on top of' are measured using *contact* and *above_proportion*; 'above' and 'over' are measured using *above_proportion* and *horizontal_distance*; 'below' and 'under' are measured using *below_proportion* and *horizontal_distance*; 'against' is measured using *contact* and *horizontal_distance*. For the full specification see Table 4.1.

4.2.2 Best Guess Model

The Best Guess Model is a copy of the Simple Model except we add in functional features for 'in', 'on' and 'against' and for 'over' we change *horizontal_distance* to *f_covers_g* and for 'under' we change *horizontal_distance* to *g_covers_f*. 'inside', 'on top of', 'above' and 'below' are the same as in the Simple Model. For the full specification see Table 4.2.

In general, for the Simple Relation Models the given prototypical feature values are limit values i.e. 0 or 1. However, in the case of 'against' the prototypical value of

	Salient Features	Prototypical Value
'in' & 'inside'	*containment*	1
'on' & 'on top of'	*contact*	1
	above_proportion	1
'above' & 'over'	*above_proportion*	1
	horizontal_distance	0
'below' & 'under'	*below_proportion*	1
	horizontal_distance	0
'against'	*contact*	1
	horizontal_distance	0

Table 4.1: Prototype feature values in the Simple Model

	Salient Features	Prototypical Value
'in'	*containment*	1
	location_control	1
'on'	*contact*	1
	above_proportion	1
	support	1
'over'	*above_proportion*	1
	f_covers_g	1
'under'	*below_proportion*	1
	g_covers_f	1
'against'	*contact*	1
	horizontal_distance	0
	location_control	0.5

Table 4.2: Prototype feature values in the Best Guess Model

location_control is presumed to be 0.5. This decision was motivated as if one imagines a typical instance of 'against', e.g. a bike leaning against a wall, it does not appear that the ground object fully constrains movement of the figure. We expect that movements of the wall towards and away from the bike significantly impact the position of the bike and that movements of the wall in other directions have a minimal impact. For this reason, the value of 0.5 was used. This particular case highlights the difficulties in general that exist in assigning prototype values.

4.2.3 Proximity Model

Finally, as a baseline we have created a Proximity Model which judges typicality based solely on *shortest_distance* — the closer two objects are, the higher the measure of typicality.

For every preposition the semantic distance to the prototype is given as:

$$d(c, P) = 0 - shortest\_distance(c) \tag{4.4}$$

This model is included based on the preliminary study [5] which indicated that judgements based solely on proximity may be relatively successful in interpreting referring expressions for some prepositions.

4.3 Data-driven Models

So far we have considered models which rely on expert knowledge. In this section we consider models which are trained on data from the Preposition Selection Task. Each of these data-driven models represent concepts in a feature space and we use all of the features given in Section 3.2.3 to represent each preposition.

4.3.1 Baseline Prototype Model

The main focus of this chapter is the Baseline Prototype Model. This model is based on Prototype Theory and typicality is calculated by considering the semantic distance to a prototype using Equation 4.3. Such a representation seems intuitively plausible for spatial prepositions, particularly if we are to follow the thesis that the meaning of spatial prepositions is structured around some sort of ideal meanings. If we suppose that the variety of usages of a spatial preposition arise from deviations from some ideal

meaning, then it makes sense to assess typicality based on its semantic distance from the ideal meaning or prototype.

As discussed in Section 2.2.2, ideal meanings may be deviated from in a variety of ways involving commonsense and pragmatic reasoning. Modelling the deviation from a single prototype in a feature space is essentially modelling only the 'tolerance shift'-type deviations and not the 'sense shifts' and is therefore a significant simplification of human usage of these terms. However, modelling such commonsense reasoning is notoriously difficult and would require, arguably, genuine intelligence. Improvements to this conceptual model which model some 'sense shift'-type deviations are, however, discussed in Chapter 5.

The underlying conceptual model and usage of Prototype Theory is not a new proposal for spatial language and follows [64, 67, 73, 147]. Of particular interest is the work of Mast et al. [64] where a pragmatic model is developed to tackle problems involving referring expressions. Mast et al. focus on projective prepositions (in particular, 'left of', 'right of', 'in front of' and 'behind') and as a result, the challenge of assigning parameters to the model is simpler and appears to be achieved via the researchers' intuition. We extend the approach taken by [64] to model a set of spatial prepositions whose semantics are not so clear and show that model parameters can be automatically determined from a small dataset using a simple regression-based methodology. By automatically generating model parameters we are able to include a wider variety of features in our models and provide support for a novel inclusion of functional features. The automatic generation of parameters will also be useful in Chapter 5 when we distinguish separate polysemes and attempt to model their semantics.

Learning Prototypes and Weights

The prototype, exemplar and conceptual space models in this chapter each rely on calculating a weighted semantic distance to some central instance or instances. However, it is not often discussed how the weights should be determined in practice, and methods for determining a prototype which rely on centrality within the concept appear unsatisfactory.

In order to generate prototypes and weights, firstly a 'Selection Ratio' is generated for each configuration (and preposition) based on how often participants would label the configuration with the given preposition in the Preposition Selection Task.

The weights in the semantic distance ought to represent how influential or salient each feature is in making typicality judgements. To determine the salience of each feature the selection ratio is plotted against the feature values. Using off-the-shelf multiple Linear Regression [148] we obtain coefficients for each feature which indicate how the selection ratio varies with changes in the feature. The feature weights are then assigned by taking the absolute value of the coefficient given by this linear regression model.

The method we propose for determining prototypes in the Baseline Prototype Model is based on a simple idea — that, rather than being *central* members of a category, prototypes should be learnt by extrapolation based on confidence in categorisation. It is hoped that this accounts for the possibility that many concept instances in the data will not be an ideal prototype. For example, there may be many instances for 'in' where the degree of containment is not 100% and in fact there may be no such instance of 'in' with 100% containment. However, if containment is a salient feature for 'in' and 'in' implies higher containment we ought to see that the higher the degree of containment, the more likely the instance is to be labelled 'in'.

In order to find the prototypical value of a given feature for a preposition we plot the feature against the selection ratio, then using simple off-the-shelf Linear Regression modelling [148] the feature value is predicted when the selection ratio is 1. Figure 4.1 shows the linear regression plot for some features in the case of 'on'. The blue cross denotes the prototype generated by the Conceptual Space model and the orange asterisk denotes the mean value of exemplars in the Exemplar model.

On inspection of the plots it is clear that the simple linear regression model is not well-suited to represent the data. This is in part because the individual features alone cannot sufficiently capture the semantics of the terms. For example, in the case of the feature *above_proportion* for the preposition 'on', there are clearly many possible cases where *above_proportion* is high but it is not an admissible instance of 'on' and vice versa (this can be seen by the line of instances along both axes in Figure 4.1). As a result, there is significant deviation from the linear regression. The linear regression, however, provides a simple and effective method for generating feature prototypes — we can see in Figure 4.1 that all salient features appear to be assigned appropriate prototypical values.

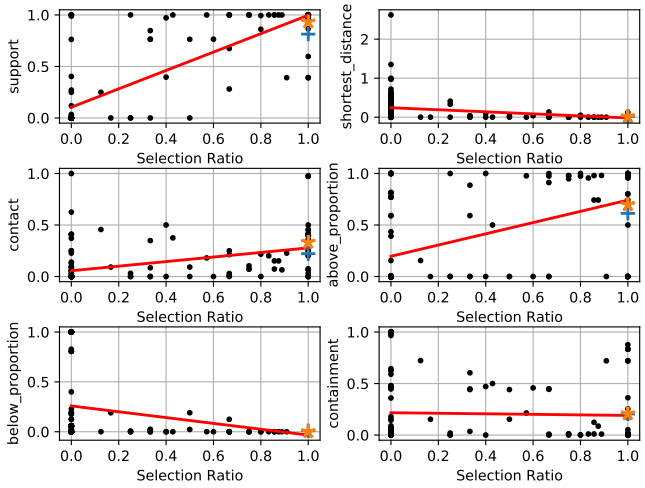


Figure 4.1: Finding prototypical feature values for 'on'.

4.3.2 Conceptual Space Model

Often Conceptual Spaces are constructed beginning with concept prototypes as a starting point, e.g. [149]. In such work, prototypes are usually presumed to be central within the concept representation and the conceptual space is accordingly constructed outwards. However, in this work we are concerned with finding prototypes which may be uncertain.

In order to replicate the Conceptual Space approach, we take the set of all possible instances of a given preposition (all configurations labelled at least once with the preposition) to provide an approximation of the conceptual region. Then, supposing that the prototype is central to the concept [150], we take the prototype to be the barycentre of the set of instances.

This is a very simple way to construct a Conceptual Space and there are other more involved approaches to constructing Conceptual Spaces. For example, a more common approach is to use Multi-Dimensional Scaling [151]. Such an approach, however, has the limitation that it cannot generalise to unseen inputs.

Regarding the calculation of similarity within the space, we assign feature weights using the weights calculated for the Baseline Prototype Model.

4.3.3 Exemplar Model

For the Exemplar model we first have to decide which datapoints can act as exemplars for a given preposition. Rather than considering all possible instances, we consider only instances that were always labelled with the preposition, these instances act as *typical exemplars*. In the absence of such instances, we take the next best instances as typical exemplars.

Typicality of a given point, $T(x)$, is then calculated by considering the similarity of the point to the given exemplars [71, 72]:

$$T(x) = \sum_{e \in \mathbb{E}} s(e, x) \tag{4.5}$$

where $\mathbb{E}$ is the set of exemplars. This is still reliant on having appropriate feature weights and again we assign feature weights using the weights calculated for the Baseline Prototype Model.

4.3.4 Training Paradigm

As we have outlined earlier in Section 4.3, the cognitive models are each trained on categorisation data from the Preposition Selection Task and are then tested on the Comparative Task — in effect the models are trained using *transfer learning*. In the case of the Exemplar and Conceptual Space models, disregarding feature weights for a moment, this is necessary as they are constructed from concept instances. The Baseline Prototype Model, however, simply requires a prototype to be constructed along with the feature weights and there may be many ways to achieve this. In particular, a supervised learning approach may be taken which refines the prototypes and weights based on performance in the Comparative Task.

It seems that such an approach, however, may find solutions which do not align with human usage of these terms. For example, it is possible to find a prototype and weights for the preposition 'inside' which performs perfectly on the Comparative Task and where the feature weight for *containment* is 0. Similarly, there are solutions which perform better than any of the models in this thesis for 'in', 'on' and 'above' where feature weights for the seemingly most important features are set to 0: *containment* and *location_control* for 'in', *support* for 'on' and *above_proportion* for 'above'.

4.4 Evaluation

In this section the performance of the models is evaluated using the metric described in Section 3.2.6.

4.4.1 Initial Results

As a preliminary insight, we generate models as described above using *all* the data from the Preposition Selection Task (139 configurations) and then evaluate the models as described in Section 3.2.6 using all data from the Comparative Task. As the tasks use the same scenes, some of the same configurations will be used for both training and testing and we therefore cannot be confident that the models are not over-fitted. Nonetheless, as discussed in Chapter 6, there may be a distinction between the kind of categorisation judgements made in the Preposition Selection Task and typicality judgements made in the Comparative Task and it is interesting to consider how well the models translate this categorical data into typicality rankings. See Figure 4.2 for initial scores.

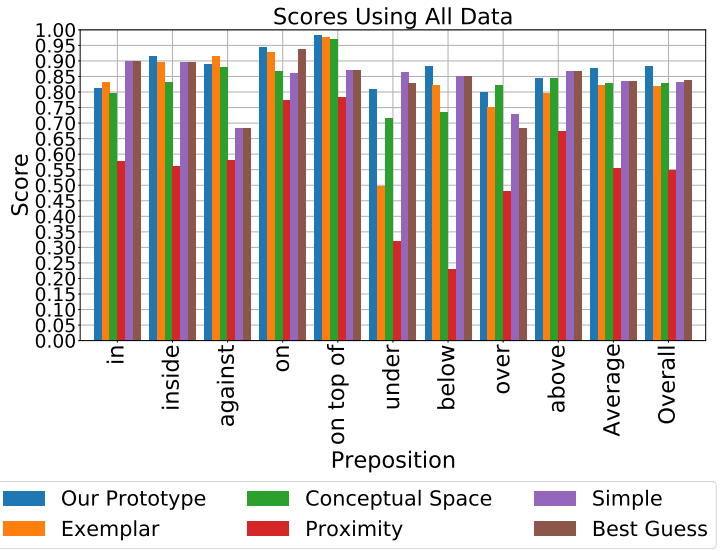


Figure 4.2: Initial Results: Scores using all scenes for both training and testing.

Regarding the Simple Relation models, the Best Guess and Simple models are quite similar, with the Best Guess model performing slightly better overall — adding functional features has significantly improved results for 'on' but has not changed 'in' or

'against'. In the case of 'in' this may be the case because, though *location_control* does influence the usage of 'in', it is difficult to generate situations where an object is the *most* 'in' another object without exhibiting any containment.

Though the 'topological' prepositions usually indicate proximity, we can see that proximity alone does not provide a reasonable measure of typicality for any of the prepositions.

Of the data driven models, the Exemplar model and Conceptual Space model have similar results overall while the Baseline Prototype Model appears to perform significantly better than all the other models. We however need to test how robust the models are to changes in training data.

4.4.2 K-Fold Testing

In order to test the ability of the models to generalise to unseen configurations of objects and compare robustness of the models, we created train-test scenes using K-fold cross-validation. We generate the models based on data from the training scenes given in the Preposition Selection Task and test the models using constraints generated from the testing scenes in the Comparative Task. We repeated this process 100 times and averaged the results, shown in Figure 4.3.

For the cross-validation $K = 2$ is used to provide the largest contrast with Section 4.4.1 and also simplify the testing process as for large values of k it may be common to generate folds which do not contain constraints to test.[30] $K = 2$ is generally preferable as it provides the largest testing sets while also maintaining disjoint training sets [152]. Note that the results with $K = 3$ are similar.

Firstly, the results show that our model is robust to reducing the training data. From $\sim$70 training configurations we can generate a model which on average outperforms all other models. Moreover, our model still performs very well when generalising to unseen configurations (overall score: 0.861) compared to the score when all data is given (overall score: 0.884).

This seems promising — that from roughly 70 tested configurations in the Preposition Selection Task we can generate a model which outperforms other cognitive models.

[30] Whenever a set of folds is generated, the folds are checked to verify that each fold has at least one constraint to test for each preposition. If not, a new set of folds is generated.

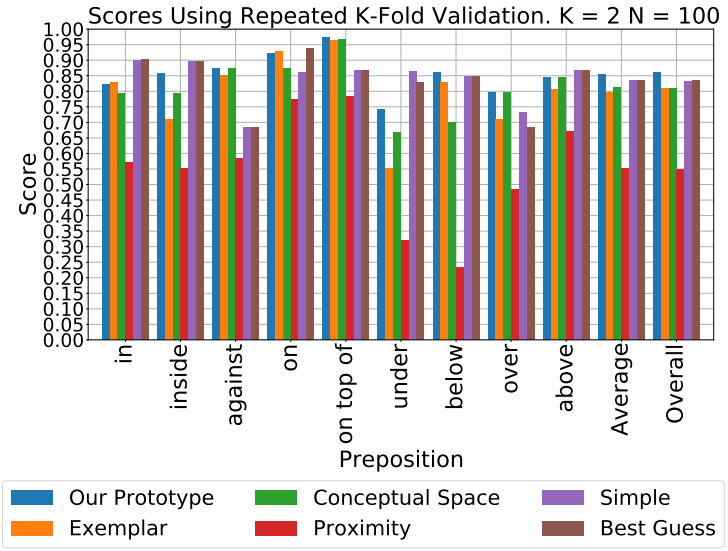


Figure 4.3: K-Fold Test Results (K=2, N=100).

If we consider each pair of 'functional' prepositions and their geometric counterparts, we can see that each of the data-driven models perform better for the geometric counterparts in both test cases (see Figures 4.2 and 4.3), with the single exception of the Exemplar Model and Conceptual Space Model and 'in'/'inside' in the repeated 2-fold cross validation. This possibly suggests that these terms have a simpler semantics which is easier to represent.

Significance

To assess whether the improvement shown by our model over the others is significant, we use a one-sided Wilcoxon signed-rank test to compare the overall score given by the models on each fold. The Wilcoxon test is used here as it doesn't assume that the distribution of the difference between scores of the models is normally distributed, as in the commonly used paired Student's t-test and is therefore statistically safer [153].[31] The null hypothesis is that the models perform equally well (with respect to the overall score) and alternative hypothesis that the Baseline Prototype Model performs better. The p-value is calculated using SciPy's implementation of the Wilcoxon signed-rank test [154].

[31] Note that this differs from the method reported in [1] where the Sign test was used. The Wilcoxon test is preferable however due to its greater statistical power.

The largest calculated p-value when comparing the Baseline Prototype Model with each of the other models is 3.5×10^{-18}, when comparing with the Best Guess Model. We may therefore conclude that our model does genuinely outperform the others.

It should be noted that here the training and testing data is split by assigning training and testing *scenes* as this ensures that particular configurations don't appear in both testing and training. As a result, individual participants may appear in both training and testing e.g. if they provide a response in the Preposition Selection Task for one of the training scenes and in the Comparative Task for one of the testing scenes. This does not mean that the same data appears in both training and testing — the configurations differ as well as the task, and participant judgements are aggregated in both tasks to generate the 'selection ratio' for training and constraints for testing, as described in Section 3.2.6. However, it is plausible that the data-driven models are over-fitted for the participants in our study.

In order to confirm that this is not the case, we have also run the same K-Fold cross validation (K=2, N=100) where folds are created by separating participants rather than scenes. In this case, each of the models performs marginally worse than when all data is used for training and testing (the largest drop in performance however is 0.009) but this is to be expected for the data-driven models as the training data has been reduced. Moreover, the overall result does not change and the largest calculated p-value when comparing the Baseline Prototype Model with each of the other models is 7.2×10^{-35}, when comparing with the Simple Model. It does not appear that the presence of individual participants in both training and testing has a significant impact on performance of the models or on the overall results.

4.4.3 Functional Features

As previously discussed, we have included features representing the functional notions of *support* and *location control* in the models. As these are novel and unexplored in computational models of spatial prepositions, in this section we briefly analyse their usefulness in the semantic model.

We will do this in two ways, firstly by considering the weights and values given to features by our model when trained on all available data. Secondly, by comparing performance of our model when functional features are removed.

Model Parameters

Firstly, *support* correlates strongly with 'on' (weight = 0.32) while *location_control* correlates strongly with 'in' (weight = 0.06). Though not as strong as the case with *support* and 'on', *location_control* is the second highest weighted feature for 'in'. This indicates that the way we have quantified these notions is appropriate.

In general, geometric features are weighted higher and have a more extreme value for the geometric counterparts. This can be seen with 'on' and 'on top of' where 'on top of' has a higher weight and value for *above_proportion* and similarly for 'in' and 'inside' with *containment*. Also, comparing 'above' with 'over' and 'below' with 'under', *above_proportion* and *below_proportion* are both given higher weights for the former while *f_covers_g* and *g_covers_f* are given higher weights for the latter.

It is not the case, however, that the functional features are more exaggerated for the more functional prepositions. In fact, it is the opposite — *support* is higher for 'on top of' than 'on' and *location_control* is higher for 'inside' than 'in'. This is unsurprising, however, as it is very often the case that being *geometrically* 'on' or 'in' implies being *functionally* 'on' or 'in' e.g. *containment* often implies *location control*.

Removing Features

In order to assess how the inclusion of these functional features affects model performance, we compared performance of our model with no features removed against our model with *support* removed and with *location_control* removed. Similarly to how we compared each model earlier, we ran 100 repetitions of K-fold cross-validation with $K = 2$. The results are shown in Table 4.3.

As we can see, in most cases our model performs better with the functional features included. Again, we can calculate the significance of this result as in Section 4.4.2, and find that the model performs significantly better overall when *location_control* (p value $= 1.28 \times 10^{-23}$) and *support* (p value $= 1.07 \times 10^{-28}$) are included. Most notably, the model performs much worse for 'in' when *location_control* is removed and for 'on' when *support* is removed. Also, we can see that the model performs worse for 'inside' when *location_control* is removed and for 'on top of' when *support* is removed. This further supports the idea discussed in Section 2.2.3 that though these prepositions are strongly influenced by geometric features they are also influenced by functional interactions.

	Location control	Support	None removed
in	0.784	0.813	**0.814**
inside	0.830	**0.868**	0.856
against	0.845	**0.883**	0.874
on	0.911	0.855	**0.913**
on top of	0.971	0.932	**0.973**
under	0.721	0.730	**0.737**
below	0.861	0.866	**0.867**
over	0.792	0.749	**0.796**
above	0.838	**0.849**	0.842
Average	0.839	0.839	**0.852**
Overall	0.848	0.841	**0.859**

Table 4.3: K-Fold Test Results (K=2, N=100), with changing feature set

4.5 Discussion

In this chapter we have explored some approaches to modelling spatial prepositions and provided a method for generating prototypes and weights for a prototype model which appears to perform well in the Comparative Task and will be useful going forward.

The overall picture that is painted of typicality in cognitive accounts is that typicality is related to *centrality* within a concept model generated from concept instances. For certain concepts, e.g. abstract concepts with *idealised* meanings, this is problematic as the notion of typicality which is useful for processing referring expressions is detached from frequency of occurrence.

Consider the spatial preposition 'inside'. Suppose that we do not know what 'inside' means but have some data representing instances of 'inside' and would like to generate a typicality measure for 'inside'. 'inside' is generally understood to have an ideal meaning represented by the notion of containment [29], where the more containment being expressed in an instance the more typical it is of 'inside'. However, as can be seen in our data, full containment is not always present for typical instances of 'inside'.[32]

Figure 4.4 compares the containment exhibited by configurations and the likelihood they are labelled with 'inside' in the Preposition Selection Task. The blue cross denotes

[32]Overall, it is for this reason we do not use 'typical' in the usual way to mean frequency of occurrence.

the prototype generated by the Conceptual Space model and the orange asterisk denotes the mean value of exemplars in the Exemplar model. The red line is a simple linear regression, and its value when the selection ratio is 1 is the prototype given by our generated model. We can see there are only two configurations labelled with 'inside' which express full containment and that there is a configuration which was always labelled with 'inside' that exhibits a very low degree of containment. As a result, the Exemplar and Conceptual Space models will provide greater typicality to configurations with containment values around 0.5, rather than 1. The same applies to our model to some degree, though it is a clear improvement. We believe that the overall improvement shown by the Baseline Prototype Model over the Conceptual Space and Exemplar approaches is primarily a result of this.

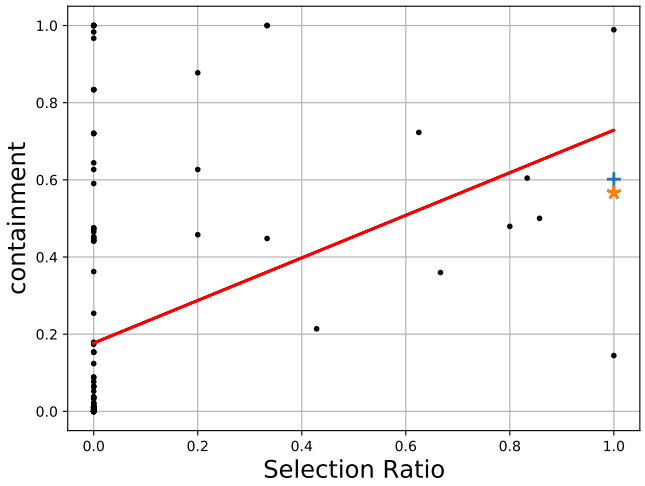


Figure 4.4: Instances of 'inside'.

As discussed in Section 2.2, many features can influence the usage of spatial prepositions and should be accounted for in the computational model. For example, 'over' is often characterised by the figure being located higher than the ground and within some region of influence. However, as discussed in [30], 'over' may also indicate *contact* between figure and ground.

Moreover, considering the polysemy exhibited by spatial prepositions, some features which may not seem to be salient for the preposition in general may be very important for determining the typicality for particular polysemes. For example, in some cases 'on' may indicate that the figure is in contact with some region of influence surrounding

the ground rather than the ground itself [34], and *shortest_distance* rather than *contact* becomes more salient in this case. For this reason we wanted to explore models which go beyond expressing spatial prepositions with one or two hand-picked features. Moreover, by automatically generating weights and prototypes for concepts we provide a method for modelling concepts where the semantics are less clear e.g. for polysemes and this is explored further in Chapter 5.

With regards to the functional relationships that spatial prepositions appear to encode, as far as I am aware, the models described in this chapter are the first to include such features. Moreover, we have shown that our Baseline Prototype Model performs significantly worse overall when either functional feature is removed. This is quite a strong result as one would expect that the functional features are not salient for some prepositions, e.g. 'above'. However, as is expected *location_control* is particularly useful for modelling 'in' and *support* is particularly useful for 'on'.

Overall, we have shown that it is possible to generate a model of typicality which (1) includes limited prior knowledge of the semantics of the concepts and (2) includes a greater range of features than most 'Simple Relation' models and outperforms them in doing so.

Using the collected data and semantic model discussed in this chapter there are a number of further issues related to spatial language use that we are interested in exploring. In the following section, we will consider the limitations and errors of the Baseline Prototype Model which suggests some directions for further improvements.

4.6 Improvements

We have seen that the Baseline Prototype Model provides a general improvement over the other models outlined in this chapter. However, there is clearly some room for improvement and in this section we will consider errors made by the model and how it may be improved.

4.6.1 Motivating Examples

In order to suggest where the Baseline Prototype Model may be improved and give further insight into the complexity of modelling spatial prepositions, we examine some unsatisfied constraints from testing in Section 4.4.1 where the model is trained on all

the data.

The highest weighted unsatisfied constraint for both ‘in’ and ‘inside’ arises from configurations (pear, cup) and (cube, cup) where to satisfy the constraint the model should assign a higher typicality to (pear, cup). Both these configurations are non-ideal instances of ‘in’ or ‘inside’, exhibiting a high degree of *location_control* but only partial *containment*. Interestingly, *containment* is actually lower for (pear, cup) than (cube, cup) but with a quick glance at an image of the configurations (Figure 4.5) we can see why (pear, cup) may be a better instance of both ‘in’ and ‘inside’ than (cube, cup).

Figure 4.5: Where the Baseline Prototype Model fails for ‘in’ and ‘inside’.

Such instances may frequently occur as not all types of containment are the same, see [37, 38]. Measuring the overlap of bounding boxes is a crude simplification of the notion of containment and, as in the case of (cube, cup), may represent containment in a very weak sense. Using the terminology of [37], in (pear, cup) the pear is partially contained in the *containable inside* of the cup, while in (cube, cup) the cube is partially geometrically inside the cup. This provides another good example of why simple relation models fail, as discussed in Section 2.2.1.

One solution to this may be to distinguish and assign differing priority to these

different types of containment. However, extracting such information from a 3D scene is a difficult computational challenge which, as far as I am aware, has not been addressed.

In such instances, one would expect that the cup offers a higher degree of *location_control* to the pear than the cube and that as a result the model assigns higher typicality to (pear, cup) than (cube, cup) for both 'in' and 'inside'.[33] However, the calculated value of *location_control* is in fact slightly higher for (cube, cup) than (pear, cup); it may therefore be possible to improve performance in these cases by refining how this feature is calculated.

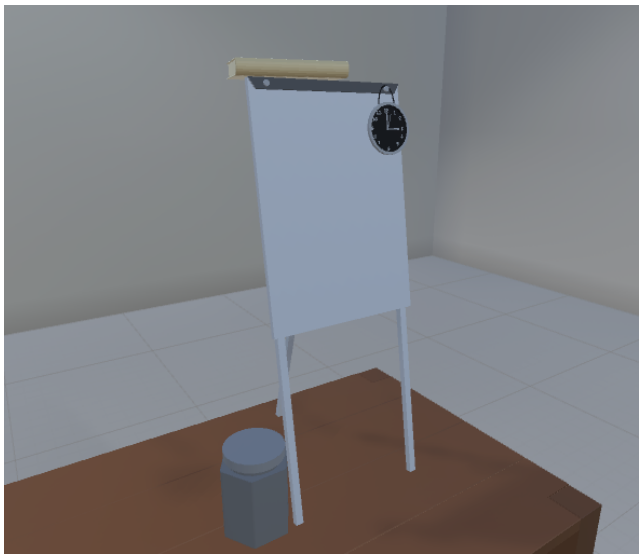


Figure 4.6: Different senses of 'on'.

The highest weighted unsatisfied constraint for 'on' arises from (clock, board) and (book, board) (see Figure 4.6) where to satisfy the constraint the model should assign a higher typicality to (clock, board). The (book, board) configuration is an instance of 'on' with a high value of *above_proportion* and *support*, but as it is precariously balanced on top *contact* is low. The (clock, board) configuration is an instance of 'on' with a high value of *contact* and *support* but not *above_proportion*. It appears that both configurations are instances of different senses of 'on' and are examples of the polysemy exhibited by this spatial preposition, and as the meaning of 'on' is learnt as a single sense by the model, and *above_proportion* is generally present for 'on', (clock, board) is assigned a low typicality score even though it is a very good instance of this 'attached

[33]The importance of *location_control* for 'inside' can be seen in Table 4.3.

to the side' sense of 'on'.

In Figure 4.7 feature values of configurations tested in the Preposition Selection Task are displayed along with their selection ratio for 'on'. We can see that configurations like (clock, board) where *above_proportion* is low are not uncommon, and that there are many with a high selection ratio. In this plot the prototypes for the Conceptual Space model and Baseline Prototype Model are given along with the mean value of exemplars in the Exemplar model. We can see that the generated models favour values of *above_proportion* around 0.7, when for the canonical sense of 'on' we would expect a prototypical value of 1. It may therefore be possible to improve the models by accounting for the various senses that spatial prepositions exhibit, and this is discussed further in the following chapter.

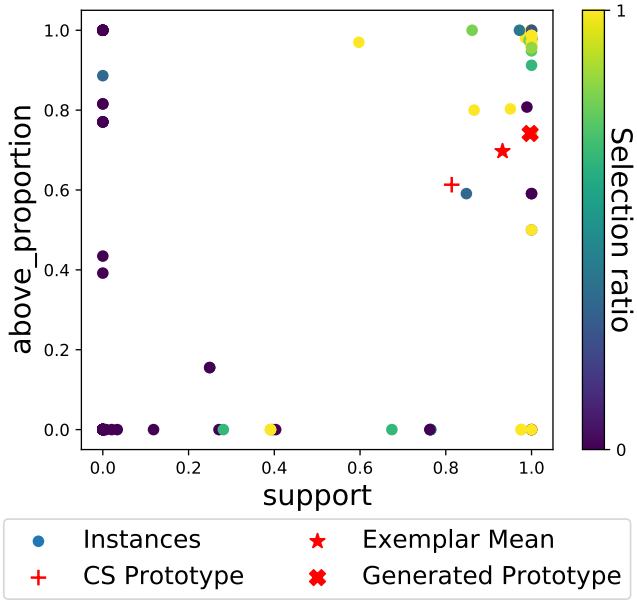


Figure 4.7: Instances of 'on'.

Chapter 5

Handling Polysemy

In the previous chapter we outlined a Baseline Prototype Model for automatically generating typicality measures for spatial prepositions in grounded settings and introduced methods for learning its parameters from data. However, though there is much to suggest that spatial prepositions exhibit polysemy (see Section 2.2.4), each term was treated as exhibiting a single sense.

In this chapter we will explore how to model the polysemy that spatial prepositions appear to exhibit and refine the previous Baseline Prototype Model by accounting for polysemy. We will provide novel methods for distinguishing separate polysemes, modelling the semantics of these polysemes and incorporating these into models of typicality for each preposition.

The previous chapter has shown that spatial prepositions are amenable to being modelled using prototypes and we now have methods for determining suitable prototypes in a feature space and feature weights for measuring typicality. In order to utilise these methods in modelling polysemy, we must first devise a method for differentiating polysemes such that when given a configuration labelled with a preposition we are able to determine which polyseme the configuration exemplifies.

Once methods for differentiating polysemes and determining the typicality of a given configuration for a specific polyseme have been defined, we must consider how typicality of a configuration is assigned for prepositions in general. This consideration gives rise to the notion of a 'polyseme hierarchy'. The Polysemy Model is then generated by combining the methods for differentiating polysemes with this notion of a 'polyseme hierarchy'.

We will also explore a more data-driven approach to modelling polysemy which relies on a clustering algorithm — we call the resulting model the k-Means Model. The performance of the Polysemy Model is initially tested by comparing it to the Baseline Prototype Model and the k-Means Model and we find that our method for incorporating polysemy into the Baseline Prototype Model provides significant improvement.

The main approach explored in this chapter relies on the author's intuition and evidence from the literature. However, once the performance of the Polysemy Model has been displayed we will explore refinements of the model and simple methods for reducing the reliance on intuition to build the model. The improved model will be called the Refined Model.

Finally, we will analyse the properties and behaviour of the generated Polysemy

Model and Refined Model, providing some insight into the improvement in performance over the Baseline Prototype Model, as well as justification for the given methods.

The main contributions of this chapter are:

1. a method of identifying polysemes based on 'ideal meanings' [29] and a modification of the 'principled polysemy' framework [30]
2. a notion of a 'polyseme hierarchy' which allows polysemes to be compared and aids typicality judgements

5.1 Which Prepositions?

The main motivations of this chapter are to deepen the understanding of the semantics of spatial prepositions and the polysemy they appear to exhibit, as well as provide a method which accounts for the apparent polysemy that helps to overcome some of the limitations discussed in Section 4.6.1. As such, the methods developed here (and the initial report of [2]) are initially focused on those spatial prepositions for which there is evidence in the literature that they exhibit polysemy at the kind of room-scales we are considering. We consider these to be 'in' [57], 'under' [155], 'over' [30, 155] and 'on' [6].[34]

Each of these prepositions may be considered 'functional' spatial prepositions and are considered semantically more complex than their geometric counterparts. Given the extra semantic complexity, it is unsurprising that these prepositions may more readily express a larger variety of senses. Furthermore, the scores obtained by the Baseline Prototype Model in Section 4.4 for each of these prepositions are worse than for their respective geometric counterparts; suggesting a greater need for a more detailed semantic model for these prepositions. In Section 5.5 we will, however, explore the applicability of the approach developed in this chapter to these 'non-polysemous' prepositions and discuss the extent to which they are actually non-polysemous.

[34] Though not explicitly studying polysemy, Bowerman and Choi [6] provide various examples of object configurations which are labelled simply with 'on' in English but are distinguished with multiple prepositions in other languages, see Figure 2.1.

5.2 Polysemy Models Based on Ideal Meanings

In this section we will specify how semantic models can be trained using data from the Preposition Selection Task which account for polysemy based on a notion of 'ideal meanings' and we will also outline the motivations for this approach.

5.2.1 Identifying Polysemes

The first challenge is to identify the different polysemes that may be expressed by a preposition and this issue is explored in this section.

For each preposition the goal is to construct a meaningful set of polysemes where, given a configuration in a scene, there is a method for determining which polysemes the configuration could represent. Once this has been achieved, the model can then be trained treating each polyseme separately, which is described in a later section.

Clustering

In order to potentially distinguish polysemes, suggest distinguishing features and support the approach taken to finding polysemes, we begin by attempting to cluster preposition instances. Data from the Preposition Selection Task is clustered using off-the-shelf clustering algorithms provided by scikit-learn [148]. In the remainder of this chapter, where the k-means clustering algorithm is used, all tested configurations are used and are weighted by their 'selection ratio' for the given preposition. Where Hierarchical Agglomerative Clustering (HAC) is used, only 'good' instances of the preposition (where the 'selection ratio' is greater than or equal to 0.5) are used. Though features which do not directly influence typicality of a preposition may help to distinguish polysemes, e.g. whether or not the ground is a container, currently only the relational features given in Section 3.2.3 are considered.

Due to the vagueness they exhibit, spatial prepositions are difficult to cluster and it may not be clear when meaningful clusters have been established. For example, when generating clusters using the k-Means algorithm, where the number of clusters k must be specified in advance, one may employ the 'Elbow' Method to determine how many clusters should be generated. This involves running the algorithm with varying values for k and plotting the inertia (within-cluster sum-of-squares) of each of the generated models against k. A distinct kink in the plot signifies the optimal value of k. When

we apply this to our data no such kink is discernible, possibly with the exception of 'under', see Figure 5.1 for the case of 'on'. It may be that, though to humans there are meaningful distinctions between polysemes, the clusters representing polysemes significantly overlap and finding well-defined significant clusters is a computational challenge.

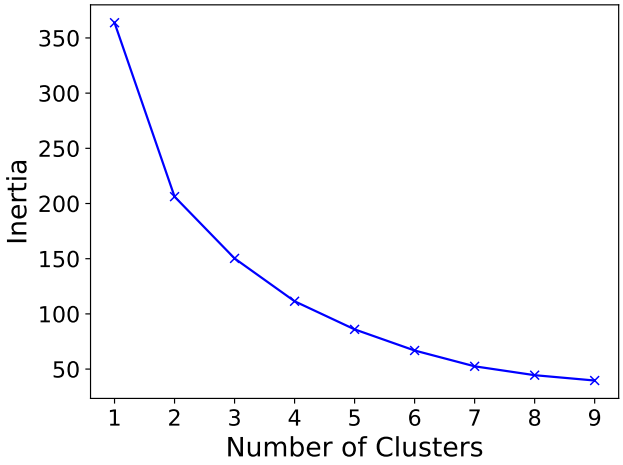


Figure 5.1: Inertia from k-means clustering of 'on'.

In order to get a better understanding of the data, we cluster the data using HAC with the Nearest Point Algorithm, and use the provided dendrograms for analysis.

In Figure 5.2 we see the clusters generated by the HAC algorithm for 'on'. We can see a large grouping (in red) which appears to represent the ideal/canonical meaning of 'on' — instances in the group have a high degree of *support*, *above_proportion* and *contact*. These are most sharply distinguished from the group in green (24, 35, 38) where *support* and *contact* are high but *above_proportion* is 0. In the turquoise group *support* and *contact* are generally apparent but *above_proportion* is low. Finally the clade (43) represents an instance where *above_proportion* and *support* are near 0 and there is some *contact*. These generated clusters appear to represent and support the distinctions (Sense 1: Red and Turquoise), (Sense 2: Green) and (Sense 3: Blue) given for 'on' in the Section 2.2.4.

In general, the clustering appears to show that for each preposition there is a cluster representing canonical examples of the preposition and that other clusters may be distinguished by their lack of a particular salient feature. We explore representations

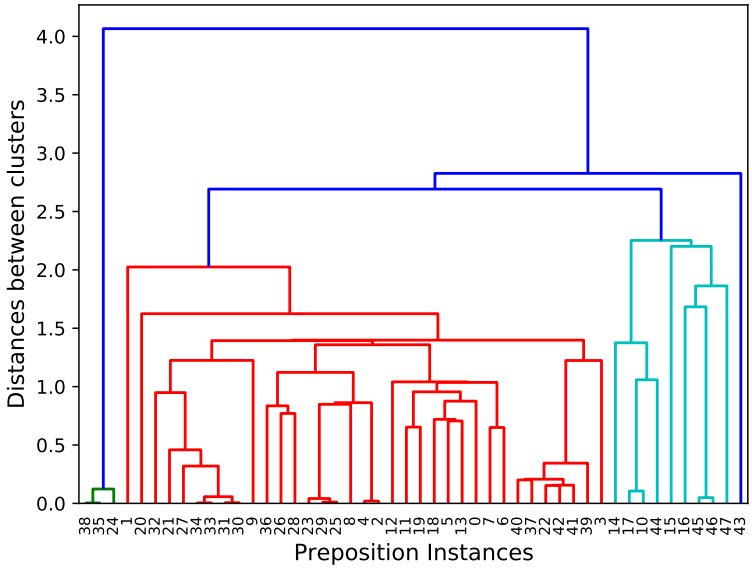


Figure 5.2: Dendrogram from HCA for 'on'.

of these canonical meanings in the following sections.

Ideal Meanings

Herskovits [29] argues that the meanings of spatial prepositions should be understood as *ideal meanings* from which other uses of the prepositions are derived, see Section 2.2.2. Clearly the ideal meaning of a preposition represents a polyseme that should be represented in our model and so we begin by defining these.

In order to represent each ideal meaning, salient features, threshold values and ordering relations are assigned to each preposition such that a configuration is considered an ideal instance of the preposition if the values are greater than (or less than, depending on the ordering relation) or equal to the threshold for each salient feature. Initially, we draw on the existing literature to determine which features should be considered salient, and rely on the author's intuition to assign threshold values such that the instances will likely be close to what we may consider an ideal meaning and also that training instances will likely be apparent for each generated polyseme (see Table B.1 in the Appendix for the precise definitions).

The polysemy model reported in [2] uses these intuitive definitions, however a more robust implementation would automatically generate these threshold values from the

training data. This has in fact been achieved, which we will discuss in Section 5.6.1.

Representations of the ideal meaning of each preposition are described below. For the prepositions 'in' and 'on' we follow [8] and assume that the underlying representations comprise both geometric and functional components.

In Following [8], 'in' expresses geometric containment as well as the functional notion of *location control*. We define the ideal meaning of 'in' by a high value of two features: *containment* and *location_control*.

On In [8] various accounts and definitions of 'on' are listed and the recurring features are *contiguity* and *support*. We also believe that the canonical representation of *support* supposes that an object is supported from below, as is discussed in [34] and is seen in the *support* image schema provided in [9]. We therefore define the ideal notion of 'on' as having a high value of three features: *support*, *above_proportion* and *contact*.

Under Herskovits [29] gives the ideal meaning of 'under' as 'partial inclusion of a geometrical construct in the lower space defined by some surface, line or point'. We therefore define the ideal meaning of 'under' by a high value of two features: *below_proportion* and g_covers_f.

Over Work on the semantics of 'over' often considers moving objects and the path taken by the figure. When we only consider static objects, 'over' appears to have two central notions — that the figure is above the ground and that the figure covers the ground [30, 46]. We therefore define the ideal meaning of 'over' by a high value of: *above_proportion* and f_covers_g.

Meaning Shifts

Once the ideal meanings are understood, the derived uses of a spatial preposition are then achieved via what Herskovits calls 'sense' and 'tolerance' shifts. In tolerance shifts the ideal meaning may be deviated from in a continuous manner — e.g. 'in' may be used to express partial containment rather than full containment. Sense shifts appear in a discontinuous manner where the relations expressed by the ideal meaning are substituted for conceptually similar relations — Herskovits gives the instructive example

of 'the muscles in his leg' where the relation being expressed by 'in' is no longer containment but parthood.

How sense shifts and their associated language conventions may arise relies on the complex interactions of commonsense reasoning and the evolution of language. We do not attempt to fully characterise how these processes occur. However, in the case of both sense and tolerance shifts, the meaning expressed by a preposition generally violates a condition of the ideal meaning but is still closely related to it.

This relates to the 'principled polysemy' approach set out in [30] which aims to provide a more objective footing for determining when preposition instances represent genuinely distinct senses. The principled polysemy framework assumes a 'primary sense', similar to the notion of 'ideal meaning' and comprises two criteria for a sense to count as distinct:

1. The sense must include a non-spatial component which distinguishes it from other senses and/or where the spatial configuration is meaningfully different from other senses

2. There must be instances of the sense where its meaning cannot simply be derived from the context along with knowledge of the other senses

With regards to the first criterion, we do not distinguish spatial and functional features. The second criterion is rather subjective and would rely on an advanced model of commonsense in order to automate. We condense the criteria to:

Criterion 1 *A sense may be considered distinct if the sense meaningfully differs from other senses with regards to some spatial or functional features*

We suppose that whether a sense satisfies or violates one of the conditions of the ideal meaning constitutes a meaningful distinction. Following this, the ideal meaning of a preposition can be considered to be a distinct polyseme and every other polyseme is represented by some non-ideal meaning.

The various ways that the conditions of the ideal meaning may be violated provide a method of grouping non-ideal meanings and we take these groupings to represent distinct polysemes. For example, in the case of 'on' each non-ideal sense is generated by negating at least one of the three conditions, giving eight potential senses for 'on'. So, for example, there is a sense of 'on' where the figure is supported by and in contact

with the ground but not above it and this sense is distinguished from the sense where the figure is above, in contact with and supported by the ground.

Clearly, it may be the case that a non-ideal meaning constructed in this way encompasses more than one genuine polyseme, however the distinctions would then become very fine-grained and a larger dataset would be required for training. This is a potential avenue for further work.

For each preposition we now have a set of polysemes each with a set of conditions that a configuration must satisfy in order to be a potential polyseme instance.

5.2.2 Determining Typicality

Given that we have outlined how polysemes may be distinguished, how do we translate this into a semantic model? Firstly, we construct models for each polyseme such that, given a particular configuration, we can assign a value representing how typical the configuration is for the polyseme.

In order to construct such models we treat each polyseme as if it were a distinct term and employ the same method, underlying model and feature space used in the Baseline Prototype Model. To train each polyseme separately and ensure that the polyseme is only trained on polyseme instances, the training datasets are modified. This is achieved simply by removing potential preposition instances that are not examples of the given polyseme i.e. configurations which have been labelled with the preposition but which do not fit the polysemes conditions. For example, for the ideal sense of 'on' we would use the 'on' dataset and remove instances of 'on' where one of the ideal conditions does not hold. In this way, the model is trained on instances of a particular polyseme and so the generated prototype and weights reflect properties of the distinct polyseme rather than the preposition in general. In Equation 5.1, the typicality, $typicality_p(c)$, assigned by a polyseme, p, to a configuration, c, is specified by these prototypes and weights.

5.2.3 Polyseme Hierarchy

Given that we have a model which assigns a typicality score to any given configuration for a given polyseme, how can we exploit this to answer the kind of referring expressions which appear in the Comparative Task e.g. 'the object on the board'?

In some cases, given a preposition and ground, only one polyseme of the preposition may be applicable to all potential figure-ground pairs in the scene. In this case we can

just compare the typicality for each figure-ground pair, with respect to that polyseme, and the most typical is the one selected.

However, in many cases there will be multiple possible figures each potentially fitting a different polyseme. For example, there may be a scene with a book on a table — Sense 1 from Section 2.2.4 — as well as a box on the floor but touching the table — Sense 3 from Section 2.2.4. It may be the case that the typicality Sense 1 assigns to (book, table) is slightly less than Sense 3 assigns to (box, table). If we are to simply select objects based on raw typicality, 'the object on the table' may be interpreted as 'box'. This would clearly be a mistake as Sense 3 is a weaker sense of 'on'. We must therefore somehow account for this apparent hierarchy of senses.

The notion of sense hierarchies is not in itself new; however hierarchies are usually based on inheritance and generality e.g. the hierarchies in WordNet [53] capture knowledge such as 'a car is a vehicle'. In the case of prepositions, [156] create a hierarchical taxonomy of preposition 'supersenses' which may be used to annotate text. These 'supersenses' group together 'fine-grained' preposition senses which are then ordered into an inheritance hierarchy. However, the apparent hierarchy of the polysemes we are considering is less related to inheritance and more related to a perceived applicability of the polyseme — in the above example Sense 1 is a better sense of 'on' than Sense 3. Furthermore, we aim to somehow quantify the hierarchy so that polysemes may be compared.

In order to account for this apparent hierarchy, the typicality scores are adjusted based on the likelihood that a participant uses the given preposition to denote the given polyseme. To determine how the scores should be adjusted, using data from the Preposition Selection Task we generate a *rank* for each polyseme. The rank for a polyseme is calculated by taking the average value of the selection ratio for all configurations that fit the conditions of the polyseme.

For a given preposition, the polysemy models calculate the typicality of a configuration, c, using Equation 5.1. P is the set of polysemes of the preposition which may apply to c, $typicality_p(c)$ is the typicality of c with respect to a polyseme p and r_p is the rank of polyseme p.

$$typicality(c) = \max_{p \in P}(typicality_p(c) \times r_p) \tag{5.1}$$

By adjusting the typicality assigned by polysemes by their rank, configurations

fitting weaker senses, e.g. Sense 3, should only be selected if there are no good examples present of stronger senses, e.g. Sense 1.

5.2.4 Specification

The polysemy model described in this section, which will be called the Distinct Prototype Model, is defined for each preposition as a set of polysemes where each polyseme is in turn defined by:

- A set of conditions under which the polyseme may be applicable
- A set of feature weights and a prototype allowing for typicality measurement
- A rank which represents the preference for the polyseme

and the overall typicality of a configuration for a given preposition is given by Equation 5.1.

It is possible that when the data is split into train/test sets, there will be cases where a polyseme is not given any positive instances to train on. In this case, the polyseme is assigned prototype and weights equal to those assigned by the Baseline Prototype Model for the associated preposition. The rank for the polyseme, instead of being 0 is then taken as the average value of the selection ratio for all training configurations.

We can see that overall the resulting model is a collection of prototypes with associated weights, organised around a central ideal meaning. This has some similarity with the radial category approach [157] in so far as each sense is linked to a central sense, though the radial category approach is aimed at distinguishing less fine-grained distinctions than we consider here where senses do not share the same underlying representations and are created through schematic transformations (similar to the 'sense shifts' discussed in Section 2.2.4).

In order to explore the nature of polysemy and how it may impact semantic representations we will also consider a similar model, the Shared Prototype Model, where the polysemes share a prototype. The Distinct Prototype Model and Shared Prototype Model are the same except that in the former each polyseme learns its own prototype while in the latter each polyseme uses the same prototype which is assigned using the prototype from the Baseline Prototype Model.

By comparing these two models we may test whether polysemes should share a prototype or be organised into multiple prototypicality centres. For example, Senses 1, 2 and 3 for 'on' from Section 2.2.4 may assign varying salience to *support*, *contact* and *aboveness* but within each sense *more support*, *contact* or *aboveness* may increase typicality i.e. if the prototype for each sense is the canonical one and is shared.

5.3 *k*-Means Model

The polysemy models we have so far described rely on the intuition of the authors and evidence from the literature to generate ideal meanings. In order to provide a more thorough analysis and explore other methods for handling polysemy, we also generate a model which requires no such expert knowledge and relies on a clustering algorithm to find polysemes. We call this model the k-Means Model and in this section we describe how it is generated and how it assigns typicality to configurations.

5.3.1 Typicality

The parameters defining the k-Means Model are:

- A set of feature weights for measuring semantic distance and similarity
- A set of clusters each defined by a cluster centre
- A rank associated with each cluster

Given these parameters, the k-Means Model assigns typicality to a given configuration, x, by first finding the cluster, C, which is semantically most similar to x. Semantic similarity of x to a cluster is calculated using Equation 4.3 where the centre of the cluster acts as the prototype P. The typicality of x is then given as the semantic similarity of x to C multiplied by the rank assigned to C.

5.3.2 Generation

Here we describe how, for a given preposition, the parameters of the k-Means Model are assigned when given training data.

Firstly, the feature weights for the k-Means Model are trained in the same way as the Baseline Prototype Model, giving a measure of feature salience for the preposition in

general. Semantic similarity can then be calculated using a weighted Euclidean metric, as in Equation 4.3.

In order to find an appropriate set of clusters for the model, we begin with a fixed number of clusters, k, to be generated. We set k to be the number of polysemes generated by the polysemy models — 'on':8, 'in':4, 'under':4, 'over':4. We then cluster the configurations in the training data using k-Means clustering to generate k clusters defined by the centre of the cluster. For the algorithm the configurations are weighted by their associated selection ratio for the preposition.

Finally we must determine a rank for each cluster. This is calculated by finding the average selection ratio of configurations in each cluster. Before this is calculated, each cluster is first modified to account for feature salience so that the given clusters are more internally coherent with respect to semantic similarity. Where previously each configuration is assigned to the cluster with the closest centre, now each configuration is assigned to the cluster with the centre that it is semantically most similar to. Finally, the rank of a given cluster is then calculated by taking the mean value of the selection ratio for configurations in the cluster.

5.4 Model Performance

As in Section 4.4, the performance of the models is evaluated as described in 3.2.6.

In Tables 5.1 and 5.2, the scores given to each preposition are the sum of weights of the satisfied constraints involving the preposition divided by the total weight of constraints involving the preposition. The average score is simply the average score for each preposition and the overall score is the sum of weights of all satisfied constraints divided by the total weight of all constraints. Higher scores imply better agreement with participants in general.

5.4.1 Initial Results

To provide an initial insight into model performance and how well the models translate categorical data into typicality judgements, we compare the models when training and testing using all the data from both tasks. Results for each preposition are given in Table 5.1.

The Distinct Prototype Model outperforms the Shared Prototype Model with 'un-

	Distinct Prototype	Shared Prototype	Baseline Model	k-Means Model
in	**0.864**	**0.864**	0.814	0.814
on	0.951	0.951	0.945	**0.957**
under	**0.908**	0.752	0.809	0.894
over	**0.824**	0.765	0.800	0.812
Average	**0.887**	0.833	0.842	0.869
Overall	**0.902**	0.842	0.857	0.891

Table 5.1: Initial Results: Training & testing on all scenes. Scores represent agreement with participants in the Comparative Task

der' and 'over' and the models draw with 'in' and 'on'. This suggests that learning a distinct prototype for each polyseme is advantageous and supports the notion that these terms ought to be represented by several distinct prototypicality centres.[35] From here on we discard the Shared Prototype Model and refer to the Distinct Prototype Model as the **Polysemy Model**.

5.4.2 K-Fold Testing

In order to test and compare robustness of the models, we split the data into training and testing scenes using K-fold cross-validation with $K = 10$. We then generate the models based on data from the training scenes given in the Preposition Selection Task and test the models using constraints generated from the testing scenes in the Comparative Task. We repeated this process 10 times and averaged the results, shown in Table 5.2.[36]

Here $K = 10$ is used as opposed to in Section 4.4.2 where $K = 2$ is used as the polysemy models require a larger dataset for training.

	Polysemy Model	Baseline Model	k-Means Model
in	**0.837**	0.828	0.810
on	0.937	0.929	**0.947**
under	**0.891**	0.766	0.886
over	**0.802**	0.798	0.694
Average	**0.867**	0.830	0.835
Overall	**0.891**	0.846	0.867

Table 5.2: K-Fold Test Results (K=10, N=10). Scores are averaged results of the cross-validation

Results

The Polysemy Model has improved on the Baseline Prototype Model for each preposition and both the models which have accounted in some way for polysemy have in general improved on the Baseline Prototype Model. In the case of 'in', the Baseline Prototype Model outperforms the k-Means Model. We believe that this is partly because the k-Means Model will in general require more data for training and 'in' is a particularly difficult preposition to collect large amounts of data for — there are only eight 'good' instances of 'in' in the data from the Preposition Selection Task.

Though the k-Means Model has under-performed for 'in' and 'over', it may provide a useful method for handling the polysemy of terms which do not have such clear ideal meanings.

Significance

In order to assess whether the improvement shown by the Polysemy Model over the other models is significant, we again use a one-sided Wilcoxon signed-rank test as in Section 4.4.2. The calculated p-value is 1.3×10^{-6} when comparing to the Baseline Prototype Model and 0.0043 when compared with the k-Means Model. We may therefore

[35] We also test the Shared Prototype Model in the following, where it performs significantly worse than the Distinct Prototype Model, but omit its results for readability and brevity.

[36] As in Chapter 4, each fold is checked to verify that it contains a constraint to test for each preposition. Also, each training set of K-1 folds must contain enough training instances for the k-Means Model — training scenes must contain at least k preposition instances, where k is the number of clusters generated for the given preposition (given in Section 5.3.2).

conclude that the Polysemy Model does genuinely outperform the others. The improvement shown by the k-Means Model over the baseline also appears to be significant (p value = 0.013).

5.5 Non-Polysemous Prepositions

So far, the study in this chapter has been based on those prepositions which, according to existing literature, appear to exhibit polysemy at room/table-top scales and also have an uncontroversial ideal meaning. It may be the case, however, that the 'non-polysemous' prepositions considered in this thesis ('inside', 'on top of', 'above', 'below' and 'against') are actually polysemous or that the approach developed in this chapter also performs well when modelling these non-polysemous concepts.

Consider the example of 'inside', it would seem that we may confidently assign an ideal meaning to 'inside' represented by full *containment* and that this single sense appropriately captures the semantics of 'inside'. This is supported by the performance of the Simple Relation models in Section 4.4. There are, however, instances of 'inside' where *location_control* seems to be just as important as *containment*, see the examples in Section 4.6.1.

It may be the case that accounting for polysemy is beneficial, or at least not several detrimental, in modelling the 'non-polysemous' prepositions. We therefore extend the original approach of [2] to include these prepositions and will discuss the degree to which they exhibit polysemy.

The main approach of this chapter relies on modelling the 'ideal meaning' of each preposition and these are described below for the additional prepositions.

5.5.1 Ideal Meanings

Supposing that 'inside', 'on top of', 'above' and 'below' are purely geometric versions of their functional counterparts, their ideal meanings are simplified, from Section 5.2.1, as follows:

- 'inside' is defined simply by a high value of *containment*
- 'on top of' is defined by a high value of *above_proportion* and *contact*

- 'above' is defined by a high value of *above_proportion* and a low value of *horizontal_distance*
- 'below' is defined by a high value of *below_proportion* and a low value of *horizontal_distance*

We attempt to include 'against' in the following, though it is not clear that such an ideal meaning or a reliable way to model it exists.

'against' is quite a confusing preposition to define which has not been much treated in the literature and in Section 4.4.1 we can see the poor performance of the Simple, Best Guess and Proximity models. It would seem that 'against' usually denotes some degree of proximity and/or contact. For example, in work by Doore et al. [158] the preference for different spatial prepositions is assessed in different contexts and 'against' is preferred to 'next to', 'touching' or 'along' when describing *contact* relations. Also, it is apparent that 'against' expresses a functional relationship where a force being exerted by the figure is resisted by the ground [25].

We take the ideal meaning of 'against' to be expressed by a high value of *contact* and *location_control* and a low value of *horizontal_distance*. Again, the precise values assigned to the ideal meanings are given in Table B.1 in the Appendix.

5.5.2 Results

Again, the models are first tested when training and testing using data from all the given scenes and the results are given in Table 5.3. We can see in general the Polysemy Model still performs well, even for the 'non-polysemous' prepositions — outperforming the Baseline Prototype Model and k-Means Model for 'in', 'inside', 'under', 'over' and 'above' and achieving a perfect score for 'on top of'; and outperforming the Baseline Prototype Model in all cases except 'against' and 'below'.

To provide a more thorough test of the models, again the models are tested using K-Fold cross validation (K=10, N=10) and the results are given in Table 5.4.

In both Tables 5.3 and 5.4 the models appear in general to perform better for the 'non-polysemous' prepositions than the polysemous prepositions. In both tables, with the exception of the Polysemy Model in Table 5.4, the models achieve a higher score on average for 'inside', 'against', 'on top of', 'below' and 'above' than 'in', 'on', 'under' and 'over'. This reinforces what is seen in Section 4.4.2 where the data-driven models all

	Polysemy Model	Baseline Model	K-Means Model
in	**0.864**	0.814	0.814
inside	**0.958**	0.917	0.896
against	0.855	**0.889**	0.838
on	0.951	0.945	**0.957**
on top of	**1.000**	0.985	0.992
under	**0.908**	0.809	0.894
below	0.776	0.884	**0.905**
over	**0.824**	0.800	0.812
above	**0.892**	0.843	0.880
Average	**0.892**	0.876	0.887
Overall	0.892	0.884	**0.899**

Table 5.3: Testing the models on all prepositions. Initial Results: Training & testing on all scenes

perform better for the non-polysemous prepositions than their functional counterparts.

The addition of 'inside', 'against' and 'below', for which the Baseline Prototype Model appears to perform better than the Polysemy Model, means that there is no significant improvement given by the Polysemy Model when considering all these terms.[37]

Nevertheless, the polysemy models perform surprisingly well on the 'non-polysemous' terms. This could simply be answered by saying that their semantics are simpler and easier to model, however this does raise a couple of questions. Firstly, are these terms actually not polysemous? Secondly, supposing these 'non-polysemous' terms *are* not polysemous, is the performance of the Polysemy Model a result of something other than the model actually capturing polysemy?

Are these prepositions non-polysemous?

In order to explore this in more detail we will consider some examples of each of these prepositions.

[37] We will however see in Section 5.6 that it is possible to refine the Polysemy Model such that it does provide a significant improvement.

	Polysemy Model	Baseline Model	K-Means Model
in	0.827	**0.843**	0.814
inside	0.893	**0.923**	0.861
against	0.815	**0.862**	0.837
on	**0.949**	0.932	0.945
on top of	**0.989**	0.975	0.969
under	**0.889**	0.782	0.874
below	0.769	0.850	**0.885**
over	**0.811**	0.793	0.684
above	**0.843**	0.828	0.824
Average	0.865	**0.865**	0.855
Overall	**0.875**	0.872	0.874

Table 5.4: Testing the models on all prepositions. K-Fold Test Results (K=10, N=10)

Inside Figure 5.3 shows some configurations which appear in the study scenes. In the Preposition Selection Task when labelling the (pear, cup) configuration, all tested participants gave 'inside'. In the Comparative Task, when selecting the object referred to by 'the object inside the cup', participants selected the pear — the Polysemy Model agrees with participants here but the Baseline Prototype Model doesn't. This appears to be similar to the often cited instances of objects being 'in' other objects when there is little or no containment and is usually explained by the presence of location control. Following the previously discussed Criterion 1 for distinguishing polysemes, this appears to be a non-ideal sense of 'inside' and provides some support that 'inside' is in fact polysemous.

On top of Again, Figure 5.4 shows some configurations from the study. In the Preposition Selection Task, half of the tested participants labelled the (pencil, lamp) configuration with 'on top of', and participants would select the pencil when given the description 'the object on top of the lamp'. Both the Polysemy Model and Baseline Prototype Model pick the pencil in this case, as the other possible objects are not very plausible instances. The Polysemy Model, however, gives a more marked distinction between (pencil, lamp) and other configurations in the scene where the lamp is the

Figure 5.3: ‘inside the cup’

ground.

This instance of ‘on top of’ may be explained by synecdoche, where the noun ‘lamp’ is being used to refer to the base of the lamp which the pencil is on top of in a canonical sense. Following this we may argue that the meaning of ‘on top of’ here is unchanged from the canonical one and that this is not evidence of ‘on top of’ exhibiting polysemy.

This relates to precisely how polysemy is defined, as we may say that being on top of an object as a whole and being on top of some salient part of an object are distinct senses of ‘on top of’ and that in this particular instance both synecdoche and polysemy are occurring. However, regardless of the precise definition of polysemy, such instances should be accounted for somehow.

One approach to modelling these phenomena would be to iterate over sections or ‘salient parts’ of objects, for example checking whether the pencil is on top of the lamp as a whole, or some important section of the lamp e.g. its base. This is the approach taken for ‘on’ in [78]. Automating such a process would require an ability to automatically demarcate and label salient parts of objects and this is a significant research problem. The method proposed in this chapter instead deals with these synecdochal instances

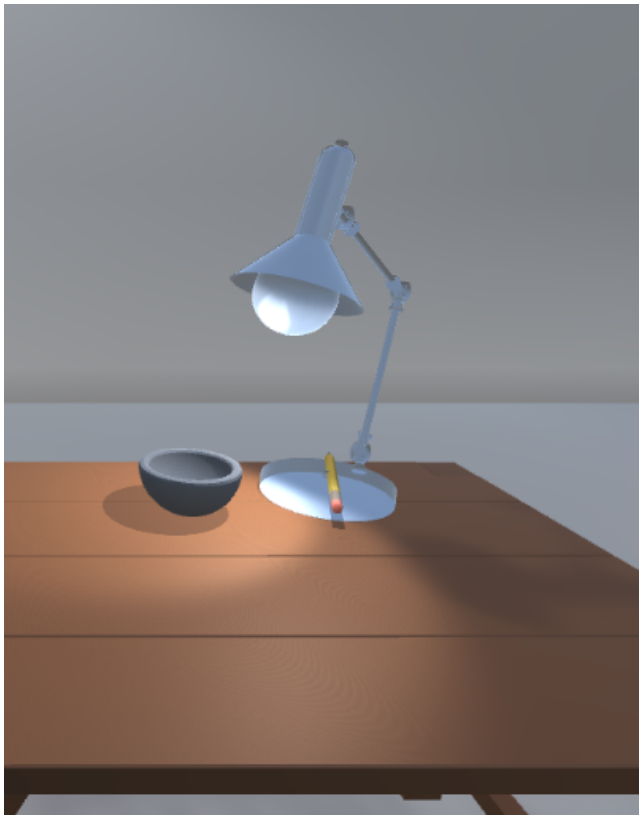


Figure 5.4: 'on top of the lamp'

in a simpler way by modelling a distinct sense of 'on top of' where *above_proportion* is low, and this potentially explains the good performance of the Polysemy Model for 'on top of'.

Above For the configuration (table, box) in Figure 5.5, all tested participants selected 'above', and there are many similar examples of this. This may seem uncontroversial, however a large proportion of the table is not actually above the box and the value of *above_proportion* is 0.77. Similarly to the example of 'on top of' discussed above, this instance may be explained by synecdoche — 'table' may be conceptualised as the horizontal part of the table. However, it is interesting to note the existence of seemingly unambiguous instances of 'above' where *above_proportion* is not 1, and following Criterion 1, we may suppose that this is a distinct sense of 'above' which is similar to the 'covering' sense of 'over'.

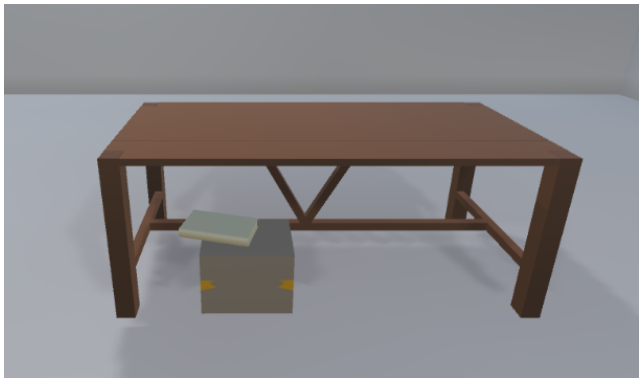


Figure 5.5: 'above the box'

Below For the (jar, board) configuration shown in Figure 5.6, four out of five tested participants selected 'below' in the Preposition Selection Task. This is similar to the example above given for 'above' (the value of *below_proportion* is 0.19), however it is even more striking as the board does not cover the jar (the value of *g_covers_f* is 0.15).

Against In both Figures 5.7 and 5.8 the configuration (box, table) was labelled with 'against' by every tested participant. These appear to be distinct senses of 'against', in one the box is leaning against the table and in the other box is simply next to it. This distinction can be drawn with either of the functional features — there is a higher degree of *support* and *location_control* in (1) than in (2).

Overall, it appears that the 'non-polysemous' prepositions may in fact exhibit polysemy to some degree and this may explain the reasonable performance of the Polysemy Model for these prepositions.

Is the model capturing polysemy?

Above we have provided some evidence that these 'non-polysemous' terms may in fact be polysemous. However, the question still stands whether the methods used for the Polysemy Model are actually capturing polysemy or are just making training more effective by partitioning the data. We do not believe the latter to be the case, and will provide some evidence for this view here.

In order to test this, we generated the Partition Model which partitions the data using defined 'ideal meanings' in the same way as the Polysemy Model, but where the 'ideal meanings' are generated in an arbitrary fashion. To achieve this, for each

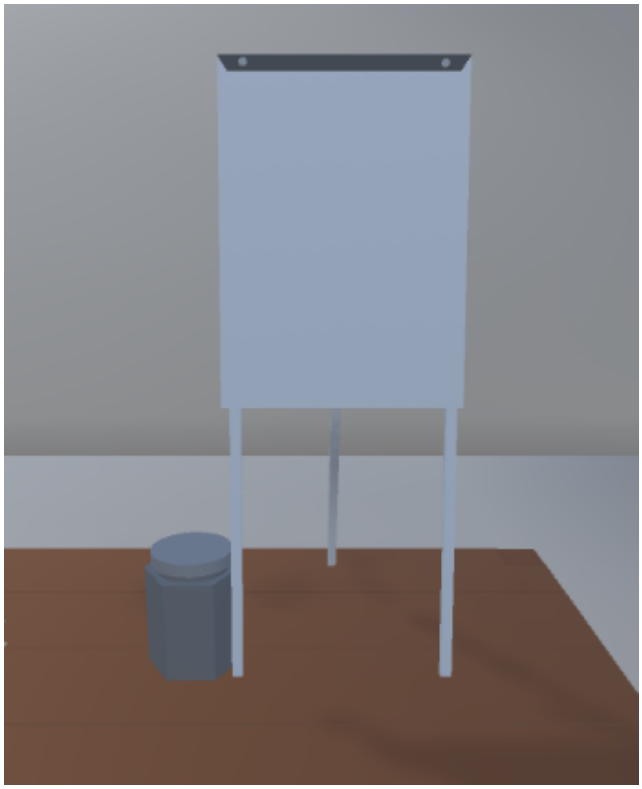


Figure 5.6: 'below the board'

preposition, we begin with the ideal meaning given by the polysemy model which is defined by a set of features and threshold values. For each feature appearing in the original ideal meaning, we randomly select a new 'non-salient' feature which does not appear in the original ideal meaning. Then to determine threshold values for each of the new features, we take the median values of the features in the training data. In this way, there will always be training instances for the ideal meaning as well as the other polysemes (provided there are at least as many training instances as polysemes). To ensure that the ideal meaning is still represented by 'good' instances of the preposition, we use the median feature values of 'good' training instances here (where the selection ratio is ≥ 0.5).

As we can see from Tables 5.5 and 5.6, the Partition Model performs worse overall than both the Baseline Prototype Model and Polysemy Model in both the initial and K-Fold testing. Moreover, in the K-Fold test, both the Baseline Prototype Model and Polysemy Model perform significantly better than the Partition Model (largest p-value

Figure 5.7: 'against the table' (1)

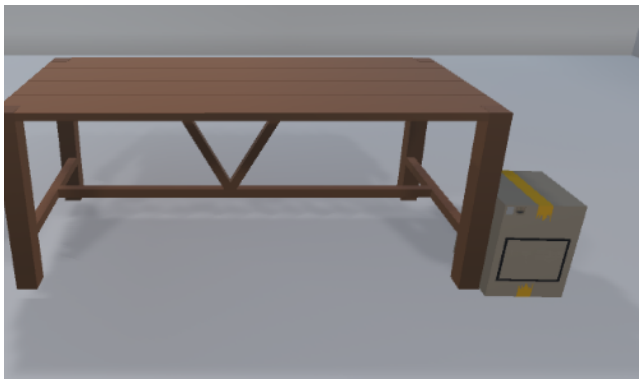


Figure 5.8: 'against the table' (2)

1.3×10^{-5} calculated using one-sided Wilcoxon signed-rank test as in Section 4.4.2). The Partition Model only consistently performs better than the Baseline Prototype Model (in both the initial and K-Fold tests) for the preposition 'under'.

These results suggest that the improvement shown by the Polysemy Model over the Baseline Prototype Model does not simply result from partitioning the data.[38] This indicates that the Polysemy Model is genuinely capturing the polysemy exhibited by these terms and, moreover, that it is important to appropriately define the ideal meanings used in the Polysemy Model. In the following section we will explore how these definitions may be refined.

[38] We will also see a similar result in Section 5.7 after exploring how the model can be refined.

	Polysemy Model	Baseline Model	Partition Model
in	**0.864**	0.814	0.593
inside	**0.958**	0.917	**0.958**
against	0.855	**0.889**	0.778
on	**0.951**	0.945	0.933
on top of	**1.000**	0.985	0.954
under	**0.908**	0.809	**0.908**
below	0.776	**0.884**	0.837
over	**0.824**	0.800	0.624
above	**0.892**	0.843	0.771
Average	**0.892**	0.876	0.817
Overall	**0.892**	0.884	0.839

Table 5.5: Testing a partition model. Initial Results: Training & testing on all scenes

5.6 Refining the Model

Though in general expert knowledge has been minimised in the creation of the Polysemy Model, some aspects of the proposed model, such as the underlying conceptual framework, feature space and definition of the ideal meanings, have relied on motivation from the literature or the intuition of the author. In particular, in defining how the ideal meanings should be represented, discussed in Section 5.2.1, we have initially relied on expert intuition where biases may be allowed to influence the results. In order to overcome this and provide a more robust evaluation of the models, in this section we outline and test methods for generating the ideal meanings from the training data.

5.6.1 Refining Ideal Meanings

The notion of an 'ideal meaning', defined by a set of salient features and threshold values, underpins the proposal of this chapter, however it is not clear exactly how it should be represented in the model. Consider the example of 'in' whose ideal meaning ought to express a high degree of both *containment* and *location_control* and a perfect representation ought to require that both of these features have a value of 1. However, configurations with these values are extremely rare (out of 616 configurations from our

	Polysemy Model	Baseline Model	Partition Model
in	0.829	**0.836**	0.751
inside	0.869	**0.892**	0.773
against	0.802	**0.862**	0.824
on	**0.940**	0.927	0.917
on top of	**0.983**	0.975	0.955
under	**0.897**	0.775	0.856
below	0.760	**0.860**	0.803
over	**0.816**	0.779	0.728
above	**0.864**	0.850	0.774
Average	**0.862**	0.862	0.820
Overall	**0.875**	0.874	0.847

Table 5.6: Testing a partition model. K-Fold Test Results (K=10, N=10)

scenes, only one configuration fits these conditions and it is not a labelled instance of 'in'). As a result, assigning such a perfect representation for the ideal meaning will likely not yield good results as the model cannot be appropriately trained. It appears that a balance should be struck in how strict the definitions of the ideal meanings are such that the representation is both meaningful and gives training instances for the model.

Below two methods for refining the definitions of the ideal meanings of the Polysemy Model are presented.

Median Values

A simple way to determine threshold values in the definitions of the ideal meanings is to take the median values of the features in the training data. In this way, there will always be training instances for the ideal meaning as well as the other polysemes (provided there are at least as many training instances as polysemes). To ensure this provides a good representation of the ideal meaning we use the median values of 'good' training instances here (where the selection ratio is ≥ 0.5).

When the Polysemy Model has its ideal meanings defined in this way we will refer to the resulting model as the Median Model.

Refining Performance

So far parameters for the models proposed in this thesis have been trained indirectly, in what may be considered *transfer learning* — using categorical data from the Preposition Selection Task models are trained to create representations which perform well in the Comparative Task. Training in this way has various benefits, not least because categorical data is easier to collect and draw conclusions from.

As it is unclear how best to represent the ideal meanings and how exactly they can be refined on the categorical data, we have implemented a simple algorithm to refine the ideal meanings which trains the model based on performance in the Comparative Task on the training scenes.

Now, to refine the ideal meanings of the models half of the training scenes are kept as training scenes and half are used as validation scenes. The threshold values of the given salient features are then varied while the model is retrained on the training scenes and tested on the validation scenes (and vice versa). The model is updated with the values that produce the best performance and then retrained on all the original training scenes. For salient features with the $\leq$ relation the tested threshold values are $0.1, 0.2, 0.3, 0.4, 0.5$ and for features with the $\geq$ relation the tested threshold values are $0.5, 0.6, 0.7, 0.8, 0.9$.[39] This is obviously a very simple way to achieve this refinement, and could be expanded on, but displays the potential of the model and appears to be effective.

When the Polysemy Model has its ideal meanings defined in this way we will refer to the resulting model as the Refined Model.

5.6.2 Evaluation

As the Refined Model is being tuned using performance of the training scenes, training and testing the models using data from all scenes is unreliable so we omit these results here, though as expected the Refined Model does perform better than any other model in this case (average: 0.913, overall: 0.918). The results of 10 runs of the K-fold cross validation with K=10 are given in Table 5.7.

The Baseline Prototype Model has again performed very well for 'inside' and 'against' and the Refined Model has in general performed well except for the preposi-

[39] With the exception of *contact* where $0.1, 0.2, 0.3, 0.4, 0.5$ are used in both as values over 0.5 are very rare and *horizontal_distance* where $0.05, 0.1, 0.15, 0.2$ are used in both.

	Polysemy Model	Baseline Model	K-Means Model	Refined Model	Median Model
in	0.817	0.849	0.805	0.764	**0.878**
inside	0.869	**0.889**	0.823	0.825	0.829
against	0.804	**0.878**	0.861	0.811	0.863
on	0.947	0.928	**0.948**	0.937	0.933
on top of	**0.984**	0.981	0.977	0.984	0.854
under	0.895	0.785	0.876	0.890	**0.912**
below	0.741	0.857	0.881	**0.911**	0.794
over	0.811	0.804	0.728	**0.842**	0.681
above	**0.852**	0.817	0.820	0.832	0.833
Average	0.858	0.865	0.858	**0.866**	0.842
Overall	0.872	0.877	0.881	**0.889**	0.854

Table 5.7: Testing refined models. K-Fold Test Results (K=10, N=10)

tion 'in' — possibly due the Refined Model requiring larger amounts of data to train sufficiently and 'in' being a difficult preposition to gather a lot of data for. Nevertheless, the Refined Model performs better than the other models overall, and performs best on two out of the nine prepositions. Moreover, the overall improvement compared to the other models, barring the k-Means Model, is significant — the largest calculated p value when comparing the Refined Model with the other models (following the same procedure as in Section 4.4.2) is 0.0222 when the Refined Model is compared with the Baseline Prototype Model. The Refined Model has been refined in a rather crude way, but this demonstrates the possibility of determining the parameters of the ideal meanings from data and could be further improved.

The Median Model does not fare so well and performs worse overall than the given models. This result suggests that the models are sensitive to the definition of the ideal meaning and that any improvements of the polysemy models over the Baseline Prototype Model do not simply arise from an arbitrary partitioning of the data.

5.7 Model Properties

In this section, we will consider the Polysemy Model and Refined Model when they are trained on all the available data and analyse their properties and behaviour. This will give some insight into the functioning of the models and provide evidence that the models are genuinely capturing the polysemy exhibited by these terms.

5.7.1 Typicality Values

Firstly, to illustrate how the models assign typicality to configurations and how this compares to the baseline we consider an example. In Figure 5.9 we can see configurations of objects that appeared in the test scenes. The typicality scores of some of the configurations given by the models for 'on' are shown in Table 5.8.

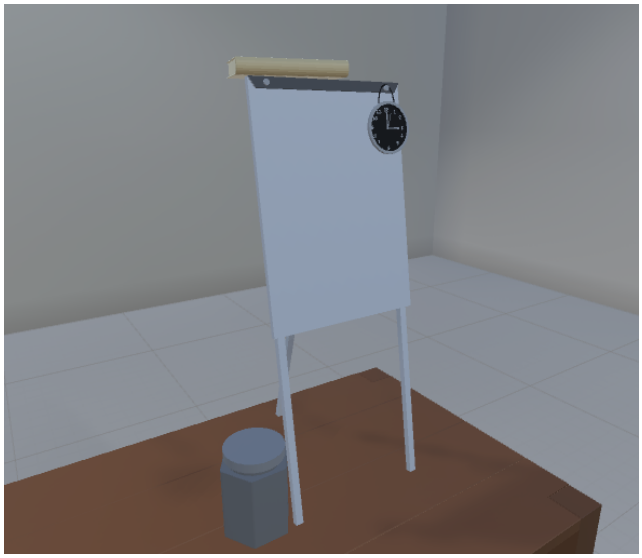


Figure 5.9: Example polyseme instances for 'on'.

The (book, board) configuration is an instance of 'on' with a high value of *above_proportion* and *support*, but as it is precariously balanced on top *contact* is low. The (clock, board) configuration is an instance of 'on' with a high value of *contact* and *support* but not *above_proportion*. The (jar, board) configuration was not labelled with 'on' by any participants and has low values of *contact*, *support* and *above_proportion*.

Clearly, (book, board) is closer to the canonical meaning of 'on' than (clock, board)

Configuration	Polysemy Model ($typicality \times rank$)	Refined Model ($typicality \times rank$)	Baseline Model
(book, board)	$0.779 \times 0.811 = 0.632$	$0.804 \times 0.785 = 0.631$	0.626
(clock, board)	$0.615 \times 0.776 = 0.477$	$0.615 \times 0.776 = 0.477$	0.204
(jar, board)	$0.477 \times 0.088 = 0.042$	$0.448 \times 0.087 = 0.039$	0.219

Table 5.8: Typicality scores assigned to some configurations for 'on'

and this appears to be represented in the values assigned by the Baseline Prototype Model as well as the ranks from the polysemy models. However, (clock, board) and (book, board) are both good examples of the respective senses of 'on' which they represent and we should expect (clock, board) to be assigned a reasonable typicality value. Moreover, (jar, board) is a mediocre example of its respective polyseme and this polyseme is far from the canonical notion of 'on', so we ought to expect this configuration to be assigned a low typicality value.

We expect (clock, board) and (book, board) to have similar typicality values and for these values to be higher than for (jar, board). This is roughly coherent with the collected testing data — when selecting the object described as 'the object on the board' participants are more likely to select the clock than either the book or jar and are more likely to select the book than the jar.

The polysemy models appear to deal with this better than the Baseline Prototype Model. Though they do assign a higher value to (book, board) than (clock, board) these values are similar, compared to the Baseline Prototype Model which assigns a very low value to (clock, board). The Baseline Prototype Model, in fact, assigns a higher value to (jar, board) than (clock, board) and therefore does not agree very well with participants in this scenario.

5.7.2 Generated Polysemes

We will now consider properties of the polysemes that the models have created.

Ranks and Ideal Meanings

Each preposition has been assigned an ideal meaning, defined by a set of conditions, and a collection of non-ideal meanings where at least one of the ideal conditions is

negated. For each polyseme, we have then assigned a rank from the data which should represent semantically how close the polyseme is to the ideal meaning and a sense of typicality *among* senses. We therefore expect, for each preposition, the rank assigned to the ideal meaning to be the highest and that as more of the ideal conditions are negated the rank should decrease.

For both models this is exactly what we observe for the 'polysemous' prepositions, with two small exceptions for the Polysemy Model.[40] This also holds in general for the 'non-polysemous' prepositions. This result suggests that we have appropriately assigned ideal meanings to the prepositions and that the semantics of the terms are indeed centred around such ideal meanings.

Clustering

To test how well the polysemy models partition the data into polysemes, here we estimate how well the polysemes cluster the given data. In the following we take polyseme instances to be any configuration that has been labelled with the preposition in the Preposition Selection Task and which fits the polysemes conditions.

In order to cluster the data with the generated polysemes, for a given preposition, first the mean feature values of the instances of each polyseme are calculated. This set of mean values then act as cluster centres and the inertia given by this clustering is measured (a point is assigned to the cluster with the nearest cluster centre[41]). These inertia values are then compared to inertia values given by a k-means clustering algorithm, see Figure 5.10 for the case of 'on'. The lower the value of the inertia the more internally coherent the clusters are. As we can see, the clustering using the polysemes performs quite well, equivalent to using the algorithm with $k = 4$.

In general, though the polysemes cluster the data worse than the k-means clustering algorithm, the polysemes appear to cluster the data reasonably well — see the github repository[42] for the respective plots.

[40] For 'in' the rank of the polyseme where both ideal conditions are negated is 0.0206 and the rank of the polyseme defined by high *containment* and low *location_control* is 0.02. For 'on' the polyseme defined by a high value of *support* and *contact* and low value of *above_proportion* has a higher rank than the ideal sense of 'on'.

[41] Note that here to be consistent with the inertia measure given by the k-means algorithm regular Euclidean distance is used rather than the weighted Euclidean metric used in Section 5.3.2.

[42] https://github.com/alrichardbollans/spatial-preposition-annotation-tool-unity3d/tree/master/Analysis/extra%20thesis%20results

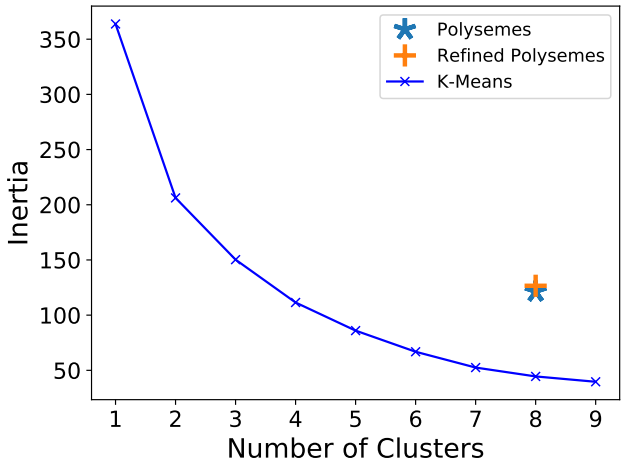


Figure 5.10: Inertia from k-means clustering vs. Polyseme clustering for 'on'.

5.8 Discussion

In this chapter we have explored how semantic models may be improved to account for polysemy when processing referring expressions involving spatial prepositions. Primarily, we have provided methods which distinguish meaningful clusters within categorical data on spatial prepositions. By simplifying the 'principled polysemy' criteria [30] for distinguishing polysemes, an approach has been developed which can be exploited by semantic models more generally. In sufficiently similar datasets where configurations are annotated with these terms, the annotations can be separated into distinct polysemes such that semantic models trained on the data can learn the semantics of individual polysemes.

We have also introduced a notion of a 'polyseme hierarchy' — a value which corresponds to how strongly a particular polyseme is associated with the given preposition — as well as methods for determining its value. In combining this with the generated polysemes, we have provided a semantic model which significantly improves on the given baseline when interpreting a particular class of referring expressions.

As we have discussed in Section 2.2.4, the kind of senses modelled here may or may not constitute polysemy in a given theoretical framework, however we have assumed that this kind of semantic variability is important to model if agents are to reliably use and interpret spatial language. That we have significantly improved on the Baseline Prototype Model by accounting for polysemy appears to support this assumption.

The initial report of [2] only considered those prepositions which according to existing literature appear to be polysemous, however the methods outlined in this chapter also appear to be applicable to some 'non-polysemous' prepositions. Moreover, we have provided some evidence that these 'non-polysemous' prepositions may be considered polysemous.

While we have initially relied somewhat on intuition to generate ideal meanings for the models, we have also shown that further improvements can be made to the model using the training data to refine the ideal meanings.

Finally, by analysing some of the behaviour and properties of the generated model we have provided evidence that the models do capture polysemy and create meaningful distinctions of the data.

As well as the polysemy models based on ideal meanings, we have created a model based on a k-Means clustering algorithm. When testing on the 'polysemous' prepositions, the k-Means Model provides significant improvement over the Baseline Prototype Model and, along with the reasonable performance of the Polysemy Model, provides further evidence that the selected prepositions do exhibit polysemy.

Throughout this current and previous chapter we have evaluated various models of spatial prepositions and focused on developing a general model which works well for all the terms considered in this thesis. However, at each stage of evaluation we have seen that no single method provides the best model of all the prepositions. For example, in the K-Fold testing of Sections 5.4.2 and 5.6.2 the k-Means model outperforms the other models for 'on' and the Baseline Prototype Model performs consistently well for 'inside' and 'against'. It may be the case that further improvements can be made by developing models separately for each preposition.

Chapter 6

Categorisation, Typicality and Object-Specific Features in Spatial Reference

A functional aspect of spatial prepositions which again is widely recognised but has not been accounted for, either so far in this thesis or the wider field, is the influence of object-specific features — related to object properties and affordances, see Section 2.2.3. In order to include these features in semantic models it is important to understand the influence these features have when interpreting and generating utterances.

So far existing studies only relate object-specific features to utterance generation but not to utterance interpretation [5, 31, 32] where the notion of typicality is often more salient. It would appear that object-specific features are mostly salient when making category rather than typicality decisions, suggesting that decisions made in the previously discussed Preposition Selection Task fundamentally differ to those made in the Comparative Task. Moreover, object-specific features are in general concerned with properties of the ground, so the role of these features was limited when assessing performance of the models in the Comparative Task as the ground in this task is fixed. For this reason, it was not clear that object-specific features were required in the models we have so far generated.

As discussed in Section 2.3.5, various accounts of cognition and semantic representations have highlighted that, for some concepts, different factors may influence category and typicality judgements [26, 27]. In particular, some features may be more salient in categorisation tasks while other features are more salient when assessing typicality. In this chapter we explore whether this distinction exists for spatial prepositions based on varying salience of object-specific features.

In existing models of spatial language (and semantic models more generally), it is generally assumed that the underlying semantics of categorisation and typicality are essentially the same. However, as we will discuss in Section 6.4, appropriately modelling categorisation and typicality judgements is important when generating and processing referring expressions.

The main hypothesis of this chapter is that object-specific features are more salient in categorisation, while geometric and physical relationships between objects are more salient in typicality judgements. In this chapter we test this hypothesis using data from the study described in Section 3.3.

Based on the collected data we cannot verify the hypothesis and will conclude that object-specific features appear to be salient in both category and typicality judgements, further evidencing the need to include these types of features in semantic models. We

will then propose how such features may be incorporated into semantic models.

6.1 Indications of a Distinction

During analysis of the dataset on which Chapters 4 and 5 are based, there were some indications that the notions of categorisation and typicality may diverge for spatial terms and we will outline these here.

Firstly, there were 18 pairs of configurations where for some preposition, one of the configurations was more likely to be labelled with the preposition in the Preposition Selection Task and the other configuration was more likely to be selected in the Comparative Task.[43] An example of this is shown in Figure 6.1: the (cube, cup) configuration was more often labelled with 'on' than (pencil, cup), but the pencil was more often selected when asked to select 'the object on the cup'.

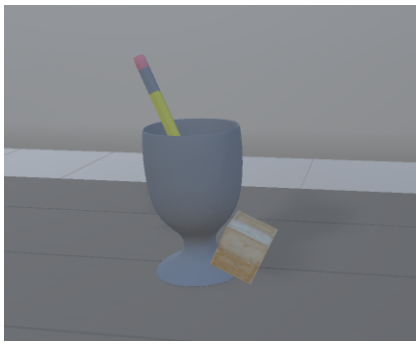


Figure 6.1: A motivating example of disagreement.

One possible reason for this disagreement is that the pencil is inside the containing part of the cup, which is a container, and therefore 'in' rather than 'on' is the preferred preposition when labelling (pencil, cup). One can't use 'in' in the same way for (cube, cup) and so 'on' is more likely to be used when labelling the configuration. However, the pencil is physically 'on' the cup — arguably more so than the cube — and therefore when comparing (pencil, cup) and (cube, cup) to select the best candidate for 'the object on the cup', the pencil is a justifiable selection.

Secondly, in the previous chapter we outlined the Refined Model (Section 5.6) which appears to assign a measure of typicality to configurations which agrees very well with

[43]Note that in the Comparative Task the judgements made are not a direct comparison of distinct configurations — the configurations share the same ground and the most appropriate figure is selected.

participant judgements in the Comparative Task. If we train this model on all the available data and plot the selection ratio (likelihood of categorisation in the Preposition Selection Task) of configurations against the typicality assigned by the model, we can see numerous pairs of configurations where one configuration is more typical and the other is more likely to be labelled with the preposition. Figure 6.2 provides such a plot in the case of 'on' (the model achieved a very high score for this preposition in the K-Fold evaluation).

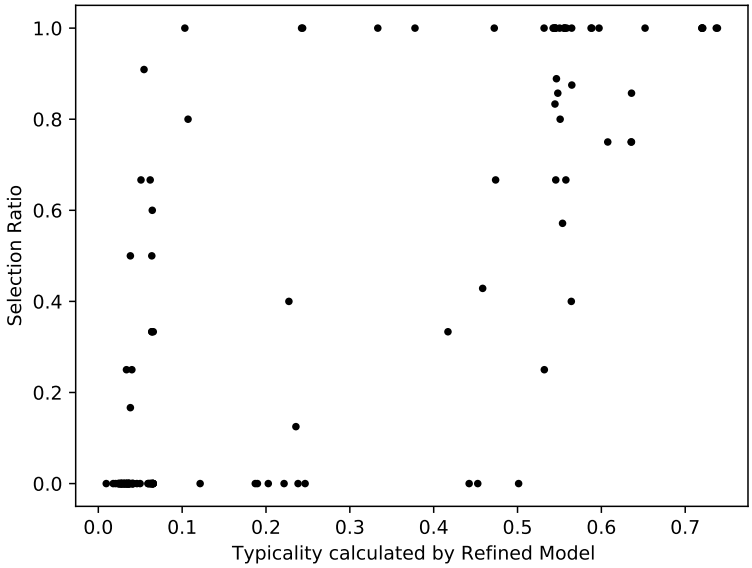


Figure 6.2: Plotting typicality calculated by the Refined Model for 'on' against the selection ratio.

From this scatter plot it is clear that these measures of categorisation and typicality are strongly correlated. To confirm this, we have also calculated the Spearman rank-order correlation coefficient comparing these measures for each preposition, see Table 6.1. We can see that in general these values are strongly correlated, which should be unsurprising. However, 'on' has the second highest correlation of any of the prepositions and even so we can see many pairs of configurations where one is more typical and the other is a better category member.

We therefore believe that the notions of categorisation and typicality may be distinct for spatial prepositions. However, it is clear that these insights may simply be the result of noise in the data and this chapter is focused on investigating this issue more

Preposition	Correlation
in	0.34
inside	0.42
against	0.56
on	0.77
on top of	0.78
under	0.66
below	0.72
over	0.49
above	0.74

Table 6.1: Spearman rank-order correlation coefficient of selection ratio and typicality calculated by the Refined Model

thoroughly.

6.2 Hypothesis

The main hypothesis being tested is that categorisation and typicality judgements may differ for spatial prepositions in the following manner: given configurations c_1, c_2 and preposition P, participants may be more likely to categorise c_1 with P yet more likely to select c_2 as a better instance of P. We hypothesise that this may arise in part because particular features are salient in category judgements which become less salient in typicality judgements. Note that under this hypothesis the relationship between categorisation and typicality is non-monotonic, making this in a sense stronger than the findings related to 'red' discussed by Osherson and Smith [114] where two entities are unambiguous members of a concept yet one is more typical than the other.

This hypothesis is grounded in the idea that features can be separated into *defining* and *characteristic* features [26]; and that object-specific features are defining features, while the physical relationships are characteristic features. Therefore, in general we expect that object-specific features will be more salient in categorisation while geometric and physical relationships, such as *containment* or *support*, will be more salient in typicality judgements.

We expect this to be the case as conventions, relating to object-specific features, appear to strongly influence and constrain which prepositions are used with particular figure or ground objects i.e. conventions strongly influence category decisions. For example, as discussed in the context of schematization in Section 2.2.2, one usually says 'on the bus' but 'in the car' even though the geometric and functional relationships present are very similar. To say '*on* the car' immediately invokes an image of an object on top of the roof of the car, while '*in* the train' suggests an object within the walls or some working part of the train.[44]

As previously discussed, we suppose that the semantics of spatial prepositions arise from *ideal meanings* which are primarily geometric in nature and may be represented as some geometric relation between abstract objects. Object-specific features are therefore not represented in the ideal meanings, while physical relationships are. In contrast to the task of categorisation, we suspect that similarity to these ideal meanings is more salient in assessing typicality. As a result, physical relationships would then be more salient than object-specific features when assessing typicality.

Moreover, I believe that the influence of object-specific features may be more salient for the functional prepositions than for their geometric counterparts and that the distinction in category and typicality judgements related to object-specific features will therefore be more pronounced. This would be somewhat a corollary of the assumption that functionality in general is more salient for the functional prepositions, for which tentative evidence is provided in Section 4.4.3 for these prepositions and in [10] in the case of 'over/above' and 'under/below'. For example, we can imagine that usage of 'inside' relies more heavily on an ideal notion of containment compared to 'in' whose usage relies on this ideal notion as well as various object-specific features.

6.3 Updating the Experimental Set Up

As previously mentioned, the tentative indications of a distinction in category and typicality judgements given in Section 6.1 may simply be the result of noise in the data. Therefore we conducted a new study, described in detail in Section 3.3, which would allow a more rigorous comparison of the two notions.

[44] It appears that flouting the convention of using 'in'/'on' with 'car'/'train' conveys that the figure is related to the ground in an unusual way.

Thus far, we have been considering typicality judgements elicited in the Comparative Task. The judgements are made in a restricted setting where the aim has been to determine an appropriate figure when given a description with a fixed ground, e.g. 'Select the object on the table'. This restricted set up causes object-specific features (in particular of the ground) to be less salient. This arises as, if the ground object is fixed and known to the participants then any reasoning regarding specific properties of the ground is likely to become unnecessary.[45]

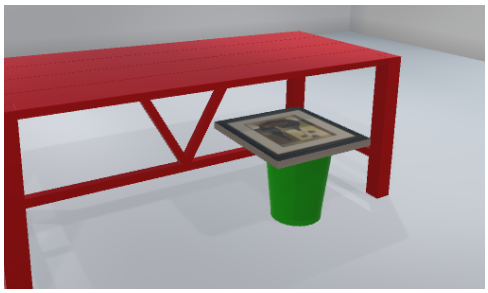
The Typicality Task used in this chapter improves on the Comparative Task in this regard by asking participants to compare pairs of distinct configurations. In this way the influence of different ground objects may be assessed. To make the given description applicable to both configurations, the figure and ground object are simply named 'green object' and 'red object'. Recalling the description of the data collection environment given in Section 3.3, the participants are shown two configurations and asked which one best fits a description of the form 'the green object [p] the red object' (where [p] is substituted for one of the prepositions being tested).

In order for the question phrasing to be the same in the Categorisation Task, again the objects are named 'green object' and 'red object' and participants are asked to select all prepositions which could fill the blank in 'the green object __ the red object'.

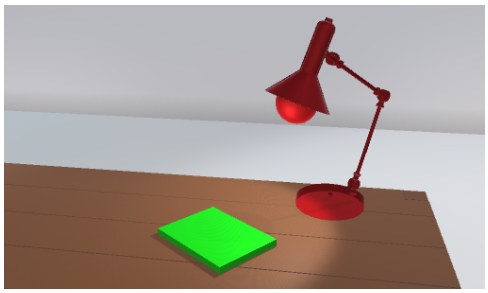
As per the hypothesis in Section 6.2, the main aim is to compare the influence of object-specific features and physical relationships. As such, scenes in the study were constructed for each preposition which varied the presence of salient object-specific features and the suitability of the physical relationships for particular prepositions. For example, the scenes shown in Figure 6.3 were created for the preposition 'under', for a collection of all used scenes see Appendix C. In (a) the bin is very much physically 'under' the table and the role of object-specific features is limited. In (b) the notepad is not physically under any part of the lamp, but there is a functional interaction present of the lamp (a light source) illuminating the notepad (an object which one often needs illuminating). In (c) the pencil is somewhat under the shelf and there are no particular object-specific features present. In (d) the sink is under some part of the tap and there is a functional interaction between the tap (providing water) and the sink (catching water from the tap). So, we have two scenes with salient object-specific features present

[45] In general it appears that object-specific features of the ground influence spatial prepositions more than that of the figure, as discussed in Section 2.2.3.

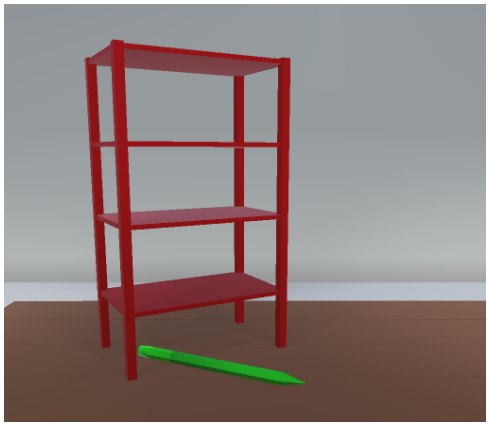
and two scenes without, and there is an apparent gradation of the physical presence of 'under' in the scenes (a>c>d>b).



(a) The bin under the table



(b) The notepad under the lamp



(c) The pencil under the shelf

(d) The sink under the tap

Figure 6.3: Scenes used for 'under'.

6.4 Categorisation and Typicality in Referring Expressions

Before further investigating this issue we will briefly highlight the possible implications for REG/C of a significant distinction between category and typicality judgements. If we have so far generated successful models, what does it matter if categorisation and typicality differ?

Firstly, clearly categorisation must be modelled appropriately in order to *produce* rather than interpret utterances. Unlike what we have considered so far, this is often reliant on selectional restrictions which constrain when a term is applicable. For example,

[92] suggest that the figure should be smaller than the ground for 'in' to apply.

Secondly, as outlined in Section 6.3, the evaluation of typicality judgements provided by the Comparative Task is somewhat restricted. However, in more open tasks involving referring expressions it is likely that both speakers and listeners need to model both category and typicality judgements.

Suppose we have a speaker and listener, intended referent, r, set of distractor objects,[46] O, and suppose that the speaker is generating a description, D.

When the speaker generates an utterance in order to refer to r, there are various semantic and pragmatic considerations they must make. A naive model of such a speaker may simply find a concept within its vocabulary which is most suitable for r. This would clearly be a flawed strategy, however, as there may be other entities in the scene which better fit the concept — an expression may be true but not satisfy the speaker's communicative goals and so pragmatics must be considered. A more refined speaker model may find a description which best distinguishes r from the distractor objects in O, similar to the algorithm of [159] which aims to maximise the *discriminatory power* of a description while minimising superfluous information.

It appears that the speaker must model how well a description fits an object (category judgement) *and* how well an object fits a description compared to other objects (typicality judgement). Moreover, it is apparent that humans will reason recursively about possible intentions of speakers and possible interpretations of listeners [63], making these judgements also necessary for listeners.

To consider a more concrete example, suppose we have a scenario as in Figure 6.4 where a bowl, b, is on a table and there is one cube, c_{red}, in it and one cube, c_{blue}, next to and not touching it. It seems plausible that humans are more likely to categorise the configuration (c_{blue}, b) with the preposition 'near' than the configuration (c_{red}, b) even though when comparing the configurations humans may agree that c_{red} is more 'near' the bowl than c_{blue}.

Suppose a speaker gives an utterance 'the cube near the bowl' in order to refer to c_{blue} which an agent must interpret. As may be expected, semantic models, e.g. [78], are likely to assign a better score for 'near' to (c_{red}, b) than (c_{blue}, b). If the system has a crude strategy for interpretation which simply selects the configuration with the highest semantic score, then such a system would erroneously select c_{red}.

[46] A distractor object is an object in the scene which the speaker is not intending to refer to.

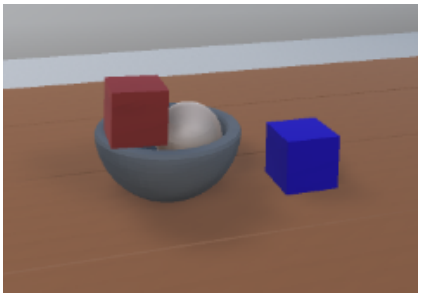


Figure 6.4: An example of possible confusion when not accounting for object-specific features.

Many systems with more sophisticated pragmatic strategies, e.g. [89], have been developed which aim to take into account and reason with the possible utterances available to the speaker. In this case, such a system may correctly select c_{blue} if it recognises that other better utterances would have been available to the speaker if the intended referent was c_{red}. 'the cube in the bowl' would be a clear example of such an utterance which seems to clearly identify c_{red} over c_{blue}. However, supposing that our hypothesis is correct, we contend that this marked distinction is not simply a matter of typicality and the fact that it would be unusual to categorise (c_{red}, b) with 'near' is more salient than any distinction in typicality between the two configurations. As c_{red} is not 'in' the bowl in an ideal sense we can imagine that a semantic model based simply on the physical relationships between the objects may provide a more marked distinction between the configurations for 'near' than for 'in'. In this case, 'the cube in the bowl' wouldn't necessarily seem like a better utterance to identify c_{red} than 'the cube near the bowl'.

To see that understanding and modelling the differences between categorisation and typicality is also important for producing utterances in REG, suppose a speaker creates an utterance where r is more typical for the concept, C say, in D than any of the distractor objects. If categorisation is not aligned with typicality, such a strategy may produce unusual utterances where, though r is typical for C, it is uncommon to categorise r with C. Such unnatural utterances may trigger unwanted conversational implicatures and be a source of confusion. For example, the utterance 'the ball on the bowl' in the context of Figure 6.4 would be an unusual way to describe the ball as the preposition 'in' is often used with objects such as bowls. From this unconventional usage of 'on' a listener may imply that for some reason 'in' was not suitable e.g. if the

speaker is actually referring to another unseen ball.

The issue of producing natural utterances has been recognised by others in the field and is an important challenge to overcome if we are to develop more sophisticated REG systems. For example, Krishnaswamy and Pustejovsky [137] recognise that though their system is able to interpret spatial terms well, it cannot fluently use them to refer to objects.

6.5 Results

In this section we analyse the data collected in the updated experiment in order to test the hypothesis and discuss the insights gained from the study as well as the implications for semantic models.

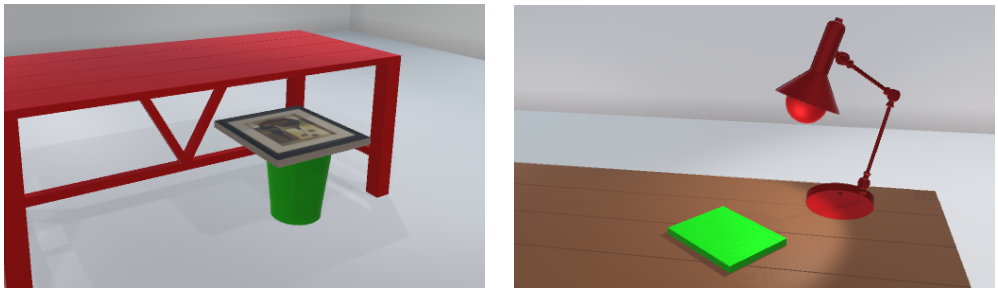
6.5.1 Comparing Categorisation and Typicality

To analyse the collected data, for each preposition we consider pairs of tested configurations and evaluate the degree to which category and typicality judgements differ. For each pair, (c_1, c_2), we analyse whether c_1 is a genuinely better category instance than c_2 or if c_1 is more typical than c_2. To do this a simple hypothesis test is used with significance level 10%[47] and null hypothesis that the given configurations are equally likely to be labelled with the preposition (in the category case) or equally likely to be selected (in the typicality case). In the category case, the p-value is calculated using the one-tailed version of Fisher's exact test using SciPy's implementation [154]. In the typicality case, the p-value is simply: $p = \sum_{k=C_{1,2}}^{N} \binom{N}{k} \times 0.5^N$, where N is the number of times the pair is tested and $C_{1,2}$ is the number of times c_1 is selected over c_2. In 22 out of the 49 given pairs, one of the configurations is a significantly better category instance or is more typical than the other. Full results for each pair of configurations are given in the Appendix, see Tables C.1 and C.2.

Considering the somewhat trivial case, similar to the case of 'red' discussed in Section 2.3.5, where two entities are both unambiguous cases of a concept but one of the entities is more typical than the other; there is one instance of this in our dataset. For the preposition 'under' the configurations shown in Figures 6.5(a) and 6.5(b) were

[47] A lenient significance level of 10% is used here to highlight that even in this case no positive instances are found to confirm the hypothesis.

both always labelled with the preposition, out of seven tests in the former and ten tests in the latter, but the (bin, table) configuration in Figure 6.5(a) is significantly more typical than the (notepad, lamp) configuration in Figure 6.5(b). As previously discussed, however, this is an unsurprising result. Moreover, such results are unreliable as there is always a degree of uncertainty in judging whether an entity is an unambiguous case of a concept e.g. if the (notepad, lamp) configuration was tested an extra time then the participant may have not labelled it with 'under'.



(a) The bin under the table

(b) The notepad under the lamp

Figure 6.5: An example from 'under'.

This example does not support the main hypothesis of this chapter as the (notepad, lamp) configuration is not a significantly better category member than (bin, table). However, it appears that the functional interaction of the lamp illuminating the notepad causes (notepad, lamp) to be categorised on a par with (bin, table) even though the physical relationships between (notepad, lamp) do not very well represent 'under'. That (notepad, lamp) is judged to be significantly less typical than (bin, table) provides some indication that the physical relationships become more salient in typicality judgements.

Regarding the main hypothesis of this chapter, there are no pairs of configurations in our dataset where one of the configurations is a significantly better category member and the other is significantly more typical. Moreover, in only nine pairs is there any possible disagreement in which one of the configurations is more often labelled with the preposition and the other configuration is more often selected in the Typicality Task. In all but one of these cases of disagreement neither configuration is a significantly better category member or is significantly more typical. We therefore cannot conclude that our hypothesis is correct based on the collected data.

Clearly we have only tested a small number of features and there is a vast array

of salient features for each preposition for which the hypothesis may still be correct. However, our results suggest that object-specific features are salient in both categorisation and typicality judgements — in some cases the object-specific features appear to have a stronger influence than the physical relationships. It therefore appears that object-specific features are defining and characteristic features. Interestingly, there is some tentative evidence that this extends in general to the geometric counterparts and suggests that these prepositions are not purely spatial — supporting findings in [45].

6.5.2 Importance of Object-Specific Features

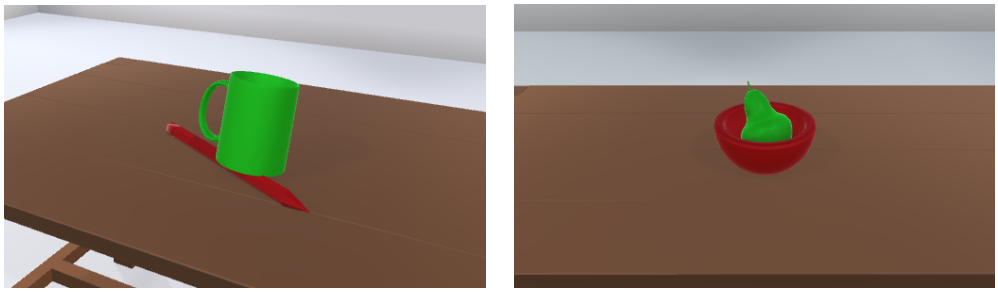
In the following we provide some collected examples which highlight the importance of object-specific features.

On/On top of For the preposition 'on', the (mug, pencil) configuration in Figure 6.6(a) is both a significantly better category member and is significantly more typical than the (pear, bowl) configuration in Figure 6.6(b). Regarding the physical relationships, (pear, bowl) appears to be a better example of 'on' than (mug, pencil). If we consider the usual salient features for 'on':

- The pear is fully *supported* by the bowl, while the mug is leaning on both the pencil and the table
- There is a high degree of *contact* between the pear and the bowl compared to the mug and pencil
- The entirety of the pear is *above* some part of the bowl, while the bottom of the mug is level with the bottom of the pencil

We therefore believe that the preference for (mug, pencil) is not due to the physical relationships of (mug, pencil) better representing 'on' than (pear, bowl) and that this result arises primarily because 'on' is generally not used for containers. One may have expected this result if the objects in the experiments were named — 'on the bowl' sounds strange while 'on the pencil' seems more plausible. It is therefore even more surprising given that the objects were not named in a way that influences the decisions.

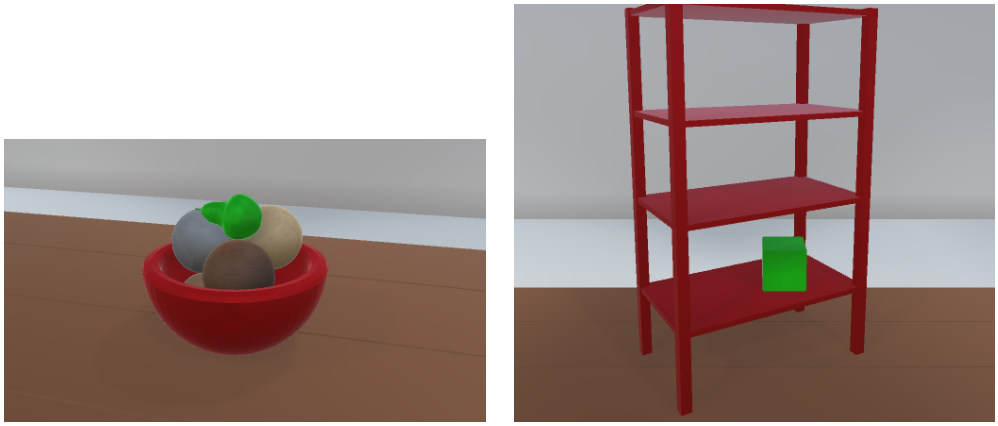
We also see the same result for 'on top of'. However, whether the physical relationships of (pear, bowl) better represent 'on top of' than (mug, pencil) is less clear than



(a) The mug on the pencil (b) The pear in/on the bowl

Figure 6.6: A comparison of configurations for 'on'/'on top of'.

in the case of 'on'. Being more geometrically biased, the degree to which the figure is above the ground is more important for 'on top of' than 'on'.[48] Though part of the mug is level with the lowest part of the pencil, the majority of the mug is above the highest part of the pencil and so *above_proportion* is high compared with (pear, bowl). It is therefore plausible that the preference for (mug, pencil) *is* due to the physical relationships of (mug, pencil) better representing 'on top of' than (pear, bowl).



(a) The pear in the bowl (b) The cube in/on the shelf

Figure 6.7: A comparison of configurations for 'in'/'inside'.

In/Inside For the prepositions 'in' and 'inside', the (pear, bowl) configuration in Figure 6.7(a) was more likely to be selected in the Typicality Task than the (cube,

[48]This is supported by a higher regression weight and prototype value for *above_proportion* in the Baseline Prototype Model.

shelf) configuration in Figure 6.7(b).

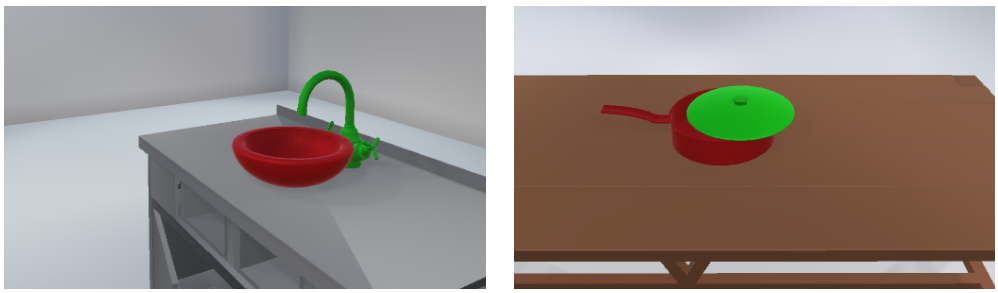
There is no apparent geometric *containment* of the pear by the bowl (barring the kind of schematizations discussed in Section 2.2.2). The degree to which the shelf contains the cube is debatable as there are no panels/sides to the shelf which obscure it from the room. However, the cube is fully contained in the convex hull of the shelf and so is contained in some sense.

From a modelling perspective, in grounded semantic models containment is usually measured as the degree to which one object is contained in the bounding box of another, e.g. [77, 78, 88]. Using this measure, the cube would be fully contained by the shelf whereas the pear is not even partially contained by the bowl.

Regarding the functional relationship of *location_control* which appears to be salient for 'in', (cube, shelf) and (pear, bowl) appear to be similar in this regard.

It therefore appears that the physical relationships of (cube, shelf) better represent both 'in' and 'inside' than (pear, bowl). As a result, it appears that the role of the bowl as a type of container is influencing typicality judgements for both these prepositions.

Over/Above For the preposition 'over', the (tap, sink) configuration in Figure 6.8(a) is both a significantly better category member and is significantly more typical than the (lid, pan) configuration in Figure 6.8(b). The same is true for the preposition 'above', though in this case the results are not significant.



(a) The tap over the sink

(b) The lid over the pan

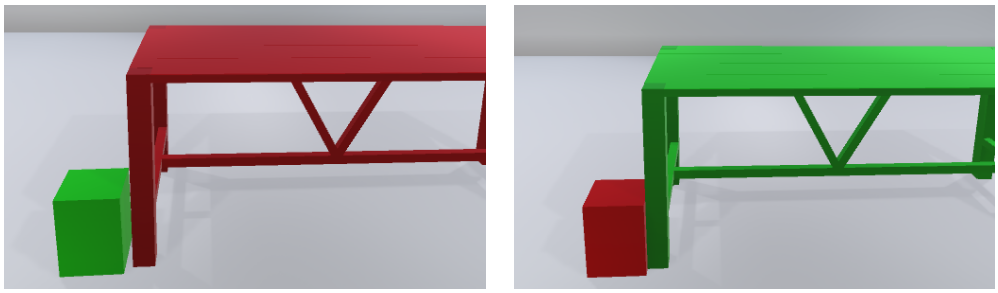
Figure 6.8: A comparison of configurations for 'over'/'above'.

Again, the physical relationships of (lid, pan) appear to better capture the geometric meanings of 'over' and 'above' than (tap, sink). There is also some functional interaction between the objects in both cases — lids are used to cover pans and sinks are

placed below taps to catch the water. The preference for the (tap, sink) configuration is therefore somewhat surprising.

However, the lid does not appear to be properly fulfilling its functional role, as it is not fully covering the container part of the pan. This may explain the preference for (tap, sink) and further highlight the importance of considering functional interactions based on usual object usages.

Against For the preposition 'against' the (box, table) configuration in Figure 6.9(a) is significantly more typical than the (table, box) configuration in Figure 6.9(b). Again, the physical relationships of (table, box) seem to better represent 'against' than (box, table) as in the former the two objects are in contact while in the latter they are not. We believe that this is a result of the preference for 'against' where the figure is mobile and the ground is fixed.



(a) The box against the table (b) The table against the box

Figure 6.9: A comparison of configurations for 'against'.

6.6 Discussion

Though this chapter has been focused on a specific hypothesis which has not been successfully proven, the insights from the collected data provide some considerations for future attempts at modelling spatial prepositions.

With regards to existing models of spatial language, it is generally assumed that the underlying semantics of categorisation and typicality are essentially the same. For example, in the PRAGR mechanism proposed in [64] a pragmatic strategy is presented which aims to maximise both the *acceptability* and *discriminatory power* of a description. Acceptability is calculated using similarity to a prototype based on physical rela-

tionships while the discriminatory power is calculated considering the acceptability of the description for the referent compared to other distractor objects.

We have not been able to provide significant evidence that typicality and category judgements differ for spatial prepositions. Therefore, it may be possible to continue in this fashion and treat the semantics of categorisation and typicality equally. However, regardless of the hypothesis, these results highlight the importance of including object-specific features in semantic models of spatial language. These types of features are rarely included in semantic models and many systems are developed in block-world type environments, e.g. [64, 160, 161], where these features are not needed. Platonov and Schubert [78] provide a possible exception to this, as their model aims to account for synecdoche by tagging and iterating over 'salient parts' of objects.

It is understandable that the role of object-specific features is often neglected as if an object exhibits some specific function or role which is associated with a preposition it generally lends itself to the geometric notion associated with the preposition. For example, it is easy for things to be geometrically contained in containers or to be covered by lids. One may even suppose that the apparent salience of object-specific features is simply a consequence of them being highly correlated with the relevant physical relationships e.g. geometric containment co-occurs with the ground being a container. However, as we have seen in Section 6.5.2 object-specific features appear to be salient even in the absence of the usual physical relationships.

Moreover, it is not immediately clear how or when object-specific features should be accounted for in typicality judgements. The type of typicality judgements occurring in the Typicality Task are uncommon when processing locative expressions. In dialogue humans usually name both figure and ground and moreover, humans select ground objects in such a way as to reduce ambiguity. As a result, the kinds of typicality judgements made often resemble the Comparative Task i.e. where the ground the speaker is intending to use is made clear to the listener and the figure object is named e.g 'pass me the box under the table'. In such scenarios the utility of considering object-specific features is unclear as any plausible distractor objects will share object-specific features with the referent and similarly for ground objects e.g. when interpreting 'the pencil in the mug' the set of distractor objects are the pencils in the scene and possible grounds are mugs in the scene. However, it is clear that object-specific features are necessary to account for in categorisation and also it may be the case that the existence of cer-

tain object-specific features modify how typicality should be calculated, e.g. *covering* becomes more salient for 'over' compared to *above_proportion* when the ground is a container and the figure is a lid.

In this way, object-specific features may indicate distinct senses, similar to the work of Rodrigues et al. [57] where some object-specific features, e.g. whether or not the ground is a type of container, are treated as features which distinguish separate polysemes. Following this idea, it may be possible to use the methods for modelling polysemy discussed in Chapter 5 to incorporate object-specific features. Currently, these methods rely on distinguishing polysemes based on the similarity to an ideal meaning but this approach may be extended by also first distinguishing polysemes based on salient object-specific features. So for example, for the preposition 'in' we currently have four distinct polysemes, in_1 in_2 in_3 and in_4 say, defined by the degree of *containment* and *location_control* present. We may include a feature, *ground_container* (representing whether or not the ground is a type of container), by using it to further distinguish different senses of 'in' e.g. p_1 p_2 p_3 p_4 and p'_1 p'_2 p'_3 p'_4 where p_i uses the definition of in_i as well as the condition that *ground_container* is true, while p'_i uses the definition of in_i as well as the condition that *ground_container* is false.

Chapter 7

Discussion

7.1 Conclusion

The main aim of this thesis has been to extend semantic models of spatial prepositions to better account for their semantic variability. In particular, we aimed to model spatial prepositions in a more developed feature space which includes functional features and also to account for the fine-grained polysemy exhibited by spatial prepositions in grounded settings.

Regarding the first point, the motivation for extending the set of features used to model each term was in part to be able to include functional features which are widely acknowledged as being salient, but also, as discussed in Section 2.2, many 'non-salient' features can influence the usage of spatial prepositions. In Chapter 4 we considered models based on cognitive accounts of concept representations which incorporate an extended range of features compared to most 'Simple Relation' models. Parameters for these cognitive models were trained on a relatively small set of categorisation data and a model based on Prototype Theory appeared to significantly outperform the other models. Moreover, we showed that this Baseline Prototype Model performed significantly worse when the functional features of *support* and *location_control* were not included, supporting the novel inclusion of these features in the model and further evidencing their salience.

By automatically generating weights and prototypes for concepts from data we provide a method for modelling concepts where the semantics are less clear and where limited prior knowledge is required. This has allowed us to tackle the second point, in Chapter 5, and model the semantics of distinct senses of spatial prepositions. In order to achieve this it was necessary to have some method of identifying instances of distinct senses and we have provided a method for achieving this based on Herskovits' notion of 'ideal meanings' as well as a modification of the 'principled polysemy' framework of Tyler and Evans. Furthermore, in order to incorporate these senses into a useful semantic model, we have introduced the notion of 'polyseme hierarchy' and methods for quantifying this notion.

Finally, a functional aspect of spatial prepositions which again is widely recognised but has not been accounted for, either in this thesis or the wider field, is the influence of object-specific features. In order to include these features in semantic models it is important to understand the influence these features have when interpreting and generating utterances.

So far existing studies only relate object-specific features to categorisation tasks but not to utterance interpretation [5, 31, 32] where the notion of typicality is often more salient. It would appear that object-specific features are mostly salient when making category rather than typicality decisions, suggesting that decisions made in the Preposition Selection Task fundamentally differ to those made in the Comparative Task. Moreover, object-specific features are in general concerned with properties of the ground, so the role of these features was limited when assessing performance of the models in the Comparative Task as the ground in this task is fixed. For this reason, it was not clear that object-specific features were required in the models we have so far generated.

However, it is plausible that the influence of object-specific features is not limited to categorisation and that they also have some influence in typicality judgements. In Chapter 6 we set out to show that object-specific features in fact were mostly influential in categorisation and not in typicality judgements while providing evidence that these two notions were distinct for spatial prepositions. However, based on the collected data we were not able to show this and it appeared that object-specific features were salient in both tasks. Though we have not accounted for these features in our semantic models, we have discussed how the methods developed in this thesis may be used to incorporate these features in semantic models.

7.2 Limitations

7.2.1 Standpoints

As with most work on the semantics of natural language, the models we have constructed accommodate a single interpretation of the language and assume that any variation among humans in how they understand the semantics of these terms is negligible. However, humans may hold different *standpoints* [162] and have different conceptualisations of what these terms mean.

In general, it appears that these types of disagreements are not strong enough to severely limit communication and semantic models can perform well even when making inevitable generalisations about peoples' standpoints, for example see the performance of the Refined Model in Section 5.6.2. Moreover, there are pragmatic strategies which may help to overcome these disagreements [65, 67].

7.2.2 Synecdoche

An important aspect of spatial language that has not been considered in this thesis is the influence of synecdoche. For example, in the utterance 'the box under the table' it is plausible that 'table' is being used to refer to one of its parts i.e. the tabletop, as the box is physically under the tabletop but not the table legs. Platonov and Schubert [78] have provided a possible solution to this issue, by iterating over salient parts of objects and checking to what extent these salient parts fit the given preposition. Their computational model is able to do this as they labelled salient parts by hand, however it may be possible to automate this labelling process such that it could be carried out by an autonomous agent [163].

7.2.3 Applications to Other Languages

The discussions in this thesis have been been limited to spatial prepositions in English and the extent to which these methods will apply to other languages is unclear. Though different languages partition different spatial situations differently, it may still be the case that the semantics of spatial prepositions are built on prelinguistic concepts that are shared across languages and on which we have attempted to base our models. For example, the treatment of polysemy relies on ideal meanings which are arguably prelinguistic, though this is difficult to ascertain. In work on this topic Bowerman and Choi [6] found that spatial semantic categories are highly influenced by the language and that prelinguistic concepts do not directly influence the formation of these categories in children. As a result, it should not be taken for granted that the analysis in this thesis can be translated to other languages.

7.2.4 Application to Real-World Settings

In this thesis we have not attempted to outline an end-to-end system which translates real situated perception to language, as the focus has been on a rich semantic analysis. As a result, in experiments utterances have been constrained such that the semantic content of single lexical items can be analysed and, moreover, experiments have been carried out in 3D virtual environments such that a rich feature set could be reliably extracted. Nevertheless the semantic model may be used in real world environments providing similar features to the ones we have used can be extracted from real world scenarios and there are a variety of recent techniques which may allow for this. One way

this can be achieved by a situated agent is to generate 3D models from the surrounding environment. For example, the system of Tulsiani et al. [164] is able to generate 3D scenes from 2D RGB images with a good level of accuracy, further detail and accuracy can be attained in grounded settings using RGB-D images such as in [165, 166].

7.2.5 Benchmarking

As discussed in Section 2.5 we haven't been able to compare performance of our models against others in a standard benchmark task. Nevertheless, it may be possible to compare model performance by testing our model on datasets provided in other research. However, where feature values for configurations in scenes are provided in similar datasets, some features, e.g. *support* and *location_control*, are non-existent and where equivalent features exist they are not precisely the same measures, meaning that our models would have to be retrained with the new features in order to test their performance. In general this would be possible, but would require a significant amount of work for each dataset and instead a well-established benchmark would be desirable.

7.3 Future Work

In Chapter 5 we have modelled the fine-grained polysemy exhibited by spatial prepositions and have based our study on those prepositions which, based on existing literature, appear to exhibit polysemy at room/table-top scales. It may be the case that this approach can be extended to other concepts which are also organised around ideal meanings. However, in order to extend and test the models on other terms it would be ideal to have some well-defined criteria and a procedure for assessing when a term is polysemous.

It is clear that object-specific features must be accounted for in semantic models of spatial prepositions and, following the suggestions in Section 6.6, further work is necessary in order to achieve this. Firstly, further investigations must be carried out in order to identify a set of salient object-specific features for each preposition. Various restricted studies have been conducted providing evidence that certain features influence certain prepositions, e.g. [31, 167], however a comprehensive study exploring this would be ideal. Such a study would face various challenges, e.g. the salience of particular object-specific features may change with changing contexts and the source

of potentially salient features may be very large, it may nevertheless be possible to isolate sets of particularly salient features in restricted contexts. Ontologies providing important object-specific features such as AfNet [168] may be helpful in this regard by highlighting object properties which are salient in many contexts.

Secondly, any implementation must be able to extract object-specific features from the scene. Assuming that the implementation is able to correctly label objects (e.g. 'bowl', 'table' etc...) in the scene, one approach to this would be to leverage information from knowledge bases such as ConceptNet [144]. For example, from ConceptNet one can determine that lids are used for covering and that bowls are containers. Another approach is to leverage affordance detection systems, e.g. [169], which use information from the scene to predict object affordances. Moreover, recent semantic representations such as VoxML [170] which allow for the specification of object affordances may lead to better systems for extracting this type of information from scenes in future.

Finally, in order to sufficiently train and test semantic models which include object-specific features, an extensive dataset is required which provides instances of prepositions representing a large variety of object-specific features.

In this thesis the generated semantic models have been tested on typicality judgements in the Comparative Task where the ground object is fixed and relational features are used to determine how well a figure object fits the given preposition-ground pair. However, in many pragmatic strategies for REG, e.g. [62], it is considered important to be able to assess how appropriate or acceptable a preposition is for a given figure-ground pair, i.e. we must also model categorisation. Though we were unable to provide evidence in this thesis that measures of categorisation and typicality differ it may be the case that the semantic models in this thesis under-perform in categorisation and this could be tested in future work.

Appendix A

Feature Extraction

This section provides precise details of how features are extracted for pairs of objects in Unity3D scenes. The code for the feature extraction process is provided in the feature extraction folder of the data collection software provided in the data archive[21] and updated versions are provided in the github repository.[49]

shortest_distance and *contact* are calculated by considering the distance of vertices of the figure, given by its mesh, to an approximation of the ground.[50] The shortest distance between F and G is taken to be the shortest distance from any vertex on the mesh of F to the approximation of G; as seen in Figure A.1.

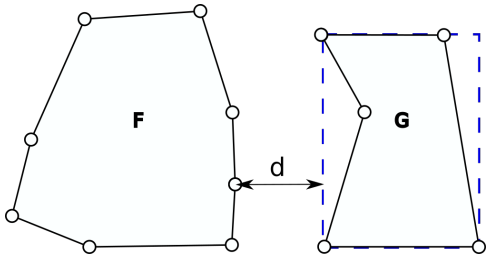


Figure A.1: Shortest distance.

The degree of contact between F and G is the number of vertices of F which are under a threshold distance (the default offset in Unity3D) to the approximation of G, divided by the total number of vertices of F. In Figure A.2 red vertices represent vertices of F which are in contact with G, *contact* is therefore $3 \div 8 = 0.375$.

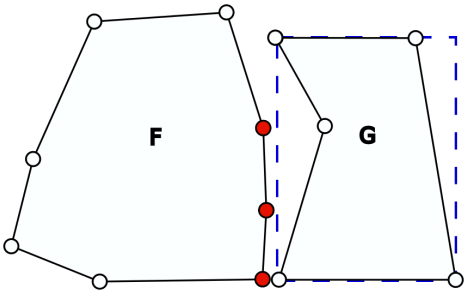


Figure A.2: Contact.

above_proportion and *below_proportion* are calculated by counting the number of

[49] https://github.com/alrichardbollans/spatial-preposition-annotation-tool-unity3d/tree/master/Unity3D%20Annotation%20Environment/Assets

[50] A collider is used to represent the ground which for simple objects is given by the object mesh, but for complex objects is approximated by a box, sphere or collection of boxes.

vertices of F which are above/below the highest/lowest point of G. In Figure A.3, the vertices above G are given in red and the vertices below G are given in blue. *above_proportion* is therefore $2 \div 8 = 0.25$ and *below_proportion* is $3 \div 8 = 0.375$.

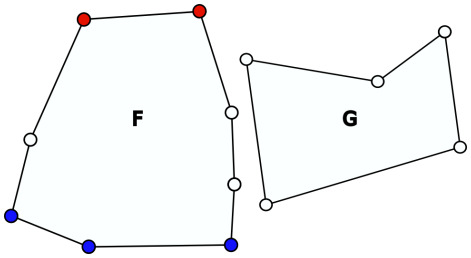


Figure A.3: Above/Below.

containment is calculated as the proportion of the axis-aligned bounding box of the figure which overlaps with the axis-aligned bounding box of the ground. In Figure A.4, *containment* is equal the volume of the purple shaded area divided by the total area of the bounding box of F.

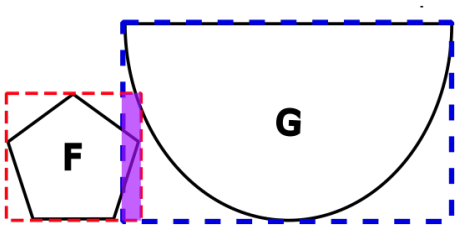


Figure A.4: Containment.

horizontal_distance calculates the horizontal distance between the centres of mass of the figure and ground, as given (in 2D) in Figure A.5.

f_covers_g aims to represent the degree to which the figure covers the ground. This is calculated by considering the degree to which the horizontal areas of the objects overlap. The greater the height separation of the figure from the ground, the larger the figure must be in order to provide effective covering. Therefore the effective area of the ground to be covered is extended (given by the blue dashed line in Figure A.6) taking into account the height separation, h: the area of G is extended on each side by $h \times \tan(5°)$. The effective overlap (given by the purple line in Figure A.6) is divided by the extended area of G in order to give the value of *f_covers_g*.

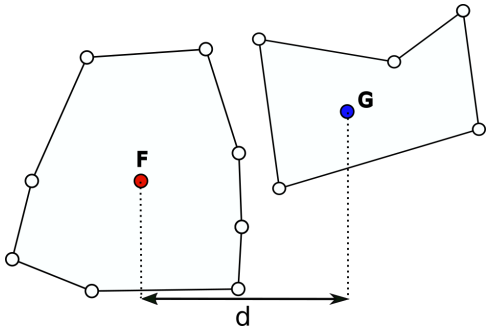


Figure A.5: Horizontal distance.

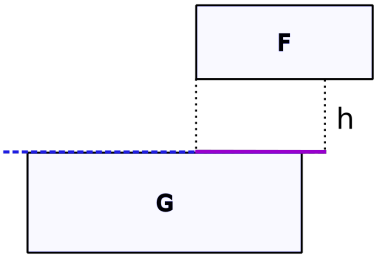
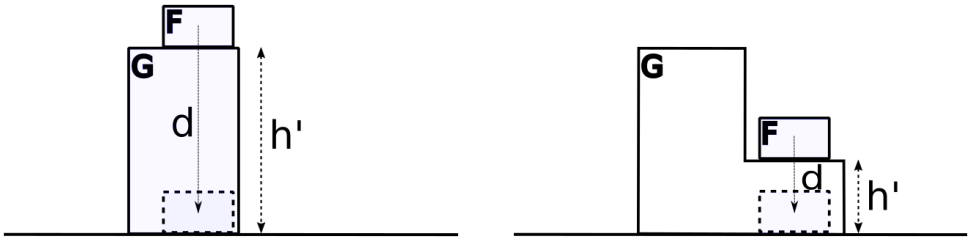


Figure A.6: F covers G.

To calculate *support*, the distance fallen by the centre of mass of the figure is assessed when the ground is removed from the scene, given by d in the following diagrams. The distance in then divided by a normalising height h'.

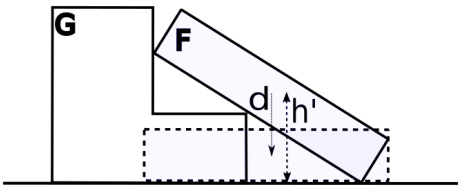
In the canonical case shown in Figure A.7(a), h' is simply equal to the height of the ground and *support* $sim1$. In the case shown in Figure A.7(b) which commonly occurs where an object is attached to the side of another, h' is equal to the height difference from the bottom of the figure to the bottom of the ground and *support* $sim1$. In the case shown in Figure A.7(c) which commonly occurs where an object is leaning on another, h' is equal to the height difference from the centre of mass of the figure to the bottom of the ground and *support* is often less than 1. In all other cases not accounted for here, h' is set as the height of the ground.

To calculate *location_control* a force is applied to the ground in a particular direction and the distance moved by the centre of mass of the figure is divided by the distance moved by the centre of mass of the ground (in the given direction). An example is given in Figure A.8, here location control is being assessed in the positive x direction and the



(a) Canonical Support: defined by the bottom of the figure being above the top of the ground.

(b) Support by attachment: defined by not being a canonical case and the bottom of the figure being above the bottom of the ground.



(c) Leaning support: defined by not being a canonical or attachment case and the centre of mass of the figure being above the bottom of the ground.

Figure A.7: Calculations of h' for support calculation.

given value is $L_{x+} = d \div n$. This is repeated for each of the other cardinal directions and the final value of *location_control* is the average, given by $\frac{L_{x+}+L_{x-}+L_{y+}+L_{y-}}{4}$.

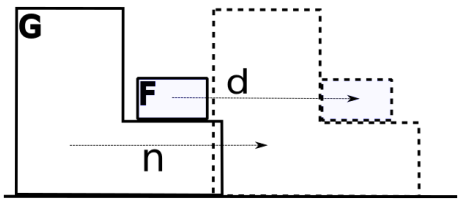


Figure A.8: Location Control.

Appendix B

Handling Polysemy

B.1 Ideal Meanings

Table B.1 provides the definitions of the ideal meanings used in Chapter 5. Each ideal meaning is defined by a set of salient features which are each assigned a threshold value, τ say, and ordering, $R(x, y)$ say. For a given configuration, the condition for the salient feature is satisfied if the feature value of the configuration, f, satisfies $R(f, \tau)$ and the configuration is an instance of the ideal meaning if the configuration satisfies each condition of the features.

Preposition	Feature	Threshold	ordering relation
on	*above_proportion*	0.9	$\geq$
	support	0.9	$\geq$
	contact	0.3	$\geq$
in	*containment*	0.7	$\geq$
	location_control	0.75	$\geq$
under	*g_covers_f*	0.9	$\geq$
	below_proportion	0.9	$\geq$
over	*f_covers_g*	0.9	$\geq$
	above_proportion	0.7	$\geq$
on top of	*above_proportion*	0.9	$\geq$
	contact	0.3	$\geq$
inside	*containment*	0.7	$\geq$
below	*horizontal_distance*	0.1	$\leq$
	below_proportion	0.9	$\geq$
above	*horizontal_distance*	0.1	$\leq$
	above_proportion	0.7	$\geq$
against	*horizontal_distance*	0.1	$\leq$
	location_control	0.25	$\geq$
	contact	0.3	$\geq$

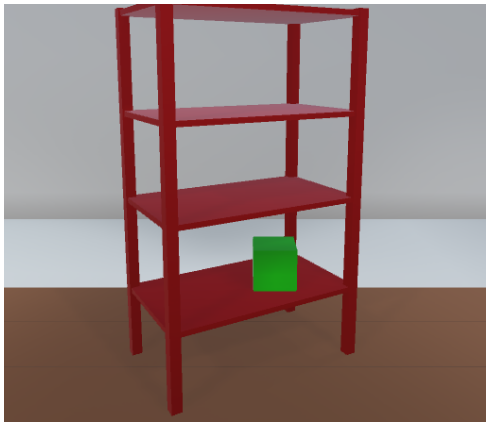
Table B.1: Initial definitions of Ideal Meanings

Appendix C

Categorisation & Typicality

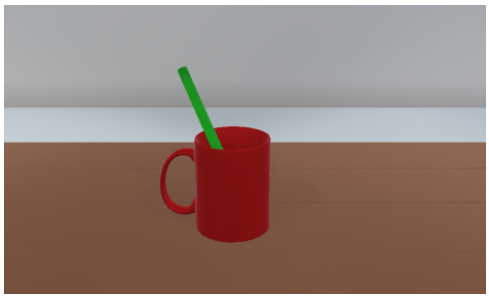
C.1 Scenes

The following Figures C.1 - C.5 are the scenes used in the data collection described in Section 3.3.

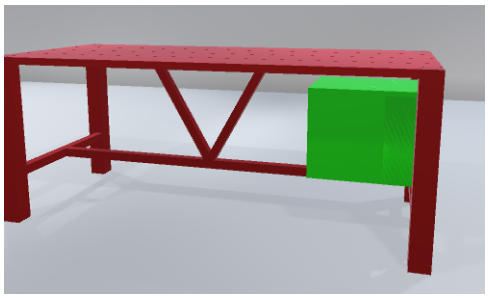


(a) The cube in the shelf.

(b) The pear in the bowl.

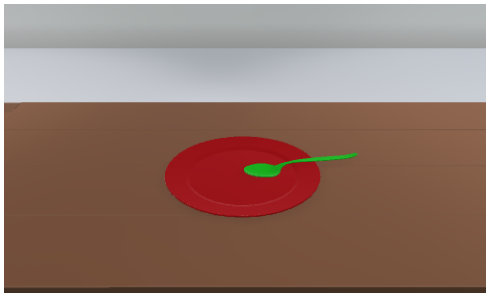


(c) The pencil in the mug.

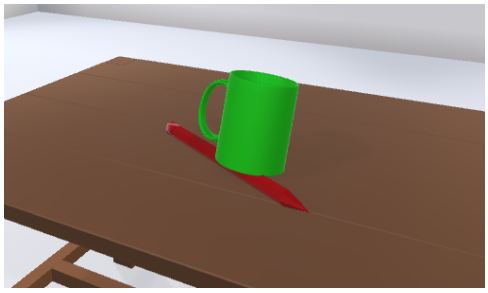


(d) The box in the table.

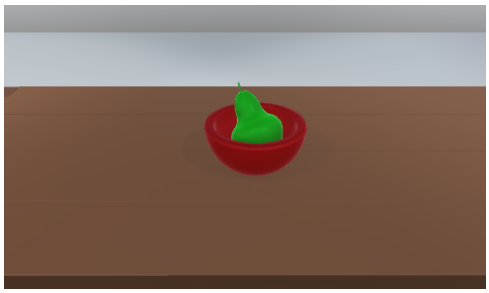
Figure C.1: Scenes used for 'in'.



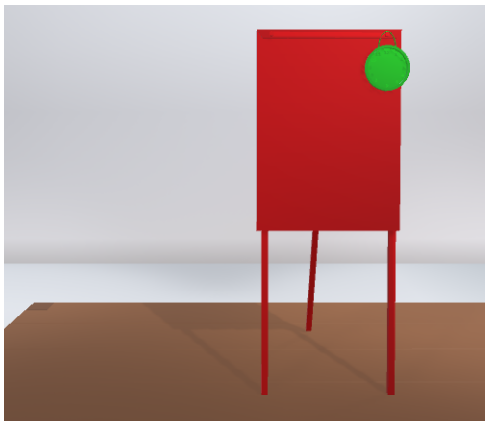
(a) The spoon on the plate.



(b) The mug on the pencil.

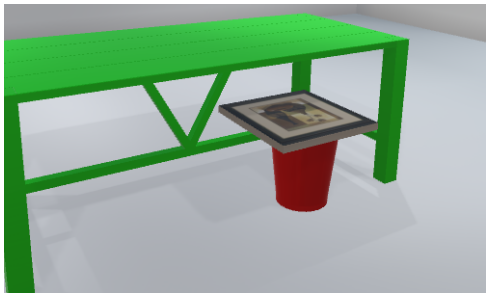


(c) The pear on the bowl.

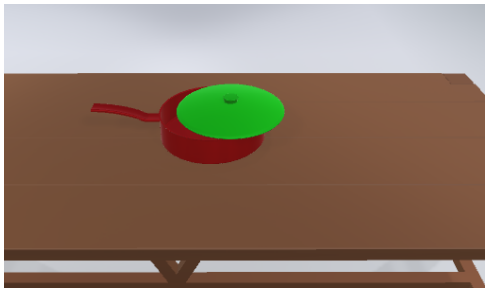


(d) The clock on the board.

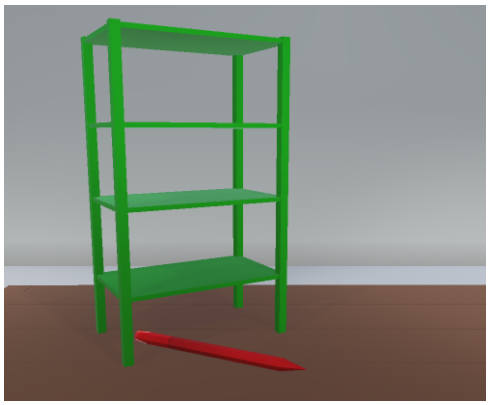
Figure C.2: Scenes used for ‘on’.



(a) The table over the bin.



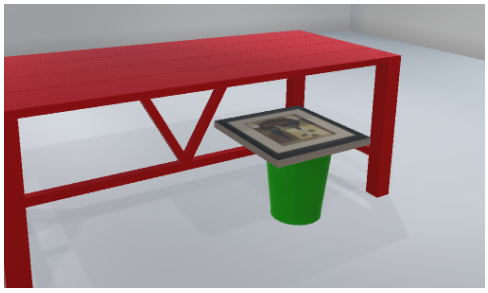
(b) The lid over the pan.



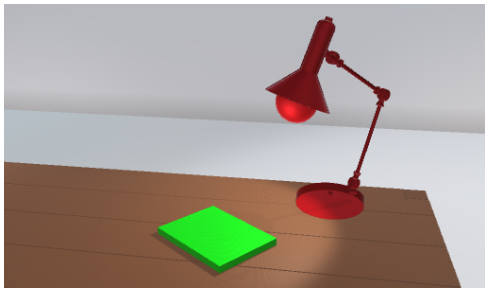
(c) The shelf over the pencil.

(d) The tap over the sink.

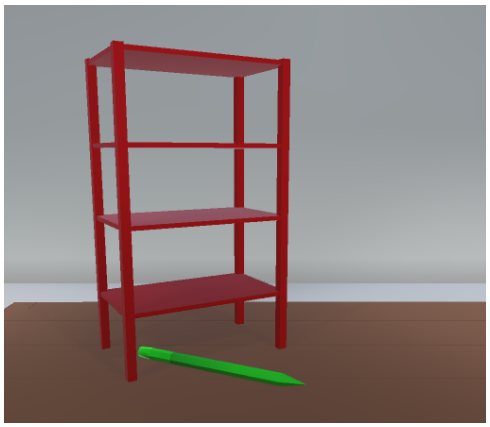
Figure C.3: Scenes used for ‘over’.



(a) The bin under the table.



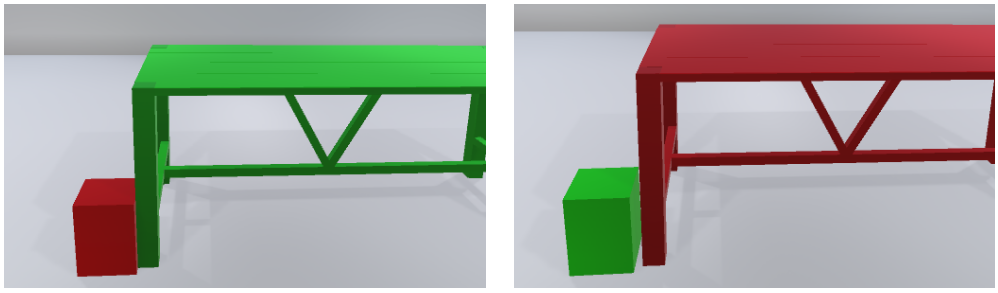
(b) The notepad under the lamp.



(c) The pencil under the shelf.

(d) The sink under the tap.

Figure C.4: Scenes used for ‘under’.



(a) The table against the box.

(b) The box against the table.

Figure C.5: Scenes used for ‘against’.

C.2 Results

Tables C.1 & C.2 provide results from the analysis discussed in Section 6.5 for each tested pair of configurations. For each pair, the better category member is the configuration which was more likely to be labelled with the given preposition in the Categorisation Task and the more typical configuration is the configuration which was more often selected when the pair was tested in the Typicality Task. * indicates a significant case with 10% significance level.

Preposition	Configuration 1	Configuration 2	Better Category Member	More Typical Configuration
in	[box,table]	[pencil,mug]	[pencil,mug]*	[pencil,mug]*
	[box,table]	[cube,shelf]	[box,table]	[cube,shelf]
	[box,table]	[pear,bowl]	[pear,bowl]	[pear,bowl]
	[pencil,mug]	[cube,shelf]	[pencil,mug]*	[pencil,mug]*
	[pencil,mug]	[pear,bowl]	[pencil,mug]*	[pencil,mug]*
	[cube,shelf]	[pear,bowl]	[pear,bowl]	[pear,bowl]
inside	[box,table]	[pencil,mug]	[pencil,mug]	[pencil,mug]*
	[box,table]	[cube,shelf]	[box,table]	[cube,shelf]
	[box,table]	[pear,bowl]	[box,table]	[box,table]*
	[pencil,mug]	[cube,shelf]	[pencil,mug]	[pencil,mug]*
	[pencil,mug]	[pear,bowl]	[pencil,mug]*	[pencil,mug]*
	[cube,shelf]	[pear,bowl]	None	[pear,bowl]
on	[spoon,plate]	[clock,board]	[spoon,plate]	[spoon,plate]
	[spoon,plate]	[mug,pencil]	[spoon,plate]	[mug,pencil]
	[spoon,plate]	[pear,bowl]	[spoon,plate]*	[spoon,plate]*
	[clock,board]	[mug,pencil]	[clock,board]	[mug,pencil]
	[clock,board]	[pear,bowl]	[clock,board]*	[clock,board]*
	[mug,pencil]	[pear,bowl]	[mug,pencil]*	[mug,pencil]*
on top of	[spoon,plate]	[clock,board]	[spoon,plate]*	[spoon,plate]*
	[spoon,plate]	[mug,pencil]	[mug,pencil]	[mug,pencil]
	[spoon,plate]	[pear,bowl]	[spoon,plate]*	[spoon,plate]*
	[clock,board]	[mug,pencil]	[mug,pencil]*	[mug,pencil]*
	[clock,board]	[pear,bowl]	[pear,bowl]	[pear,bowl]
	[mug,pencil]	[pear,bowl]	[mug,pencil]*	[mug,pencil]*

Table C.1: Pairwise results for 'in', 'inside', 'on' & 'on top of'
* - Indicates a significant case.

Preposition	Configuration 1	Configuration 2	Better Category Member	More Typical Configuration
over	[lid,pan]	[table,bin]	[table,bin]*	[table,bin]
	[lid,pan]	[tap,sink]	[tap,sink]*	[tap,sink]*
	[lid,pan]	[shelf,pencil]	[shelf,pencil]	[shelf,pencil]
	[table,bin]	[tap,sink]	[table,bin]	None
	[table,bin]	[shelf,pencil]	[table,bin]	[table,bin]*
	[tap,sink]	[shelf,pencil]	[tap,sink]	[tap,sink]
above	[lid,pan]	[table,bin]	[table,bin]*	[table,bin]
	[lid,pan]	[tap,sink]	[tap,sink]	[tap,sink]
	[lid,pan]	[shelf,pencil]	[shelf,pencil]	[shelf,pencil]
	[table,bin]	[tap,sink]	[table,bin]	[table,bin]
	[table,bin]	[shelf,pencil]	[table,bin]	[table,bin]
	[tap,sink]	[shelf,pencil]	[tap,sink]	[tap,sink]
under	[sink,tap]	[pencil,shelf]	[pencil,shelf]	[sink,tap]
	[sink,tap]	[bin,table]	[bin,table]	[bin,table]*
	[sink,tap]	[notepad,lamp]	[notepad,lamp]	[sink,tap]
	[pencil,shelf]	[bin,table]	None	[bin,table]
	[pencil,shelf]	[notepad,lamp]	None	[pencil,shelf]
	[bin,table]	[notepad,lamp]	None	[bin,table]*
below	[sink,tap]	[pencil,shelf]	[pencil,shelf]	[pencil,shelf]*
	[sink,tap]	[bin,table]	[bin,table]	None
	[sink,tap]	[notepad,lamp]	[notepad,lamp]	None
	[pencil,shelf]	[bin,table]	[pencil,shelf]	[bin,table]
	[pencil,shelf]	[notepad,lamp]	[pencil,shelf]	[pencil,shelf]
	[bin,table]	[notepad,lamp]	[notepad,lamp]	[bin,table]
against	[table,box]	[box,table]	[table,box]	[box,table]*

Table C.2: Pairwise results for 'over', 'above', 'under', 'below' & 'against'
* - Indicates a significant case.

Bibliography

[1] Adam Richard-Bollans, Brandon Bennett, and Anthony G. Cohn. Automatic generation of typicality measures for spatial language in grounded settings. In *Proceedings of 24th European Conference on Artificial Intelligence*, 2020. doi: https://doi.org/10.3233/FAIA200341.

[2] Adam Richard-Bollans, Lucía Gómez Álvarez, and Anthony G. Cohn. Modelling the polysemy of spatial prepositions in referring expressions. In *Proceedings of 17th International Conference on Principles of Knowledge Representation and Reasoning*, 2020. doi: https://doi.org/10.24963/kr.2020/72.

[3] Adam Richard-Bollans, Lucía Gómez Álvarez, and Anthony Cohn. Categorisation, typicality & object-specific features in spatial referring expressions. In *Proceedings of the Third International Workshop on Spatial Language Understanding*, pages 39–49. Association for Computational Linguistics, 2020. doi: http://doi.org/10.18653/v1/2020.splu-1.5.

[4] Adam Richard-Bollans. Towards a cognitive model of the semantics of spatial prepositions. In *ESSLLI Student Session Proceedings*. Springer, 2018.

[5] Adam Richard-Bollans, Lucía Gómez Álvarez, Brandon Bennett, and Anthony G. Cohn. Investigating the dimensions of spatial language. In *Proceedings of Speaking of Location 2019: Communicating about Space*. CEUR Workshop Proceedings, 2019.

[6] Melissa Bowerman and Soonja Choi. Shaping meanings for language: universal and language-specific in the acquisition of semantic categories. In *Language acquisition and conceptual development*, pages 475–511. Cambridge University Press, 2001.

[7] Annette Herskovits. Semantics and pragmatics of locative expressions. *Cognitive Science*, 9(3):341–378, 1985.

[8] Simon Garrod, Gillian Ferrier, and Siobhan Campbell. In and on: investigating the functional geometry of spatial prepositions. *Cognition*, 72(2):167–189, 1999. doi: https://doi.org/10.1016/s0010-0277(99)00038-4.

[9] Jean M Mandler. How to build a baby: II. Conceptual primitives. *Psychological review*, 99(4):587, 1992.

[10] Kenny R. Coventry, Mercè Prat-Sala, and Lynn Richards. The interplay between geometry and function in the comprehension of over, under, above, and below. *Journal of Memory and Language*, 44(3):376–398, 2001. doi: https://doi.org/10.1006/jmla.2000.2742.

[11] Lucía Gómez Álvarez. *Standpoint logic: a logic for handling semantic variability, with applications to forestry information.* PhD Thesis, University of Leeds, 2019.

[12] Lucía Gómez Álvarez. Ambiguity: What is it that needs representing and what needs resolving? In *Ambiguity: Perspectives on Representation and Resolution*, 2018.

[13] James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. SemEval-2015 task 8: SpacEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 884–894, Denver, Colorado, 2015. Association for Computational Linguistics. doi: 10.18653/v1/S15-2149. URL http://aclweb.org/anthology/S15-2149.

[14] Hannah Rohde, Scott Seyfarth, Brady Clark, Gerhard Jäger, and Stefan Kaufmann. Communicating with cost-based implicature: A game-theoretic approach to ambiguity. In *Proceedings of 16th Workshop on the Semantics and Pragmatics of Dialogue*, pages 107–116, 2012.

[15] Kais Dukes. SemEval-2014 task 6: supervised semantic parsing of robotic spatial commands. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 45–53, 2014. doi: https://doi.org/10.3115/v1/s14-2006.

[16] Emanuele Bastianelli, Giuseppe Castellucci, Danilo Croce, Luca Iocchi, Roberto Basili, and Daniele Nardi. HuRIC: a human robot interaction corpus. In *LREC*, pages 4519–4526, 2014.

[17] Jette Viethen and Robert Dale. GRE3D7: a corpus of distinguishing descriptions for objects in visual scenes. In *UCNLG+Eval: Language Generation and Evaluation Workshop*, pages 12–22. Association for Computational Linguistics, 2011.

[18] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[19] Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Turney, and Daniel Khashabi. Combining retrieval, statistics, and inference to answer elementary science questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016. Issue: 1.

[20] Hongyu Gong, Suma Bhat, and Pramod Viswanath. Embedding syntax and semantics of prepositions via tensor decomposition. In *Proceedings of NAACL-HLT*, pages 896–906, 2018.

[21] Robyn Carston. The semantics/pragmatics distinction: A view from relevance theory. In Ken Turner, editor, *The semantics/pragmatics interface from different points of view*, pages 85–125. Elsevier, Oxford, UK, 1999. URL http://www.phon.ucl.ac.uk/publications/WPL/98papers/carston.pdf.

[22] Daniel R Montello. Scale and multiple psychologies of space. In *European conference on spatial information theory*, pages 312–321. Springer, 1993.

[23] Anna-Katharina Lautenschütz, Clare Davies, Martin Raubal, Angela Schwering, and Eric Pederson. The influence of scale, context and spatial preposition in linguistic topology. In *International Conference on Spatial Cognition*, pages 439–452. Springer, 2006.

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

[25] Leonard Talmy. Force dynamics in language and cognition. *Cognitive science*, 12 (1):49–100, 1988. doi: https://doi.org/10.1207/s15516709cog1201_2.

[26] Edward E Smith, Edward J Shoben, and Lance J Rips. Structure and process in semantic memory: A featural model for semantic decisions. *Psychological review*, 81(3), 1974. doi: https://doi.org/10.1037/h0036351. Publisher: American Psychological Association.

[27] Lance J . Rips. Similarity, typicality, and categorization. In Stella Vosniadou and Andrew Ortony, editors, *Similarity and Analogical Reasoning*. Cambridge University Press, 1st edition, 1989. doi: 10.1017/CBO9780511529863.004.

[28] Eleanor Rosch. Principles of categorization. In Eleanor Rosch and Barbara B. Lloyd, editors, *Cognition and Categorization*, volume 1, pages 27–78. Lawrence Erlbaum Associates, Hillsdale, NJ, 1978.

[29] Annette Herskovits. *Language and spatial cognition*. Cambridge University Press, 1987.

[30] Andrea Tyler and Vyvyan Evans. Reconsidering prepositional polysemy networks: the case of over. *Language*, 77(4):724–765, 2001. doi: https://doi.org/10.1353/lan.2001.0250.

[31] Kenny R. Coventry, Richard Carmichael, and Simon C. Garrod. Spatial prepositions, object-specific function, and task requirements. *Journal of Semantics*, 11 (4):289–309, 1994. doi: https://doi.org/10.1093/jos/11.4.289.

[32] Michele I Feist and Derdre Gentner. On plates, bowls, and dishes: Factors in the use of English IN and ON. In *Proc 20th annual meeting of the cognitive science society*, pages 345–349, 1998.

[33] Stephen C Levinson. Frames of reference and Molyneux's question: Crosslinguistic evidence. *Language and space*, 109:169, 1996.

[34] George A Miller and Philip N Johnson-Laird. *Language and perception*. Belknap Press, 1976.

[35] Stacy A Doore. *Spatial relations and natural-language semantics for indoor scenes*. PhD thesis, University of Maine, 2017.

[36] Gloria S. Cooper. A semantic analysis of english locative prepositions. Technical report, Defense Technical Information Center, Fort Belvoir, VA, January 1968.

[37] Anthony G Cohn, David A Randell, and Zhan Cui. Taxonomies of logically defined qualitative spatial relations. *Journal of human-computer studies*, 43(5-6): 831–846, 1995.

[38] Torsten Hahmann and Boyan Brodaric. Kinds of full physical containment. In *Spatial Information Theory*, volume 8116, pages 397–417. Springer International Publishing, Cham, 2013. doi: 10.1007/978-3-319-01790-7_22.

[39] Renée Baillargeon. The acquisition of physical knowledge in infancy: A summary in eight lessons. In Goswami Usha, editor, *Blackwell handbook of childhood cognitive development*, pages 47–83. Blackwell, 2002.

[40] Annette Herskovits. Language, spatial cognition, and vision. In Oliviero Stock, editor, *Spatial and Temporal Reasoning*, pages 155–202. Kluwer Academic Publishers, Dordrecht, 1997.

[41] Leonard Talmy. How language structures space. In *Spatial orientation*, pages 225–282. Springer, 1983.

[42] Simon C. Garrod and Anthony J. Sanford. Discourse models as interfaces between language and the spatial world. *Journal of Semantics*, 6(1):147–160, 1988. doi: 10.1093/jos/6.1.147.

[43] John A Bateman, Joana Hois, Robert Ross, and Thora Tenbrink. A linguistic ontology of space for natural language processing. *Artificial Intelligence*, 174(14): 1027–1071, 2010.

[44] Kenny R Coventry and Simon C Garrod. *Saying, seeing and acting: The psychological semantics of spatial prepositions.* Psychology Press, 2004.

[45] Simon Dobnik and Mehdi Ghanimifard. Spatial descriptions on a functional-geometric spectrum: the location of objects. In *Proc German Conference on Spatial Cognition*, pages 219–234. Springer, 2020. doi: 10.1007/978-3-030-57983-8_17.

[46] Souma Mori. A cognitive analysis of the preposition *over*: image-schema transformations and metaphorical extensions. *Canadian Journal of Linguistics*, 64(3): 444–474, 2019. doi: 10.1017/cnj.2018.43.

[47] Peter Gärdenfors. The geometry of preposition meanings. *Baltic International Yearbook of Cognition, Logic and Communication*, 10(1), 2015.

[48] Agustín Vicente, Ingrid L. Falkum, Agustín Vicente, and Ingrid L. Falkum. Polysemy. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press, 2017. doi: 10.1093/acrefore/9780199384655.013.325.

[49] Advaith Siddharthan and Ann Copestake. Generating referring expressions in open domains. In *Association for Computational Linguistics*, 2004. doi: 10.3115/1218955.1219007.

[50] Thomas Wasow, Amy Perfors, and David Beaver. The puzzle of ambiguity. *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe*, pages 265–282, 2005. Publisher: CSLI Publications, Stanford.

[51] Devorah E Klein and Gregory L Murphy. Paper has been my ruin: conceptual relations of polysemous senses. *Journal of Memory and Language*, 47(4):548–570, 2002.

[52] Barbara Lewandowska-Tomaszczyk. Polysemy, prototypes, and radial categories. In *The Oxford Handbook of Cognitive Linguistics*, pages 139–169. Oxford University Press, 2007. doi: 10.1093/oxfordhb/9780199738632.013.0006.

[53] George A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[54] George Lakoff. *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago press, 1987.

[55] Roberto Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69, February 2009. doi: 10.1145/1459352.1459355.

[56] Fieke Van der Gucht, Klaas Willems, and Ludovic De Cuypere. The iconicity of embodied meaning. Polysemy of spatial prepositions in the cognitive framework. *Language Sciences*, 29(6):733–754, 2007. doi: 10.1016/j.langsci.2006.12.027.

[57] Edilson J Rodrigues, Paulo E Santos, Marcos Lopes, Brandon Bennett, and Paul E Oppenheimer. Standpoint semantics for polysemy in spatial prepositions. *Journal of Logic and Computation*, 2020. doi: 10.1093/logcom/exz034.

[58] A. Muller, C. Roch, T. Stadtfeld, and T. Kiss. Annotating spatial interpretations of german prepositions. In *2011 IEEE Fifth International Conference on Semantic Computing*, pages 459–466, Palo Alto, CA, USA, September 2011. IEEE. ISBN 978-1-4577-1648-5. doi: 10.1109/ICSC.2011.46. URL http://ieeexplore.ieee.org/document/6061501/.

[59] Vyvyan Evans. A unified account of polysemy within LCCM Theory. *Lingua*, 157:100–123, 2015. Publisher: Elsevier.

[60] Kees van Deemter. *Computational models of referring: a study in cognitive science*. MIT Press, 2016.

[61] Kees van Deemter. Generating referring expressions that involve gradable properties. *Computational Linguistics*, 32(2):195–222, 2006. doi: 10.1162/coli.2006.32.2.195.

[62] M. C. Frank and N. D. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012. doi: 10.1126/science.1218633.

[63] Noah D. Goodman and Michael C. Frank. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829, November 2016. doi: 10.1016/j.tics.2016.08.005.

[64] Vivien Mast, Zoe Falomir, and Diedrich Wolter. Probabilistic reference and grounding with PRAGR for dialogues with robots. *Journal of Experimental & Theoretical Artificial Intelligence*, 28(5):889–911, 2016. doi: 10.1080/0952813X.2016.1154611.

[65] Helmut Horacek. Generating referential descriptions under conditions of uncertainty. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*, 2005.

[66] Judith Degen, Robert D Hawkins, Caroline Graf, Elisa Kreiss, and Noah D Goodman. When redundancy is useful: A Bayesian approach to" overinformative" referring expressions. *Psychological review*, 2020.

[67] Michael Spranger and Simon Pauw. Dealing with perceptual deviation: vague semantics for spatial language and quantification. In *Language Grounding in Robots*, pages 173–192. Springer US, Boston, MA, 2012. doi: 10.1007/978-1-4614-3064-3_9.

[68] Gordon D. Logan and Daniel D. Sadler. A computational analysis of the apprehension of spatial relations. In Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, editors, *Language and space*, Language, speech, and communication., pages 493–529. MIT Press, 1996.

[69] Igor Douven. Putting prototypes in place. *Cognition*, 193, 2019. doi: 10.1016/j.cognition.2019.104007.

[70] Eleanor Rosch and Carolyn B. Mervis. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4):573 – 605, 1975. ISSN 0010-0285. doi: http://dx.doi.org/10.1016/0010-0285(75)90024-9. URL http://www.sciencedirect.com/science/article/pii/0010028575900249.

[71] Robert M Nosofsky. Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: learning, memory, and cognition*, 14(4), 1988.

[72] W. Voorspoels, W. Vanpaemel, and G. Storms. Exemplars and prototypes in natural language concepts: A typicality-based evaluation. *Psychonomic Bulletin & Review*, 15(3):630–637, 2008. doi: 10.3758/PBR.15.3.630.

[73] Peter Gärdenfors. Conceptual spaces as a framework for knowledge representation. *Mind and Matter*, 2(2):9–27, 2004.

[74] Antonio Lieto, Antonio Chella, and Marcello Frixione. Conceptual Spaces for cognitive architectures: a lingua franca for different levels of representation. *Biologically Inspired Cognitive Architectures*, 19:1–9, 2017. doi: 10.1016/j.bica.2016.10.005.

[75] Martin Raubal. Formalizing conceptual spaces. In *Formal Ontology in Information Systems*, volume 114, pages 153–164, 2004.

[76] Robert L Goldstone, Alan Kersten, and Paulo F Carvalho. Concepts and categorization. In *Handbook of Psychology*, pages 597–621. Wiley Online Library, 2003.

[77] Alicia Abella and John R Kender. Qualitatively describing objects using spatial prepositions. In *IEEE Workshop on Qualitative Vision*, pages 33–38. IEEE, 1993.

[78] Georgiy Platonov and Lenhart Schubert. Computational models for spatial prepositions. In *Proc 1st International Workshop on Spatial Language Understanding*, pages 21–30, 2018. doi: https://doi.org/10.18653/v1/w18-1403.

[79] Fethiye Irmak Doğan, Sinan Kalkan, and Iolanda Leite. Learning to generate unambiguous spatial referring expressions for real-world environments. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4992–4999. IEEE, 2019.

[80] Mandar Haldekar, Ashwinkumar Ganesan, and Tim Oates. Identifying spatial relations in images using convolutional neural networks. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3593–3600. IEEE, 2017.

[81] Michael Barclay and Antony Galton. A scene corpus for training and testing spatial communication systems. In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, 2008.

[82] Ankit Goyal, Kaiyu Yang, Dawei Yang, and Jia Deng. Rel3D: a minimally contrastive benchmark for grounding spatial relations in 3D. *Advances in Neural Information Processing Systems*, 33, 2020.

[83] Adam Richard-Bollans, Lucía Gómez Álvarez, and Anthony G. Cohn. The role of pragmatics in solving the Winograd Schema Challenge. In *Proceedings of 13th International Symposium on Commonsense Reasoning*. CEUR Workshop Proceedings, 2017. URL http://eprints.whiterose.ac.uk/122937/.

[84] Hal Daumé III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126, 2006.

[85] Sameer S. Pradhan, Wayne Ward, and James H. Martin. Towards robust semantic role labeling. *Computational linguistics*, 34(2):289–310, 2008.

[86] Kaiyu Yang, Olga Russakovsky, and Jia Deng. SpatialSense: an adversarially crowdsourced benchmark for spatial relation recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2051–2060, Seoul, Korea (South), 2019. IEEE. doi: 10.1109/ICCV.2019.00214.

[87] Muhannad Alomari, Paul Duckworth, Majd Hawasly, David C Hogg, and Anthony G Cohn. Natural language grounding and grammar induction for robotic manipulation commands. In *First Workshop on Language Grounding for Robotics*, pages 35–43, 2017.

[88] Angel Chang, Manolis Savva, and Christopher D. Manning. Learning spatial knowledge for text to 3d scene generation. In *Proc EMNLP*, pages 2028–2038. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1217.

[89] Dave Golland, Percy Liang, and Dan Klein. A game-theoretic approach to generating spatial descriptions. In *Proc EMNLP*, pages 410–419, 2010.

[90] Peter Gorniak and Deb Roy. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470, 2004.

[91] John D. Kelleher and Fintan J. Costello. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2): 271–306, 2009.

[92] Jugal K Kalita and Norman I Badler. Interpreting prepositions physically. In *Proc AAAI*, pages 105–110, 1991.

[93] Driss Kettani and Bernard Moulin. A spatial model based on the notions of spatial conceptual map and of object's influence areas. In *Proc COSIT*, pages 401–416. Springer, 1999.

[94] Terry Regier and Laura A Carlson. Grounding spatial language in perception: an empirical and computational investigation. *Journal of experimental psychology: General*, 130(2), 2001.

[95] David A. Randell, Zhan Cui, and Anthony G. Cohn. A spatial logic based on regions and connection. *KR*, 92:165–176, 1992.

[96] Yavor Neychev Nenov. *Computability of euclidean spatial logics*. PhD thesis, Citeseer, 2011.

[97] Bruce Porter, Vladimir Lifschitz, and Frank Van Harmelen, editors. *Handbook of knowledge representation*. Foundations of artificial intelligence. Elsevier, Amsterdam ; Boston, 1st edition, 2008.

[98] Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.

[99] Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, 2015.

[100] John A Bateman, Bernardo Magnini, and Giovanni Fabris. The generalized upper model knowledge base: Organization and use. *Towards very large knowledge bases*, pages 60–72, 1995.

[101] Gérard Ligozat, Jakub Nowak, and Didier Schmitt. From language to pictorial representations. In *Language and Technology Conference*, 2007.

[102] James Pustejovsky, Jessica L. Moszkowicz, and Marc Verhagen. Using ISO-Space for annotating spatial information. In *Proceedings of the International Conference on Spatial Information Theory*, 2011.

[103] Parisa Kordjamshidi, Joana Hois, Martijn van Otterlo, and Marie-Francine Moens. Learning to interpret spatial natural language in terms of qualitative spatial relations. In *Representing Space in Cognition*, Explorations in Language and Space. Oxford University Press, 2013.

[104] Brandon Bennett and Claudia Cialone. Corpus Guided Sense Cluster Analysis: a methodology for ontology development (with examples from the spatial domain). In *Formal Ontology in Information Systems*, volume 267, pages 213–226. IOS Press, 2014.

[105] Max J Egenhofer and M Andrea Rodríguez. Relation algebras over containers and surfaces: An ontological study of a room space. *Spatial Cognition and Computation*, 1(2):155–180, 1999. Publisher: Springer.

[106] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. The IAPR TC-12 benchmark: a new evaluation resource for visual information systems. In *LREC*, 2006.

[107] Parisa Kordjamshidi, Taher Rahgooy, Marie-Francine Moens, James Pustejovsky, Umar Manzoor, and Kirk Roberts. CLEF 2017: Multimodal spatial role labeling (mSpRL) task overview. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 367–376. Springer, 2017.

[108] Parisa Kordjamshidi, Steven Bethard, and Marie-Francine Moens. SemEval-2012 task 3: spatial role labeling. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 365–373. Association for Computational Linguistics, 2012.

[109] Oleksandr Kolomiyets, Parisa Kordjamshidi, Marie Francine Moens, and Steven Bethard. Semeval-2013 task 3: Spatial role labeling. In *Proceedings of the seventh international workshop on semantic evaluation*, pages 255–262, 2013.

[110] H. Paul Grice. Logic and conversation. In *Syntax and Semantics, Vol. 3, Speech Acts*, pages 41–58. Academic Press, New York, 1975.

[111] James A. Hampton. Conceptual combination: Conjunction and negation of natural concepts. *Memory & Cognition*, 25(6):888–909, November 1997. doi: 10.3758/BF03211333.

[112] Michael E. McCloskey and Sam Glucksberg. Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, 6(4):462–472, 1978. doi: 10.3758/BF03197480.

[113] Ludwig Wittgenstein. *Philosophical investigations*. Macmillan, 1953.

[114] Daniel Osherson and Edward E Smith. On typicality and vagueness. *Cognition*, 64(2):189–206, 1997. ISSN 00100277. doi: 10.1016/S0010-0277(97)00025-5.

[115] Brent Berlin and Paul Kay. *Basic color terms: Their universality and evolution*. Univ of California Press, 1991.

[116] Lotfi A Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965.

[117] Klaus-Peter Gapp. Basic meanings of spatial relations: Computation and evaluation in 3d space. In *AAAI*. Universität des Saarlandes, 1994.

[118] Reinhard Moratz and Thora Tenbrink. Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. *Spatial cognition and computation*, 6(1):63–107, 2006.

[119] Klaus-Peter Gapp. An empirically validated model for computing spatial relations. In *KI-95: Advances in Artificial Intelligence*, volume 981, pages 245–256. Springer Berlin Heidelberg, Berlin, Heidelberg, 1995. ISBN 978-3-540-60343-6 978-3-540-44944-7. doi: 10.1007/3-540-60343-3_41.

[120] Konstantinos Zampogiannis, Yezhou Yang, Cornelia Fermüller, and Yiannis Aloimonos. Learning the spatial semantics of manipulation actions through preposition grounding. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 1389–1396. IEEE, 2015.

[121] Joana Hois and Oliver Kutz. Natural language meets spatial calculi. In Christian Freksa, Nora S. Newcombe, Peter Gärdenfors, and Stefan Wölfl, editors, *Spatial Cognition VI. Learning, Reasoning, and Talking about Space*, volume 5248, pages 266–282. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[122] Zoe Falomir and Thomas Kluth. Qualitative spatial logic descriptors from 3D indoor scenes to generate explanations in natural language. *Cognitive Processing*, pages 1–20, 2017.

[123] Fujian Yan, Dali Wang, and Hongsheng He. Robotic understanding of spatial relationships using neural-logic learning. In *International Conference on Intelligent Robots and Systems*, 2020.

[124] Severin Fichtl, Andrew McManus, Wail Mustafa, Dirk Kraft, Norbert Krüger, and Frank Guerin. Learning spatial relationships from 3D vision using histograms. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 501–508. IEEE, 2014.

[125] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. Publisher: Springer.

[126] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proc NAACL*. ACL, 2010.

[127] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and others. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Journal of Computer Vision*, 123(1):32–73, 2017.

[128] Myung Jin Choi, Joseph J. Lim, Antonio Torralba, and Alan S. Willsky. Exploiting hierarchical context on a large database of object categories. In *Proc CVPR*, pages 129–136. IEEE, 2010.

[129] Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, and Song-Chun Zhu. Holistic 3d scene parsing and reconstruction from a single rgb image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 187–203, 2018.

[130] Yixin Chen, Siyuan Huang, Tao Yuan, Yixin Zhu, Siyuan Qi, and Song-Chun Zhu. Holistic++ scene understanding: single-view 3D holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8647–8656, Seoul, Korea (South), October 2019. IEEE. doi: 10.1109/ICCV.2019.00874.

[131] Muhannad Alomari, Paul Duckworth, David C. Hogg, and Anthony G. Cohn. Natural language acquisition and grounding for embodied robotic systems. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 4349–4356, 2017.

[132] Sergio Guadarrama, Lorenzo Riano, Dave Golland, Daniel Go, Yangqing Jia, Dan Klein, Pieter Abbeel, and Trevor Darrell. Grounding spatial relations for human-robot interaction. In *Proc IROS*, pages 1640–1647. IEEE, 2013.

[133] Kais Dukes. Train robots: A dataset for natural language human-robot spatial interaction through verbal commands. In *Proc International Conference on Social Robotics*, 2013.

[134] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. IQA: visual question answering in interactive environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4089–4098, 2018.

[135] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: an interactive 3D environment for visual AI. *arXiv:1712.05474 [cs]*, 2019. URL http://arxiv.org/abs/1712.05474.

[136] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, Honolulu, HI, July 2017. IEEE. doi: 10.1109/CVPR.2017.215.

[137] Nikhil Krishnaswamy and James Pustejovsky. Generating a novel dataset of multimodal referring expressions. In *Proceedings of the 13th International Conference on Computational Semantics - Short Papers*, pages 44–51. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-0507.

[138] Georgiy Platonov, Benjamin Kane, Aaron Gindi, and Lenhart K. Schubert. A spoken dialogue system for spatial question answering in a physical blocks world. *arXiv:1911.02524 [cs]*, 2019. URL http://arxiv.org/abs/1911.02524.

[139] Michael Spranger, Martin Loetzsch, and Luc Steels. A perceptual system for language game experiments. In *Language Grounding in Robots*, pages 89–110. Springer US, Boston, MA, 2012. doi: 10.1007/978-1-4614-3064-3_5.

[140] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. Clevr-ref+: diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4185–4194, 2019.

[141] Melissa Bowerman and Erik Pederson. Topological relations picture series. In *Space stimuli kit 1.2*, page 51. Max Planck Institute for Psycholinguistics, 1992.

[142] David M. Mark and Max J. Egenhofer. Topology of prototypical spatial relations

between lines and regions in English and Spanish. In *Proc Auto Carto*, pages 245–254, 1995.

[143] Thora Tenbrink, Elena Andonova, Gesa Schole, and Kenny R. Coventry. Communicative success in spatial dialogue: the impact of functional features and dialogue strategies. *Language and Speech*, 60(2):318–329, 2017.

[144] Robert Speer and Catherine Havasi. Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686, 2012.

[145] Igor Douven, Lieven Decock, Richard Dietz, and Paul Égré. Vagueness: a conceptual spaces approach. *Journal of Philosophical Logic*, 42(1):137–160, 2013. doi: 10.1007/s10992-011-9216-0.

[146] Maria M. Hedblom, Oliver Kutz, Till Mossakowski, and Fabian Neuhaus. Between contact and support: introducing a logic for image schemas and directed movement. In *Proc IAAI*, volume 10640, pages 256–268. Springer, 2017.

[147] Henrietta Eyre and Jonathan Lawry. Language games with vague categories and negations. *Adaptive Behavior*, 22(5):289–303, 2014. doi: 10.1177/1059712314547318.

[148] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12:2825–2830, 2011.

[149] Martha Lewis and Jonathan Lawry. Hierarchical conceptual spaces for concept combination. *Artificial Intelligence*, 237:204–227, 2016. ISSN 00043702. doi: 10.1016/j.artint.2016.04.008.

[150] Antonio Lieto, Daniele Radicioni, Valentina Rho, and Enrico Mensa. Towards a unifying framework for conceptual represention and reasoning in cognitive systems. *Intelligenza Artificiale*, 11(2):139–153, 2017.

[151] Lucas Bechberger and Kai-Uwe Kühnberger. Generalizing Psychological Similarity Spaces to Unseen Stimuli. *arXiv:1908.09260 [cs, stat]*, 2020. URL http://arxiv.org/abs/1908.09260.

[152] Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998. Publisher: MIT Press.

[153] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7:1–30, 2006.

[154] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[155] Jordan Zlatev. A study of perceptually grounded polysemy in a spatial microdomain. Technical Report TR-92-048, International Computer Science Insitute, Berkley, California, 1992.

[156] Nathan Schneider, Vivek Srikumar, Jena D. Hwang, and Martha Palmer. A Hierarchy with, of, and for Preposition Supersenses. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 112–123. Association for Computational Linguistics, 2015. doi: 10.3115/v1/W15-1612.

[157] Claudia Brugman and George Lakoff. Cognitive topology and lexical networks. In *Lexical Ambiguity Resolution*, pages 477–508. Elsevier, 1988. doi: 10.1016/B978-0-08-051013-2.50022-7.

[158] Stacy Doore, Kate Beard, and Nicholas Giudice. Spatial prepositions in natural-language descriptions of indoor scenes. In *Proceedings of Workshops at COSIT*, pages 255–260. Springer, 2017. doi: https://doi.org/10.1007/978-3-319-63946-8_41.

[159] Robert Dale. Cooking up referring expressions. In *Proceedings of 27th Annual*

Meeting of the association for Computational Linguistics, pages 68–75. Association for Computational Linguistics, 1989. doi: 10.3115/981623.981632.

[160] Ian Perera, James Allen, Choh Man Teng, and Lucian Galescu. Building and learning structures in a situated blocks world through deep language understanding. In *Proc 1st International Workshop on Spatial Language Understanding*, pages 12–20, 2018. doi: 10.18653/v1/W18-1402.

[161] Michael Spranger. Grounded lexicon acquisition - case studies in spatial language. *2013 IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pages 1–6, 2013. doi: 10.1109/DevLrn.2013.6652534.

[162] B. Bennett. Standpoint semantics: a framework for formalising the variable meaning of vague terms. In P. Cintula, C. Fermüller, L. Godo, and P. Hájek, editors, *Understanding Vagueness - Logical, Philosophical and Linguistic Perspectives*, volume 36 of *Studies in Logic*, pages 261 – 278. College Publications, 2011.

[163] Fu-Jen Chu, Ruinian Xu, and Patricio A Vela. Learning affordance segmentation for real-world robotic manipulation via synthetic images. *IEEE Robotics and Automation Letters*, 4(2):1140–1147, 2019. Publisher: IEEE.

[164] Shubham Tulsiani, Saurabh Gupta, David Fouhey, Alexei A. Efrosefros, and Jitendra Malik. Factoring shape, pose, and layout from the 2D image of a 3D scene. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 302–310, Salt Lake City, UT, June 2018. IEEE. doi: 10.1109/CVPR.2018.00039.

[165] Peter Henry, Michael Krainin, Evan Herbst, Xiaofeng Ren, and Dieter Fox. RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments. *The International Journal of Robotics Research*, 31(5):647–663, 2012.

[166] Saurabh Gupta, Pablo Arbelaez, Ross Girshick, and Jitendra Malik. Aligning 3D models to RGB-D images of cluttered scenes. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4731–4740, Boston, MA, USA, 2015. IEEE. doi: 10.1109/CVPR.2015.7299105.

[167] Michele I Feist and Dedre Gentner. Factors involved in the use of in and on. *Proc Annual Meeting of the Cognitive Science Society*, page 7, 2003.

[168] Karthik Mahesh Varadarajan and Markus Vincze. Afnet: The affordance network. In *Asian Conference on Computer Vision*, pages 512–523. Springer, 2012. doi: https://doi.org/10.1007/978-3-642-37331-2_39.

[169] Thanh-Toan Do, Anh Nguyen, and Ian Reid. AffordanceNet: an end-to-end deep learning approach for object affordance detection. In *Proceedings of 2018 IEEE international conference on robotics and automation*. IEEE, 2018.

[170] James Pustejovsky and Nikhil Krishnaswamy. VoxML: a visualization modeling language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4606–4613, 2016.