Proceedings Paper:
Richard-Bollans, A orcid.org/0000-0003-1980-0107, Gomez Alvarez, L and Cohn, AG orcid.org/0000-0002-7652-8907 (2020) Modelling the Polysemy of Spatial Prepositions in Referring Expressions. In: Calvanese, D, Erdem, E and Thielscher, M, (eds.) Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning. 17th International Conference on Principles of Knowledge Representation and Reasoning, 12-18 Sep 2020, Rhodes, Greece. IJCAI Organization , pp. 703-712. ISBN 978-0-9992411-7-2

https://doi.org/10.24963/kr.2020/72

Modelling the Polysemy of Spatial Prepositions in Referring Expressions

Adam Richard-Bollans , Lucía Gómez Álvarez , Anthony G. Cohn
University of Leeds, UK
{mm15alrb, sc14lga, a.g.cohn}@leeds.ac.uk

Abstract

In previous work exploring how to automatically generate typicality measures for spatial prepositions in grounded settings, we considered a semantic model based on Prototype Theory and introduced a method for learning its parameters from data. However, though there is much to suggest that spatial prepositions exhibit polysemy, each term was treated as exhibiting a single sense. 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. In this paper we address this problem by analysing the issue of reference using spatial language and examining how the polysemy exhibited by spatial prepositions can be incorporated into semantic models for situated dialogue. We support our approach on theoretical developments of Prototype Theory, which suggest that polysemy may be analysed in terms of radial categories, characterised by having several prototypicality centres. After providing a brief overview of polysemy in spatial language and a review of the related work, we define the Baseline Model and discuss how polysemy may be incorporated to improve it. We introduce a method of identifying polysemes based on 'ideal meanings' and a modification of the 'principled polysemy' framework. In order to compare polysemes and aid typicality judgements we then introduce a notion of 'polyseme hierarchy'. Subsequently, we test the performance of the extended Polysemy Model by comparing it to the Baseline Model as well as a data-driven model of polysemy which we derive with a clustering algorithm. We conclude that our method for incorporating polysemy into the Baseline Model provides significant improvement. Finally, we analyse the properties and behaviour of the generated Polysemy Model, providing some insight into the improvement in performance, as well as justification for the given methods.

1 Introduction

In previous work (Richard-Bollans, Bennett, and Cohn 2020) exploring how to automatically generate typicality measures for spatial prepositions in grounded settings, we considered a semantic model based on Prototype Theory (Rosch 1978) and introduced a method for learning its parameters from data. However, though there is much to suggest that spatial prepositions exhibit polysemy, each term was treated as exhibiting a single sense.

The ability for terms to represent distinct but related meanings is unexplored in the work on grounded semantics and referring expressions, where even homonymy[1] is rarely considered, as noted in (Siddharthan and Copestake 2004).

In this paper we address this problem by analysing the issue of reference using spatial language and examining how the polysemy exhibited by spatial prepositions can be incorporated into semantic models for situated dialogue. We support our approach on theoretical developments of Prototype Theory, which suggest that polysemy may be analysed in terms of radial categories (Lakoff 1999, Lewandowska-Tomaszczyk 2007), characterised by having several prototypicality centres.

After providing a brief overview of polysemy in spatial language and a review of the related work, we define the Baseline Model (based on (Richard-Bollans, Bennett, and Cohn 2020)) and discuss how polysemy may be incorporated to improve it. We introduce a method of identifying polysemes based on 'ideal meanings' (Herskovits 1987) and a modification of the 'principled polysemy' framework (Tyler and Evans 2001). In order to compare polysemes and aid typicality judgements we then introduce a notion of 'polyseme hierarchy'. Subsequently, we test the performance of the extended Polysemy Model by comparing it to the Baseline Model as well as a data-driven model of polysemy which we derive with a clustering algorithm. We conclude that our method for incorporating polysemy into the Baseline Model provides significant improvement. Finally, we analyse the properties and behaviour of the generated Polysemy Model, providing some insight into the improvement in performance, as well as justification for the given methods.

In this paper we consider those spatial prepositions which appear to both have an 'ideal meaning' and to exhibit polysemy at the kind of room-scales we are considering. We consider these to be 'in' (Rodrigues et al. 2020), 'under' (Zlatev 1992), 'over' (Tyler and Evans 2001, Zlatev 1992) and 'on' (Bowerman and Choi 2001)[2].

[1] Homonymy denotes the capacity of a sign to convey two or more unrelated meanings

[2] Though not explicitly studying polysemy, Bowerman & Choi provide various examples of object configurations which are labelled simply with 'on' in English but are distinguished with multiple prepositions in other languages

2 Background

With regards to terminology, in the following 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**'.

2.1 Spatial Language & Polysemy

As opposed to homonymy where senses are semantically distinct, we say that a term exhibits polysemy if it denotes multiple *related* senses and we call these distinct senses *polysemes*. For instance, a figure may be 'on' a ground if it is (1) resting on top of it e.g. 'a book on a table' (2) attached to the side of it e.g. 'a clock on a wall' (3) simply in contact with it e.g. 'a balloon on the ceiling'. This phenomenon is well known in linguistics and it is pervasive in natural language (Vicente et al. 2017). As the senses of polysemous terms are so closely intertwined, the theoretical and computational treatment of polysemy presents a difficult challenge for semantic models.

The polysemy of spatial prepositions is well recognised in the literature (Herskovits 1987, Van der Gucht, Willems, and De Cuypere 2007) which includes both detailed analysis of the semantic variation of spatial prepositions, e.g. (Tyler and Evans 2001), and attempts to provide a formal treatment of them, such as (Rodrigues et al. 2020). However, polysemy is rarely, if ever, accounted for in computational models for situated dialogue. We discuss these issues further in Section 3.

2.2 Referring Expressions

A particular aspect of situated dialogue which we explore is the processing of *referring expressions* — noun phrases which serve to identify entities e.g. 'the book under the table'. Referring Expression Generation and Comprehension (REG/C) situations provide useful scenarios for analysing the semantics of lexical items and how they are used to achieve communicative success. A lot of work has been done in creating computational models for REG/C, see (van Deemter 2016) for an overview. However, most of this work avoids expressions involving *vague* language where the extensions of lexical items (sets of entities the term may refer to) are uncertain.

In situations involving vague descriptions, binary classifications of possible referents are problematic as the problem becomes over-simplified and semantic information is lost. In place of categorisation, *typicality* becomes a central notion i.e. how *well* a potential referent fits the description (van Deemter 2016). Note that here we use typicality to denote similarity to some ideal *prototypical* notion of a concept, rather than simply frequency of occurrence.

When vagueness is explored in REG/C, it is usually with respect to *gradable* properties whose parameters are clearly defined e.g. (Gatt et al. 2018, van Deemter 2006). We explore the issue of reference using spatial language, where the semantics are not so clear and terms may be used to denote various distinct senses. Continuing the example of 'on' given in Section 2.1, imagine a scene where a book is resting on top of a table (Sense 1), a sheet of paper is attached to the side of the table (Sense 2) and a box is on the floor but touching the table (Sense 3). Each of these objects represents a particular sense in which 'on' may be used to described the relationship between the object and the table. A semantic model must be able to recognise that each of these configurations fit the term 'on' to some degree and also be able to discern which is the *best* instance of 'on' in this scenario.

3 Related Work

The possibility for a lexical item to represent multiple distinct meanings is rarely treated in the work on REG/C. This is in part because of the nature of polysemy and homonymy. For example, when collecting data an annotator may label something as 'on', but it is difficult to isolate automatically which polyseme is being intended. Moreover, in the case of spatial language, the kind of detailed semantic distinctions encountered when dealing with polysemy requires a rich dataset where the meanings of the terms are contextualised.

In studies related to situated dialogue generally, the domains of discourse are often restricted to avoid the kind of ambiguity that may arise from homonymy. Moreover, when homonymy does arise one may draw on the plethora of techniques and resources from the field of Word Sense Disambiguation (WSD) to deal with the ambiguity, as is the case in (Siddharthan and Copestake 2004). The case of polysemy, however, is less clear as (1) the set of polysemes for a given term is not clearly defined (2) it is not clear how the semantics of distinct polysemes differ or how they should be treated pragmatically (3) given a polysemous term in a referring expression context, multiple polysemes of the term may be simultaneously acceptable.

In an attempt to provide a logical framework for handling polysemy Rodrigues et al. (2020) 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 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, as discussed in (Bateman et al. 2010). Herskovits (1987) 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.

4 Data

In order to train and test typicality measures of spatial language, we collected data on spatial prepositions using 3D virtual environments, which we described in (Richard-Bollans, Bennett, and Cohn 2020). Collected data and details of the framework[3] along with more recent code and figures used in the current analysis[4] can both be found in the Leeds research data repository. The latest version of the data collection environment and code for analysis can be found on the GitHub repository[5].

Two tasks were used in the data collection process — a Preposition Selection Task and a Comparative Task. In the Preposition Selection Task participants are shown a figure-ground pair (highlighted and with text description) and asked to select all prepositions in a list which fit the configuration. In the Comparative Task a description is given with a single preposition and ground object where the figure is left ambiguous and participants are asked to select an object in the scene which *best fits* the description. The former task allows for the collection of categorical data and provides a dataset of object configurations along with a ratio, which we call the 'selection ratio', measuring the likelihood a participant would label the configuration with a given preposition. The models are trained on this selection data and the Comparative Task provides typicality judgements on which the models are tested.

4.1 Features

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 extracted features which comprise the feature space used by the semantic models. Motivation for inclusion of each of the features and further detail is given in (Richard-Bollans, Bennett, and Cohn 2020).

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

Functional Features There are two particular functional notions that appear repeatedly in the literature on spatial language: *support* and *location control*. We take *support* to express that the ground impedes motion of the figure due to gravity, while *location control* expresses that moving the ground moves the figure. Rather than attempting to formally define these notions, as in (Hedblom et al. 2017, Kalita and Badler 1991), we quantified these notions via *simulation* using Unity3D's built-in physics engine.

5 Baseline Model

The underlying model follows from previous work described in (Richard-Bollans, Bennett, and Cohn 2020). Given a preposition and a configuration (figure-ground pair) in a scene, the model assigns a value of typicality for how well the configuration fits the preposition. Following (Eyre and Lawry 2014, Gärdenfors 2004, Mast, Falomir, and Wolter 2016, Spranger and Pauw 2012), typicality in our model is calculated by considering the semantic distance to a prototype.

The model is 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

Typicality of a configuration, c, is then calculated as the semantic similarity to the prototype:

$$\textit{typicality}(c) = e^{-d_W(c,P)} \tag{1}$$

where $d_W(x, y)$ is a weighted Euclidean metric using weights W.

This model currently represents prepositions as a single sense and we use this as the baseline. In this work we aim to extend the model such that a preposition is associated with a *set* of prototypes and weights.

6 Identifying Polysemes

The first challenge is to identify the different polysemes that may be expressed by a preposition and in this section we explore how this may be achieved.

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 be trained treating each polyseme separately, which we describe in a later section.

6.1 Clustering

In order to potentially distinguish polysemes, support the polysemes we choose and suggest distinguishing features, we attempt to cluster preposition instances. We cluster the

[3] https://doi.org/10.5518/764

[4] https://doi.org/10.5518/825

[5] https://github.com/alrichardbollans/spatial-preposition-annotation-tool-unity3d

data from the Preposition Selection Task using off-the-shelf clustering algorithms provided by scikit-learn (Pedregosa et al. 2011). In the remainder of this paper, where the k-means algorithm is used we use all configurations, which are then weighted by their selection ratio for the given preposition. Where we use Hierarchical Agglomerative Clustering (HAC) we only consider 'good' instances of the preposition (where the selection ratio is greater than or equal to 0.5). 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, we currently only consider the relational features given in Section 4.1.

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 3 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.

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 1 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 0 and there is some *contact*. These generated clusters appear to represent and support the distinctions given for 'on' in Section 2.1: (Sense 1: Red & Turquoise), (Sense 2: Green) and (Sense 3: Blue).

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 of these canonical meanings in the following sections.

6.2 Ideal Meanings

Herskovits (1987) argues that the meanings of spatial prepositions should be understood as *ideal meanings* from which other uses of the prepositions are derived. Clearly the ideal meaning of a preposition represents a polyseme that should be represented in our model and so we begin by defining these. We draw on intuition and the literature to assign representations of the ideal meaning to each preposition below.

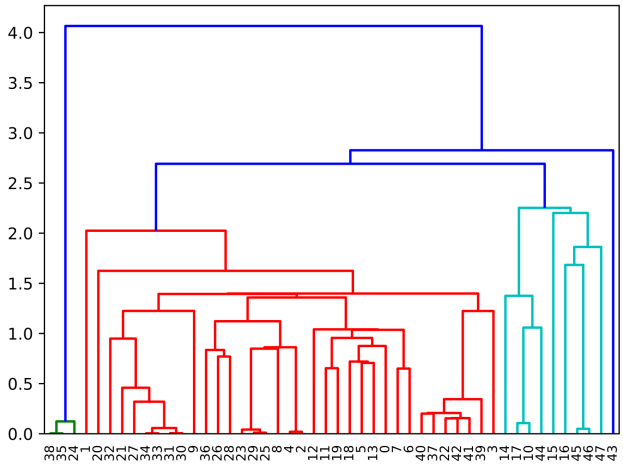

Figure 1: Dendrogram from HCA for 'on'

For the prepositions 'in' and 'on' we follow (Garrod, Ferrier, and Campbell 1999) and assume that the underlying representations comprise both geometric and functional components.

In Following (Garrod, Ferrier, and Campbell 1999), '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 (Garrod, Ferrier, and Campbell 1999) 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 in the *support* image schema provided in (Mandler 1992). We therefore define the ideal notion of 'on' as having a high value of three features: *support*, *above_proportion* and *contact*.

Under Herskovits (1987) 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 (Mori 2019, Tyler and Evans 2001). 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 part 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 (Tyler and Evans 2001) 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.

7 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 and underlying model used in the Baseline 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. 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 2, the typicality, $\textit{typicality}_p(c)$, assigned by a polyseme, p, to a configuration, c, is specified by these prototypes and weights.

7.1 Sharing Prototypes

In order to explore the nature of polysemy and how it may impact semantic representations we initially consider two separate polysemy models, which we call the Distinct Prototype Model and Shared Prototype Model. The models 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 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 2.1 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.

8 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 2.1 — as well as a box on the floor but touching the table — Sense 3 from Section 2.1. 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 (Miller 1995) capture knowledge such as 'a car is a vehicle'. In the case of prepositions, (Schneider et al. 2015) 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 2. P is the set of polysemes of the preposition which may apply to c, $\textit{typicality}_p(c)$ is the typicality of c with respect to a polyseme p and r_p is the rank of polyseme p.

$$\textit{typicality}(c) = \max_{p \in P}(\textit{typicality}_p(c) \times r_p) \qquad (2)$$

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.

Our polysemy models are then 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

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 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.

9 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.

9.1 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 1 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.

9.2 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 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 1.

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.

10 Model Performance

While the Preposition Selection Task provides categorical data from each participant, the Comparative Task provides

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

	Polysemy Model	Baseline Model	K-Means Model
in	0.801	**0.813**	0.790
on	0.94	0.924	**0.952**
under	**0.898**	0.764	0.882
over	**0.814**	0.800	0.685
Average	**0.863**	0.825	0.827
Overall	**0.893**	0.845	0.869

Table 2: K-Fold Test Results (K=10, N=10). Scores are averaged results of the cross-validation

qualitative judgements regarding which configurations of objects better fit a given description. The testing scenario is restricted in such a way that the involved pragmatics is simple and ideally the only judgement occurring is related to typicality. We suppose that the configuration 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 (Richard-Bollans, Bennett, and Cohn 2020)) 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 the Comparative Task, 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. We turn the collection of obtained rankings into inequalities, or *constraints*, which the models should satisfy. Weights are assigned to each constraint representing the evidence supporting the constraint. This set of weighted constraints provides a metric for testing the models and is described in more detail in (Richard-Bollans, Bennett, and Cohn 2020).

In Tables 1 & 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.

10.1 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 1.

The Distinct Prototype Model outperforms the Shared Prototype Model with 'under' 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[6]. From here on we discard the Shared Prototype Model and refer to the Distinct Prototype Model as the **Polysemy Model**.

10.2 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 2.

We use $k = 10$ here as opposed to in (Richard-Bollans, Bennett, and Cohn 2020) where $k = 2$ is used as the Polysemy Model requires a larger dataset for training.

Results The Polysemy Model has significantly improved on the Baseline Model for the prepositions 'on', 'under' and 'over'. In the case of 'in', the Baseline Model outperforms the Polysemy Model. We believe that this is partly because the Polysemy 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.

Both the models which have accounted in some way for polysemy have in general improved on the Baseline Model and, 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 baseline is significant, we assume a null hypothesis that both the models are equally likely to perform better than the other (with respect to the overall score) for a given random fold. The Polysemy Model performs better on 64 out of 100 repetitions. Assuming the null hypothesis, the probability of one model outperforming the other on at least 64 of the repetitions is very small: $(P(\geq 64) = \sum_{r\geq 64}^{100} {}^{100}C_r 0.5^{100} = 0.0033$. We may therefore conclude that the Polysemy Model has offered significant improvement over the baseline.

[6]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

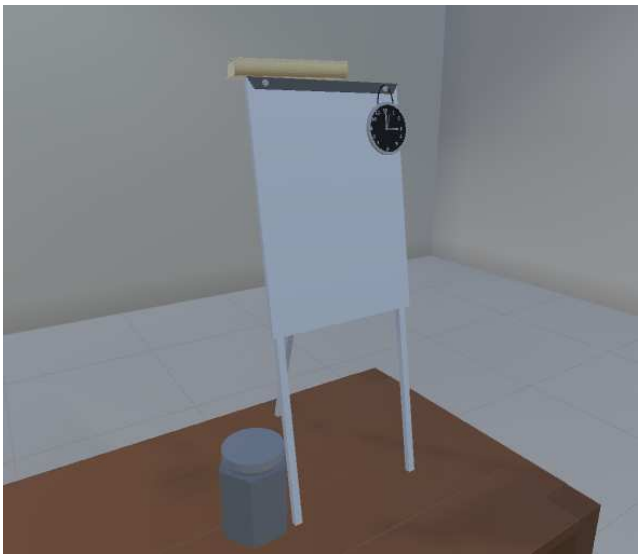

Figure 2: Example polyseme instances for 'on'

Configuration	Polysemy Model (*typicality* × *rank*)	Baseline Model
(book, board)	$0.779 \times 0.811 = 0.632$	0.626
(clock, board)	$0.615 \times 0.776 = 0.477$	0.204
(jar, board)	$0.477 \times 0.088 = 0.042$	0.219

Table 3: Assigned typicality scores for 'on'

The improvement shown by the k-Means Model over the baseline, however, does not appear to be significant. Also, though the Polysemy Model improves on the k-Means Model it is not a clearly significant improvement.

11 Generated Model

Here we consider the Polysemy Model when it is trained on all the available data and analyse its properties and behaviour.

11.1 Typicality Values

Firstly, to illustrate how the model assigns typicality to configurations and how this compares to the baseline we consider an example. In Figure 2 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 3.

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) and this appears to be represented in the values assigned by the Baseline Model as well as the ranks from the Polysemy Model. 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 Model appears to deal with this better than the Baseline Model. Though it does assign a higher value to (book, board) than (clock, board) these values are similar, compared to the Baseline Model which assigns a very low value to (clock, board). The Baseline 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.

11.2 Generated Polysemes

Ranks & 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. With one small exception[7], this is exactly what we observe. 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 Model partitions the data into polysemes, we estimate how well the polysemes cluster the data. In the following we take polyseme instances to be any configuration that has been labelled with the preposition and which fits the polysemes conditions.

In order to cluster the data with the generated polysemes, for a given preposition, we first calculate the mean feature values of the instances of each polyseme. We then take this set of means to act as cluster centres and measure the inertia given by this clustering (a point is assigned to the cluster with the nearest cluster centre[8]). We compare this to inertia values given by a k-means clustering algorithm, see Figure 3 for the case of 'on'. The lower the value of the inertia the more internally coherent the clusters are. As we can

[7] 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

[8] Note that here to be consistent with the inertia measure given by the k-means algorithm we use regular Euclidean distance rather than the weighted Euclidean metric used in Section 9.2

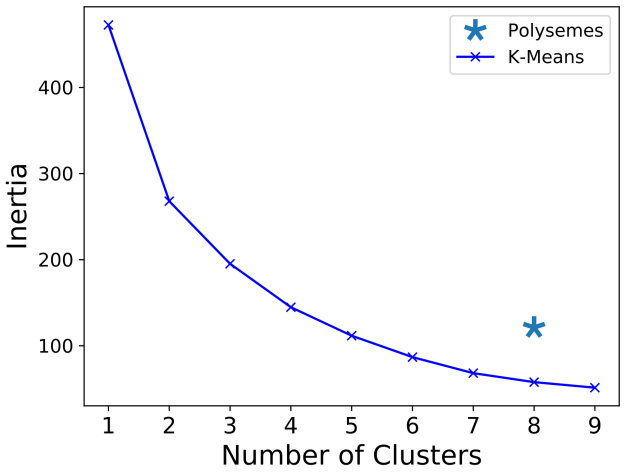

Figure 3: Inertia from k-means clustering vs. Polyseme clustering

see, the clustering using the polysemes performs quite well, equivalent to using the algorithm with $k = 5$. This result is similar for the other prepositions we consider, see the data archive[4] for the respective plots.

12 Discussion

In this paper 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 (Tyler and Evans 2001) for distinguishing polysemes, we have developed an approach which we hope can be incorporated in semantic models more generally. While we have relied somewhat on intuition to generate ideal meanings, it may be possible to generalise our approach to concepts whose ideal meaning is not so clear e.g. by analysing which features are most salient.

We have also introduced a notion of '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 baseline when interpreting a particular class of referring expressions. Moreover, the model we have generated is transparent, with parameters that are clearly interpretable.

As well as a polysemy model based around Criterion 1 and the authors' intuition, we also created a model based on a k-Means clustering algorithm. Though we cannot judge the k-Means Model to be significantly better than the baseline, that it performs reasonably in our tests suggests that with further refinement it may provide another method for accounting for polysemy in semantic models. Furthermore, the reasonable performance of the k-Means Model as well as the significantly improved performance of the Polysemy Model provides further evidence that the selected prepositions do exhibit polysemy and, moreover, that accounting for polysemy is important when processing referring expressions.

13 Future Work

In this paper we have been considering the role of polysemy in typicality judgements related to spatial language 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/C, e.g. (Frank and Goodman 2012), it is considered important to be able to assess how appropriate or acceptable a preposition is for a given figure-ground pair. Though related, this is a different challenge and provides extra information on the possible utterances that a speaker could make. Unlike what we have considered so far, this is often reliant on particular properties of ground objects (e.g. for 'in' whether or not the ground is a type of *container* (Richard-Bollans et al. 2019)). We intend to extend the work we have done so far and explore how polysemy may be accounted for in these types of acceptability judgements.

We have based our study on those prepositions which, based on existing literature, appear to exhibit polysemy at room/table-top scales. 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. This does not appear to have been addressed in the existing work and is something we would like to work towards.

Acknowledgments

The first author is supported by an EPSRC Studentship; the final author is partially supported by an Alan Turing Institute Fellowship. The partial support of the EU under the Horizon 2020 research and innovation programme grant agreement 825619 is also gratefully acknowledged.

References

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

Bowerman, M., and Choi, S. 2001. Shaping meanings for language: universal and language-specific in the acquisition of semantic categories. In *Language acquisition and conceptual development*. Cambridge University Press. 475–511.

Eyre, H., and Lawry, J. 2014. Language games with vague categories and negations. *Adaptive Behavior* 22(5):289–303.

Frank, M. C., and Goodman, N. D. 2012. Predicting pragmatic reasoning in language games. *Science* 336(6084):998–998.

Garrod, S.; Ferrier, G.; and Campbell, S. 1999. In and on: investigating the functional geometry of spatial prepositions. *Cognition* 72(2):167–189.

Gatt, A.; Marín, N.; Rivas-Gervilla, G.; and Sánchez, D. 2018. Specificity measures and reference. In *Proc 11th International Conference on Natural Language Generation*, 492–502.

Gärdenfors, P. 2004. Conceptual spaces as a framework for knowledge representation. *Mind and Matter* 2(2):9–27.

Hedblom, M. M.; Kutz, O.; Mossakowski, T.; and Neuhaus, F. 2017. Between contact and support: introducing a logic for image schemas and directed movement. In *Proc IAAI*, volume 10640, 256–268. Springer.

Herskovits, A. 1987. *Language and spatial cognition*. Cambridge University Press.

Kalita, J. K., and Badler, N. I. 1991. Interpreting prepositions physically. In *Proc AAAI*, 105–110.

Lakoff, G. 1999. Cognitive models and prototype theory. *Concepts: Core Readings* 391–421. Publisher: MIT Press Cambridge, MA.

Lewandowska-Tomaszczyk, B. 2007. Polysemy, prototypes, and radial categories. In *The Oxford Handbook of Cognitive Linguistics*. Oxford University Press. 139–169.

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

Mast, V.; Falomir, Z.; and Wolter, D. 2016. Probabilistic reference and grounding with PRAGR for dialogues with robots. *Journal of Experimental & Theoretical Artificial Intelligence* 28(5):889–911.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM* 38(11):39–41.

Mori, S. 2019. A cognitive analysis of the preposition *over*: image-schema transformations and metaphorical extensions. *Canadian Journal of Linguistics* 64(3):444–474.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; and others. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12:2825–2830.

Richard-Bollans, A.; Bennett, B.; and Cohn, A. G. 2020. Automatic generation of typicality measures for spatial language in grounded settings. In *Proceedings of 24th European Conference on Artificial Intelligence*.

Richard-Bollans, A.; Gómez Álvarez, L.; Bennett, B.; and Cohn, A. G. 2019. Investigating the dimensions of spatial language. In *Proceedings of Speaking of Location 2019: Communicating about Space*. CEUR Workshop Proceedings.

Rodrigues, E. J.; Santos, P. E.; Lopes, M.; Bennett, B.; and Oppenheimer, P. E. 2020. Standpoint semantics for polysemy in spatial prepositions. *Journal of Logic and Computation*.

Rosch, E. 1978. Principles of categorization. In Rosch, E., and Lloyd, B. B., eds., *Cognition and Categorization*, volume 1. Hillsdale, NJ: Lawrence Erlbaum Associates. 27–78.

Schneider, N.; Srikumar, V.; Hwang, J. D.; and Palmer, M. 2015. A Hierarchy with, of, and for Preposition Supersenses. In *Proceedings of The 9th Linguistic Annotation Workshop*, 112–123. Association for Computational Linguistics.

Siddharthan, A., and Copestake, A. 2004. Generating referring expressions in open domains. In *Association for Computational Linguistics*.

Spranger, M., and Pauw, S. 2012. Dealing with perceptual deviation: vague semantics for spatial language and quantification. In *Language Grounding in Robots*. Boston, MA: Springer US. 173–192.

Tyler, A., and Evans, V. 2001. Reconsidering prepositional polysemy networks: the case of over. *Language* 77(4):724–765.

van Deemter, K. 2006. Generating referring expressions that involve gradable properties. *Computational Linguistics* 32(2):195–222.

van Deemter, K. 2016. *Computational models of referring: a study in cognitive science*. MIT Press.

Van der Gucht, F.; Willems, K.; and De Cuypere, L. 2007. The iconicity of embodied meaning. Polysemy of spatial prepositions in the cognitive framework. *Language Sciences* 29(6):733–754.

Vicente, A.; Falkum, I. L.; Vicente, A.; and Falkum, I. L. 2017. Polysemy. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Zlatev, J. 1992. A study of perceptually grounded polysemy in a spatial microdomain. Technical Report TR-92-048, International Computer Science Insitute, Berkley, California.