ECAI 2020
G.D. Giacomo et al. (Eds.)

doi:10.3233/FAIA200341

Automatic Generation of Typicality Measures for Spatial Language in Grounded Settings

Adam Richard-Bollans[1] and **Brandon Bennett**[2] and **Anthony G. Cohn**[3]

Abstract.

In cognitive accounts of concept learning and representation three modelling approaches provide methods for assessing typicality: *rule-based*, *prototype* and *exemplar* models. The prototype and exemplar models both rely on calculating a weighted semantic distance to some central instance or instances. However, it is not often discussed how the central instance(s) or weights should be determined in practice. In this paper we explore how to automatically generate prototypes and typicality measures of concepts from data, introducing a prototype model and discussing and testing against various cognitive models. Following a previous pilot study, we build on the data collection methodology and have conducted a new experiment which provides a case study of spatial language for the current proposal. After providing a brief overview of cognitive accounts and computational models of spatial language, we introduce our data collection environment and study. Following this, we then introduce various models of typicality as well as our prototype model, before comparing them using the collected data and discussing the results. We conclude that our 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.

1 Introduction

In cognitive accounts of concept learning and representation, three modelling approaches provide methods for assessing typicality: *rule-based*, *prototype* and *exemplar* models. The prototype and exemplar models both rely on calculating a weighted semantic distance to some central instance or instances. However, it is not often discussed how the central instance(s) or weights should be determined in practice. In this paper we explore how to automatically generate prototypes and typicality measures of concepts from data, introducing a prototype model and discussing and testing against various cognitive models. Following a previous pilot study [29], we build on the data collection methodology and have conducted a new experiment which provides a case study of spatial language for the current proposal. After providing a brief overview of cognitive/computational models of spatial language, we introduce the data collection environment[4] and study[5]. Following this we then introduce various models of typicality as well as our prototype model, before comparing them using the collected data and discussing the results. We conclude that our 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.

The primary motivation for this work is to explore semantic issues of spatial language in order to provide a robust, cognitively-aligned semantic model that can be applied to natural language understanding and generation in the context of human-robot interaction. Through exploring semantics in situated dialogue we also aim to provide analysis and data which furthers the theoretical work on spatial language and cognition as well as cognitive models of concepts more generally.

In this paper we investigate the semantics of spatial prepositions, in particular those considered to have a functional component as well as those prepositions that seem to act as a geometric counterpart. In English, we consider these to be: 'in', 'inside', 'on', 'on top of', 'against', 'over', 'above', 'under' and 'below'.

2 Background & Related Work

A particular aspect of situated dialogue we explore is the processing of *referring expressions* — noun phrases which serve to identify entities e.g. 'the book under the table'.

2.1 Referring Expressions

Referring Expression Generation and Comprehension (REG & REC) situations provide useful scenarios for analysing the semantics of lexical items and how terms are used to achieve communicative success. A lot of work has been done in creating computational models for REG and REC, see [34] for an 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. [33]. 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.

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 [34]. Note that here we use typicality to denote similarity to some ideal *prototypical* notion of 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

[1] University of Leeds, United Kingdom, email: mm15alrb@leeds.ac.uk
[2] University of Leeds, United Kingdom, email: b.bennett@leeds.ac.uk
[3] University of Leeds, United Kingdom, email: a.g.cohn@leeds.ac.uk
[4] Details of the software can be found here:
`https://github.com/alrichardbollans/spatial-preposition-annotation-tool-unity3d`
[5] Collected data and archive of software can be found here:
`https://doi.org/10.5518/764`

the Gricean Maxim of Manner [13], 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'.

2.2 Modelling Prototypes & Typicality

Cognitive accounts of concept learning and representation present three separate approaches to modelling typicality — rule-based, exemplar and prototype models.

Rule-based models of typicality largely rely on expert intuition to generate rules, for example [1, 25] in the context of spatial language. This can prove successful where the semantics of terms involve a small number of well-understood features; however due to the semantic complexity of spatial language it is not clear how one would build a robust rule-based model. In [16] these models are referred to as *Simple Relation* models and many examples of the ways they fail are given.

Based on Rosch's Prototype Theory [30], prototype models assess typicality of an instance by measuring its semantic distance to the *prototype*, where the prototype is the most central member of the category. In geometric representations this usually takes the form of the geometric centre of the category [5]. In feature-models this takes the form of family resemblance [30],where prototypical members of a category are those members with the most properties in common with other members of the category.

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 [23, 35].

A more recent approach that has been considered as a unification of both the prototype and exemplar view is that of Conceptual Spaces [14]. 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 [21, 26].

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. The current paper explores this issue and proposes that for certain classes of concepts a different notion of typicality may be suitable.

2.3 Computational Models

There have been many attempts to model the semantics of spatial prepositions in grounded settings. The majority of this work, however, focuses on modelling projective prepositions. We believe the underlying semantics of these terms to be simpler than those prepositions with a functional component and that the problem of interpretation is more pragmatic in nature. Of the models which tackle the spatial prepositions that we are currently considering, there are some simple rule-based models, e.g. [1, 25], as well as trained models, e.g. [2, 11], whose representations do not provide a clear insight into the semantic of the terms. We desire a detailed trained model which also provides semantic insights.

Of particular interest is the work of Mast et al. [22] where a pragmatic model is developed to tackle problems involving referring expressions. In order to model the semantics of the terms involved, following [7, 14, 31], they use the notion of semantic distance in a feature space, where graded category membership can be determined by calculating the semantic distance to a prototype in the space. Mast et al. also 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 [22] 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. By automatically generating model parameters we hope to provide support for the inclusion of functional features in our model and also to aid future work regarding the polysemy[6] exhibited by spatial prepositions.

2.4 Existing 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 give a brief overview here of features that appear in existing computational models, outlining geometric and functional relations that are used to model the above terms.

Unsurprisingly, geometric features have been well covered in the field. We list the principal and most commonly occurring geometric features here:

- Contact [25]
- Distance [1, 2, 3, 11, 12, 19, 25]
- Overlap with projection from objects [1, 3, 25]
- Height difference [1, 25]
- Object alignment [1, 2, 11, 12, 19]
- Containment [1, 3, 11, 25]

Various subtle differences may exist between the features in these models e.g. distance between objects may be calculated between object bounds or centres of mass. Also, simplifications are often made for computational reasons e.g. calculations are often made using bounding boxes of objects.

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 [4, 8, 10, 17]. The use of spatial prepositions is determined by geometric *and* functional features, as evidenced in [4, 8, 10].

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*, [18] does 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 [12, 19, 20, 27] which could be considered as something like a proxy for location control, but other than this the feature is non-existent in existing work.

3 Data Collection

In order to investigate typicality measures and compare models, we extend our previous work [29] and collect data on spatial prepositions

[6] A word is said to exhibit *polysemy* where the word has multiple related senses. Each of these senses is called a *polyseme*.

using 3D virtual environments. To do this we set up a data collection framework which we describe below. Collected data and details of the framework can be found in the Leeds research data repository[5].

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**'.

3.1 Environment & Tasks

The data collection framework is built on the Unity3D[7] game development software, which provides ample functionality for the kind of tasks we implement. Two tasks were created for our study — a Preposition Selection Task and a Comparative Task. The former allows for the collection of categorical data while the latter provides typicality judgements.

In the Preposition Selection Task participants are shown a figure-ground pair (highlighted and with text description, see Figure 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.

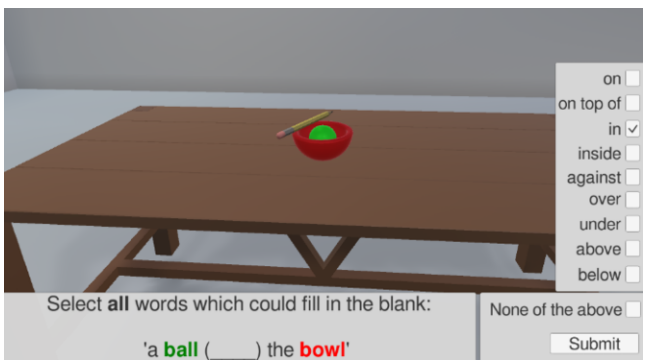

Figure 1. Preposition Selection Task

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 [6], 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. 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.

In the Comparative Task a description is given with a single preposition and ground object where the figure is left ambiguous, see Figure 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.

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.

[7] https://unity.com/

Figure 2. Comparative Task

3.2 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 features extracted from scenes and used in our analysis. Exact details of how each feature is calculated are given in the data archive[5].

In our analysis we have represented in some form each relational feature discussed in Section 2.4, which we believe accounts for the majority of features given in computational models of spatial prepositions.

3.2.1 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
- *g_covers_f*: the proportion of the figure which is covered by the ground, either above or below
- *f_covers_g*: the proportion of the ground which is covered by the figure, either above or below

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 to an approximation of the ground.

3.2.2 Functional Features

Building on our previous work [29], we explore the relationship between spatial prepositions and functional features and consider how to extend existing semantic models to account for them.

There are two particular functional notions that appear over and over 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 [15, 18], we quantified these notions via *simulation* using Unity3D's built-in physics engine.

To assess the degree to which a ground gives *support* to a given figure, we first measure the distance fallen by the figure when the ground is removed from the scene. This is then divided by an appropriate normalising distance which creates a value of 1 if the ground fully supports the figure and 0 if no support is offered. To measure *location_control* we apply horizontal forces to the ground and measure how much the figure is moved.

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[8]. For the study 67 separate scenes were created in order to capture a variety of tabletop configurations. 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.

4 Models

In this section we introduce the models we tested and provide details of how they were generated. We set up three Simple Relation models relying on expert intuition of the first author, all the other models were generated using data from the Preposition Selection Task.

4.1 Standardising Features

Firstly, it is necessary to standardise the features such that the calculated feature weights are meaningful and can be compared. As in [26], 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:

$$z = \frac{x - \bar{x}}{\sigma} \tag{1}$$

where $\bar{x}$ is the mean of the given feature and σ is the standard deviation.

4.2 Distance & Semantic Similarity

The models that are trained on the data rely on a notion of semantic distance and, following [7, 14, 22, 31], typicality in our proposed model is calculated by considering the semantic distance to a prototype.

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

Following much of the existing literature, e.g. [23], semantic similarity between two points x and y in a feature space is measured as a decaying function of the distance, $d(x,y)$:

$$s(x,y) = e^{-c \cdot d(x,y)} \tag{2}$$

where c is the *specificity* of the category which denotes how sensitive the concept is to changing values. Note that we are not currently concerned with this value and set c equal to 1 for the remainder. We take the distance, $d(x,y)$, to be the weighted Euclidean metric:

$$d(x,y) = \sqrt{w_1(x_1 - y_1)^2 + \cdots + w_n(x_n - y_n)^2} \tag{3}$$

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 2:

$$T(x) = s(x,P) = e^{-d(x,P)} \tag{4}$$

4.3 Simple Relation Models

For the Simple Relation models we replicate rule-based models given in the literature and have chosen salient features and their typical values for each preposition. Typicality is then calculated using the above formulae where the weight is 1 for each salient feature and 0 for non-salient features. We set up a simple geometric model and an intuitive best guess model as a benchmark and for comparison.

The Simple Model is based on what can be found in most 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*.

The Best Guess Model is a copy of the Simple Model except we add in functional features for 'in', 'on' and 'against' — *location_control* for 'in' and 'against' and *support* for 'on' — 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.

Finally, as a baseline we created a Proximity Model which judges typicality based solely on *shortest_distance* — the closer two objects are, the higher the measure of typicality. We include this model based on our previous study [29] which indicated that judgements based solely on proximity may be relatively successful in interpreting referring expressions for some prepositions.

4.4 Our Prototype Model

Our 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'.

Firstly, we generate a 'Selection Ratio' for each configuration and preposition based on how often participants would label the configuration with the given preposition in the Preposition Selection Task.

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 [24] we predict the feature value when the selection ratio is 1. Figure 3 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, which are described below.

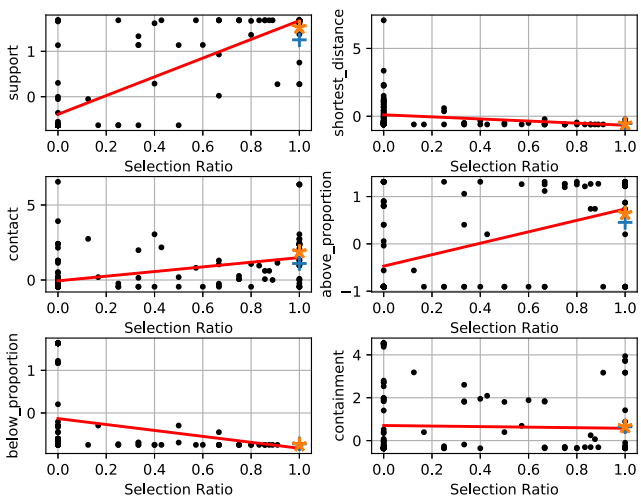

Figure 3. Finding prototypical feature values for 'on'

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 3). 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 3 that all salient features appear to be assigned appropriate prototypical values.

In order to find the salience of each feature we plot the selection ratio against the feature values. Using multiple linear regression we obtain coefficients for each feature which indicate how the selection ratio varies with changes in the feature. We can therefore assign feature weights by taking the absolute value of the coefficient given by this linear regression model.

4.5 Conceptual Space Model

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. To calculate the prototype in this space we calculate the geometric centre of all these points. We assign feature weights using the weights calculated for the Prototype model.

4.6 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 [23, 35]:

$$T(x) = \sum_{e \in \mathbb{E}} s(e,x) \tag{5}$$

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

5 Results

Firstly, to assess whether it is sensible to try and capture a generally agreed notion of typicality for spatial prepositions we calculate and compare annotator agreement in both tasks.

5.1 Annotator Agreement

In order to assess annotator agreement we calculate Cohen's Kappa for each pair of annotators in each task, Table 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[5] for details.

Task	Shared Annotations	Average Expected Agreement	Average Observed Agreement	Average Cohen's Kappa
Preposition Selection	11880	0.757	0.878	0.684
Comparative	1325	0.566	0.766	0.717

Table 1. Summary of annotator agreements

The observed agreement is 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 Task is therefore higher than in the Comparative Task and when we account for this using Cohen's Kappa we get higher agreement for the Comparative Task. We therefore conclude that it is reasonable to attempt to construct a model which represents a generally agreed notion of typicality for spatial prepositions.

5.2 Evaluation Set Up

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 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, 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. 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 $\mathbb{M}_p(f_1,g) > \mathbb{M}_p(f_2,g)$ and $\mathbb{M}_p(f_2,g) > \mathbb{M}_p(f_3,g)$. Let's say that $x_1 = 10, x_2 = 1, x_3 = 0$. It is more important that the model satisfies the first 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 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.

5.3 Initial Model Testing

As a preliminary insight, we generate models as described above using *all* the data from the Preposition Selection Task (∼140 configurations) and then evaluate the models as described above using all data from the Comparative Task. As the tasks use the same scenes, some of the same configurations will be used for both learning and testing and we therefore cannot be confident that the models are not over-fitted. Nonetheless it is interesting to consider how well the models translate categorical data into typicality rankings. See Figure 4 for initial scores.

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.

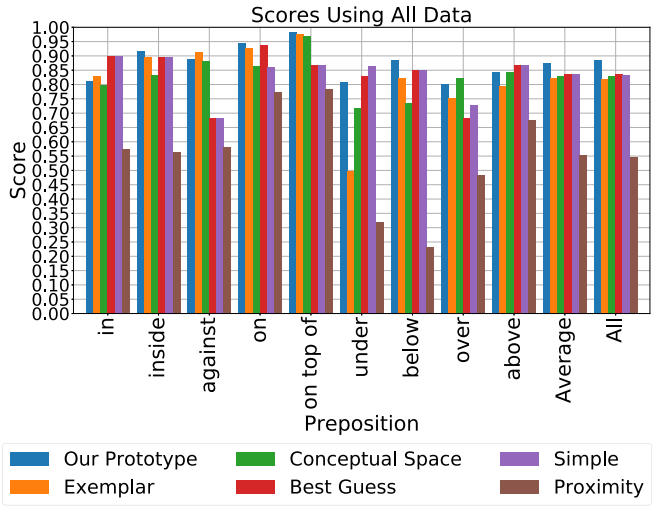

Figure 4. Scores using all scenes for both training and testing

Though most of the 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 with our model appearing to perform significantly better. We however need to test how robust the models are to changes in training data, as there is a possibility these models are over-fitted.

5.4 Restricting the Training Set

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 with $k = 2$. 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 100 times and averaged the results, shown in Figure 5. The results with $k = 3$ are similar.

Firstly, initial results show that our model is robust to reducing the training data. From ∼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: 0.863) 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.

To assess whether the improvement shown by our model over others 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, as described above. Over 200 repetitions, for any given model in a minimum of 155 repetitions our Prototype model performs better. Assuming the null hypothesis, the

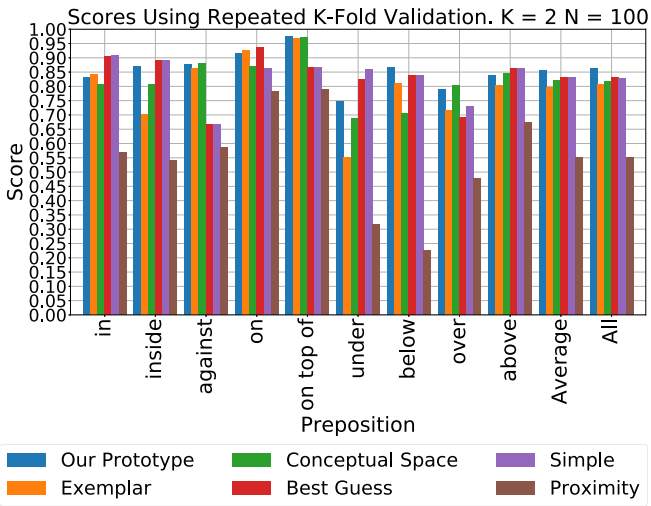

Figure 5. Scores with 100 repetitions of 2-fold cross validation

probability of one model outperforming the other on at least 155 of the repetitions is minuscule ($P(\geq 155) = \sum_{k\geq 155}^{200} {}^{200}C_k 0.5^{200} = 1.2 \times 10^{-15}$). We may therefore conclude that our model does genuinely outperform the others.

5.5 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.

5.5.1 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*. Similarly for *containment* with 'in' and 'inside'. 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*.

5.5.2 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 Figure 6.

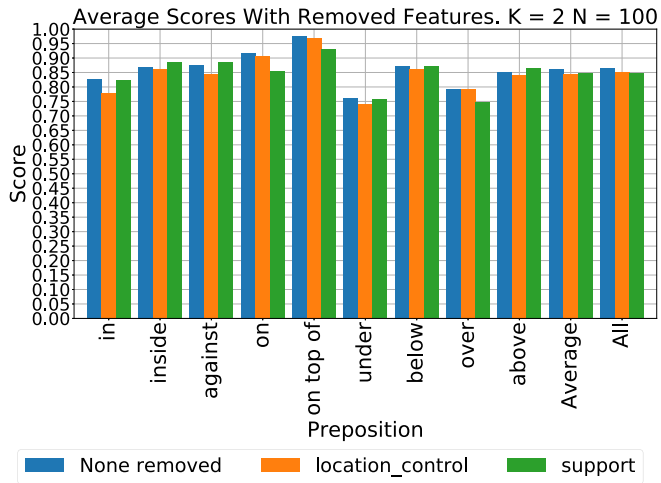

Figure 6. Scores with 100 repetitions of 2-fold cross validation, with changing feature set

As we can see in most cases our model performs better with the functional features included. In 169 of the tests the score is higher for the model with no features removed compared to when *location_control* is removed, likewise for *support* this number is 175. Again, this is significant in both cases ($P(\geq 169) = \sum_{k\geq 169}^{200} {}^{200}C_k 0.5^{200} = 1.7 \times 10^{-24}$). Our model does therefore perform better with these features included. In particular, note that 'in' is significantly affected when removing *location_control* and 'on' is affected when removing *support*.

6 Discussion

We believe that the improvement shown by our model over the Conceptual Space and Exemplar approaches is mostly due to these models being reliant on having very good exemplars in the data; which is not always practical, in particular when modelling abstract concepts with *idealised* meanings.

Consider the spatial preposition 'in'. Suppose that we do not know what 'in' means but have some data representing instances of 'in' and would like to generate a typicality measure for 'in'. 'in' is generally understood to have an ideal meaning represented by the notion of containment [17], where the more containment being expressed in an instance the more typical it is of 'in'. However, as can be seen in our data, full containment is not always present for typical instances of 'in'. Therefore, the most typical instance in the Exemplar and Conceptual Space models is likely to display less than full containment.

As discussed in [28], 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 [32], 'over' may also indicate *contact* between figure and ground. For this reason we wanted to explore models

which go beyond expressing spatial prepositions with one or two hand-picked features. Moreover, in doing so we show the potential applicability of our method to concepts which do not have small set of known salient features.

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 'Simple Relation' models and outperforms them in doing so.

7 Future Work

Using the semantic model and data collection environment that we have developed there are a number of further issues related to spatial language use that we are interested in exploring. Firstly, we would like to explore how polysemes can be automatically identified in grounded settings and how polysemy can be appropriately accounted for in our model.

Secondly, we have been considering 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, e.g. [9], 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* [29]). This issue is also related to polysemy and is something we intend to explore further.

Finally, we would like to explore pragmatic issues related to spatial language use and how our model can provide semantic input for pragmatic strategies. To explore the pragmatic issues we intend to collect data using a similar environment to the current work where participants will have to communicate with a prototype dialogue system in order to complete a simple task, e.g. collecting objects in a scene.

REFERENCES

[1] Alicia Abella and John R Kender, 'Qualitatively describing objects using spatial prepositions', in *IEEE Workshop on Qualitative Vision*, pp. 33–38. IEEE, (1993).

[2] Muhannad Alomari, Paul Duckworth, Majd Hawasly, David C Hogg, and Anthony G Cohn, 'Natural Language Grounding and Grammar Induction for Robotic Manipulation Commands', in *Proceedings of the First Workshop on Language Grounding for Robotics*, pp. 35–43, (2017).

[3] Angel Chang, Manolis Savva, and Christopher D. Manning, 'Learning Spatial Knowledge for Text to 3d Scene Generation', in *Proc EMNLP*, pp. 2028–2038. Association for Computational Linguistics, (2014).

[4] 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).

[5] Igor Douven, 'Putting prototypes in place', *Cognition*, **193**, (2019).

[6] Igor Douven, Lieven Decock, Richard Dietz, and Paul Égré, 'Vagueness: A Conceptual Spaces Approach', *Journal of Philosophical Logic*, **42**(1), 137–160, (2013).

[7] Henrietta Eyre and Jonathan Lawry, 'Language games with vague categories and negations', *Adaptive Behavior*, **22**(5), 289–303, (2014).

[8] 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*, pp. 345–349, (1998).

[9] M. C. Frank and N. D. Goodman, 'Predicting Pragmatic Reasoning in Language Games', *Science*, **336**(6084), 998–998, (2012).

[10] Simon Garrod, Gillian Ferrier, and Siobhan Campbell, 'In and on: investigating the functional geometry of spatial prepositions', *Cognition*, **72**(2), 167–189, (1999).

[11] Dave Golland, Percy Liang, and Dan Klein, 'A Game-Theoretic Approach to Generating Spatial Descriptions', in *Proc EMNLP*, p. 10, (2010).

[12] Peter Gorniak and Deb Roy, 'Grounded semantic composition for visual scenes', *Journal of Artificial Intelligence Research*, **21**, 429–470, (2004).

[13] H. Paul Grice, 'Logic and conversation', in *Syntax and Semantics, Vol. 3, Speech Acts*, 41–58, Academic Press, New York, (1975).

[14] Peter Gärdenfors, 'Conceptual spaces as a framework for knowledge representation', *Mind and Matter*, **2**(2), 9–27, (2004).

[15] 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, pp. 256–268. Springer, (2017).

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

[17] Annette Herskovits, *Language and spatial cognition*, Cambridge University Press, 1987.

[18] Jugal K Kalita and Norman I Badler, 'Interpreting prepositions physically', in *AAAI*, pp. 105–110, (1991).

[19] John D. Kelleher and Fintan J. Costello, 'Applying computational models of spatial prepositions to visually situated dialog', *Computational Linguistics*, **35**(2), 271–306, (2009).

[20] 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*, pp. 401–416. Springer, (1999).

[21] 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).

[22] 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).

[23] Robert M Nosofsky, 'Exemplar-based accounts of relations between classification, recognition, and typicality.', *Journal of Experimental Psychology: learning, memory, and cognition*, **14**(4), 700, (1988).

[24] 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).

[25] Georgiy Platonov and Lenhart Schubert, 'Computational Models for Spatial Prepositions', in *Proc 1st International Workshop on Spatial Language Understanding*, pp. 21–30, (2018).

[26] Martin Raubal, 'Formalizing conceptual spaces', in *Proc FOIS*, volume 114, pp. 153–164, (2004).

[27] Terry Regier and Laura A Carlson, 'Grounding spatial language in perception: an empirical and computational investigation.', *Journal of experimental psychology: General*, **130**(2), (2001).

[28] Adam Richard-Bollans, 'Towards a Cognitive Model of the Semantics of Spatial Prepositions', in *ESSLLI Student Session Proceedings*. Springer, (2018).

[29] Adam Richard-Bollans, Lucía Gómez Álvarez, Brandon Bennett, and Anthony G. Cohn, 'Investigating the Dimensions of Spatial Language', in *Proc Speaking of Location 2019: Communicating about Space*, (2019).

[30] Eleanor Rosch and Carolyn B. Mervis, 'Family resemblances: Studies in the internal structure of categories', *Cognitive Psychology*, **7**(4), 573 – 605, (1975).

[31] Michael Spranger and Simon Pauw, 'Dealing with Perceptual Deviation: Vague Semantics for Spatial Language and Quantification', in *Language Grounding in Robots*, 173–192, Springer US, Boston, MA, (2012).

[32] Andrea Tyler and Vyvyan Evans, 'Reconsidering Prepositional Polysemy Networks: The Case of over', *Language*, **77**(4), 724–765, (2001).

[33] Kees van Deemter, 'Generating Referring Expressions that Involve Gradable Properties', *Computational Linguistics*, **32**(2), 195–222, (2006).

[34] Kees van Deemter, *Computational models of referring: a study in cognitive science*, MIT Press, 2016.

[35] 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).