Cognitive Systems Research 77 (2023) 45–61

Contents lists available at ScienceDirect

Cognitive Systems Research

journal homepage: www.elsevier.com/locate/cogsys

Identifying and modelling polysemous senses of spatial prepositions in referring expressions

Adam Richard-Bollans [a,*], Lucía Gómez Álvarez [b], Anthony G. Cohn [a,c,d,e]

[a] School of Computing, University of Leeds, LS2 9JT, United Kingdom
[b] TU Dresden, Germany
[c] Luzhong Institute of Safety, Environmental Protection Engineering and Materials, Qingdao University of Science & Technology, Zibo 255000, China
[d] Department of Computer Science and Technology, Tongji University, Shanghai, China
[e] School of Civil Engineering, Shandong University, Jinan, China

ARTICLE INFO

Action editor: Amir Aly

Keywords:
Semantics
Spatial language
Polysemy
Referring expressions

ABSTRACT

In this paper we analyse the issue of reference using spatial language and examine how the polysemy exhibited by spatial prepositions can be incorporated into semantic models for situated dialogue. After providing a brief overview of polysemy in spatial language and a review of related work, we describe an experimental study we used to collect data on a set of relevant spatial prepositions. We then establish a semantic model in which to integrate polysemy (the Baseline Prototype Model), which we test against a Simple Relation Model and a Perceptron Model. To incorporate polysemy into the baseline model we introduce two methods of identifying polysemes in grounded settings. The first is based on 'ideal meanings' and a modification of the 'principled polysemy' framework and the second is based on 'object-specific features'. In order to compare polysemes and aid typicality judgements we then introduce a notion of 'polyseme hierarchy'. Finally, we test the performance of the polysemy models against the Baseline Prototype Model and Perceptron Model and discuss the improvements shown by the polysemy models.

1. Introduction

Referring Expression Generation and Comprehension (REG/C) situations provide useful scenarios for analysing the semantics of lexical items and how terms are used to achieve communicative success. However, despite the large body of work on creating computational models for REG/C – see (van Deemter, 2016) for an overview – the ability for terms to represent distinct but related meanings remains largely unexplored in the literature, where even homonymy is rarely considered (Siddharthan & Copestake, 2004). Though there does exist work exploring vagueness in REG/C, the vagueness considered is usually with respect to *gradable* properties whose parameters are clearly defined, e.g. height (van Deemter, 2006). In contrast, modelling polysemy requires dealing with distinct senses, which presents a more thorough challenge for semantic representations. There is wide agreement in philosophy of language and linguistics that many natural language terms display some degree of polysemy (Klein & Murphy, 2002; Wasow, Perfors, & Beaver, 2005) and one would expect that semantic models could benefit from accounting for this phenomenon, both in terms of cognitive alignment and performance.

We address this gap in the research by examining how the polysemy exhibited by spatial prepositions can be incorporated into semantic models for situated dialogue. We focus on a set of English spatial prepositions for which there is evidence in the literature that they exhibit polysemy at the kind of room scales we are considering. We consider these to be 'in' (Rodrigues, Santos, Lopes, Bennett, & Oppenheimer, 2020), 'under' (Zlatev, 1992), 'over' (Tyler & Evans, 2001; Zlatev, 1992) and 'on' (Bowerman & Choi, 2001).[1]

All of these 'polysemous' prepositions may also be considered to be 'functional' prepositions in that object affordances and functional relationships, such as *support* and *location control*, appear to be salient (Coventry, Prat-Sala, & Richards, 2001; Garrod, Ferrier, & Campbell, 1999). Moreover, each of these spatial prepositions appears to have what we call a *geometric counterpart* which has a weaker functional influence, these are: 'inside', 'below', 'above' and 'on top of'. We also consider 'against' to be a functional preposition (Talmy, 1988) which is not clearly polysemous and also does not have a clear geometric counterpart (though there are possible candidates e.g. 'next

* Correspondence to: Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AE, UK.
E-mail address: a.richard-bollans@kew.org (A. Richard-Bollans).

[1] Though not explicitly studying polysemy, Bowerman and Choi (2001) provide various examples of object configurations which are labelled simply with 'on' in English but are distinguished with multiple prepositions in other languages.

https://doi.org/10.1016/j.cogsys.2022.09.004
Received 4 April 2022; Received in revised form 6 August 2022; Accepted 23 September 2022
Available online 10 October 2022

to', 'near', 'by' or 'at'). We will refer to the geometric counterparts along with 'against' as 'non-polysemous' prepositions; however, we will explore the applicability of our approaches to these 'non-polysemous' prepositions and discuss the extent to which they are actually non-polysemous. The prepositions analysed in this paper are therefore 'in', 'inside', 'on', 'on top of', 'over', 'above', 'under', 'below' and 'against'.

It is apparent that contextual factors relating to scale (Lautenschütz, Davies, Raubal, Schwering, & Pederson, 2006; Montello, 1993) and domain Klippel, Xu, Li, and Yang (2011) influence the usage of spatial prepositions. In this paper we limit the analysis to a specific context – the usage of spatial prepositions in single rooms containing objects on or around a tabletop – and will discuss how our work may be adapted for differing contexts.

Our paper is organised as follows. After discussing the background and current work in Section 2, we present our first contribution, a framework intended to facilitate the collection of rich data for spatial prepositions (Section 3) and we describe an experimental study we used to collect data on our set of prepositions. There is much discussion regarding how the semantics of these spatial terms are shaped and understood, and though there is general agreement that non-geometric aspects play a significant role in preposition usage, there is a lack of available data for these semantic aspects to be modelled. Our framework is aimed at facilitating the acquisition of data that supports theoretical analysis and helps understand the extent to which different kinds of features play a role in the semantics of spatial prepositions in different contexts.

Our second contribution comes in Section 4, while establishing a suitable semantic model in which to integrate polysemy (the Baseline Prototype Model). This semantic model relies on Prototype Theory, which is interesting in this context as its models are interpretable. However, these types of models rely on calculating a weighted semantic distance to some central instance or instances, and how the central instance(s) or weights should be determined is not often discussed. Our main contribution in this section is a method for automatically generating prototypes and typicality measures of concepts from data. By automating this process, we lay the foundation for an efficient strategy for the modelling of the possibly many polysemes of terms. We introduce a Baseline Prototype Model and test it against a Simple Relation Model and a Perceptron Model, the three of which will later establish the baseline for our polysemy models.

Finally, our third and main contribution is introduced in Section 5, where we incorporate polysemy into the baseline model produced in Section 4. To achieve this, we introduce two methods of identifying polysemes in grounded settings: the first is based on 'ideal meanings' and a modification of the 'principled polysemy' framework, and the second is based on 'object-specific features'. Moreover, in order to compare polysemes and aid typicality judgements we introduce a notion of 'polyseme hierarchy'. We test the performance of the polysemy models against the Baseline Prototype Model and Perceptron Model and we observe that our method for incorporating polysemy into the Baseline Prototype Model provides significant improvement. In Section 6, 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. Finally, we conclude in Section 7 with the results of this work and discuss avenues for future work in Section 8.

2. Background and related work

In this section we provide the background and literature review for this work. First, we discuss the nature of polysemy and how it relates to the spatial prepositions of interest. Given the gap in the research on this topic, we will then discuss the state of the art in semantic models of spatial prepositions, paying specific attention to approaches that are suitable for the integration of polysemy. Finally, we recall that rich spatial data is necessary to characterise the nuances of different polysemes, and thus we conclude the section with a review of the features that are considered relevant for the semantic characterisation of spatial prepositions.

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

2.1. Polysemy and spatial language

The polysemy of spatial prepositions is well recognised in the literature (Herskovits, 1987; Van der Gucht, Willems, & 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 (Muller, Roch, Stadtfeld, & Kiss, 2011; Rodrigues et al., 2020). However, polysemy is rarely, if ever, accounted for in computational models for situated dialogue.

As opposed to homonymy where a term may express semantically distinct senses, a term is considered to exhibit polysemy if it denotes multiple *related* senses (i.e. the *polysemes*). As the senses of polysemous terms are so closely intertwined, the theoretical and computational treatment of polysemy presents a difficult challenge for semantic models.

To provide some examples in the context of spatial language, a figure may be 'on' a ground if it is (Sense 1) resting on top of it e.g. 'a book on a table' (Sense 2) attached to the side of it e.g. 'a clock on a wall' (Sense 3) simply in contact with it e.g. 'a balloon on the ceiling'. These distinctions are particularly fine-grained and are not concerned with the wider usages of spatial prepositions which may provide better examples of polysemy. For example, the phrase 'John is on TV' has little concrete spatial sense as, presumably, we are talking about a projection of an image representing John which is made by the TV. Furthermore, there also appear to be senses which are not so clearly derived from the spatial senses. For example, 'on' may be used to relate an entity with some state e.g. 'To be on alert' (Evans, 2015).

The definition of polysemy is the subject of much debate in cognitive linguistics (Lewandowska-Tomaszczyk, 2007), and the notion of polysemy overlaps with vagueness and ambiguity which may result in a varied theoretical treatment (Gómez Álvarez, 2018). Regardless of whether senses (1)–(3) given above constitute distinct polysemes in any particular theoretical framework, the semantic variability that arises from these senses will be important for semantic models to capture if they are to reliably use and interpret spatial language and it is these kinds of distinctions that are tackled in this paper.

2.2. Models

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

One approach to modelling the semantics of spatial prepositions has been to generate rules which capture their meaning, for example (Abella & Kender, 1993; Platonov & Schubert, 2018). One advantage of rule-based models is the ability to precisely explore and incorporate a particular aspect of spatial language. For example, (Platonov & Schubert, 2018) provide a rule-based computational model of spatial prepositions that encodes various senses of the terms and also aims to account for synecdoche[2] by tagging and iterating over 'salient parts' of

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

objects. As an example, the canonical sense of the preposition 'on' is measured by the extent to which the figure is above and touching the ground and the model then checks if this sense of 'on' applies better to any of the 'interactive parts' of the ground.

These models, however, largely rely on expert intuition to generate rules and, as a result, such approaches often lead to over-simplified representations which are susceptible to the pitfalls associated with the 'simple relations' model of spatial prepositions, as discussed in Herskovits (1987). For example, in Platonov and Schubert (2018) 'in' is simply measured using geometric containment.

Another common approach to modelling spatial prepositions, largely popularised in Logan and Sadler (1996), has been to define (spatial) regions for which the prepositions unambiguously apply and measure deviations from these acceptable regions. These regions provide natural gradations of the applicability of spatial prepositions centred around the 'good' acceptability region and a focus of subsequent work (Gapp, 1995; Kelleher & Costello, 2009; Moratz & Tenbrink, 2006; Regier & Carlson, 2001) has been to quantify the deviation in acceptability.

In Mast, Falomir, and Wolter (2016) this deviation in acceptability of spatial prepositions is modelled using Prototype Theory (Rosch, 1978), where a prototypical point is given in a feature space and the acceptability is measured by the distance from the prototype. This approach is taken to develop a semantic component of a dialogue system to tackle problems involving referring expressions. The use of a prototype in a feature space rather than spatial template means that the semantics are not constrained to simple geometric features. However, as with the majority of work (e.g. Falomir & Kluth, 2017; Gapp, 1995; Hois & Kutz, 2008; Moratz & Tenbrink, 2006; Zampogiannis, Yang, Fermüller, & Aloimonos, 2015) on computational models of spatial prepositions (Mast et al., 2016) focus on modelling projective prepositions[3] (in particular, 'left of', 'right of', 'in front of' and 'behind'). Clearly this is relevant to this paper as we are attempting to model 'over', 'under', 'above' and 'below'. However, often this work only considers a simple geometric representation of the terms and is focused on the pragmatic and/or grammatical complexities that arise. In Section 4 we will extend the approach of Mast et al. (2016) to model our set of prepositions and also provide an empirical method for generating the model parameters from data.

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

For example, Doğan, Kalkan, and Leite (2019) consider the problem of grounding spatial prepositions for human–robot interaction in scenarios where a robot must identify an object on a tabletop given a locative expression. To model the semantics of spatial terms, Doğan et al. (2019) train a 'Relation Presence Network', a multilayer perceptron that takes as inputs the feature values of a configuration and that outputs the probability that each spatial preposition is present in the configuration. Similar work has been carried out for 3D block's world environments in Yan, Wang, and He (2020). It may be the case that with enough training data a perceptron based model creates an internal representation that is closely aligned with a satisfactory cognitive model. However, such models are likely to be highly context sensitive and subject to dataset bias (Daumé III & Marcu, 2006; Pradhan, Ward, & Martin, 2008; Yang, Russakovsky, & Deng, 2019) as well as being uninterpretable by humans. Part of the intention of this paper is to better understand the nature of spatial language; and due to the black-box nature of neural networks this is an unattractive approach.

[3] Projective terms convey information about the direction that an object is located in relative to another e.g. 'the light *above* the desk'.

2.3. Features

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

2.3.1. Geometric features

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

- Contact (Platonov & Schubert, 2018)
- Distance (Abella & Kender, 1993; Alomari, Li, Hogg, & Cohn, 2022; Chang, Savva, & Manning, 2014; Golland, Liang, & Klein, 2010; Gorniak & Roy, 2004; Kelleher & Costello, 2009; Platonov & Schubert, 2018)
- Overlap with projection from objects (Abella & Kender, 1993; Chang et al., 2014; Platonov & Schubert, 2018)
- Height difference (Abella & Kender, 1993; Platonov & Schubert, 2018)
- Object alignment (Abella & Kender, 1993; Alomari, Duckworth, Hawasly, Hogg, & Cohn, 2017; Golland et al., 2010; Gorniak & Roy, 2004; Kelleher & Costello, 2009)
- Containment (Abella & Kender, 1993; Chang et al., 2014; Golland et al., 2010; Platonov & Schubert, 2018)

Various subtle differences may exist between the implementation of these features in different semantic models e.g. the distance between objects may be calculated between object bounds or centres of mass. Similarly, these features may be encoded in a more or less general form; for example, in Mast et al. (2016), object alignment is measured by two separated features: 'centre point angular deviation' and 'bounding box angular deviation'. Also, simplifications are often made for computational reasons, e.g. the bounding boxes of objects are commonly used for calculations.

2.3.2. Functional features

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

- Location control (Garrod & Sanford, 1988)
- Support (Garrod et al., 1999)
- Covering/Protection (Coventry et al., 2001; Tyler & Evans, 2001)

Location control is the ability for one object to constrain the movement of another. Location control arising through some form of enclosure of one object inside another, referred to as 'fcontainment' in Garrod et al. (1999), is seen to be salient for the preposition 'in'. The notion of *support* may be considered as a particular type of location control which is constrained to the vertical direction — X supports Y if X resists the acceleration of Y due to gravity. Support is most often associated with the preposition 'on'.

The prepositions 'over' and 'under' appear in some instances to encode a functional relationship of *covering* or *protection*. This sense of covering does not simply reflect a geometric relationship but is also concerned with properties and affordances of the figure and ground objects in a given context.

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

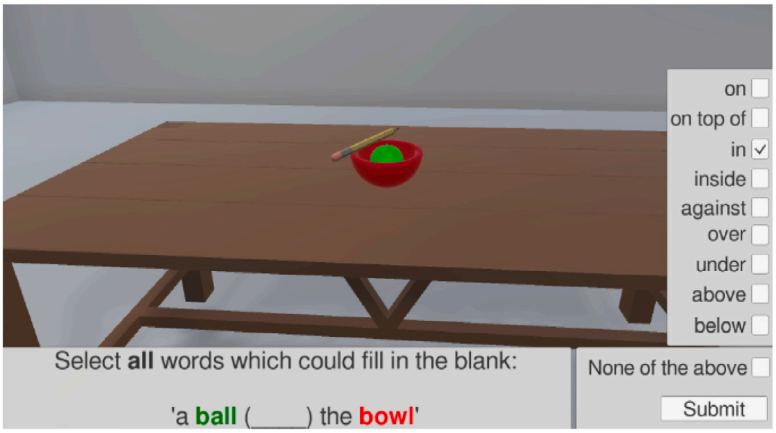

Fig. 1. Preposition Selection Task.

prepositions 'on' and 'in'; however there is very little with regards to how these features should be modelled. Regarding *support*, (Kalita & Badler, 1991) 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 (Gorniak & Roy, 2004; Kelleher & Costello, 2009; Kettani & Moulin, 1999; Regier & Carlson, 2001) which could be considered as something like a proxy for location control, but other than this, the feature does not appear in existing work.

As opposed to functional relationships concerned with the physical interactions between objects, it is also apparent that various object properties and affordances influence the usage of spatial prepositions (Coventry, Carmichael, & Garrod, 1994; Feist & Gentner, 1998). For example, the animacy of the figure object may influence a decision to use 'in' or 'on' (Feist & Gentner, 1998).[4] Following Coventry et al. (1994), we call these kinds of features 'object-specific' features. Again, though recognised as salient, these types of features are not included in semantic models of spatial prepositions for situated dialogue.

To conclude, in general there are various issues which are not covered so far in computational accounts of spatial prepositions. Firstly, functional features such as *support* and *location control* are not represented, though they are often cited as important. Secondly, the sorites vagueness exhibited by spatial prepositions is well recognised and is captured by most models. However, the conceptual vagueness, or *polysemy*, exhibited by spatial prepositions has not been addressed, with the possible exception of the preposition 'on' in Platonov and Schubert (2018). Finally, in general features in the models are *relational* and features specific to the figure or ground are not taken into account.

2.4. Datasets

There are various datasets relating to spatial language in referring expressions; however we find that there is a lack of detailed geometric and functional data which hinders the capacity to properly investigate the semantic complexity of spatial prepositions. As a result, we have created a new data collection environment and experimental study, described in Section 3.

For example, many datasets are based on scenes containing simple geometric objects such as blocks and balls (Gorniak & Roy, 2004; Kelleher & Costello, 2009; Liu, Liu, Bai, & Yuille, 2019; Platonov, Kane, Gindi, & Schubert, 2019; Viethen & Dale, 2011). Such datasets may be easy to generate and allow researchers to test specific pragmatic or semantic issues of spatial language; however, as we would like to explore the influence of object-specific features such datasets are not appropriate. As a result of this lack of data, similar research has relied on constructing small scale datasets using household objects (Golland et al., 2010; Platonov & Schubert, 2018), however these datasets lack the functional features and prepositions we aim to model.

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

3. Data collection

In order to train and test typicality measures of spatial language, we collected data on spatial prepositions using 3D virtual environments. The data collection framework is built on the Unity3D[5] game development software and details of the framework can be found on our GitHub repository for the annotation tool.[6] The data collected from the study has been archived in the Leeds research data repository[7] and details of the data analysis can be found on our GitHub repository.[8]

3.1. Tasks

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

3.1.1. Preposition Selection Task

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

Often concepts are viewed in an antagonistic manner; for example work on Conceptual Spaces is often concerned with comparison of categories, e.g. partitioning a feature space (Douven, Decock, Dietz, & Égré, 2013). 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.

[5] https://unity.com/

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

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

[8] https://github.com/alrichardbollans/semantic-analysis-spatial-prepositions

Fig. 2. Comparative Task.

3.1.2. Comparative Task

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

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

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

3.2. Scenes

For the study, 67 separate scenes were created in the Unity3D editor in order to capture a variety of tabletop configurations. Each scene is a small collection of objects which provide test configurations for each task. The scenes contained 22 different kinds of objects in total and were selected to provide natural and varied indoor scenes. There are convex objects (e.g. pencils) and non-convex objects (e.g. mugs), objects at different scales e.g. books vs tables and objects with various functional associations (e.g. lamps, bowls, chairs).

For the Preposition Selection Task, configurations to test in each scene are predetermined and when a participant is tested on a scene they are tested on each of these configurations. For the Comparative Task, ground objects to test are predetermined and when a participant is tested on a scene they are tested on each ground object with a randomly selected preposition. All salient objects are made to be visible from the initial view of the camera.

3.3. Feature extraction

The use of virtual 3D environments allows for the extraction of a wide range of features that would not be immediately available in real-world or image-based studies. In this section we describe the features extracted from scenes and used in our analysis. Exact details of how each feature is calculated are given in the repository for the annotation tool.[8]

In our analysis we have represented in some form each relational feature discussed in Section 2.3, which we believe accounts for the majority of features given in computational models of spatial prepositions.

3.3.1. 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
- *f_covers_g*: this feature takes the area of the figure and ground in the horizontal plane and measures the proportion of the area of the ground which overlaps with the area of the figure (with some adjustments made with respect to vertical separation)
- *g_covers_f*: As above, with figure and ground reversed

Some simplifications have been made in the calculations of these features. For example, we measured *contact* as the proportion of the vertices of the figure mesh which are under a threshold distance[9] to an approximation of the ground.

3.3.2. Functional features

Building on a preliminary investigation (Richard-Bollans, Gómez Álvarez, Bennett, & Cohn, 2019), we explore the relationship between spatial prepositions and the functional features *support* and *location control* and consider how to extend existing semantic models to account for them.

We take *support* to express that the ground impedes motion of the figure due to gravity, while *location control* expresses that a horizontal movement of the ground causes a movement of the figure. As discussed in Section 2.3.2, useful methods of quantifying these notions in a given scene are not apparent. Rather than attempting to formally define these notions, as in Hedblom, Kutz, Mossakowski, and Neuhaus (2017), Kalita and Badler (1991), we quantified these notions via *simulation* using Unity3D's built-in physics engine.

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

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

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

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

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

h' is calculated as follows:

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

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

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

3.3.3. Object-specific features

The scenes in our study contain a number of household objects with associated object-specific features which appear to be salient for the considered prepositions. The features we consider for each object are the properties of being a container or being a light source, and we extract these features from the relational knowledge base ConceptNet (Speer & Havasi, 2012). In order to capture a notion of mobility of objects, we also consider the ratio of objects sizes as an indication of this. Extracting mobility directly is difficult as it is often a comparative judgement (e.g. a chair is mobile compared to a table, but a table is mobile compared to a wall), and a comparison of sizes appears to be a good proxy for this.

3.3.4. Standardising features

In order for the feature weights calculated in the following sections to be meaningful and comparable, it is necessary to standardise the feature values. As in Raubal (2004), 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:

$$z = \frac{x - \bar{x}}{\sigma} \tag{2}$$

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

Table 1
Annotator agreement for each preposition in both tasks.

Preposition	Preposition Selection Task	Comparative Task
in	0.954	0.806
inside	0.841	0.822
against	0.679	0.608
on	0.735	0.786
on top of	0.648	0.761
under	0.772	0.656
below	0.600	0.693
over	0.644	0.653
above	0.589	0.622
average	0.718	0.717

3.4. Study

The study was conducted online and participants from the university were recruited via internal mailing lists along with recruitment of friends and family.[10] 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 min. Some scenes were removed towards the end of the study to make sure each scene was completed at least three 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. In order to continue, the participant must identify the correct objects in both scenes to show that they have understood the premise, and are able to correctly use the software.

3.5. Annotator agreement

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

A breakdown of annotator agreements is provided in Table 1. Overall the annotator agreements appear to be similar for both tasks — for the Preposition Selection Task the average Cohen's Kappa is 0.718, while in the Comparative Task the average Cohen's Kappa is 0.717. There are some clear differences among prepositions e.g. 'in' and 'inside' have very high agreements compared to other prepositions. This may be an indication that people in general agree more on the semantics of these terms, however this may also be a result of natural variation resulting from manual construction of the scenes for the study. We will discuss this further in Section 6.2.

3.5.1. Preposition co-occurrences

A potential source of disagreement between participants, and a particular challenge for analysing the semantics of spatial prepositions, is that multiple prepositions may often be simultaneously applicable. In order to visualise this we have generated the co-occurrence matrix given in Fig. 3. The given values represent the proportion of configurations labelled with both prepositions in the Preposition Selection Task.

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

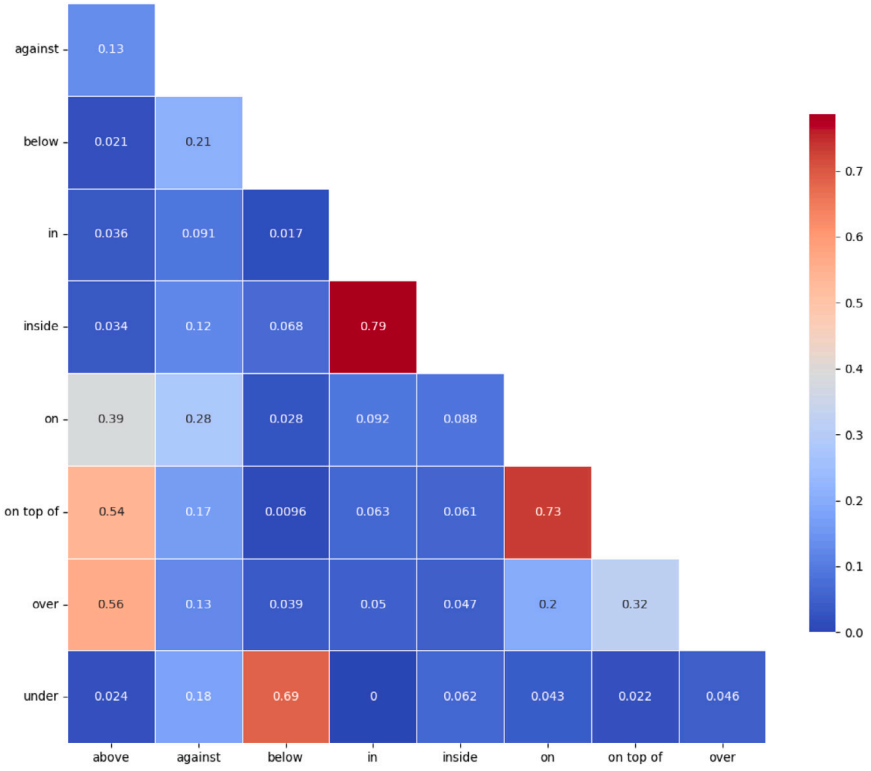

Fig. 3. Preposition co-occurrence matrix.

As expected, the prepositions 'in', 'on', 'over' and 'under' are strongly related to their geometric counterparts; while 'against' is weakly related to all the prepositions. We can also see that 'on' and 'on top of' are moderately related to 'above' and 'over', which may be a result of 'on' and 'on top of' commonly being used to express that a figure object is above the ground object.

3.6. Model evaluation

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

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

Taking the aggregate of participant judgements for a particular preposition-ground pair in a given scene, we can order possible figures in the scene by how often they were chosen. This creates a ranking of configurations within a scene from most to least typical for a given preposition and ground. We turn the collection of obtained rankings into inequalities, or *constraints*, which the models should satisfy. This provides a metric for testing the models.

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

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

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

In this way we generate a set of weighted constraints to be satisfied. For each preposition, the score given to the models is then equal to the sum of weights of satisfied constraints divided by the total weight of the constraints.[11] A higher score then implies better agreement with participants in general.

4. Establishing a baseline

In this section we aim to establish a baseline model which will have two purposes. Firstly, we aim to create a baseline which performs reasonably well in the Comparative Task and can be used to compare with the performance of the polysemy models in Section 5. Secondly, in Section 5 we will attempt to model various distinct senses for each preposition and this will require methods for determining model parameters from the given data. The Baseline Prototype Model in this section will provide such methods and, though initially modelling the semantics of the given prepositions as a single sense, will be extended in Section 5 to account for polysemy. To demonstrate the suitability of this approach, this model will be initially compared to a 'simple relation'

[11] In the case that the model assigns equal typicality scores to both configurations in a given constraint, half the weight of the constraint is added.

model as well as a neural network. The models in this section will use the 'relational' features given in Section 3.3 and object-specific features will be included in Section 5 when we consider polysemy.

4.1. Prototype model

Based on Rosch's Prototype Theory (Rosch, 1978), prototype models assess typicality of an instance by measuring its semantic similarity to a *prototype*. Such a representation seems intuitively plausible for spatial prepositions, particularly if we are to follow the thesis that the meaning of spatial prepositions is structured around some sort of ideal meanings.

Following much of the existing literature, e.g. Nosofsky (1988), semantic similarity between two points x and y in a feature space is measured as a decaying function of the distance, $d(x, y)$:

$$s(x, y) = e^{-d(x,y)} \tag{3}$$

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

$$d(x, y) = \sqrt{w_1(x_1 - y_1)^2 + \cdots + w_n(x_n - y_n)^2} \tag{4}$$

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.

The model is then defined, for each preposition, by a prototype and set of feature weights:

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

where typicality of a configuration, x, is then calculated as the semantic similarity to the prototype using Eq. (3):

$$T(x) = s(x, P) = e^{-d(x,P)} \tag{5}$$

The underlying conceptual model and usage of Prototype Theory is not a new proposal for spatial language and follows (Eyre & Lawry, 2014; Gärdenfors, 2004; Mast et al., 2016; Spranger & Pauw, 2012). Of particular interest is the work of Mast et al. (2016) where a pragmatic model is developed to tackle problems involving referring expressions. Mast et al. (2016) 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 Mast et al. (2016) to model a set of spatial prepositions whose semantics are not so clear and show that model parameters can be automatically determined from a small dataset using a simple regression-based methodology. By automatically generating model parameters we are able to include a wider variety of features in our models and provide support for a novel inclusion of functional features while avoiding human biases in model construction. The automatic generation of parameters will also be useful in Section 5 when we distinguish separate polysemes and attempt to model their semantics. In the following section we describe how model parameters are determined.

4.1.1. Learning prototypes and weights

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

The weights in the semantic distance ought to represent how influential or salient each feature is in making typicality judgements. To determine the salience of each feature the selection ratio is plotted against the feature values. Using off-the-shelf multiple Linear Regression (Pedregosa et al., 2011) we obtain coefficients for each feature which indicate how the selection ratio varies with changes in the feature. The feature weights are then assigned by taking the absolute value of the coefficients given by this linear regression model.

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

In order to find the prototypical value of a given feature for a preposition we plot the feature against the selection ratio, then using simple off-the-shelf Linear Regression modelling (Pedregosa et al., 2011) the feature value is predicted when the selection ratio is 1. Fig. 4 shows the linear regression plot for some features in the case of '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. 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 Fig. 4 that all salient features appear to be assigned appropriate prototypical values.

4.2. Simple Relation Model

It is apparent that some spatial prepositions encode basic general notions such as 'in' expressing containment and 'on' expressing contact or support. Herskovits (1985) refers to the assumption that spatial prepositions simply encode basic relations between objects as the *simple relation* model of spatial prepositions.

The Simple Relation Model we create here is based on what can be found in many computational models of spatial prepositions, with the addition of the functional features *support* and *location_control*. For each preposition we provide basic simple relation models which are based on various rule-based accounts given in the literature (Abella & Kender, 1993; Cooper, 1968; Kalita & Badler, 1991; Platonov & Schubert, 2018; Regier & Carlson, 2001).

For readability in the following model definitions, the models are specified by salient features and prototypical feature values for each preposition — see Table 2. The typicality in the models is then calculated using Eq. (5), where the feature weights are 1 for each of the given salient features and 0 for non-salient features. Note that the given prototypical feature values, e.g. 1 for *containment* when specifying 'in' and 'inside', are values prior to standardisation, but these values are standardised, as discussed in Section 3.3.4, in the actual models.

In general, for the Simple Relation Model the given prototypical feature values are limit values i.e. 0 or 1. However, in the case of 'against' the prototypical value of *location_control* is presumed to be 0.5. This decision was motivated as if one imagines a typical instance of 'against', e.g. a bike leaning against a wall, it does not appear that the ground object fully constrains movement of the figure. We expect that movements of the wall towards and away from the bike significantly impact the position of the bike and that movements of the wall in other directions have a minimal impact. For this reason, the value of 0.5 was used. This particular case highlights the difficulties in general that exist in assigning prototype values.

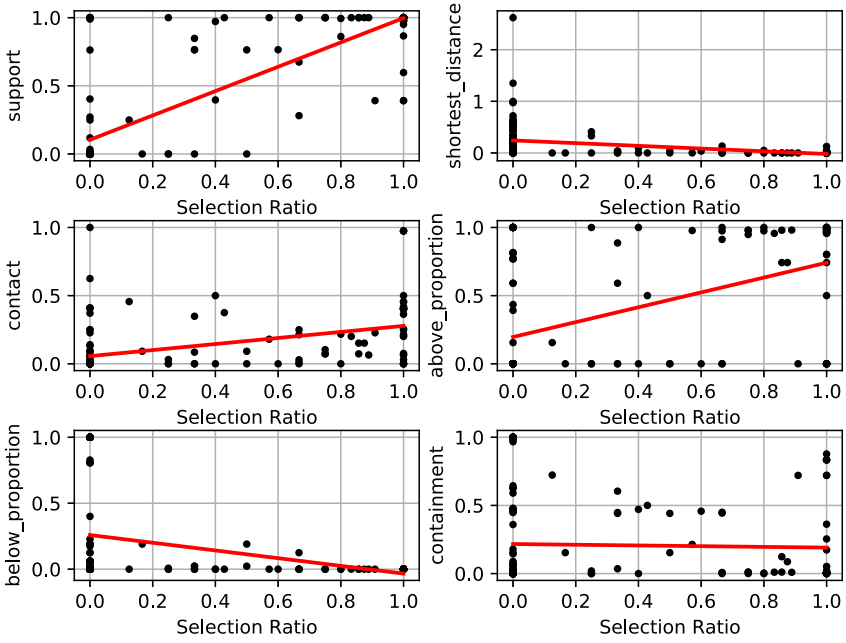

Fig. 4. Finding prototypical feature values for 'on'.

Table 2
Prototype feature values in the simple relation model.

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

4.3. Perceptron model

Neural networks are biologically inspired and so at a lower level may be more cognitively valid than the abstract representations provided above. Though neural networks have proven extremely popular for many AI applications, this approach has not received much attention for modelling the semantics of spatial language in situated contexts (Doğan et al., 2019; Haldekar, Ganesan, & Oates, 2017). This is possibly due to a lack of suitable datasets for training (Barclay & Galton, 2008; Bastianelli et al., 2014; Goyal, Yang, Yang, & Deng, 2020).

However, similar to Doğan et al. (2019), we have implemented simple multilayer perceptrons to recognise the presence of spatial prepositions. A single neural network is generated for each preposition. The network has a single input layer which is given the (relational) feature values of a given configuration and outputs a probability of the given preposition. The network has a single hidden layer of 6 neurons with ReLU activation. These models are trained on data from the Preposition Selection Task to predict the selection ratio for given configurations and plots given in the data repository[8] display the accuracy of the generated models.

4.4. Evaluation

In this section the performance of the models is evaluated using the metric described in Section 3.6. 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=10. We generate the models based on data from the training scenes given in the Preposition Selection Task and test the models using constraints generated from the testing scenes in the Comparative Task.[12] We repeated this process 10 times and averaged the results, shown in Table 3 (the standard deviation is given for the average scores of each fold).

Firstly, in Table 3, we can see that the Baseline Prototype Model and Perceptron Model perform consistently better than the Simple Relation Model. As discussed in Section 2.3, many features can influence the usage of spatial prepositions. 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 Tyler and Evans (2001), 'over' may also indicate *contact* between figure and ground. We believe this partially explains the poor performance of the Simple Relation Model.

On average the Baseline Prototype Model performs better than the Perceptron Model (performing better on six out of the nine prepositions), and has a more consistent score. It appears that the Baseline Prototype Model provides a reasonable baseline and in the following section we will explore how polysemy may be incorporated in order to improve the model.

Considering the polysemy exhibited by spatial prepositions, some features which may not seem to be salient for the preposition in general may be important for determining the typicality for particular polysemes. For example, in some cases 'on' may indicate that the figure is in contact with some region of influence surrounding the ground rather than the ground itself (Miller & Johnson-Laird, 1976), and *shortest_distance* rather than *contact* becomes more salient in this case. For this reason we wanted to explore models which go beyond

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

Table 3
K-Fold test results (K=10, N=10).

	Prototype model	Simple relation model	Perceptron model
in	0.860	**0.924**	0.844
inside	**0.899**	0.888	0.818
against	**0.877**	0.667	0.860
on	0.931	0.933	**0.964**
on top of	**0.976**	0.880	0.954
under	0.771	0.834	**0.923**
below	0.853	0.814	**0.871**
over	**0.780**	0.670	0.718
above	0.843	**0.856**	0.816
Average	**0.866** (SD: 0.061)	0.830 (SD: 0.062)	0.863 (SD: 0.081)

expressing spatial prepositions with one or two hand-picked features. Moreover, by automatically generating weights and prototypes for concepts we provide a method for modelling concepts and senses where the semantics are less clear.

5. Incorporating polysemy

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

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

The main contributions of this section are:

1. a method of identifying polysemes based on 'ideal meanings' (Herskovits, 1987) and a modification of the 'principled polysemy' framework (Tyler & Evans, 2001)
2. a notion of a 'polyseme hierarchy' which allows polysemes to be compared and aids typicality judgements

5.1. Identifying polysemes

The first challenge is to identify the different polysemes that may be expressed by a preposition and this issue is explored in this section. For each preposition the goal is to construct a meaningful set of polysemes where, given a configuration in a scene, there is a method for determining which polysemes the configuration could represent. Once this has been achieved, models can be trained to model each polyseme separately.

5.1.1. 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. Ideal meanings are to be understood as geometric abstractions which represent something similar to a prototypical notion of a concept. For example, the ideal meaning of the preposition 'in' is 'inclusion of a geometric construct within another geometric construct'. In this section we describe how ideal meanings may be defined for each preposition, and how distinct senses may be distinguished using these definitions.

In order to represent each ideal meaning, salient features, threshold values and ordering relations are assigned to each preposition such that a configuration is considered an ideal instance of the preposition if the values are greater than (or less than, depending on the ordering relation) the threshold for each salient feature. The choice of salient features relies on the authors' intuition and supporting literature, and the threshold values will be learnt from training data (described in detail in Section 5.5).

Representations of the ideal meaning of each 'polysemous' preposition are described below. For the prepositions 'in' and 'on' we follow Garrod et al. (1999) and assume that the underlying representations comprise both geometric and functional components.

In. Following Garrod et al. (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 et al. (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 is discussed in Miller and Johnson-Laird (1976) and is seen 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*.

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 & Evans, 2001). We therefore define the ideal meaning of 'over' by a high value of: *above_proportion* and *f_covers_g*.

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*.

Non-polysemous prepositions. Supposing that 'inside', 'on top of', 'above' and 'below' are purely geometric versions of their functional counterparts, their ideal meanings are simplified as follows:

- 'inside' is defined simply by a high value of *containment*
- 'on top of' is defined by a high value of *above_proportion* and *contact*
- 'above' is defined by a high value of *above_proportion* and a low value of *horizontal_distance*
- 'below' is defined by a high value of *below_proportion* and a low value of *horizontal_distance*

We attempt to include 'against' in the following, though it is not clear that such an ideal meaning or a reliable way to model it exists. 'against' is quite a confusing preposition to define which has not been much treated in the literature. It would seem that 'against' usually denotes some degree of proximity and/or contact. For example, in Doore, Beard, and Giudice (2017) the preference for different spatial prepositions is assessed in different contexts and 'against' is preferred to 'next to', 'touching' or 'along' when describing *contact* relations. Also,

it is apparent that 'against' expresses a functional relationship where a force being exerted by the figure is resisted by the ground (Talmy, 1988).

We take the ideal meaning of 'against' to be expressed by a high value of *contact* and *location_control* and a low value of *horizontal_distance*.

5.1.2. Meaning shifts

Following Herskovits' account of the semantic variability of spatial prepositions, once the ideal meanings are understood, the derived uses of a spatial preposition are then achieved via what Herskovits calls 'sense' and 'tolerance' shifts. In tolerance shifts the ideal meaning may be deviated from in a continuous manner — e.g. 'in' may be used to express partial containment rather than full containment. Sense shifts appear in a discontinuous manner where the relations expressed by the ideal meaning are substituted for conceptually similar relations — Herskovits gives the 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. We will call the model that distinguishes senses in this way the *Polysemy Model*.

5.1.3. Object-specific features

The methods given above rely on distinguishing polysemes by considering physical relationships between objects. However, as discussed in Section 2.3.2, it is apparent that various object properties and affordances also influence the usage of spatial prepositions.

One way to include these features would be to simply include these features in the feature space used by the Baseline Prototype Model. However, such a treatment is unlikely to reflect how object-specific features influence semantic decisions, for example the influential feature for 'in' of the ground being a container (*ground_container*) is likely to be assigned a prototype value of 1 but a weight of ~ 0 as cases of 'in' almost always include a ground which is a type of container. We may instead suppose that object-specific features distinguish senses of the prepositions, as in Rodrigues et al. (2020), and treat these distinct senses as polysemes as in the Polysemy Model. In this section we will outline such a model, which we call the OS Feature Model.

Similar to the senses generated by ideal meanings, for the OS Feature Model salient object-specific features distinguish senses for each preposition. Similar to the Polysemy Model, senses will be distinguished by the presence of the given object-specific features. So, for example, an instance of 'in' where the ground is a container is considered to be distinct from an instance where the ground is not a container. Salient object-specific features for each preposition are given below.

In. As 'in' expresses a notion of *containment*, the ability of the ground to contain the figure is often salient whether or not the ground does in fact contain the figure in a geometric sense. Therefore, whether the ground is a type of container appears to be salient for 'in' (Coventry et al., 1994; Feist & Gentner, 1998) and this may be considered a salient object-specific feature.

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

As discussed above, the preposition 'in' is often preferred when the ground object is a container; 'on' is therefore used less frequently in these scenarios (Feist & Gentner, 1998), even though the physical relationships between the objects often fulfil the requirements for 'on'. As a result, whether or not the ground is a container appears to be a salient object-specific feature for 'on'.

Finally, 'on' may be used to denote attachment of the figure to the ground. It is therefore plausible that, similarly to 'against', 'on' is more applicable in situations where the ground object is fixed relative to the figure. Extracting mobility directly is difficult as it is often a comparative judgement (e.g. a chair is mobile compared to a table, but a table is mobile compared to a wall), and a comparison of object sizes appears to be a good proxy for this i.e. we say the figure is mobile compared to the ground if the figure is smaller than the ground (*size_ratio* < 1). Clearly this is a crude method as mobility does not depend on size e.g. a small object may be fixed in place. However, for the objects in the scenes from our study this provides a reasonable measure of mobility.

Over/under. There appears to be a 'covering' sense of 'over' (Tyler & Evans, 2001) (and similarly of 'under') which is closely related to the functions of the figure and ground (Mori, 2019). For example, a covering object like a lid may exhibit this sense of 'over' when covering a container. There is also a non-covering sense where a specific functional interaction exists between part of the figure and ground. For example, a tap may be 'over' a sink if only the spout of the tap is above the sink. Similarly, an object may be 'under' a lamp when the object is not under the lamp in a geometric sense but the light from the lamp shines on the object.

We expect 'over' and 'under' to have similar salient object-specific features, but where the roles of figure and ground are reversed. For

'over' we consider whether the ground is a type of container,[13] and for 'under' we consider whether the figure is a type of container and whether the ground is a light source.

Non-polysemous prepositions. We expect the geometric counterparts of each of the above prepositions to be less influenced by object-specific features as they are generally less influenced by functionality, for example (Coventry et al., 2001) evidence that functional interactions have a stronger effect on 'over' and 'under' than on 'above' and 'below'. However, if they do exhibit polysemy based on object-specific features, we would expect them to share the same features with their functional counterparts and this is how they are encoded in the model.

'Against' is commonly used to denote contact between two objects and, as argued in Herskovits (1987), is more applicable in situations where the ground object is fixed and the figure is mobile. For example, one may describe a chair as being 'against a wall' but it would be odd to describe a wall as being 'against a chair'. We therefore consider mobility of the figure compared to the ground as a salient object-specific feature (in the same way as 'on').

5.2. Determining typicality

Now that we have outlined two methods for distinguishing polysemes, how do we translate these into semantic models? Firstly, we construct models for each polyseme such that, given a particular configuration, we can assign a value representing how typical the configuration is for the polyseme.

In order to construct such models we treat each polyseme as if it were a distinct term and employ the same method, underlying model and feature space used in the Baseline Prototype Model. To train each polyseme separately and ensure that the polyseme is only trained on polyseme instances, the training datasets are modified. This is achieved simply by removing potential preposition instances that are not examples of the given polyseme i.e. configurations which have been labelled with the preposition but which do not fit the polyseme's conditions. For example, for the ideal sense of 'on' in the Polysemy Model 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 Eq. (6), the typicality, $typicality_p(c)$, assigned by a polyseme, p, to a configuration, c, is specified by these prototypes and weights.

5.3. Polyseme hierarchy

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

In some cases, given a preposition and ground, only one polyseme of the preposition may be applicable to all potential figure–ground pairs in the scene. In this case we can just compare the typicality for each figure–ground pair, with respect to that polyseme, and the most typical is the one selected.

However, in many cases there will be multiple possible figures each potentially fitting a different polyseme. For example, there may be a scene with a book on a table — Sense 1 from Section 2.1 — 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, Srikumar, Hwang, and Palmer (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 Eq. (6). P is the set of polysemes of the preposition which may apply to c, $typicality_p(c)$ is the typicality of c with respect to a polyseme p and r_p is the rank of polyseme p.

$$typicality(c) = \max_{p \in P}(typicality_p(c) \times r_p) \tag{6}$$

By adjusting the typicality assigned by polysemes by their rank, configurations fitting weaker senses, e.g. Sense 3, should only be selected if there are no good examples present of stronger senses, e.g. Sense 1.

5.4. Specification

The polysemy models described in this section are defined for each preposition as a set of polysemes where each polyseme is in turn defined by:

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

and the overall typicality of a configuration for a given preposition is given by Eq. (6).

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 a prototype and weights equal to those assigned by the Baseline Prototype Model for the associated preposition. The rank for the polyseme, instead of being 0 is then taken as the average value of the selection ratio for all training configurations.

We can see that overall the resulting models are collections of prototypes with associated weights, organised around a central ideal meaning. This has some similarity with the radial category approach (Brugman & Lakoff, 1988) in so far as each sense is linked to a central sense, though the radial category approach is aimed at distinguishing less fine-grained distinctions than we consider here where senses do not share the same underlying representations and are created through schematic transformations.

5.5. Learning threshold values

We have now provided a method for distinguishing senses for each preposition. However, we have not yet outlined how threshold values should be assigned.

[13] We would also consider whether the figure is a light source, though there are no such training instances in our dataset.

Table 4
K-Fold test results (K=10, N=10).

	Prototype model	Perceptron model	Polysemy model	OS Feature model
in	0.829	0.847	0.784	**0.916**
inside	0.920	0.863	0.899	**0.925**
against	0.882	**0.902**	0.846	0.886
on	0.929	**0.946**	0.932	0.934
on top of	0.974	0.968	**0.990**	0.972
under	0.758	**0.922**	0.915	0.834
below	0.876	0.895	**0.918**	0.833
over	0.816	0.774	**0.863**	0.821
above	0.864	0.850	**0.881**	0.858
Average	0.872 (SD: 0.06)	0.885 (SD: 0.058)	**0.892** (SD: 0.048)	0.887 (SD: 0.062)

As it is not clear how to intuitively assign threshold values to the ideal meanings, we have implemented a simple algorithm which refines these parameters based on performance in the Comparative Task on the training scenes. In order to achieve this, the model is trained and tested on the given training scenes while varying the threshold values for each salient feature. The model is updated with the values that produce the best performance and then retrained on all the original training scenes.[14] This is obviously a very simple way to achieve this refinement, and could be expanded on, but displays the potential of the model and appears to be effective.

5.6. Evaluation

Again, to test the models, we created train-test scenes using K-fold cross-validation with K=10. We generate the models based on data from the training scenes given in the Preposition Selection Task and test the models using constraints generated from the testing scenes in the Comparative Task. We repeated this process 10 times and averaged the results, shown in Table 4 (the standard deviation is given for the average scores of each fold).

Both the Polysemy Model and OS Feature Model have improved on the Baseline Prototype Model and Perceptron Model, with the Polysemy Model performing best and with lower standard deviation than the other models.

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

5.6.1. Is the model capturing polysemy?

It is plausible that the methods used for the Polysemy Model are just making training more effective by partitioning the data rather than actually capturing polysemy. We do not believe this to be the case, and will provide some evidence for this view here.

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

Table 5
K-Fold test results (K=10, N=10).

	Prototype model	Partition model
in	**0.844**	0.786
inside	**0.929**	0.836
against	**0.871**	0.839
on	**0.931**	0.921
on top of	**0.981**	0.971
under	0.757	**0.835**
below	**0.849**	0.829
over	**0.783**	0.729
above	**0.840**	0.772
Average	**0.865** (SD: 0.064)	0.835 (SD: 0.062)

As we can see from Table 5, the Partition Model performs worse than the Baseline Prototype Model and the Partition Model only performs better than the Baseline Prototype Model for the preposition 'under'.

These results suggest that the improvement shown by the Polysemy Model over the Baseline Prototype Model do not simply result from partitioning the data. This indicates that the Polysemy Model is genuinely capturing the polysemy exhibited by these terms and, moreover, that it is important to appropriately define the ideal meanings used in the Polysemy Model.

Ranks and ideal meanings. For the Polysemy Model, 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; and similarly, for the OS Feature Model where the ideal meaning is specified with object-specific features associated with the preposition. 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.

This is exactly what we observe for both the Polysemy Model and OS Feature Model for the 'polysemous' prepositions with one small

[14] For salient features with the $\leq$ relation the tested threshold values are $0.05, 0.1, 0.15, \ldots, 0.55$ and for features with the $\geq$ relation the tested threshold values are $0.45, 0.5, 0.55, \ldots, 0.95$. With the exception of *contact* where $0.05, 0.1, 0.15, 0.2, \ldots, 0.55$ are used in both as high values are very rare and *horizontal_distance* where $0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3$ are used in both.

Fig. 5. 'Inside the cup'.

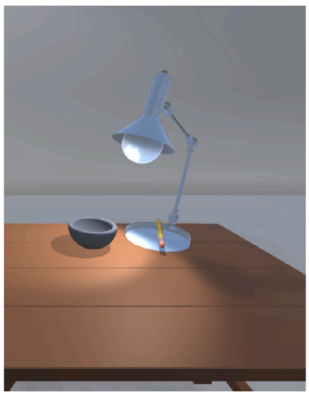

Fig. 6. 'On top of the lamp'.

exception,[15] and this also holds in general for the 'non-polysemous' prepositions. This result suggests that for the Polysemy Model we have appropriately assigned ideal meanings to the prepositions and that the semantics of the terms are indeed centred around such ideal meanings. Regarding the OS Feature Model, this result provides further evidence that the given object-specific features are salient.

6. Discussion

6.1. Are these prepositions non-polysemous?

As we have seen, the polysemy models perform surprisingly well on the 'non-polysemous' terms. Do these non-polysemous terms actually exhibit polysemy? In order to explore this in more detail we will consider some examples of each of these prepositions.

Inside. Fig. 5 shows some configurations which appear in the study scenes. In the Preposition Selection Task when labelling the (pear, cup) configuration, all tested participants gave 'inside'. In the Comparative Task, when selecting the object referred to by 'the object inside the cup', participants selected the pear — the Polysemy Model agrees with participants here but the Baseline Prototype Model does not. This appears to be similar to the often cited instances of objects being 'in' other objects when there is little or no containment and is usually explained by the presence of location control. Following the previously discussed Criterion 1 for distinguishing polysemes, this appears to be a non-ideal sense of 'inside' and provides some support that 'inside' is in fact polysemous.

On top of. Again, Fig. 6 shows some configurations from the study. In the Preposition Selection Task, half of the tested participants labelled the (pencil, lamp) configuration with 'on top of', and participants would select the pencil when given the description 'the object on top of the lamp'. Both the Polysemy Model, OS Feature Model and Baseline Prototype Model pick the pencil in this case, as the other possible objects are not very plausible instances. The Polysemy Model and OS Feature Model, however, give more marked distinctions between (pencil, lamp) and other configurations in the scene where the lamp is the ground.

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

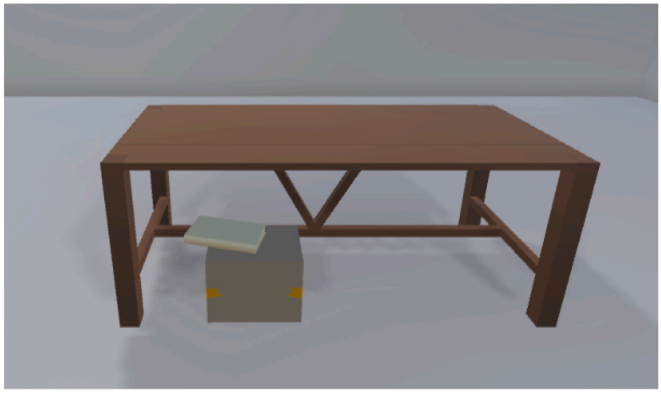

Fig. 7. 'Above the box'.

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

One approach to modelling these phenomena would be to iterate over sections or 'salient parts' of objects, for example checking whether the pencil is on top of the lamp as a whole, or some important section of the lamp e.g. its base. This is the approach taken for 'on' in Platonov and Schubert (2018). Automating such a process would require an ability to automatically demarcate and label salient parts of objects and this is a significant research problem. The method proposed in this section instead deals with these synecdochical instances in a simpler way by modelling a distinct sense of 'on top of' where *above_proportion* is low, and this potentially explains the good performance of the Polysemy Model for 'on top of'.

Above. For the configuration (table, box) in Fig. 7, all tested participants selected 'above', and there are many similar examples of this. This may seem uncontroversial, however a large proportion of the table is not actually above the box and the value of *above_proportion* is 0.77. Similarly to the example of 'on top of' discussed above, this instance may be explained by synecdoche — 'table' may be conceptualised as the horizontal part of the table. However, it is interesting to note the existence of seemingly unambiguous instances of 'above' where *above_proportion* is not 1, and following Criterion 1, we may suppose that this is a distinct sense of 'above' which is similar to the 'covering' sense of 'over'.

[15] For 'under' in the OS Feature Model the rank assigned to the sense defined by a low value of *figure_container* and low value of *ground_lightsource* has a higher rank than the sense with a low value of *figure_container* and high value of *ground_lightsource*.

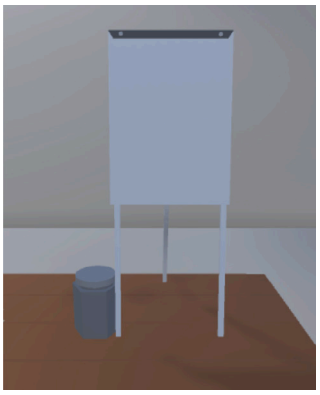

Fig. 8. 'Below the board'.

Fig. 9. 'Against the table' (1).

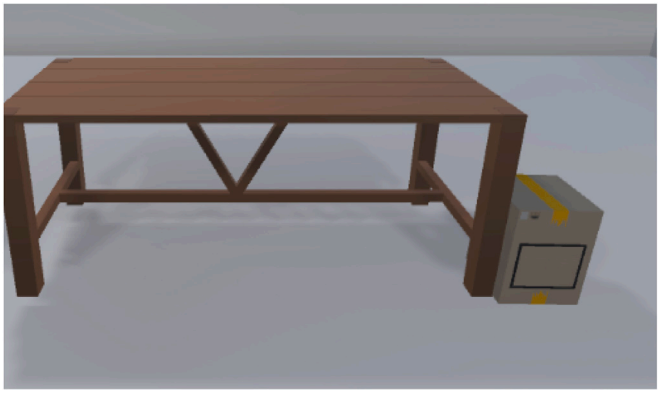

Fig. 10. 'Against the table' (2).

Below. For the (jar, board) configuration shown in Fig. 8, four out of five tested participants selected 'below' in the Preposition Selection Task. This is similar to the example above given for 'above' (the value of *below_proportion* is 0.19), however it is even more striking as the board does not cover the jar (the value of *g_covers_f* is 0.15).

Against. In both Figs. 9 and 10 the configuration (box, table) was labelled with 'against' by every tested participant. These appear to be distinct senses of 'against', in one the box is leaning against the table and in the other box is simply next to it. This distinction can be drawn with either of the functional features — there is a higher degree of *support* and *location_control* in (1) than in (2).

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

6.2. Limitations

6.2.1. Scene construction

Throughout this paper we see some variability among prepositions. In Table 1 we see that people appear to agree on the semantics of some terms more than others and when evaluating the models some prepositions appear to be harder to model e.g. the models consistently perform poorly for 'over'. Ideally we may conclude from this that 'over' is a difficult preposition to model with complex semantics compared to the other prepositions. However, such a conclusion must also recognise the influence of the particular scenes used for the studies — it may simply be that instances of 'over' were relatively ambiguous in our scenes. In order to reliably test such an assumption it would be ideal to compare annotator agreements and model performance on much larger procedurally generated scenes, and this is an avenue of further work.

6.2.2. Choice of features

The features we have included in our models have relied on supporting literature, either having been previously included in semantic models or discussed as being salient. However, due to the complex nature of the semantics of spatial prepositions, it is nevertheless possible that some salient features have not been included. Moreover, though there is much overlap in the semantics of these terms, it is likely that the model performance could be improved by using a specific set of features for each preposition. In future work it would be informative to generate a much larger feature set which is reduced for each preposition in a preprocessing step.

Similarly, the choice of salient features when defining ideal meanings has been reliant on existing literature and the authors' intuition. Though these definitions have been reasonably justified, it would be ideal to be able to generate these kinds of definitions from given data and this is another avenue of future work.

6.2.3. Context

In this paper we have restricted the study to a specific context, however the semantics of the terms will be modified when the context changes. One approach to account for changes in context, for example when training a neural network, would be to include features in the model which provide contextual information. Another approach, which our model may be incorporated into, would be to define a set of distinct contexts inside which the semantics of the terms are relatively consistent (similar to the idea discussed in McCarthy, Buvac, et al., 1997). Separate models could then be trained and utilised for each specific context, though this would require a method for automatically recognising a given context.

6.3. Categorisation

The models we have discussed in this paper have been aimed at modelling the semantics of spatial prepositions in order to select objects when given a spatial expression. However, it is unclear whether these models could be used to generate appropriate prepositions when given an object to describe. For example, when generating such utterances, *selectional restrictions* may apply which are less salient when interpreting these utterances e.g. Kalita and Badler (1991) suggest that the figure should be smaller than the ground for 'in' to apply.

7. Conclusion

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 & Evans, 2001) for distinguishing polysemes, an approach has been developed which can be exploited by semantic models more generally.

We have also introduced a notion of a 'polyseme hierarchy' – a value which corresponds to how strongly a particular polyseme is associated with the given preposition – as well as methods for determining its value. In combining this with the generated polysemes, we have provided a semantic model which significantly improves on the given baseline when interpreting a particular class of referring expressions. As well as the Polysemy Model based on ideal meanings, we have created the OS Feature Model which distinguishes senses based on a novel inclusion object-specific features.

The terms motivating this work have been those prepositions which according to existing literature appear to be polysemous, however the methods outlined in this section also appear to be applicable to some 'non-polysemous' prepositions. Moreover, we have provided some evidence that these 'non-polysemous' prepositions may be considered polysemous.

8. Future work

It is clear that object-specific features must be somehow accounted for in semantic models of spatial prepositions and that one way to achieve this is to use object-specific features to distinguish senses of the prepositions. We have included some object-specific features in our models, however there are many more features which may be salient for the given prepositions and we feel that further investigations must be carried out in order to identify a comprehensive set of salient object-specific features for each preposition.

Various restricted studies have been conducted providing evidence that certain features influence certain prepositions, e.g. Coventry et al. (1994), Feist and Gentner (2003), however a comprehensive study exploring this would be ideal. Such a study would face various challenges, e.g. the salience of particular object-specific features may change with changing contexts and the source of potentially salient features may be very large, it may nevertheless be possible to isolate sets of particularly salient features in restricted contexts. Ontologies providing important object-specific features such as AfNet (Varadarajan & Vincze, 2012) may be helpful in this regard by highlighting object properties which are salient in many contexts.

Once salient object-specific features have been identified, any implementation must be able to extract these features from the scene. The approach we have taken in this paper is to leverage information from the knowledge base ConceptNet (Speer & Havasi, 2012). Another approach is to use affordance detection systems, e.g. Do, Nguyen, and Reid (2018), which use information from the scene to predict object affordances. Recent semantic representations, such as VoxML (Pustejovsky & Krishnaswamy, 2016), which allow for the specification of object affordances may lead to better systems for extracting this type of information from scenes in future.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

We would like to thank everyone who participated in our study. Funding: The first author has been supported by an EPSRC Studentship No. 1961029. The third author has been partially supported by the Alan Turing Institute, UK as a Turing Fellow, and by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825619 (AI4EU).

References

Abella, A., & Kender, J. R. (1993). Qualitatively describing objects using spatial prepositions. In *IEEE workshop on qualitative vision* (pp. 33–38). IEEE.

Alomari, M., Duckworth, P., Hawasly, M., Hogg, D. C., & Cohn, A. G. (2017). Natural language grounding and grammar induction for robotic manipulation commands. In *First workshop on language grounding for robotics* (pp. 35–43).

Alomari, M., Li, F., Hogg, D. C., & Cohn, A. G. (2022). Online perceptual learning and natural language acquisition for autonomous robots. *Artificial Intelligence*, *303*, Article 103637. http://dx.doi.org/10.1016/j.artint.2021.103637.

Barclay, M., & Galton, A. (2008). A scene corpus for training and testing spatial communication systems. In *AISB 2008 convention communication, interaction and social intelligence, vol. 10* (pp. 26–29).

Bastianelli, E., Castellucci, G., Croce, D., Iocchi, L., Basili, R., & Nardi, D. (2014). HuRIC: A human robot interaction corpus. In *LREC* (pp. 4519–4526).

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

Brugman, C., & Lakoff, G. (1988). Cognitive topology and lexical networks. In *Lexical ambiguity resolution* (pp. 477–508). Elsevier, http://dx.doi.org/10.1016/B978-0-08-051013-2.50022-7.

Chang, A., Savva, M., & Manning, C. D. (2014). Learning spatial knowledge for text to 3D scene generation. In *Proc EMNLP* (pp. 2028–2038). Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/D14-1217.

Cooper, G. S. (1968). *A semantic analysis of english locative prepositions*: *Technical report*, Fort Belvoir, VA: Defense Technical Information Center, http://dx.doi.org/10.21236/AD0666444.

Coventry, K. R., Carmichael, R., & Garrod, S. C. (1994). Spatial prepositions, object-specific function, and task requirements. *Journal of Semantics*, *11*(4), 289–309. http://dx.doi.org/10.1093/jos/11.4.289.

Coventry, K. R., Prat-Sala, M., & Richards, L. (2001). The interplay between geometry and function in the comprehension of over, under, above, and below. *Journal of Memory and Language*, *44*(3), 376–398. http://dx.doi.org/10.1006/jmla.2000.2742.

Daumé III, H., & Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, *26*, 101–126.

Do, T.-T., Nguyen, A., & Reid, I. (2018). AffordanceNet: An end-to-end deep learning approach for object affordance detection. In *Proceedings of 2018 IEEE international conference on robotics and automation* (pp. 5882–5889). IEEE.

Doore, S., Beard, K., & Giudice, N. (2017). Spatial prepositions in natural-language descriptions of indoor scenes. In *Proceedings of workshops at COSIT* (pp. 255–260). Springer, http://dx.doi.org/10.1007/978-3-319-63946-8_41.

Doğan, F. I., Kalkan, S., & Leite, I. (2019). Learning to generate unambiguous spatial referring expressions for real-world environments. In *2019 IEEE/RSJ international conference on intelligent robots and systems* (pp. 4992–4999). IEEE.

Douven, I., Decock, L., Dietz, R., & Égré, P. (2013). Vagueness: A conceptual spaces approach. *Journal of Physiology (Cambridge, Eng)*, *42*(1), 137–160. http://dx.doi.org/10.1007/s10992-011-9216-0.

Evans, V. (2015). A unified account of polysemy within LCCM theory. *Lingua*, *157*, 100–123.

Eyre, H., & Lawry, J. (2014). Language games with vague categories and negations. *Adaptive Behavior*, *22*(5), 289–303. http://dx.doi.org/10.1177/1059712314547318.

Falomir, Z., & Kluth, T. (2017). Qualitative spatial logic descriptors from 3D indoor scenes to generate explanations in natural language. *Cognitive Processing*, 1–20.

Feist, M. I., & Gentner, D. (1998). 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).

Feist, M. I., & Gentner, D. (2003). Factors involved in the use of in and on. In *Proc annual meeting of the cognitive science society* (p. 7).

Gapp, K.-P. (1995). An empirically validated model for computing spatial relations. In *KI-95: Advances in artificial intelligence, vol. 981* (pp. 245–256). Berlin, Heidelberg: Springer Berlin Heidelberg, http://dx.doi.org/10.1007/3-540-60343-3_41.

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

Garrod, S., Ferrier, G., & Campbell, S. (1999). In and on: Investigating the functional geometry of spatial prepositions. *Cognition*, *72*(2), 167–189. http://dx.doi.org/10.1016/s0010-0277(99)00038-4.

Garrod, S. C., & Sanford, A. J. (1988). Discourse models as interfaces between language and the spatial world. *Journal of Semantics*, *6*(1), 147–160. http://dx.doi.org/10.1093/jos/6.1.147.

Golland, D., Liang, P., & Klein, D. (2010). A game-theoretic approach to generating spatial descriptions. In *Proc EMNLP* (pp. 410–419).

Gómez Álvarez, L. (2018). Ambiguity: What is it that needs representing and what needs resolving? In *Ambiguity: Perspectives on representation and resolution*.

Gorniak, P., & Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, *21*, 429–470.

Goyal, A., Yang, K., Yang, D., & Deng, J. (2020). Rel3D: A minimally contrastive benchmark for grounding spatial relations in 3D. *Advances in Neural Information Processing Systems*, *33*.

Haldekar, M., Ganesan, A., & Oates, T. (2017). Identifying spatial relations in images using convolutional neural networks. In *2017 international joint conference on neural networks* (pp. 3593–3600). IEEE.

Hedblom, M. M., Kutz, O., Mossakowski, T., & Neuhaus, F. (2017). Between contact and support: Introducing a logic for image schemas and directed movement. In *Proc IAAI, vol. 10640* (pp. 256–268). Springer.

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

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

Hois, J., & Kutz, O. (2008). Natural language meets spatial calculi. In C. Freksa, N. S. Newcombe, P. Gärdenfors, & S. Wölfl (Eds.), *Spatial cognition VI. Learning, reasoning, and talking about space, vol. 5248* (pp. 266–282). Springer Berlin Heidelberg.

Kalita, J. K., & Badler, N. I. (1991). Interpreting prepositions physically. In *Proc AAAI* (pp. 105–110).

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

Kettani, D., & Moulin, B. (1999). A spatial model based on the notions of spatial conceptual map and of object's influence areas. In *Proc COSIT* (pp. 401–416). Springer.

Klein, D. E., & Murphy, G. L. (2002). Paper has been my ruin: Conceptual relations of polysemous senses. *Journal of Memory and Language*, *47*(4), 548–570.

Klippel, A., Xu, S., Li, R., & Yang, J. (2011). Spatial event language across domains. In *Workshop on computational models for spatial language interpretation and generation* (pp. 40–47).

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

Lewandowska-Tomaszczyk, B. (2007). Polysemy, prototypes, and radial categories. In *The oxford handbook of cognitive linguistics* (pp. 139–169). Oxford University Press, http://dx.doi.org/10.1093/oxfordhb/9780199738632.013.0006.

Liu, R., Liu, C., Bai, Y., & Yuille, A. L. (2019). Clevr-Ref+: diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4185–4194).

Logan, G. D., & Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. A. Peterson, L. Nadel, & M. F. Garrett (Eds.), *Language, speech, and communication*, *Language and space* (pp. 493–529). MIT Press.

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

Mast, V., Falomir, Z., & Wolter, D. (2016). Probabilistic reference and grounding with PRAGR for dialogues with robots. *Journal of Experimental & Theoretical Artificial Intelligence*, *28*(5), 889–911. http://dx.doi.org/10.1080/0952813X.2016.1154611.

McCarthy, J., Buvac, S., et al. (1997). Formalizing context. *Computing Natural Language, Stanford University*, 13–50.

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

Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and perception*. Belknap Press.

Montello, D. R. (1993). Scale and multiple psychologies of space. In *European conference on spatial information theory* (pp. 312–321). Springer.

Moratz, R., & Tenbrink, T. (2006). Spatial reference in linguistic human-robot interaction: Iterativ, empirically supported development of a model of projective relations. *Spatial Cognition and Computation*, *6*(1), 63–107.

Mori, S. (2019). A cognitive analysis of the preposition over: Image-schema transformations and metaphorical extensions. *Canadian Journal of Linguistics*, *64*(3), 444–474. http://dx.doi.org/10.1017/cnj.2018.43.

Muller, A., Roch, C., Stadtfeld, T., & Kiss, T. (2011). Annotating spatial interpretations of german prepositions. In *Fifth international conference on semantic computing* (pp. 459–466). IEEE, http://dx.doi.org/10.1109/ICSC.2011.46.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Platonov, G., Kane, B., Gindi, A., & Schubert, L. K. (2019). A spoken dialogue system for spatial question answering in a physical blocks world. arXiv:1911.02524 [Cs], URL: http://arxiv.org/abs/1911.02524.

Platonov, G., & Schubert, L. (2018). Computational models for spatial prepositions. In *Proc 1st international workshop on spatial language understanding* (pp. 21–30). http://dx.doi.org/10.18653/v1/w18-1403.

Pradhan, S. S., Ward, W., & Martin, J. H. (2008). Towards robust semantic role labeling. *Computational Linguistics*, *34*(2), 289–310.

Pustejovsky, J., & Krishnaswamy, N. (2016). VoxML: A visualization modeling language. In *Proceedings of the tenth international conference on language resources and evaluation* (pp. 4606–4613).

Raubal, M. (2004). Formalizing conceptual spaces. In *Formal ontology in information systems vol. 114* (pp. 153–164).

Regier, T., & Carlson, L. A. (2001). Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General*, *130*(2).

Richard-Bollans, A., Gómez Álvarez, L., Bennett, B., & Cohn, A. G. (2019). Investigating the dimensions of spatial language. In *Proceedings of speaking of location 2019: communicating about space, vol. 2455* (pp. 47–56). CEUR Workshop Proceedings.

Rodrigues, E. J., Santos, P. E., Lopes, M., Bennett, B., & Oppenheimer, P. E. (2020). Standpoint semantics for polysemy in spatial prepositions. *Journal of Logic and Computation*, http://dx.doi.org/10.1093/logcom/exz034.

Rosch, E. (1978). Principles of categorization. In E. Rosch, & B. B. Lloyd (Eds.), *Cognition and categorization, vol. 1* (pp. 27–78). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schneider, N., Srikumar, V., Hwang, J. D., & Palmer, M. (2015). A hierarchy with, of, and for preposition supersenses. In *Proceedings of the 9th linguistic annotation workshop* (pp. 112–123). Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/W15-1612.

Siddharthan, A., & Copestake, A. (2004). Generating referring expressions in open domains. In *Association for computational linguistics* (pp. 407–414). http://dx.doi.org/10.3115/1218955.1219007.

Speer, R., & Havasi, C. (2012). Representing general relational knowledge in conceptnet 5. In *LREC* (pp. 3679–3686).

Spranger, M., & Pauw, S. (2012). Dealing with perceptual deviation: vague semantics for spatial language and quantification. In *Language grounding in robots* (pp. 173–192). Boston, MA: Springer US, http://dx.doi.org/10.1007/978-1-4614-3064-3_9.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, *12*(1), 49–100. http://dx.doi.org/10.1207/s15516709cog1201_2.

Tyler, A., & Evans, V. (2001). Reconsidering prepositional polysemy networks: The case of over. *Language*, *77*(4), 724–765. http://dx.doi.org/10.1353/lan.2001.0250.

van Deemter, K. (2006). Generating referring expressions that involve gradable properties. *Computational Linguistics*, *32*(2), 195–222. http://dx.doi.org/10.1162/coli.2006.32.2.195.

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

Van der Gucht, F., Willems, K., & De Cuypere, L. (2007). The iconicity of embodied meaning. Polysemy of spatial prepositions in the cognitive framework. *Language Sciences*, *29*(6), 733–754. http://dx.doi.org/10.1016/j.langsci.2006.12.027.

Varadarajan, K. M., & Vincze, M. (2012). Afnet: The affordance network. In *Asian conference on computer vision* (pp. 512–523). Springer, http://dx.doi.org/10.1007/978-3-642-37331-2_39.

Viethen, J., & Dale, R. (2011). GRE3D7: a corpus of distinguishing descriptions for objects in visual scenes. In *UCNLG+Eval: Language generation and evaluation workshop* (pp. 12–22). Association for Computational Linguistics.

Wasow, T., Perfors, A., & Beaver, D. (2005). The puzzle of ambiguity. In *Morphology and the web of grammar: essays in memory of Steven G. Lapointe* (pp. 265–282).

Yan, F., Wang, D., & He, H. (2020). Robotic understanding of spatial relationships using neural-logic learning. In *International conference on intelligent robots and systems* (pp. 8358–8365).

Yang, K., Russakovsky, O., & Deng, J. (2019). SpatialSense: an adversarially crowdsourced benchmark for spatial relation recognition. In *2019 IEEE/CVF international conference on computer vision* (pp. 2051–2060). Seoul, Korea (South): IEEE, http://dx.doi.org/10.1109/ICCV.2019.00214.

Zampogiannis, K., Yang, Y., Fermüller, C., & Aloimonos, Y. (2015). Learning the spatial semantics of manipulation actions through preposition grounding. In *2015 IEEE international conference on robotics and automation* (pp. 1389–1396). IEEE.

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