# Interfacing Language, Spatial Perception and Cognition in Type Theory with Records

*Simon Dobnik and Robin Cooper*
Department of Philosophy, Linguistics & Theory of Science (FLOV)
Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
http://www.clasp.gu.se

## ABSTRACT

We argue that computational modelling of perception, action, language, and cognition introduces several requirements on a formal semantic theory and its practical implementations in situated dialogue agents. Using examples of semantic representations of spatial descriptions we show how Type Theory with Records (TTR) satisfies these requirements and represents a promising knowledge representation system for situated agents.

*Keywords: spatial language, Type Theory with Records (TTR), computational framework*

## 1 INTRODUCTION

In this paper we consider the treatment of spatial language from the perspective of a robot learning spatial concepts and classifying situations according to the spatial relations holding between objects while interacting with a human conversational partner. We start from our experience of building such agents and a conclusion that there is a need for a unified knowledge representation system that connects theories of meaning from formal semantics to practical implementations. We suggest that the type-theoretic notion of judgement is important and that a type theory such as TTR (Type Theory with Records) is advantageous because it can be used to model both the low level perceptual judgements of the robot as well as the conceptual judgements associ-

ated with spatial relations. This is distinct from previous approaches (discussed in Section 3) where the low level perceptual processing is carried out in an entirely different system to that used for semantic processing. An advantage we claim for our approach is that it facilitates the construction of types which have components relating to both low level and high level processing.

In Section 2 we give an overview of the problem area before describing some of the approaches that have been taken in Section 3. We then give a brief intuitive account of the tools we are using from TTR in Section 4 and give some examples of how this relates to understanding spatial descriptions (as our focus is on knowledge representation) in Section 5. Finally, in Section 6, we offer some conclusions and perspectives for future work. An implementation of examples in this paper is available on `https://github.com/GU-CLASP/pyttr/blob/master/lspc.ipynb`.

## 2 COMPUTATIONAL MODELLING OF SPATIAL LANGUAGE

We approach the study of spatial descriptions from the perspective of building computational models for situated agents which we have implemented so far, the typical problems and the ad-hoc solutions taken when representing multi-sourced information. Spatial language is central for situated agents as these must resolve their meaning and reference to visual scenes when being involved in conversations with humans. In such conversations humans would use locational information to identify objects "the chair to the left of the table", describe directed action "pick up the red cube near the green one" or give route instructions "go down this corridor nearly towards its end and then take the second door to your right". However, interfacing language and perception is not only in the domain of applications that involve language-based interaction with humans. There is an emerging trend in robotics where information represented in language is used as assistance to *visual search* (Sjöö 2011; Kunze *et al.* 2014). Robots are typically equipped with several sensors that allow creation of perceptual representations at different levels of abstraction. Creating and classifying for all representations all the time is therefore a computationally expensive task. In the domain of visual object recognition a system

would have to employ all image classifiers on every observation it makes even if most of these classifiers would not yield a match in these situations. For example, the robot is in a corridor and is applying classifiers that would recognise objects found in a kitchen. Having background knowledge about the likely distribution of objects would allow it to prioritise certain classifications. The ontology capturing this knowledge may be static or dynamically built through interaction (Dobnik and Kelleher 2016). In the latter case humans programme the robot through language (Lauria *et al.* 2002).

Cross-disciplinary research has shown that spatial language is dependent on several contextual factors that are part of an agent's interaction with the environment through perception and other agents through dialogue, for example geometrical arrangement of the scene (Regier and Carlson 2001), the type of objects referred to and their interaction (Coventry *et al.* 2001; Dobnik and Kelleher 2013, 2014), visual and discourse salience of objects (Kelleher *et al.* 2005), alignment in dialogue (Watson *et al.* 2004; Dobnik *et al.* 2015), and gesture (Tutton 2013) among others.

The geometrical arrangement of scenes is captured in *spatial templates* or *potential fields*. These can be captured experimentally by placing the target object in various locations around the landmark object and asking participants for judgements whether a particular spatial relation holds (Logan and Sadler 1996; Dobnik and Åstbom 2017). The semantics of spatial templates may be approximated to functions (Gapp 1994a,b) or expressed as a general function with trainable parameters as in the case of the *Attentional Vector Sum (AVS)* model (Regier and Carlson 2001). Figure 1 shows a spatial template for a description *in front of* relating a table and a chair. Spatial templates capture gradience of semantics of spatial descriptions in terms of angles and distances from the location and the orientation of the landmark object. There are regions where native speakers would judge the relation holds to a high degree, for example for the placement of chairs A and D, and regions where the relation holds to a lesser degree, the placement of chairs C and E or does not hold at all, the placement of chair F. A particular scene may be matched by several spatial descriptions. Spatial templates are far from being fixed or universally applicable. In addition to angle and distance, several contextual parameters can be incorporated, for example the presence of distractor

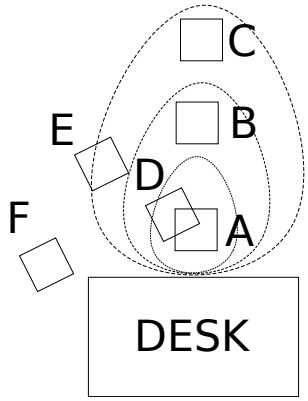



Figure 1: The chair is in front of the desk.

objects (Costello and Kelleher 2006), object occlusion (Kelleher *et al.* 2011) or the function itself can be learned from a dataset of perceptual observations and descriptions as a classifier (Roy 2002; Dobnik 2009).

Scene geometry is not the only meaning component of spatial descriptions. Spatial relations are also expressing other non-geometric aspects of how we view the relation between the landmark and the target objects. For example, a description such as "Alex is at her desk" might not only mean that Alex is proximal to her desk. Instead, we might interpret the description that she is sitting in her chair facing a computer screen and working. In literature such aspects of meaning are known as *functional aspects* (Coventry and Garrod 2005) because they are dependent on the function of interacting objects: what are they used for, how do they interact with each other and how they can be manipulated? In order to understand the interaction of objects one needs to observe what will happen to scenes. Coventry *et al.* (2005) model functional aspects of meaning as *dynamic-kinematic routines* captured by several stacked recurrent neural networks that take both visual and language input data. Modelling different takes on the scene and the relations into which the target and the landmark objects enter leads to the development of qualitative spatial ontologies (Bateman *et al.* 2010) and logics such as (Zwarts and Winter 2000; Cohn and Renz 2008) which are similar to Allen's interval algebra for temporal

**Katie:** Please tell me, where is the blue box?

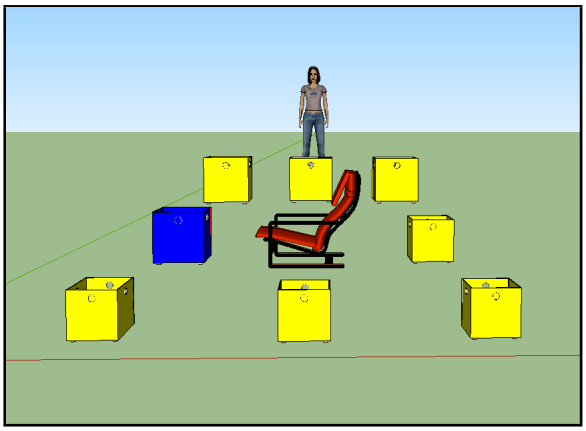


Figure 2: Assignment of FoR in dialogue

reasoning (Allen 1983).

Spatial descriptions are also sensitive to changing linguistic context that arises in linguistic interaction. One such example is coordination of referring expressions (Garrod and Doherty 1994). Projective spatial descriptions such as "to the left of" and "behind" require setting a perspective or the frame of reference (FoR) which can be modelled as a set of three orthogonal axes fixed at some point (the location of the landmark object) and oriented in a direction determined by the viewpoint (Maillat 2003). The viewpoint can be any conversational participant or object in the scene (for this reason such FoR assignment is known as *relative FoR*[1]) that has an identifiable front and back which introduces considerable referential ambiguity of projective spatial descriptions. Alternatively, a scene can also be described from a global bird's eye perspective, e.g. "North of", in which case we talk about *extrinsic FoR* assignment. The FoR may be specified overtly such as "from your point of view" but frequently it is omitted and its resolution is relied upon (among other things) the dynamics of conversation.

Figure 2 shows a virtual scene involving a conversational partner,

[1] We do not distinguish *intrinsic* FoR as this is *relative* FoR where the viewpoint is the landmark object.

Katie, facing us at the opposite side of the room. What FoR would we use to continue the conversation? How would FoR be assigned over several utterances and conversational role changes? Would conversational partners align to a particular FoR or would they tend to change it frequently – and what are conditions licensing such change? What other factors in addition to linguistic conversation contribute to the assignment of FoR? Can a system learn from human assignments of FoR and successfully demonstrate its knowledge in a new conversation with a human? We investigate the strategies of FoR assignment in dialogue, both restricted and free in (Dobnik *et al.* 2014) and (Dobnik *et al.* 2015) respectively.

The preceding discussion demonstrates that the semantics of spatial descriptions involves meaning representations at three distinct levels none of which have been so far captured in a single representational framework which could be employed with situated conversational agents: (i) geometric representations involve grounding symbols in perceptual observations (Harnad 1990), (ii) integrating of functional knowledge involves lexical and compositional semantics, and (iii) FoR assignment involves both previous steps and pragmatics of conversation. Modelling the semantics of spatial descriptions thus raises several open questions. How is an agent able to determine *sense* and *reference* (Frege 1948)[2] of spatial descriptions? The former relates to what components of lexical meaning are involved and the latter relates to how expressions relate to contextual features arising from perceptual and discourse contexts. A model of *grounding* is required: how are perceptual and conceptual domains bridged (reference) and how is information from contextual features fused into bundles of meaning representations (sense)? The resulting framework should possess sufficient *formal accuracy* and *expressiveness* of representations for modelling human language and reasoning to capture notions such as logical entailment, scoping properties, underspecification, hierarchical organisation of meaning and structure, compositionality of structure for words, sentences and utterances, recursion, feature unification and others. The framework should also include a *learning theory* concerning how an agent is able to adapt or learn its representations in new physical and conversational contexts (Cooper

[2] The paper was first published in 1892.

*et al.* 2015; Dobnik and Kelleher 2016).

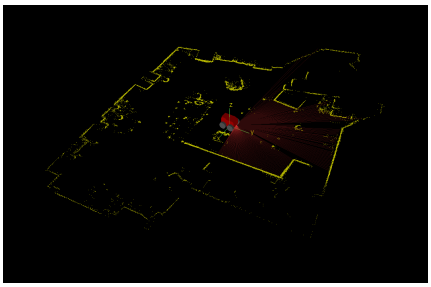
## 3 COMPUTATIONAL FRAMEWORKS

### 3.1 *A classical view of vision and language*

Figure 3 shows a typical approach to modelling language and vision. We start by building a model of the perceptual scene which captures its geometrical representation. In this example, the robot starts with a SLAM map (Dissanayake *et al.* 2001) which contains clouds of points in 3-dimensional coordinate space. The perceptual model is then connected to a formal conceptual representation of the scene which in this example is expressed in first-order logic. An important and challenging issue here is to find a mapping between a reasonably accurate geometric representation of a scene with continuous parameters (locations in the coordinate space and angles of orientation) to cognitive categories that are reflected in language. The mapping between two such domains thus results in vagueness. The formal representation is then mapped to the linguistic expression. The mapping between the layers is typically learned from datasets of collected observations with machine learning. For example, in (Dobnik 2009) we learn classifiers that map representations from SLAM maps to words, thus skipping an intermediate representational layer. Matuszek *et al.* (2012a) present a method where also the intermediate semantic representation is included: linguistic expressions are grounded in compositional semantic forms which are grounded in perception. Finally, language does not only need to be grounded in perception but also in the robotic control language (Matuszek *et al.* 2012b).

### 3.2 *Model-theoretic Montague semantics*

Classical model-theoretic or Montague semantics uses higher order logic (Montague 1974; Dowty *et al.* 1981; Blackburn and Bos 2005; Bird *et al.* 2009) which provides the required and desired formal accuracy and expressiveness of a representation system. It accounts for how meaning representations of words are composed in the form of higher order functions to form meaning representations of sentences. The functional composition of constituents allows us to translate between sentence constituent structure and its logical representation as

**1.** 

**2.** $\forall x \forall y$[supports($y$,$x$) $\wedge$ contiguous(surface($x$), surface($y$)) $\rightarrow$ on$_1$($x$,$y$)]

**3.** "The newspaper is on the table"

Figure 3: Grounding language in perception

shown in Figure 4. The final logical forms of spatial prepositions are slightly more complicated than presented in this example and due to their context dependency a single description or utterance (surface form) may resolve to several representations as discussed in (Miller and Johnson-Laird 1976; Herskovits 1986), for example on(x,y)$_1$: object(x) $\wedge$ object(y) $\wedge$ supports(y,x) $\wedge$ contiguous(surface(x),surface(y)) and on(x,y)$_2$: object(x) $\wedge$ object(y) $\wedge$ contiguous(boundary(x),y). However, dealing with these two issues separately, we are able to derive their compositional representation along the same lines as in Figure 4.

In model-theoretic semantics the expression's reference is determined by an assignment, a valuation function between linguistics strings and entities (or sets of tuples of entities) in a model. The model is agent external and fixed. The valuation returns true if an entity or a relation between entities denoted by an expression can be found in the model, otherwise it returns false. While it would be possible to represent the referential semantics of a "on" in a model by listing a set of all coordinates of locations where this spatial description applies, this referential representation of meaning is cumbersome as the model would have to include an assignment for every scale, for every spatial relation, for every pair of objects. Since angles and distances in a coordinate system are continuous measures this means that such sets would be infinite. The model also does not straightforwardly represent gradience and vagueness of spatial descriptions. In order to do that one would have to resort to the notion of possible worlds (Las-

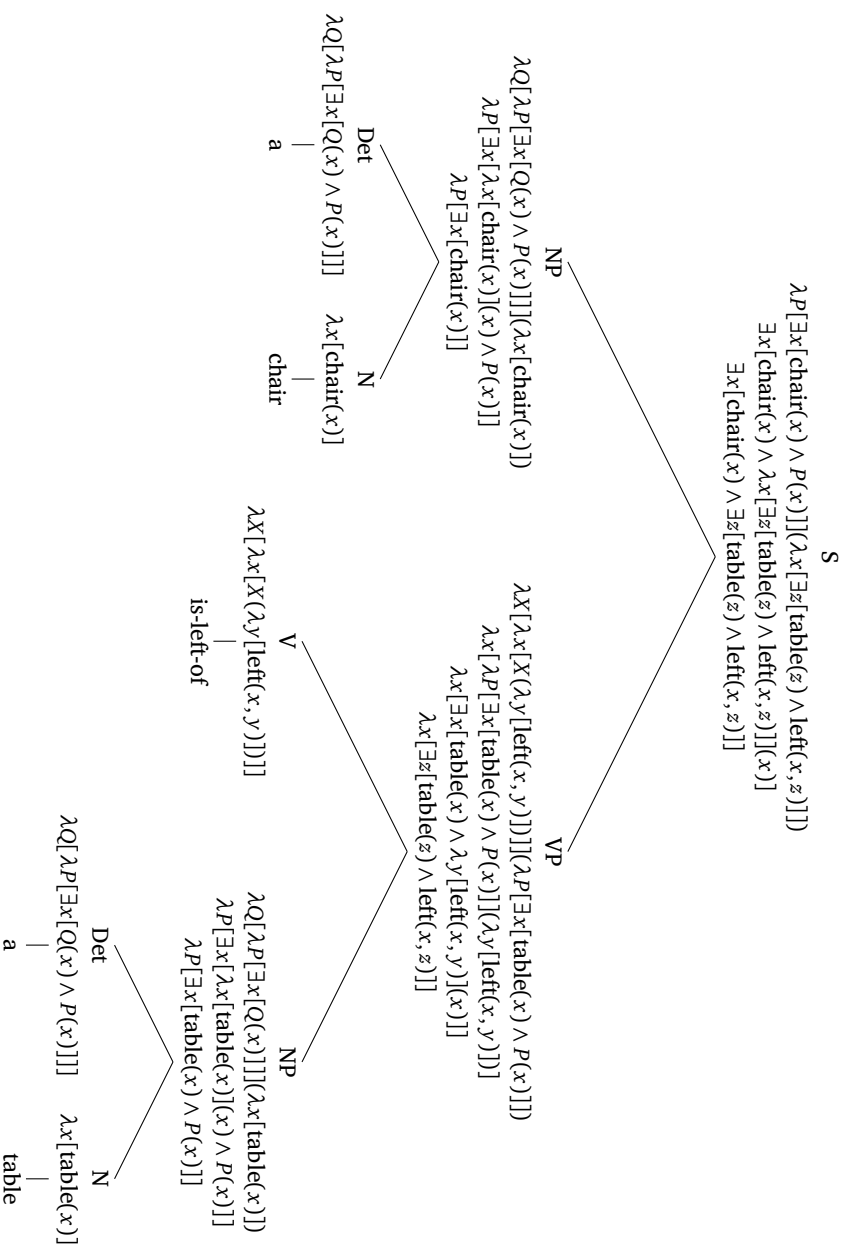



Figure 4: A chair is left of a table.

siter 2011) which introduces further computational complexity (for discussion see (Cooper *et al.* 2015, Section 1.1, p.3ff)).

As discussed earlier both vagueness and gradience of spatial language are captured in computational models as spatial templates or potential fields. While spatial templates can be thought of as referential overlays of regions induced experimentally (as a set of points where participants consider a particular spatial relation to apply), potential fields capture the notion that such regions can be generalised as functions. However, as argued in (Lappin 2013) these functions do not represent objects in a model (or extensions or the referential meaning of these descriptions) but rather they capture their *sense* or *intension* specifying in what ways a description relates to perceptual observations. Knowing this function we can check whether a particular spatial relation associated with the function applies for particular pair of objects and to what degree. The notion of applying a function from perceptual observations to words (or the other way around) representing the meaning of words is also known as *grounding* these words in perception (Harnad 1990).

The model-theoretic approach to semantics assumes that a model is derived through some external process and therefore pre-given, that it is complete and represents a state of affairs at a particular temporal snapshot (Fagin *et al.* 1995). In practice, however, complete models may be rarely observable and we must deal with partial models. We must also account for the fact that we may incrementally observe more and more of the world and we have to update the model with new observations, sometimes even correct representations that we have already built in the light of new evidence. Finally, the world is not static itself as new objects and events continually come into existence. Imagine a robot (and indeed such robots were used in the early days of robotics) with a pre-programmed static model of the world. Every minute change in the world would render it useless as there would be a discrepancy between its representation of the world and the actual world. Modern robotic models used in localisation and map building are incrementally learned or updated over time by taking into account robot's perceptional observations and motion and errors associated with both (Dissanayake *et al.* 2001). An important consequence of this is that the model of the world a robot builds is individual to a particular robot's life-span and experience. Two robots experiencing the

same world will have slightly different models. Of course, the more they experience the world, the more similar their models will be. It is conceivable that humans learn meanings in the same way. However, doing so they are equipped with yet another tool to overcome individual inconsistencies in their model. They can use linguistic dialogue interaction to resolve such inconsistencies in the form of repair (Pickering and Garrod 2004). In robotics several models that explore learning language through interaction have been built which include (Steels and Belpaeme 2005; Skočaj *et al.* 2011; Ivaldi *et al.* 2014), also related to spatial cognition (Steels and Loetzsch 2009). We describe a system that allows modelling of learning semantic concepts through dialogue interaction in (Dobnik and de Graaf 2017).

### 3.3 *Models used in robotics*

In building situated conversational agents, several systems have been proposed but none of them capture all of the requirements discussed in Section 2. For example, *semiotic schemas* (Roy 2005) represent lexical meaning of words as directed graphs composed of nodes that represent sequences of perceptual observations and classification events as shown in Figure 5. The meaning/sense of an object is defined in terms of what can be experienced with sensors and actuators of a robot. The reference is determined by embedding a semiotic schema with the actual sensory readings. For example, a cup can be experienced and classified for either through visual or haptic modalities. The location of the sensory readings determines the location of the object. Semiotic schemas represent a very attractive model of grounded lexical semantics of words, but how such semiotic schemas compose to larger linguistic structures is left unaccounted for.

Quite frequently grounded representations are arranged into layers which is related to the fact that in practical applications several distinctive sub-systems are used that are stacked into a pipeline. For example, in the layered approach of (Kruijff *et al.* 2007), here summarised in Figure 6, the lowest level consists of a feature map which directly relates to laser sensors. Here features are sets of points which can be connected to lines which represent walls. The next level is a navigation graph. As the robot moves around space it creates nodes. If the robot can move directly between two nodes, a connection is made and on the basis of these connections a navigation graph is created.

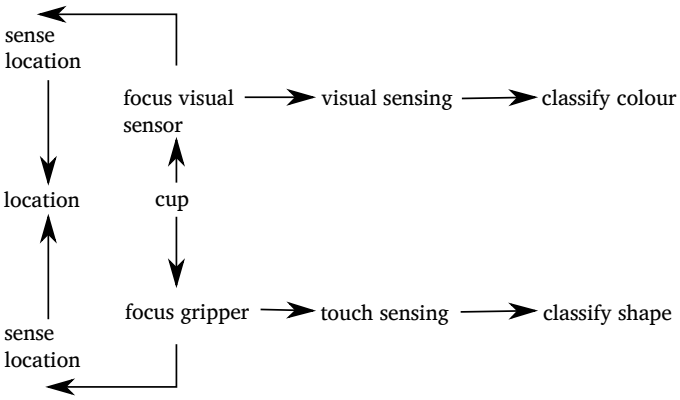



Figure 5: A simplified semiotic schema

Groups of nodes may be identified whereby two groups are only connected through a single node in each group. Such nodes are gateway nodes and indicate passages between different areas or doors. From such topology of nodes a topological map can be hypothesised which identifies enclosed spaces such as corridors, kitchens and rooms. The information about the spaces can be further augmented with linguistic information from ontology, for example what objects are found in kitchens. In this approach one needs to design interfaces between representational levels in the pipeline. Most frequently, representations and operations at each level are distinct from each other. A question we would like to explore is whether representations at different levels can be generalised by taking inspiration from the way humans assign, learn and reason with meaning. A unified meaning representation would allow interactions between modalities that are required in modelling human cognition but are difficult to implement in a layered pipeline architecture.

## 4 TYPE THEORY WITH RECORDS (TTR)

Type Theory with Records (TTR) (Cooper 2012) builds on the tradition of classical formal semantics (and therefore captures the notion of compositionality) but at the same time, drawing on insights from situation semantics, addresses the outstanding questions related to per-

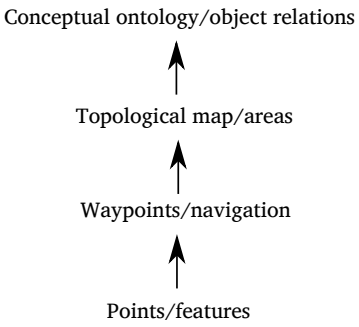



Figure 6: A layered approach

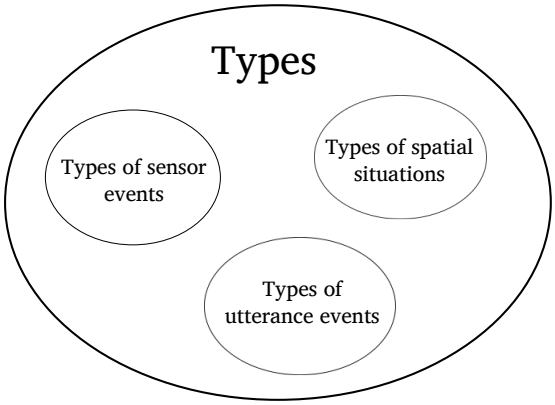



Figure 7: An unified view: types all over the place

ception discussed in the preceding paragraphs. It starts from the idea that information is founded on our ability to perceive and classify the world, that is to perceive or *judge* objects and situations as being of types. All information can be represented as types (Figure 7) which makes type assignment an abstract theory of cognition and perception. Having a single representational layer allows information fusion between perception, conceptual knowledge and linguistic communication which is an important requirement for modelling spatial descriptions.

Types are intensional – that is, there can be distinct types which

have identical extensions. For example, the type of situations in which an object, $a$, is to the left of another object, $b$, in symbols, left($a$,$b$), can have exactly the same witnesses as the type of situations in which $b$ is to the right of $a$, right($b$,$a$), without requiring that the two types are identical. For some more discussion of the intensional nature of types in TTR see (Cooper 2017). This allows us to relate linguistic propositions to types, the so-called "propositions as types" dictum which is standard in type theories deriving from the original work of Martin-Löf (Martin-Löf 1984; Nordström *et al.* 1990). The notion of truth is linked to judgements that an object $a$ is of type $T$ ($a : T$). As in standard Martin-Löf type theories, a type is "true" just in case it has some witness. Thus the type of situations 'left($a$,$b$)' is "true" just in case there is some situation where $a$ is to the left of $b$.

We can furthermore seek to operationalise the types as computable functions (Lappin 2013) or classifiers (Larsson 2015), rather than associating them with sets of witnesses as in the standard definition of TTR (Cooper in prep, 2012). Under this view we can consider an agent to have access to a particular type inventory as a resource. Different agents can have access to different type resources which can be dynamically revised, both in terms of learning new types and in modifying the witness conditions in terms of classifiers which can change as the result of the agent's experience of new situations (Dobnik *et al.* 2013; Larsson 2015). In order for communication between agents to be possible, they must converge on sufficiently similar type resources. This convergence is in part enabled by the fact that the agents exist in similar environments and have similar perceptual apparatus to classify features in the environment. But in addition it is important that the agents be able to use language to communicate with each other about their classification of features in the environment, for example an agent may receive linguistic information which provides a classification which is at variance from that given by its perceptual apparatus or in linguistic communication between agents corrective feedback might be used to express a variance in judgement by two agents.

This is perhaps a novel view in linguistic semantics and computational linguistics but it relates to a standard view in mobile robotics (Dissanayake *et al.* 2001) where a map of an environment is constructed dynamically as a robot moves around in it and features are

constructed on the basis of clouds of points in 3D space where the robot's sensors indicate that something is present. In our terms this would correspond to recognising the physical presence of an object and assigning a particular type to it.

In such a learning scenario it is natural to consider the role of probabilistic judgements, that is, the judgement that an object $a$ is of type $T$ with probability $p$ instead of the standard categorical judgements to be found in type theory. For a proposal of how this might be incorporated into TTR see (Cooper *et al.* 2015). This means that an agent can determine a degree of belief that a particular situation is of a particular type. For example, the probability that a situation is to be classified as one where an umbrella is *over* a person may vary with respect to both geometric configuration and the degree to which the umbrella is protecting the person from rain (Coventry *et al.* 2001).

In contrast to the classical Montagovian semantic framework which employs a variant of the simple theory of types, TTR introduces an extended set of basic types (for example *Ind* and *Real* that correspond to the basic conceptual categories individuals and real numbers). However, it is also *a rich type system* which, in addition to basic types, contains complex types constructed from types and other objects, among them *ptypes* constructed from predicates and their arguments, such as 'left($a$,$b$)', and *record types*, such as,

$$\begin{bmatrix} \text{x} & : & \mathit{Ind} \\ \text{y} & : & \mathit{Ind} \\ \text{e} & : & \text{left(x,y)} \end{bmatrix}$$

whose witnesses would be any record with three fields labelled by 'x', 'y' and 'e', respectively (and possibly more fields with other labels) such that the 'x'-field contains an object $a$ of type *Ind*, the 'y'-field contains an object $b$ of type *Ind* and the 'e'-field contains an object of type 'left($a$,$b$)'. For a detailed characterisation of record types in TTR see (Cooper in prep, 2012). Record types in TTR are used to model, among other things, lexical content and dialogue information states. For our present purposes the structured nature of record types allows us to combine in a single object the kind of multi-source information needed for robotics and the modelling of spatial descriptions representing a bridge between what might be thought of in other approaches as the

sub-symbolic domain of perception and the symbolic domain of high level conceptual analysis.

The structured nature of record types in TTR allows representation of several kinds of formal structural relations which has implications for inference of representations containing multi-sourced information. Record types (and the corresponding records) can be compared with each other. Consider the following example. If

$$\textit{Relation} = \begin{bmatrix} \mathrm{x} & : & \textit{Ind} \\ \mathrm{y} & : & \textit{Ind} \\ \mathrm{c}_1 & : & \mathrm{target(x)} \\ \mathrm{c}_2 & : & \mathrm{landmark(y)} \end{bmatrix}$$

and

$$\textit{Left} = \begin{bmatrix} \mathrm{x} & : & \textit{Ind} \\ \mathrm{y} & : & \textit{Ind} \\ \mathrm{c}_1 & : & \mathrm{target(x)} \\ \mathrm{c}_2 & : & \mathrm{landmark(y)} \\ \mathrm{c}_3 & : & \mathrm{left(x,y)} \end{bmatrix}$$

then *Left* $\sqsubseteq$ *Relation* where $\sqsubseteq$ denotes the *subtype* relation (Cooper 2012, p.295). Similarly, record types allow identification of dependencies using *dependent types*. The notation like target(x) within the context of the record type above is an abbreviation for a tuple of objects $\langle \lambda v{:}\textit{Ind} \,.\, \mathrm{target}(v), \langle \mathrm{x} \rangle \rangle$ where the first element is a dependent type, a function mapping objects to a type, and the second element is a sequence of paths to the arguments of this function within a record type. Finally, both *Ind* and target(x) are *component types* of record types *Relation* and *Left* which means that the latter types are representations of thematic relations between individuals and properties found in language (Lin and Murphy 2001; Estes *et al.* 2011).

## 5 TYPES OF SPATIAL DESCRIPTIONS

In the remainder of the paper we discuss how our empirical investigations of learning geometric meanings of spatial descriptions with situated robots (Dobnik 2009; Dobnik and de Graaf 2017), learning functional meanings of prepositions from collections of image descriptions (Dobnik and Kelleher 2013, 2014), and modelling of reference

frame assignment in conversation (Dobnik *et al.* 2014, 2015) can be captured in the TTR framework.

The idea is that TTR can be seen as an abstract model of cognition and perception (Cooper 2012, in prep) which can be used to model both the linguistic behaviour of humans as well as perception based on sensor readings in artificial agents. It is important to note that robots have different perceptual apparatus than humans, both in the number and the nature of sensors. It follows that their sensors will give rise to different types of information at the lowest sensory level. However, these sensory types can be related to types corresponding to concepts which are similar enough to conceptual types internalised by humans to allow communication between the two. Nevertheless, the type system an agent can acquire is constrained by the agent's perceptual apparatus. We cannot, for example, expect an agent incapable of colour perception to successfully make judgements about colour based on colour concepts available to a human, however much we may talk to the agent or train it on objects of different colours. It simply does not have the required sensors/classifiers to distinguish the appropriate situations.

There are two main aspects of theoretical interest with the approach we suggest:

1. The notion of *judgement* from type theory can be used to model both the kind of low level perceptual discrimination carried out by classifiers in robotic systems and the high level conceptual classification including the truth of propositions which are important for linguistic semantics. Thus it offers the possibility of a unified approach to both.
2. Given the kind of structured types that are proposed in a system like TTR it is not only possible to express relations between the low level and high level types but even to have a low level perceptual type and a high level conceptual type as components within a single type and even to have one type depend on the other. This gives a very different perspective on the cognitive makeup of situated agents than that given by the kind of layered approaches discussed in Section 3, where the different layers involve entirely different systems.

In the next section we will give examples which illustrate this.

### 5.1 *Types of objects*

Figure 8 shows an example of bridging between perceptual and conceptual domains for object recognition. Step 1 shows a record of type *PointMap* which is produced by SLAM (for details see (Dobnik *et al.* 2013)). The type *PointMap* is a subtype of a type that represents a list of records containing three real numbers modelling points in three-dimensional space. A point map is a list (or a set) of points that a robot is tracking in space. TTR allows function types one of which is exemplified in the object detection function in Step 2. This function maps an object of type *Pointmap* to a type that represents a set of records specifying (1) the reg(ion) occupied by the object (a subpointmap) and (2) a property which is modelled as a pfun which maps an individual to a type, in this case a ptype or a predicate type. The purpose of this function type is to associate a perceptual object and some property, thus to pair two kinds of information. The property functions take objects of type *Ind*ividuals to types of individuals having some property. The target record type of the main function type does not yet constrain any individuals that this property could be assigned to nor does this record type correspond to a situation. In Step 3 we introduce an individuation function which takes records of associated perceptual objects and properties and yields a type of situation involving an individual located at a certain location and having this property. This type therefore represents a cognitive take on a situation.

In this example, the mappings between the types are modelled with functions but in practice (some) associations would be learned. For example, Harnad (1990) argues that grounding, associating perceptual and conceptual domains, can only be accomplished through classification. In (Dobnik 2009) decision tree and Naïve Bayes classifier models are learned to classify between point clouds and spatial descriptions. Here the associating function that the classifier has learned is in the domain of the hypothesis space of each learning algorithm and is therefore quite complex. Larsson (2015) introduces a perceptron model to TTR and Cooper *et al.* (2015) give the type system, including function types, a Bayesian interpretation. The latter allows direct propagation of Bayesian probabilistic beliefs between the types while the observed type probabilities can be trained from agent's

1. A point is a record with three coordinates:

$$Point = \begin{bmatrix} x & : & Real \\ y & : & Real \\ z & : & Real \end{bmatrix}$$

A point map is a list of points: $PointMap = \text{list}(Point)$

$$[\begin{bmatrix} x & = & 34 \\ y & = & 24 \\ z & = & 48 \end{bmatrix}, \begin{bmatrix} x & = & 56 \\ y & = & 78 \\ z & = & 114 \end{bmatrix}, \ldots] : PointMap$$

2. A property is a function from individuals to a type:
$Ppty = (Ind \rightarrow Type)$
$\lambda x{:}Ind\ .\ \text{chair}(x) : Ppty$
An object detection function is a function from a point map to a set of records containing a sub-point map of the original and a property associated with it:

$$ObjectDetector = (Pointmap \rightarrow \text{set}(\begin{bmatrix} \text{reg} & : & Pointmap \\ \text{pfun} & : & Ppty \end{bmatrix}))$$

3. Individuation function

$$IndFun = (\begin{bmatrix} \text{reg} & : & Pointmap \\ \text{pfun} & : & Ppty \end{bmatrix} \rightarrow \begin{bmatrix} a & : & Ind \\ \text{loc} & : & Type \\ c & : & Type \end{bmatrix})$$

$$\lambda r{:}\begin{bmatrix} \text{reg}{:}Pointmap \\ \text{pfun}{:}Ppty \end{bmatrix}\ .\ \begin{bmatrix} a & : & Ind \\ \text{loc} & : & \text{location}(a, r.\text{reg}) \\ c & : & r.\text{pfun}(a) \end{bmatrix} : IndFun$$

Perceptual domain

Conceptual domain

Figure 8: From perceptional to conceptional domain

observations.

## 5.2 *Types of spatial situations*

Spatial descriptions, e.g. *"over"* and *"above"* are sensitive to classes of interacting objects and the contribution of such functional world-knowledge versus geometric knowledge for the semantics is different from one spatial preposition to another (Coventry *et al.* 2001; Coventry and Garrod 2005; Coventry *et al.* 2005). While previous work attempted to determine the contribution of each modality experimentally, Dobnik and Kelleher (2013, 2014) extract functional information from a large corpus of text describing images. Image descriptions are constrained by the properties of the visual scene shown in the im-

(a) $\lambda r{:}\begin{bmatrix}\text{a}{:}\mathit{Ind}\\ \text{c}{:}\text{person(a)}\end{bmatrix}.\text{organism}(r.\text{a})$

(b) If $s : \begin{bmatrix}\text{a} & : & \mathit{Ind}\\ \text{loc} & : & \text{location(a},\pi)\\ \text{c} & : & \text{person(a)}\end{bmatrix}$
then $\exists s'\ [s' : \text{organism}(s.\text{a})]$

Figure 9: Representing functional knowledge

age, both perceptual (geometric arrangement of the scene) and functional (the nature and interaction of objects shown there). Both kinds of information will be reflected in the text describing the image, in a particular choice of descriptions that annotators used. Building lexical models of word co-occurrence thus allows us to capture functional interactions between prepositions and targets and landmarks. In (Dobnik and Kelleher 2013) we capture the strength of association between a preposition and different target-landmark pairs with *log-likelihood ratio*. In (Dobnik and Kelleher 2014) we generalise the types of targets and landmarks of a particular spatial preposition by ascending in a WordNet hierarchy (Fellbaum 1998). This allows us to generate patterns of prepositional use such as the following: *person.n.01 under tree.n.01*, *shirt.n.01 under sweater.n.01*, and *person.n.01 under body of water.n.01*. Labels such as *person.n.01* indicate the labels given to the generalised synsets in the WordNet hierarchy. The patterns indicate types of spatial situations that the "under" relation applies to. Importantly, each of these patterns corresponds to quite a different arrangement of target and landmark objects and without such functional knowledge it were difficult to capture a single spatial template that would not over-generate. The functional knowledge represented in these types thus constrains on sub-sets of spatial situations for which individual spatial templates can be learned.

Figure 9(a) shows a TTR function that maps ontological knowledge from one ontological category to another. This is a similar function to *pfun* in the object detection function shown in Figure 8. It assigns the individual of the type in the domain of the function a particular property $\lambda r.\text{organism}(r)$. Figure 9(b) shows how associative reasoning is captured in TTR. Having a meaning postulate in Figure 9(a)

$$\lambda r{:}\left[\begin{array}{lcl} \mathrm{o}_1 & : & \left[\begin{array}{lcl} \mathrm{a} & : & \textit{Ind} \\ \mathrm{reg} & : & \textit{Pointmap} \\ \mathrm{c} & : & \mathrm{person(a)} \end{array}\right] \\ \mathrm{o}_2 & : & \left[\begin{array}{lcl} \mathrm{a} & : & \textit{Ind} \\ \mathrm{reg} & : & \textit{Pointmap} \\ \mathrm{c} & : & \mathrm{artefact(a)} \end{array}\right] \\ \mathrm{st} & : & \mathrm{spatial\text{-}template}_{under_1}(o_1.\mathrm{reg}, o_2.\mathrm{reg}) \end{array}\right].\mathrm{under}_1(\mathrm{r}.o_1.\mathrm{a}, \mathrm{r}.o_2.\mathrm{a})$$

$$\lambda r{:}\left[\begin{array}{lcl} \mathrm{o}_1 & : & \left[\begin{array}{lcl} \mathrm{a} & : & \textit{Ind} \\ \mathrm{reg} & : & \textit{Pointmap} \\ \mathrm{c} & : & \mathrm{person(a)} \end{array}\right] \\ \mathrm{o}_2 & : & \left[\begin{array}{lcl} \mathrm{a} & : & \textit{Ind} \\ \mathrm{reg} & : & \textit{Pointmap} \\ \mathrm{c} & : & \mathrm{body\text{-}of\text{-}water(a)} \end{array}\right] \\ \mathrm{st} & : & \mathrm{spatial\text{-}template}_{under_2}(o_1.\mathrm{reg}, o_2.\mathrm{reg}) \end{array}\right].\mathrm{under}_2(\mathrm{r}.o_1.\mathrm{a}, \mathrm{r}.o_2.\mathrm{a})$$

Figure 10: Spatial templates sensitive to object function

an agent can make a conclusion that a situation $s$ of the first type (the left hand side of the *If-then* rule) requires that there is also a situation of the second type (the right hand side of the same rule).

Each type of situation representing a spatial pattern involves a different interplay of geometric and conceptual knowledge spanning the domain of point clouds and "logical" individuals. Figure 10 shows the conceptual constraints on the target and landmark objects limiting top-down a subset of spatial situations over which individual types of spatial relations are built. Hence, the resulting spatial template spatial-template$_{under_1}$ is a distinct pytpe classifier from spatial-template$_{under_2}$. In the generation step the function in Figure 10 takes account of conceptual properties of objects that could be obtained by computing relevant hypernyms such as "person" and "furniture" and an associated spatial template that relates the point clouds associated with them. It then generates a type of situation which involves a conceptual spatial relation between individuals.

### 5.3 *Types of dialogue information states*

TTR can also be used to model dialogue by representing types of information states (IS). Agents in conversation align to the primed frame

of reference (FoR) and continue to use it (Dobnik *et al.* 2014). However, such alignment is only local and depends on the nature of the dialogue that agents are engaged in and other contextual factors of the conversation such as the perceptual properties of the scene or the task that agents are performing (Dobnik *et al.* 2015). Dobnik *et al.* (2014) study the properties of local FoR alignment over several turns of conversation in the constrained environment (Figure 2). The experiment captures participants' understanding of the agreed FoR and therefore alignment. In Game 1 a virtual conversational partner generates an unambiguous description that refers only to one of the objects. The participant must then click on that object. Here the system primes the participant for a particular FoR. In Game 2 the system generates an ambiguous description which may refer to several objects. Again the participant must click on one of the target objects but this time they must decide on a particular FoR assignment. Will this be aligned with the previous turn pair or will they assume a new strategy? Game 3 is identical to Game 2 and it tests if the priming from Game 1 is persistent over several games. In Game 4 the speaker-hearer roles reverse: the system selects an object and the participant must describe it using a particular FoR assignment. The role of this game is to test whether priming will persist if the conversational roles change.

The preceding interaction is formalised as a probabilistic model of FoR assignment over several local turns of conversation. This model is then applied in a generation experiment. Here the system is making assumptions about the human conversational partner and is trying to align to them to the extent captured in the previous experiment. In Game 1 the system chooses an object and a human primes the system by generating an unambiguous description. In Game 2 a human selects a box and the system generates a description using its FoR model. The human then confirms if the description is a good one. Game 3 is identical to Game 2. In Game 4 a human chooses a box. The system asks the user to describe it and also generates a description for itself. A match between the human description and the system-generated description is compared. The results show a good agreement between humans and the system ($\geq$ 82.76% for Game 4).

The model of FoR assignment predicts, for example, that speakers initiating conversation tend to be egocentric. Figure 11 show two types of information states (ISs). When Alex is planning the utterance "The

$$\text{(a) } s_0^{Alex}: \left[\begin{array}{l} \text{priv} : \left[\begin{array}{l} \text{objs}: \begin{bmatrix} o_0{:}\Sigma_0 \\ o_1{:}\Sigma_1 \\ o_2{:}\Sigma_2 \\ o_3{:}\Sigma_3 \end{bmatrix} \\ \text{bel}: \begin{bmatrix} c_{me} = [\text{c:me}(\Uparrow^2\text{objs.o}_0\text{.a})]{:}\textit{Type} \\ c_{left} = [\text{c:left}(\Uparrow^2\text{objs.o}_2\text{.a}, \Uparrow^2\text{objs.o}_3\text{.a})]{:}\textit{Type} \end{bmatrix} \\ \text{for-origin} = \textbf{objs.o}_0\textbf{.a}{:}\textit{Ind} \\ \text{agenda} = [\begin{bmatrix} \text{move}{:}\textit{Assertion} \\ \text{cont} = \Uparrow^2\text{bel.}c_{left}{:}\textit{Type} \end{bmatrix}]{:}\text{list}(\textit{DMove}) \end{array}\right] \\ \text{shared}: [c_{in-focus}{:}\Uparrow\text{priv.objs.o}_2\text{.a}] \end{array}\right]$$

$$\text{(b) } s_1^{Sam}: \left[\begin{array}{l} \text{priv} : \left[\begin{array}{l} \text{objs}: \begin{bmatrix} o_0{:}\Sigma_0 \\ o_1{:}\Sigma_1 \\ o_2{:}\Sigma_2 \\ o_3{:}\Sigma_3 \end{bmatrix} \\ \text{bel} : [c_{me} = [\text{c:me}(\Uparrow^2\text{objs.o}_1\text{.a})]{:}\textit{Type}\ ] \end{array}\right] \\ \text{shared}: \left[\begin{array}{l} \text{speaker} = \Uparrow\text{priv.objs.o}_0\text{.a}{:}\textit{Ind} \\ c_{in-focus}{:}\Uparrow\text{priv.objs.o}_2 \\ \text{latest-move}: \begin{bmatrix} \text{speaker} = \Uparrow^2\text{priv.objs.o}_0\text{.a}{:}\textit{Ind} \\ \text{cont} = [\text{c:left}(\Uparrow^3\text{priv.objs.o}_2\text{.a}, \Uparrow^3\text{priv.objs.o}_3\text{.a})]{:}\textit{Type} \end{bmatrix} \\ \text{for-origin} = \textbf{speaker}{:}\textit{Ind} \end{array}\right] \end{array}\right]$$

Figure 11: Types of dialogue information states

chair is to the left of the table" her information state would of the type shown in (a). Information states represent information that is private to the agent, and information that the agent believes is a part of the common ground with another conversational participant or shared. In the shared part of the IS in (a) there is a pointer to the object in focus. The object is stored in the private part of the IS as each agent builds its own objects. $\Sigma_i$ is a type returned by an individuation function on the basis of the pointmap that the agent has constructed. The agent also has a private belief that they are one of the objects and a belief that two particular objects are in the left relation. Crucially, at this stage, the FoR origin is assigned to the object corresponding to the individual having this IS. A double arrow $\Uparrow^2$ indicates that the path refers to the container-type which the current type is a dependent type of, the superscript indicates the depth of embedding. Notation such as *label = value : Type* as in *for-origin = objs.o$_0$.a : Ind* represents singleton types where the *value* stands for a manifest field.

The model of FoR assignment also predicts that hearers assume that speakers are egocentric. Figure 11(b) shows the Sam's IS accommodating the Alex's utterance. After Alex has made an utterance, the shared part of the IS is expanded through accommodation. There is information about the latest move: the speaker and the content of the move. Since Sam is a hearer of the utterance, he assumes that the FoR is identical to the speaker of the previous utterance as predicted by our probabilistic model. In this example, we assume that agents use identical labels for objects. However, it is not necessary or indeed possible that they have identified the same objects. In the future work we plan to investigate how agents resolve such differences using language in particular what mechanisms of clarification and repair are used in such cases (Purver *et al.* 2003).

## 6 CONCLUSION

In this paper we outlined an application of type theory to natural language semantics in the framework called Type Theory with Records or TTR which allows to relate semantics to action, perception and cognition. We used TTR to represent different components of analysis of spatial descriptions. TTR is naturally suited for this task as it treats meaning being based on perception and interaction. Perception and conceptual reasoning can be related within one unified approach. The framework also points to similarities between linguistic and non-linguistic learning. We will be testing practical implementations of TTR with situated agents in our forthcoming work based on the framework described in (Dobnik and de Graaf 2017). The expressiveness of the type theoretic framework is associated with high computational cost. In order to make the framework computationally more tractable, we are investigating mechanisms of attention from psychological research which allow us to contextually restrict the type judgements a situated agent has to make (Dobnik and Kelleher 2016).

One aspect of spatial meaning which we have not discussed in this paper is the gradability of types like "left($a$,$b$)". For example, $a$ would be judged to be left of $b$ with a high probability if the two objects are close to each other. However, the probability of this judgement would decrease if $a$ is much closer to the observer than $b$. This suggests exploring the use of probabilistic judgements in TTR as described in

(Cooper *et al.* 2015) and which we plan to explore this in future work.

## ACKNOWLEDGEMENTS

This paper was supported in part by the project Networks and Types (Vetenskapsrådet/Swedish Research Council project VR 2013-4873).

## REFERENCES

James F Allen (1983), Maintaining knowledge about temporal intervals, *Communications of the ACM*, 26(11):832–843.

John A. Bateman, Joana Hois, Robert Ross, and Thora Tenbrink (2010), A linguistic ontology of space for natural language processing, *Artificial Intelligence*, 174(14):1027–1071.

Steven Bird, Ewan Klein, and Edward Loper (2009), *Natural language processing with Python*, O'Reilly, `http://nltk.org/book/`.

Patrick Blackburn and Johan Bos (2005), *Representation and inference for natural language. A first course in computational semantics*, CSLI Publications.

Anthony G. Cohn and Jochen Renz (2008), Qualitative Spatial Representation and Reasoning, in Vladimir Lifschitz Frank van Harmelen and Bruce Porter, editors, *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*, chapter 13, pp. 551–596, Elsevier.

Robin Cooper (2012), Type theory and semantics in flux, in Ruth Kempson, Nicholas Asher, and Tim Fernando, editors, *Handbook of the Philosophy of Science*, volume 14 of *General editors: Dov M Gabbay, Paul Thagard and John Woods*, Elsevier BV.

Robin Cooper (2017), Adapting Type Theory with Records for Natural Language Semantics, in Stergios Chatzikyriakidis and Zhaohui Luo, editors, *Modern Perspectives in Type-Theoretical Semantics*, number 98 in Studies in Linguistics and Philosophy, pp. 71–94, Springer.

Robin Cooper (in prep), Type theory and language: from perception to linguistic communication, `https://sites.google.com/site/typetheorywithrecords/drafts`, draft of book chapters.

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson (2015), Probabilistic Type Theory and Natural Language Semantics, *Linguistic Issues in Language Technology — LiLT*, 10(4):1–43.

Fintan J. Costello and John D. Kelleher (2006), Spatial prepositions in context: the semantics of near in the presence of distractor objects, in

*Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*, Prepositions '06, pp. 1–8, Association for Computational Linguistics, Stroudsburg, PA, USA.

Kenny R. COVENTRY, Angelo CANGELOSI, Rohanna RAJAPAKSE, Alison BACON, Stephen NEWSTEAD, Dan JOYCE, and Lynn V. RICHARDS (2005), Spatial Prepositions and Vague Quantifiers: Implementing the Functional Geometric Framework, in Christian FREKSA, Markus KNAUFF, Bernd KRIEG-BRÜCKNER, Bernhard NEBEL, and Thomas BARKOWSKY, editors, *Spatial Cognition IV. Reasoning, Action, Interaction*, volume 3343 of *Lecture Notes in Computer Science*, pp. 98–110, Springer Berlin Heidelberg.

Kenny R. COVENTRY and Simon C. GARROD (2005), Spatial prepositions and the functional geometric framework. Towards a classification of extra-geometric influences, *Functional features in language and space: Insights from perception, categorisation and development*, pp. 163–173.

Kenny R. COVENTRY, Mercè PRAT-SALA, and Lynn RICHARDS (2001), The interplay between geometry and function in the apprehension of Over, Under, Above and Below, *Journal of Memory and Language*, 44(3):376–398.

M. W. M. G DISSANAYAKE, P. M. NEWMAN, H. F. DURRANT-WHYTE, S. CLARK, and M. CSORBA (2001), A solution to the simultaneous localization and map building (SLAM) problem, *IEEE Transactions on Robotic and Automation*, 17(3):229–241.

Simon DOBNIK (2009), *Teaching mobile robots to use spatial words*, Ph.D. thesis, University of Oxford: Faculty of Linguistics, Philology and Phonetics and The Queen's College, Oxford, United Kingdom, `http://www.dobnik.net/simon/documents/thesis.pdf`.

Simon DOBNIK and Amelie ÅSTBOM (2017), (Perceptual) grounding as interaction, in Volha PETUKHOVA and Ye TIAN, editors, *Proceedings of Saardial – Semdial 2017: The 21st Workshop on the Semantics and Pragmatics of Dialogue*, pp. 1–9, Saarbrücken, Germany.

Simon DOBNIK, Robin COOPER, and Staffan LARSSON (2013), Modelling Language, Action, and Perception in Type Theory with Records, in Denys DUCHIER and Yannick PARMENTIER, editors, *Constraint Solving and Language Processing: 7th International Workshop, CSLP 2012, Orléans, France, September 13–14, 2012, Revised Selected Papers*, volume 8114 of *Lecture Notes in Computer Science*, pp. 70–91, Springer Berlin Heidelberg.

Simon DOBNIK and Erik DE GRAAF (2017), KILLE: a Framework for Situated Agents for Learning Language Through Interaction, in Jörg TIEDEMANN, editor, *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa)*, volume 131 of *Linköping Electronic Conference Proceedings and NEALT Proceedings Series Vol. 29*, pp. 1–10, Northern European Association for Language Technology (NEALT), Linköping University Electronic Press, Gothenburg, Sweden.

Simon DOBNIK, Christine HOWES, and John D. KELLEHER (2015), Changing perspective: Local alignment of reference frames in dialogue, in Christine HOWES and Staffan LARSSON, editors, *Proceedings of goDIAL – Semdial 2015: The 19th Workshop on the Semantics and Pragmatics of Dialogue*, pp. 24–32, Gothenburg, Sweden.

Simon DOBNIK and John D. KELLEHER (2013), Towards an automatic identification of functional and geometric spatial prepositions, in *Proceedings of PRE-CogSsci 2013: Production of referring expressions – bridging the gap between cognitive and computational approaches to reference*, pp. 1–6, Berlin, Germany.

Simon DOBNIK and John D. KELLEHER (2014), Exploration of functional semantics of prepositions from corpora of descriptions of visual scenes, in *Proceedings of the Third V&L Net Workshop on Vision and Language*, pp. 33–37, Dublin City University and the Association for Computational Linguistics, Dublin, Ireland.

Simon DOBNIK and John D. KELLEHER (2016), A Model for Attention-Driven Judgements in Type Theory with Records, in Julie HUNTER, Mandy SIMONS, and Matthew STONE, editors, *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue*, volume 20, pp. 25–34, New Brunswick, NJ USA.

Simon DOBNIK, John D. KELLEHER, and Christos KONIARIS (2014), Priming and Alignment of Frame of Reference in Situated Conversation, in Verena RIESER and Philippe MULLER, editors, *Proceedings of DialWatt – Semdial 2014: The 18th Workshop on the Semantics and Pragmatics of Dialogue*, pp. 43–52, Edinburgh.

David R DOWTY, Robert Eugene WALL, and Stanley PETERS (1981), *Introduction to Montague semantics*, D. Reidel Pub. Co., Dordrecht, Holland.

Zachary ESTES, Sabrina GOLONKA, and Lara L JONES (2011), Thematic Thinking: The Apprehension and Consequences of Thematic Relations, in Brian ROSS, editor, *The Psychology of Learning and Motivation*, volume 54, pp. 249–294, Burlington: Academic Press.

Ronald FAGIN, Joseph Y. HALPERN, Yoram MOSES, and Moshe Y. VARDI (1995), *Reasoning about knowledge*, MIT Press, Cambridge, Mass.

Christiane FELLBAUM (1998), *WordNet: an electronic lexical database*, MIT Press, Cambridge, Mass.

Gottlob FREGE (1948), Sense and Reference, *The Philosophical Review*, 57(3):209–230.

Klaus-Peter GAPP (1994a), Basic Meanings of Spatial Relations: Computation and Evaluation in 3D Space, in Barbara HAYES-ROTH and Richard E. KORF, editors, *AAAI*, pp. 1393–1398, AAAI Press/The MIT Press.

Klaus-Peter GAPP (1994b), A computational model of the basic meanings of graded composite spatial relations in 3D space, in *Advanced geographic data*

*modelling. Spatial data modelling and query languages for 2D and 3D applications (Proceedings of the AGDM'94)*, Publications on Geodesy 40, pp. 66–79, Netherlands Geodetic Commission.

Simon GARROD and Gwyneth DOHERTY (1994), Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions, *Cognition*, 53(3):181–215.

Stevan HARNAD (1990), The symbol grounding problem, *Physica D*, 42(1–3):335–346.

Annette HERSKOVITS (1986), *Language and spatial cognition: an interdisciplinary study of the prepositions in English*, Cambridge University Press, Cambridge.

Serena IVALDI, Sao Mai NGUYEN, Natalia LYUBOVA, Alain DRONIOU, Vincent PADOIS, David FILLIAT, Pierre-Yves OUDEYER, and Sigaud OLIVIER (2014), Object Learning Through Active Exploration, *IEEE Transactions on Autonomous Mental Development*, 6(1):56–72.

John D. KELLEHER, Fintan J. COSTELLO, and Josef VAN GENABITH (2005), Dynamically Structuring Updating and Interrelating Representations of Visual and Linguistic Discourse, *Artificial Intelligence*, 167:62–102.

John D. KELLEHER, Robert J. ROSS, Colm SLOAN, and Brian MAC NAMEE (2011), The effect of occlusion on the semantics of projective spatial terms: a case study in grounding language in perception, *Cognitive Processing*, 12(1):95–108.

Geert-Jan M. KRUIJFF, Hendrik ZENDER, Patric JENSFELT, and Henrik I. CHRISTENSEN (2007), Situated dialogue and spatial organization: what, where... and why?, *International Journal of Advanced Robotic Systems*, 4(1):125–138, special issue on human and robot interactive communication.

Lars KUNZE, Chris BURBRIDGE, and Nick HAWES (2014), Bootstrapping Probabilistic Models of Qualitative Spatial Relations for Active Visual Object Search, in *AAAI Spring Symposium 2014 on Qualitative Representations for Robots*, Stanford University in Palo Alto, California, US.

Shalom LAPPIN (2013), Intensions as Computable Functions, *Linguistic Issues in Language Technology*, 9:1–12.

Staffan LARSSON (2015), Formal semantics for perceptual classification, *Journal of Logic and Computation*, 25(2):335–369.

Daniel LASSITER (2011), Vagueness as probabilistic linguistic knowledge, in *Proceedings of the international conference on vagueness in communication (ViC'09)*, pp. 127–150, Springer-Verlag, Berlin, Heidelberg.

Stanislao LAURIA, Guido BUGMANN, Theocharis KYRIACOU, and Ewan KLEIN (2002), Mobile robot programming using natural language, *Robotics and Autonomous Systems*, 38(3–4):171–181.

Emilie L. LIN and Gregory L. MURPHY (2001), Thematic relations in adults' concepts, *Journal of experimental psychology: General*, 130(1):3–28.

Gordon D. LOGAN and Daniel D. SADLER (1996), A computational analysis of the apprehension of spatial relations, in Paul BLOOM, Mary A. PETERSON, Lynn NADEL, and Merrill F. GARRETT, editors, *Language and Space*, pp. 493–530, MIT Press, Cambridge, MA.

Didier MAILLAT (2003), *The semantics and pragmatics of directionals: a case study in English and French*, Ph.D. thesis, University of Oxford: Committee for Comparative Philology and General Linguistics, Oxford, United Kingdom.

Per MARTIN-LÖF (1984), *Intuitionistic Type Theory*, Bibliopolis, Naples.

Cynthia MATUSZEK, Nicholas FITZGERALD, Luke ZETTLEMOYER, Liefeng BO, and Dieter FOX (2012a), A joint model of language and perception for grounded attribute learning, in John LANGFORD and Joelle PINEAU, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, Edinburgh, Scotland.

Cynthia MATUSZEK, Evan HERBST, Luke ZETTLEMOYER, and Dieter FOX (2012b), Learning to Parse Natural Language Commands to a Robot Control System, in *Proceedings of the 13th International Symposium on Experimental Robotics (ISER)*.

George A. MILLER and Philip N. JOHNSON-LAIRD (1976), *Language and perception*, Cambridge University Press, Cambridge.

Richard MONTAGUE (1974), *Formal Philosophy: Selected Papers of Richard Montague*, Yale University Press, New Haven, ed. and with an introduction by Richmond H. Thomason.

Bengt NORDSTRÖM, Kent PETERSSON, and Jan M. SMITH (1990), *Programming in Martin-Löf's Type Theory*, volume 7 of *International Series of Monographs on Computer Science*, Clarendon Press, Oxford.

Martin J. PICKERING and Simon GARROD (2004), Toward a mechanistic psychology of dialogue, *Behavioral and Brain Sciences*, 27(2):169–190.

Matthew PURVER, Jonathan GINZBURG, and Patrick HEALEY (2003), On the means for clarification in dialogue, in *Current and new directions in discourse and dialogue*, pp. 235–255, Springer.

Terry REGIER and Laura A. CARLSON (2001), Grounding spatial language in perception: an empirical and computational investigation, *Journal of Experimental Psychology: General*, 130(2):273–298.

Deb ROY (2002), Learning visually-grounded words and syntax for a scene description task, *Computer speech and language*, 16(3):353–385.

Deb ROY (2005), Semiotic schemas: a framework for grounding language in action and perception, *Artificial Intelligence*, 167(1-2):170–205.

Kristoffer SJÖÖ (2011), *Functional understanding of space: Representing spatial knowledge using concepts grounded in an agent's purpose*, Ph.D. thesis, KTH, Computer Vision and Active Perception (CVAP), Centre for Autonomous Systems (CAS), Stockholm, Sweden.

Danijel SKOČAJ, Matej KRISTAN, Alen VREČKO, Marko MAHNIČ, Miroslav JANÍČEK, Geert-Jan M. KRUIJFF, Marc HANHEIDE, Nick HAWES, Thomas KELLER, Michael ZILLICH, and Kai ZHOU (2011), A system for interactive learning in dialogue with a tutor, in *IEEE/RSJ International Conference on Intelligent Robots and Systems IROS 2011*, San Francisco, CA, USA.

Luc STEELS and Tony BELPAEME (2005), Coordinating Perceptually Grounded Categories Through Language: A Case Study For Colour, *Behavioral and Brain Sciences*, 28(4):469–489.

Luc STEELS and Martin LOETZSCH (2009), Perspective Alignment in Spatial Language, in Kenny R. COVENTRY, Thora TENBRINK, and John. A. BATEMAN, editors, *Spatial Language and Dialogue*, Oxford University Press.

Mark TUTTON (2013), A new approach to analysing static locative expressions, *Language and Cognition*, 5:25–60.

Matthew E. WATSON, Martin J. PICKERING, and Holly P. BRANIGAN (2004), Alignment of reference frames in dialogue, in *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, Chicago, USA.

Joost ZWARTS and Yoad WINTER (2000), Vector Space Semantics: A Model-Theoretic Analysis of Locative Prepositions, *Journal of Logic, Language and Information*, 9:169–211.