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

Simon Dobnik*

Dept. of Philosophy, Linguistics & Theory of Science
Centre for Language Technology
University of Gothenburg, Sweden
simon.dobnik@gu.se
http://flov.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. Using examples of semantic representations of spatial descriptions we show how Type Theory with Records (TTR) satisfies these requirements. The advantage of truth being based on agent-relative judgements in TTR is crucial in this but practically it comes with a computational cost. We argue that the number of type judgements an agent has to make can be minimised by incorporating a cognitive notion of judgement that is driven by perceptual attention.

Keywords: spatial language, Type Theory with Records (TTR), attention driven judgements, computational framework

In the proposed presentation we overview and connect two lines of our work related to Type Theory with Records (TTR) [2, 3]: modelling of spatial language and cognition [8] and modelling attention-driven judgement [18].

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 [28], the type of objects referred to and their interaction [5], visual and discourse salience of objects [17], alignment in dialogue [32], and gesture [31] among others. Although the contribution of these contextual factors has been well studied in psychology, computational linguistics, computer science, geo-information science and robotics, several questions relating to representing their semantics and building formal computational models for situated agents still remain. These relate to (i) how an agent is able to determine *the sense and reference* of spatial descriptions; (ii) *grounding and information fusion* of contextual features into bundles of meaning representations; (iii) *bridging of perceptual and conceptual domains*; (iv) *formal accuracy and*

* I am grateful to Robin Cooper, Staffan Larsson and John D. Kelleher for discussion which has lead to significant changes in this paper.

sufficient expressiveness of representations for modelling human reasoning; (v) their *compositionality* with meaning representations of other words, sentences and utterances; (vi) their *adaptability and learnability* by an agent in new physical and conversational contexts.

In building situated conversational agents, several systems have been proposed but none of them capture all of these requirements. For example, *semiotic schemas* [29] account for the meaning of words that define perceivable entities and performable actions but it is not straightforwardly evident how they relate to other linguistic representations. [20] adopt a layered model with distinct representations at each layer. Although there exist mechanisms by which these representational levels interact, the kinds of representations at each level are quite distinct from each other and are shaped by different operations. The question we would like to address is whether such representational levels and operations can be generalised by taking inspiration from the way humans assign, learn and reason with meaning.

Classical formal semantics based on first order logic [11, 1] provides the required formal accuracy and expressiveness of a representation system. However, it mainly on explaining how meaning representations of words are composed to form meaning representations of sentences and does not address how meaning (both *sense* or *intension* and *reference* or *extension*) is learned and assigned in perception. The analyses of spatial descriptions are represented in first order logic such as: 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), see for example [26, 15]. The analysis tell us that the meaning of spatial descriptions is composed of several geometric primitives (surface/1, contiguous/2, boundary/1) but the meaning of these primitives is left un-accounted for. In model-theoretic semantics the expression's reference is determined by an assignment in a form of a valuation function between the 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 the coordinates of the locations where this spatial description applies, this referential representation of meaning is cumbersome as the model would have to represent for every scale, for every spatial relation, for every pair of objects. Note also that angles and distances in a coordinate system are continuous measures which means that such sets would be infinite. Furthermore, the way humans refer to space is vague (an object may be "near" another object depending on several contextual factors) and there is gradience of reference (some objects are "nearer" the landmark than others. Both vagueness and gradience of spatial language is 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) [25], potential fields capture the notion that such regions can be generalised as func-

tions [13, 28]. However, these functions do not represent objects in the model (or the extension or referential meaning of these descriptions) but rather capture their *sense* or *intension*: 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. In addition to angle and distance, several contextual parameters can be incorporated, for example the presence of distractor objects [4], object occlusion [19], etc. or the function itself can be learned from the dataset of perceptual observations and descriptions as a classifier [30, 7]. The notion of applying a function representing the meaning of words to the perceptual observations is also known as *grounding* these words in perception [14]. The grounded meanings of spatial descriptions or their senses can be thought of as bundles of several distinct yet interacting sources of information organised at different levels of conceptual abstraction, ranging from sub-conceptual perceptual information to contents of entire dialogue interactions in an information state of an agent [22].

Model-theoretic approach to semantics assumes that the model is given (derived through some external process), complete and represents a state of affairs at a particular temporal snapshot [12]. However, practically complete models may be rarely observable and we must deal with partial models. We must also account for the fact that the we may incrementally observe more and more of the world and we have to update the model with new observations, sometimes even correct the representation that we have already built in light of the new evidence. Finally, the world is not static itself as new new objects and events continuously 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 perception and motion and errors associated with both [6]. 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 a slightly different models. Of course, the more they experience the world, the more similar the 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 [27].

Type Theory with Records (TTR) builds on the tradition of the 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 perception 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. Types are intensional - that is, there can be distinct types which have

identical extensions. In this way sense is derived operationally as a computable function [21] or a classifier [23]. The notion of truth is linked to judgements that an object a is of type T ($a : T$). Under this view the type inventory is internal to an agent as types are learned and continuously refined by each agent as it encounters new situations [23]. Agents converge on sufficiently similar type representations which are a requirement for successful communication because they are part of the same perceptual and discourse contexts that impose external constraints on it, for example in terms of corrective feedback or the differences of what an agent expects to perceive and what it perceives. In such a situated dialogue learning scenario the TTR system can be also given a Bayesian probabilistic interpretation [3]. Because TTR relates perception directly to higher-level conceptual reasoning in a probabilistic way which allows modelling of gradience which makes it suitable for modelling semantics of spatial descriptions.

In contrast to the classical model-theoretic framework where types are used for the purpose of compositionality (the denotations of phrases are either model objects of basic types such as entities and truth values or functions composed from these types), TTR introduces an extended set of basic types (for example *Ind* and *Real* that correspond to basic human conceptual categories such as individuals and real numbers) and *a rich type system* which contains arbitrarily complex record types which are able to, among other things, express complex lexical semantics and dialogue information states. The proof objects of record types are records. Records and record types are similar to feature structures containing label-value pairs. The information expressed in types can be compared and reasoned about as the type systems allows *subtype* and *dependent type* relations. The ability of TTR to represent hierarchically organised multi-source information fulfils another requirement for modelling spatial descriptions.

In this presentation we discuss how our empirical investigations of learning geometric meanings of spatial descriptions with situated robots [7], learning functional meanings of prepositions from collections of image descriptions [9], and modelling of reference frame assignment in conversation [10] can be formulated in the TTR framework. For examples and details of TTR formulations refer to [8]. The overall goal is to provide an account of semantics of spatial prepositions for these modalities, and over the longer term, use the framework as a knowledge representation system of a situated agent.

This leads to a question how well as a semantic framework TTR is practically suited as a semantic representation layer for embodied agents. Humans are very flexible in assigning meaning and naturally we would like to preserve the same flexibility in our framework. New types can be created or learned by an agent dynamically. Furthermore, the record types allow us to construct the following relations between types which allow us to compare and reasoning about meaning:

- Intensionality/non-exclusivity of types: an object may belong to more than one type which may be structurally (nearly) an entirely different representation. For example, a sensory reading of a particular situation in the world involving spatial arrangement of objects may be assigned several record types

of spatial relations simultaneously, each with a unique internal structure: *Left*, *Near*, *At*, *Behind*, etc.
- A type may be a subtype of another type. An object judged as being of a particular type is also of all types that this type is a sub-type of: given that *Chair* is a sub-type of *Object*, a situation of type *Chair* is also of type *Object*.
- A type may be a component of another type. An object of the first type is partially matched with the second type. For example, a situation of type *Chair* is a component of the situation of *Table-Left-Chair*.
- A type may be a dependent type of another type. For example, the type *Left* is a dependent type of *Table-Left-Chair*. In order to judge a situation to be of type *Table-Left-Chair* one has to judge it to be of type *Left*.

Since each type assignment involves a binary judgement (something is of a type T or not) for each record of situation an agent having an inventory of n types can make n assignments. Learning what types a particular situation can be assigned involves 2^n possible outcomes, hence for $n = 3$, $2^3 = 8$: $\{\}$, $\{T_1\}$, $\{T_2\}$, $\{T_3\}$, $\{T_1, T_2\}$, $\{T_1, T_3\}$, $\{T_2, T_3\}$ and $\{T_1, T_2, T_3\}$, but if types are sub-types or dependent of another the number of judgements could be reduced.

We argue that agents such as situated robots need (i) a judgement control mechanism and (ii) a method for organising their type inventory [18]. For (i) we propose the Load Theory of selective attention and cognitive control [24] to be a suitable candidate. This model of attention distinguishes between two mechanisms of selective attention: *perceptual selection* and *cognitive control*. Following this theory, we argue that type judgements can be grouped into three categories: (i) pre-attentive, (ii) task induced, and (iii) context induced judgements. An agent makes the first kind of judgements continuously, but varying strategies depending on its cognitive load. The other two kinds of judgements are primed by the task and the physical context that the agent is engaged with. We propose that agents organise their inventory of types that fall under (ii) and (iii) into subsets or bundles (computationally they can be modelled as lists) that are associated with the agent's cognitive states. The states can be modelled as Partially Observable Markov Decision Processes (POMDPs, [16]) and can be thought of as sensitivities towards certain objects, events, and situations. The states of a POMDP network are connected by actions, in this case the priming policies (ii) and (iii). The types an agent actually perceives in each state represent the observations for each state (note that here we are not dealing with learning new types). The model ensures that observing certain types at a particular state primes the agent to observe particular other types in the states following it. Hence, past experience primes the agent to observe new situations. The reward function is governed by the benefit of an agent being primed to perceive the world this way.

In this paper we outlined an application of type theory to natural language semantics and demonstrated how the TTR framework allows us to relate the semantics to action, perception and cognition. Furthermore, on the example of spatial descriptions we argued tat natural language, perception and cognition put high demands on the expressiveness of the type theoretic framework which

is associated with high computational cost. In order to counter this, we proposed a method to limit the possible type judgements of an agent by separate cognitive attentual mechanism which we will be testing in practical implementations with situated agents in our forthcoming work.

References

1. Blackburn, P., Bos, J.: Representation and inference for natural language. A first course in computational semantics. CSLI Publications (2005)
2. Cooper, R.: Type theory and semantics in flux. In: Kempson, R., Asher, N., Fernando, T. (eds.) Handbook of the Philosophy of Science, General editors: Dov M Gabbay, Paul Thagard and John Woods, vol. 14. Elsevier BV (2012)
3. Cooper, R., Dobnik, S., Lappin, S., Larsson, S.: A probabilistic rich type theory for semantic interpretation. In: Cooper, R., Dobnik, S., Lappin, S., Larsson, S. (eds.) Proceedings of the EACL 2014 Workshop on Type Theory and Natural Language Semantics (TTNLS). pp. 72–79. Association for Computational Linguistics, Gothenburg, Sweden (27 April 2014)
4. Costello, F.J., Kelleher, J.D.: Spatial prepositions in context: the semantics of near in the presence of distractor objects. In: Proceedings of the Third ACL-SIGSEM Workshop on Prepositions. pp. 1–8. Prepositions '06, Association for Computational Linguistics, Stroudsburg, PA, USA (2006)
5. Coventry, K.R., Prat-Sala, M., Richards, L.: The interplay between geometry and function in the apprehension of Over, Under, Above and Below. Journal of Memory and Language 44(3), 376–398 (2001)
6. Dissanayake, M.W.M.G., Newman, P.M., Durrant-Whyte, H.F., Clark, S., Csorba, M.: A solution to the simultaneous localization and map building (SLAM) problem. IEEE Transactions on Robotic and Automation 17(3), 229–241 (2001)
7. Dobnik, S.: 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 (September 4 2009)
8. Dobnik, S., Cooper, R., Larsson, S.: Type Theory with Records: a general framework for modelling spatial language. In: Dobnik, S., Cooper, R., Larsson, S. (eds.) Proceedings of The Second Workshop on Action, Perception and Language (APL'2). The Fifth Swedish Language Technology Conference (SLTC), Uppsala, Sweden (13 November 2014)
9. Dobnik, S., Kelleher, J.: 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 (August 2014)
10. Dobnik, S., Kelleher, J.D., Koniaris, C.: Priming and alignment of frame of reference in situated conversation. In: Rieser, V., Muller, P. (eds.) Proceedings of DialWatt - Semdial 2014: The 18th Workshop on the Semantics and Pragmatics of Dialogue. pp. 43–52. Edinburgh (1–3 September 2014)
11. Dowty, D.R., Wall, R.E., Peters, S.: Introduction to Montague semantics. D. Reidel Pub. Co., Dordrecht, Holland (1981)
12. Fagin, R., Halpern, J.Y., Moses, Y., Y. Vardi, M.: Reasoning about knowledge. MIT Press, Cambridge, Mass. (1995)
13. Gapp, K.P.: Basic meanings of spatial relations: Computation and evaluation in 3d space. In: Hayes-Roth, B., Korf, R.E. (eds.) AAAI. pp. 1393–1398. AAAI Press/The MIT Press (1994)

14. Harnad, S.: The symbol grounding problem. Physica D 42(1–3), 335–346 (June 1990)
15. Herskovits, A.: Language and spatial cognition: an interdisciplinary study of the prepositions in English. Cambridge University Press, Cambridge (1986)
16. Kaelbling, L.P., Littman, M.L., Cassandra, A.R.: Planning and acting in partially observable stochastic domains. Artificial intelligence 101(1), 99–134 (1998)
17. Kelleher, J., Costello, F., van Genabith, J.: Dynamically structuring updating and interrelating representations of visual and linguistic discourse. Artificial Intelligence 167, 62–102 (2005)
18. Kelleher, J.D., Dobnik, S.: A model for attention-driven judgements in Type Theory with Records. In: Kempson, R., Purver, M. (eds.) Proceedings of the Worshop on Interactive Meaning Construction at the International Workshop on Computational Semantics (IWCS 2015). pp. 13–14. Queen Mary University of London (14 April 2015)
19. Kelleher, J.D., Ross, R.J., Sloan, C., Namee, B.: 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 (2011)
20. Kruijff, G.J.M., Zender, H., Jensfelt, P., Christensen, H.I.: Situated dialogue and spatial organization: what, where... and why? International Journal of Advanced Robotic Systems 4(1), 125–138 (2007), special issue on human and robot interactive communication
21. Lappin, S.: Intensions as computable functions. Linguistic Issues in Language Technology 9, 1–12 (2013)
22. Larsson, S.: Issue-based Dialogue Management. Ph.D. thesis, University of Gothenburg. (2002)
23. Larsson, S.: Formal semantics for perceptual classification. Journal of Logic and Computation online, 1–35 (December 18 2013)
24. Lavie, N., Hirst, A., de Fockert, J.W., Viding, E.: Load theory of selective attention and cognitive control. Journal of Experimental Psychology: General 133(3), 339–354 (2004)
25. Logan, G.D., Sadler, D.D.: A computational analysis of the apprehension of spatial relations. In: Bloom, P., Peterson, M.A., Nadel, L., Garrett, M.F. (eds.) Language and Space, pp. 493–530. MIT Press, Cambridge, MA (1996)
26. Miller, G.A., Johnson-Laird, P.N.: Language and perception. Cambridge University Press, Cambridge (1976)
27. Pickering, M.J., Garrod, S.: Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences 27(2), 169–190 (2004)
28. Regier, T., Carlson, L.A.: Grounding spatial language in perception: an empirical and computational investigation. Journal of Experimental Psychology: General 130(2), 273–298 (2001)
29. Roy, D.: Semiotic schemas: a framework for grounding language in action and perception. Artificial Intelligence 167(1-2), 170–205 (Sep 2005)
30. Roy, D.K.: Learning visually-grounded words and syntax for a scene description task. Computer speech and language 16(3), 353–385 (2002)
31. Tutton, M.: A new approach to analysing static locative expressions. Language and Cognition 5, 25–60 (3 2013)
32. Watson, M.E., Pickering, M.J., Branigan, H.P.: Alignment of reference frames in dialogue. In: Proceedings of the 26th Annual Conference of the Cognitive Science Society. Chicago, USA (2004)