Department of Computer Science,
University of Otago


---

Technical Report OUCS-2003-04

# Grounding syntactic representations in an architecture for sensorimotor control

**Alistair Knott**
Department of Computer Science

---

Department of Computer Science,
University of Otago, PO Box 56, Dunedin, Otago, New Zealand

http://www.cs.otago.ac.nz/trseries/

# Grounding syntactic representations in an architecture for sensorimotor control

Alistair Knott
Dept of Computer Science, University of Otago

## 1 Introduction

The aim of this paper is to suggest a way of characterising the syntactic structure of a sentence as a description or trace of a sensorimotor process. My starting point is the idea (see also Hurford, to appear; Jackendoff, 2002) that the logical form of a sentence should be thought of not as a direct representation of the world, but rather as an evocation of a cognitive process which interfaces with the world via various perceptual and motor mechanisms. I will argue, departing from Hurford and Jackendoff, that the structure of this evoked cognitive process has extensive reflexes in the syntax of sentences as well as in their semantics. My specific proposal is that syntactic representations can be seen as reflecting operations in a model of sensorimotor control. This focus on motor representations is similar to that of Rizzolatti and Arbib (1998). However, the sensorimotor model I present is different from theirs in a number of respects. And I want to describe how this model relates to syntactic structures at a greater level of detail, making reference to some well-established and well-motivated syntactic representations.

## 2 Sensorimotor model

The sensorimotor model I propose (see Figure 1) is a representation of the cognitive processes involved in executing a simple transitive action, such as reaching for a target object with the hand. My main aim is to link together some recent models of these processes. Figure 1 has three parts; each part features a computational model adapted from the current literature in psychology and neuroscience.

Part 1 is a model of visual attention and its relationship to visual object categorisation. It provides a very broad-brush picture of visual processing in the dorsal and ventral visual pathways. In the first stage (arc **1**), information from the retina is converted into a set of retinotopic feature maps, with units in each map coding for the presence of simple visual features such as lines of particular orientations and spatial frequencies. These maps are used in two different ways. In the dorsal stream (arc **2**) they are used to construct (in parallel) a map of the most salient visual stimuli on the retina, essentially by summing information about the degree of local contrast at each point in the retina in each feature map (see e.g. Itti and Koch, 2001). This retinal saliency map is the input (via arc **3**) to a 'winner-take-all' network which identifies the most salient point on the retina (see again Itti and Koch). Information in the retinal feature maps also serves as input (via arc **4**) to the ventral stream, whose function is to determine the categories of objects present on the retina. Classification is assumed to occur through progressively more complex combinations of simple visual features, by units which have progressively wider receptive fields (see e.g. Le Cun and Bengio, 1995; Riesenhuber and Poggio, 1999). Importantly, while this classification system receives input in parallel from across the retina via arc **4**, this input is gated by the most salient retinal position (arc **5**). Mozer and Sitton (1996) have a computational model of this gating relationship, and there is considerable evidence for the gating relationship in neuroscience; see e.g. Moran and Desimone (1985).

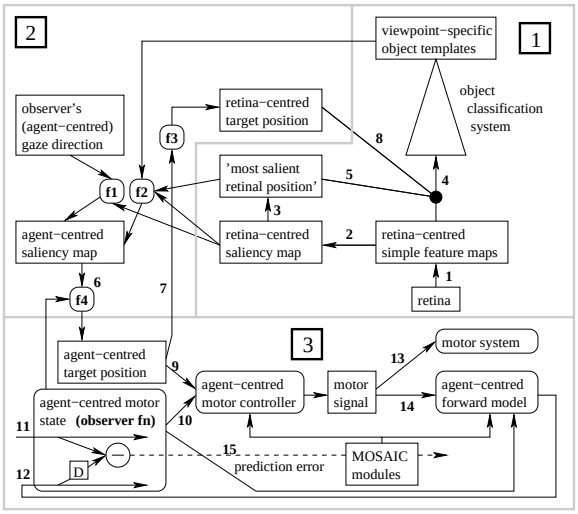


Figure 1: Sensorimotor circuit diagram. (Functions are in round boxes; data are in square boxes.)

Part 2 is a model of the creation of a saliency map using a coordinate system centred on the agent who is observing the world (more accurately, on the torso of this agent). This representation is just like the retinal saliency map, except that it is stable over the observer's eye and head movements. The function which creates this representation (**f1**) takes as input the retinal saliency map, and a model of the observer's (torso-centred) gaze direction. (See Treisman, 1998 for arguments for the existence of this representation.) This map is also the input to an attentional function (**f4**, via arc **6**), this time to select a target object in the agent's perispace. **f4** also takes as input the agent's current motor state, to deliver the kind of 'action-centred' representations described by Tipper *et al.*, 1992; 1998, in which objects compete in virtue of their closeness to the starting position of the hand which will reach for them. Finally, the selected target position is converted back into retina-centred coordinates (via **f3**, arc **7**), and the retina-centred version acts as another gate on the input to the object classification system (arc **8**). The gating scheme allows input from the retina *either* at the position of the most salient retinal stimulus, *or* at the position of the target object. This second gating mechanism provides a way of modelling the finding of Deubel *et al.* (1998) that visual object identification is speeded if the object to be identified is also the target of a reaching action.

Another possibility is that the observer can produce an agent-centred saliency map centred not on his own body, but on that of an *observed* agent. If the observer is watching this agent, then presumably the position of the agent emerges as the most salient retinal position, and the object recognition system will produce a template associated with the observed agent. We can then envisage a function (**f2**) which delivers a saliency map centred on this agent using as input the most salient retinal position, the full retinal saliency map, and a viewpoint-specific object template to provide the current orientation of the agent. Again there is considerable evidence for this idea; see e.g. Perrett *et al.* (1989); Jellema *et al.* (2000) for compelling evidence both of viewpoint-specific perceptual representations of animate agents and of frames of reference centred on observed agents.

Part 3 is a model of the motor controller proper. I draw mainly on the MOSAIC model of Wolpert and Kawato (1998; Haruno *et al.*, 2001) here. A controller function takes two inputs (a goal motor state and a current motor state) and generates a motor signal (ideally, one which causes the current motor state to become closer to the goal state at the following moment in time). In the present case, the goal input is the agent-centred target position (delivered via arc **9**), and the current motor state is the agent-centred position of the agent's arm (delivered by an **observer function**, via arc **10**). This function takes as input proprioceptive information about current arm position (arc **11**), but also information derived from an internal **forward model** of arm position and velocity (arc **12**). This internal model is kept updated by **efferent copies** of motor signals (see e.g. Wolpert *et al.*, 1995). The motor signal output by the controller is passed to the motor system, naturally (arc **13**) and an action ensues. However, it is also passed to a forward model of the motor system (arc **14**), which *predicts* what the next arm state will be, given the current motor signal and the current arm state. Forward models in fact have two roles in MOSAIC. A final important feature of the system is that it actually uses a *set* of controller-forward model pairs, rather than the single pair shown in Figure 1. An agent needs different controller functions in different circumstances, even for a given current state/goal position pair. For one thing, different controllers are required to execute different motor programmes (e.g. *touch* vs *push*) in relation to a target position. In addition, the hand to be moved to the target might already be carrying an object, considerably changing the arm's inertial properties. The solution in MOSAIC is to develop a set of **modules** for controlling action, each consisting of a controller/forward-model pair tailored to a particular circumstance. At any time, the modules which are used are those most suited to the current circumstances. Suitability is decided by computing the **prediction error** of each module—the difference between the current motor state and the module's prediction about the state made at the previous time point. The modules with the lowest prediction errors are activated for the next iteration. The influence of prediction error (which is computed within the observer) on module activations while an action is under way is reflected in arc **15**.[1]

# 3 Syntactic structure

My aim is now to relate the above sensorimotor model to syntactic representations. I will assume a

[1] This description of MOSAIC assumes the agent is the the observer. But there is good evidence that the mechanism by which we recognise actions in others uses the same representations that are used to control our own actions (see e.g. Rizzolatti *et al.*, 2000; Knoblich and Flach, 2001), and I am currently exploring the idea that the MOSAIC model can be extended to function as a mechanism for 'biological motion recognition'.

fairly standard version of GB/Minimalism, in which the syntactic structure of a transitive clause appears as in Figure 2 (the terms in bold face are syntactic constituents). The analysis I am assuming incorporates two ideas associated with Minimalism: the VP-internal subject hypothesis of Koopman and Sportiche (1991), and the $\text{Agr}_o$P hypothesis of Pollock (1989). In a Minimalist analysis, subject and object DPs are generated within VP, and raise to [Spec,TP] and [Spec,$\text{Agr}_o$P] respectively to get case. V originates within VP, and (in finite sentences) raises successively to $\text{Agr}_o$ and T.

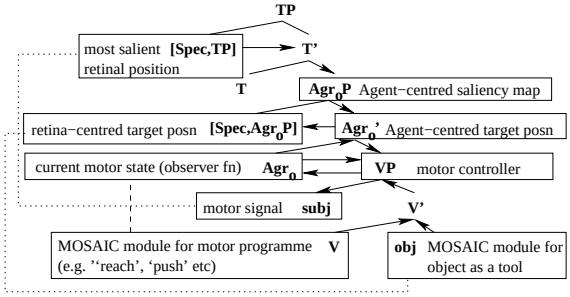


Figure 2: Syntactic structure of a transitive clause, with the proposed sensorimotor interpretation

As can be seen from Figure 2, my main suggestion is that syntactic constituents can be associated with sensorimotor representations or sensorimotor functions. Given my starting assumption that the logical form of a sentence is a cognitive process rather than a representation of the world, it seems appropriate to use the semantic term *denotation* to express this association. Specifically, I propose that a syntactic constituent denotes an episode of activity within a sensorimotor representation or function of the kind that appear in Figure 1.

In a conventional Tarskian/Montagovian compositional semantics, the denotation of a sentence is a function of the denotations of the lexical items it contains. Syntax specifies the order in which these atomic denotations are to be combined together, with the denotation of each nonterminal node in the tree being built from the denotations of its daughter nodes. Clearly if denotations are episodes of sensorimotor processing, different notions of semantic composition and hierarchical constituent structure must be supplied. There does indeed seem to be room for alternatives. Note that there is no reason why the notion of compositionality should imply that information from lower consituents is 'passed up' towards the root node of a sentence, where it is all gathered in one place. In fact, this localist assumption seems quite out of place if the aim is to ground meaning in psychological processes. A system whose visible behaviour occurs as a result of interactions between several partly-autonomous modules can also be understood as compositional in exactly the sense we need for language, without any assumption of locality.

Nonetheless, hierarchical syntactic structure clearly plays a critical role in specifying how complex meanings are built from from simpler ones, whether we understand these meanings as localised or not. We still need to find a sensorimotor interpretation for hierarchical syntactic structure which explains why this is so. My proposal, which again should already be visible in Figure 2, is that composition is grounded in a notion of *links or couplings between sensorimotor processes*. The basic idea is that nodes which are close together in the syntactic structure denote representations which directly influence one another. Accordingly, the sensorimotor denotations of syntactic nodes in the in Figure 2 are arranged so that the resulting diagram is a strict subgraph of the circuit diagram in Figure 1.

Naturally, evidence is needed (and lots of it) that the mapping between a GB-style syntactic structure and a sensorimotor wiring diagram is not just a topological coincidence—especially given how much room for manoeuvre there is in the independent specification of the syntactic and sensorimotor models which are being brought together. In the paper for which this is the abstract, I will provide sensorimotor interpretations for several other syntactic constructs, which are consistent with the sensorimotor conception of constituency just outlined. These are summarised below.

Firstly, I will give a characterisation of the notion of hierarchical position in a syntax tree in terms of the order in which representations become active during a sensorimotor action, with hierarchically high constituents becoming active before hierarchically lower ones. On the basis of this, I will give a structural account of the distinction between subject ([Spec,TP]) and object ([Spec,$\text{Agr}_o$P]) positions.

Secondly, I will characterise the phenomenon of DP-movement in relation to the sensorimotor model. The idea is basically that the syntactic positions in an A-chain (the ones connected by the dotted lines in Figure 2) denote the same sensorimotor process, and that the appearance of this representation at two points in the syntactic structure is due to the process having an influence in two distinct sensorimotor circuits.

Finally, I will look in some detail at the sensori-

motor characterisation of the $Agr_o$ node. The original motivation for $Agr_o$ (Pollock, 1989) emerges from a discussion of the factors affecting verb movement to Infl. The basic phenomenon is that verb movement out of VP occurs more readily in nonfinite sentences, and with the verbs *have* and *be*. I will suggest that movement of V to $Agr_o$ (the dashed line in Figure 2) reflects the linkage between the observer function and the set of active MOSAIC modules via which motor learning occurs. I will argue that motor learning occurs in finite sentences and not in nonfinite ones; and also that sentences using *be* and *have* do not involve the MOSAIC controller at all, which explains the additional freedom these verbs have in nonfinite sentences.

# References


Deubel, H., Schneider, W., and Paprotta, I. (1998). Evidence for a common attentional mechanism in reaching and perception. *Visual Cognition*, **5**, 81–107.

Haruno, M., Wolpert, D., and Kawato, M. (2001). MOSAIC model for sensorimotor learning and control. *Neural Computation*, **13**, 2201–2220.

Hurford, J. (to appear). The neural basis of predicate-argument structure. *Behavioral and Brain Sciences*.

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews—Neuroscience*, **2**, 1–11.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press, Oxford.

Jellema, T., Baker, C., Wicker, B., and Perrett, D. (2000). Neural representation for the perception of the intentionality of actions. *Brain and Cognition*, **44**, 280–302.

Knoblich, G. and Flach, R. (2001). Predicting the effects of actions: interactions of perception and action. *Psychological Science*, **12**(6), 467–472.

Koopman, H. and Sportiche, D. (1991). The position of subjects. *Lingua*, **85**, 211–258.

Le Cun, Y. and Bengio, Y. (1995). Convolutional networks for images, speech, and time series. In M. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 255–258. MIT Press, Cambridge, MA.

Moran, J. and Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, **229**, 782–784.

Mozer, M. and Sitton, M. (1996). Computational modeling of spatial attention. In H. Pashler, editor, *Attention*, pages ??–?? UCL Press, London.

Perrett, D., Harries, M., Bevan, R., Thomas, S., Benson, P., Mistlin, A., Chitty, A., Hiatenen, J., and Ortega, J. (1989). Frameworks of analysis for the neural representation of animate objects and actions. *Journal of Experimental Biology*, **146**, 87–113.

Pollock, J.-Y. (1989). Verb movement, universal grammar and the structure of IP. *Linguistic Inquiry*, **20**(3), 365–424.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, **2**, 1019–1025.

Rizzolatti, G. and Arbib, M. (1998). Language within our grasp. *Trends in Neurosciences*, **21**, 188–194.

Rizzolatti, G., Fogassi, L., and Gallese, V. (2000). Cortical mechanisms subserving object grasping and action recognition: a new view on the cortical motor functions. In M. Gazzaniga, editor, *The new cognitive neurosciences*, pages 539–552. MIT Press.

Tipper, S., Lortie, C., and Baylis, G. (1992). Selective reaching: Evidence for action-centred attention. *Journal of Experimental Psychology: Human Perception and Performance*, **18**, 891–905.

Tipper, S., Howard, L., and Houghton, G. (1998). Action-based mechanisms of attention. *Philosophical Transactions of the Royal Society of London B*, **353**, 1385–1393.

Treisman, A. (1998). The perception of features and objects. In R. Wright, editor, *Visual attention*, pages 26–54. Oxford University Press, Oxford.

Wolpert, D. and Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, **11**, 1317–1329.

Wolpert, D., Ghahramani, Z., and Jordan, M. (1995). An internal model for sensorimotor integration. *Science*, **269**, 1880–1882.