NOTES FOR THE CAMBRIDGE WORKSHOP ON *THE SPEECH, LANGUAGE AND HUMAN-COMPUTER INTERACTION GRAND CHALLENGE*

GERRY ALTMANN, YORK, UK.
g.altmann@psych.york.ac.uk

These notes are a little hurried (and long, lifted from other proposals I've written!), but they give a flavour of a component that I think would be worthwhile. The general theme is on how we map language onto mental representations of the external world, and my hope would be that we could work into the Grand Challenge the the work described below under 'Background' (continuation thereof), 'Modeling', and 'Imaging'. Or, at least, the background and imaging. I have provided some further details (as a kind of appendix) of the Modeling and Imaging work, but not of how the Background work could be extended – that stretches to several pages!

BACKGROUND. It has recently become apparent that there is an extremely tight coupling between language and vision, as evidenced by the speed with which eye movements around a visual scene can be mediated by language (Allopenna et al., 1998; Altmann & Kamide, 1999). For example, on hearing '*the man will taste..*' while viewing a scene portraying a man, a child, a beer, and some sweets, the eyes are directed towards the beer *during* the verb '*taste*'. When the sequence is '*the child will taste*', the eyes are directed towards the sweets instead (no such pattern is observed when the verb is replaced by '*ride*'; Kamide et al., 2003). These *anticipatory* eye movements reflect the rapid integration of various sources of information; the meaning of the verb, the meaning of its grammatical subject ('*the man/girl*'), the syntactic knowledge that determines that the man or girl will do the tasting, and the real-world knowledge about what in the scene is likely to be tasted given who in the scene is to do the tasting. Monitoring eye movements is a non-obtrusive, real-time, and direct measure of how language interpretation directs attention to the external world. By monitoring shifts in attention under different linguistic (and visual) conditions, we can infer what knowledge becomes available when, and how this knowledge is used, as the language unfolds. These shifts in attention even occur, and with the same timing, when the scene *is removed* before the sentence is heard, in which case the eyes move to where the objects *had been* (Altmann, 2004). More surprising, perhaps, is that if a preamble is presented such as *'the barman will move the beer to the other side of the bar. The man will then taste the beer..*', the eyes move after '*taste*' to where the participant has heard the beer will move to (the other side of the bar), even though the scene remains unchanged and shows the beer at its original location (Altmann & Kamide, in press; submitted). Finally, Huettig & Altmann (in press) found that eye movements during a target word were directed not simply towards the target itself but also to 'category competitors' ('*piano*' engendered more looks to a trumpet than to unrelated distractors, even when a piano was present also), to 'form competitors' ('*pen*' engendered looks to a needle), and to 'colour competitors' ('*frog*' engendered looks to a green jacket). The degree of 'conceptual overlap' (determined using McRae's published norms) predicted the proportion/duration of looks to the competitor. These observations form the basis for the research outlined below. Its aims are to understand how language is 'mapped' onto the external world (or its internal representation).

MODELING SITUATED LANGUAGE. In previous work, we have used a modified SRN architecture to model how information in one domain (e.g. language) can be mapped onto information in another (e.g. visual). These studies modeled behavioural data from both adults (Altmann et al., 1995; Dienes et al., 1999) and infants (Altmann & Dienes, 1999; Altmann, 2002). The success of these models was due to a representational substrate that was shared between the two domains. Indeed, we believe that the linguistic mediation of visual attention is a consequence of a common representational substrate serving both interpreted representations of the visual world and interpreted representations of linguistic structures. We propose to simulate, in a neural network, the relationship between linguistic and visual inputs and attentional control, and specifically, within the context of modeling the development of this relationship. Distinctions such as those between category membership and co-occurrence relations, within and across modalities, will form a central part of the modeling. The models will enable us to make precise predictions regarding the deployment at different stages of development of different kinds of information (about physical form, function, or category membership). They will also enable predictions regarding the timing with which distinct kinds of information are applied during the unfolding of individual words, thereby feeding into our work on the relationship between lexical activation and contextual integration. The aim will be to gain a deeper understanding of the computational principles which support this development. Tyler and colleagues (e.g. Randall et al., 2004), as well as McRae and others (e.g. McRae et al., 1997), have demonstrated the efficacy of computational models in respect of understanding some of the principles which underlie dissociations between different kinds of conceptual knowledge. And Elman and colleagues have demonstrated their efficacy in respect of understanding the developmental emergence of categories (e.g. Elman et al., 1996). The purpose of the modeling is to instantiate the principles we believe underlie our human empirical work on 'situated language'; and to develop testable predictions of when different kinds of information are deployed as a sentence unfolds (and under what conditions they may not be deployed).

IMAGING SITUATED LANGUAGE. We believe that the interpretation of visual scenes and the interpretation of sentences that may describe those scenes may share a common representational substrate. Human neuroimaging can shed light on aspects of this common substrate. MEG studies of manual manipulation (and the temporal dynamics of activity in Brodmann's area 44 relative to primary left and right motor cortex) suggest that BA44 is implicated in the human 'mirror-neuron' system – neurons that fire when an activity is performed but also when that activity is passively observed (Nishitani & Hari, 2000; see also Hari et al., 2000; Rizzolatti & Arbib, 1998). If the purpose of language is to cause the re-enactment of the experiential states associated with directly observing an event (cf. Barsalou et al., 2003), it would be surprising if language did not activate the mirror neuron system. We note that BA44 and BA45 constitute Broca's area, an area implicated in thematic processing (establishing on the basis of syntactic structure the roles played by the different participants in the event described by the sentence) – see Friederici (2002) for review. Below, we propose an investigation of the relationship between language processing, event perception, and the human mirror neuron system. We shall take advantage of recent work by Glenberg & Kaschak (2002), who have developed a set of sentence stimuli which exhibit strong motoric components: '*the man will turn on the tap*' has a measurable inhibitory effect on manual responses towards the body (because turning on a tap involves moving the hand *away* from the body). We describe below the use of MEG and co-registered fMRI to track the time course of the development of the (pre)motor response assumed to underpin this effect. We shall contrast these sentences with others that entail no motor component ('*the man thought about the idea*'), thereby dissociating the grammatical processing of a sentence from its motoric embodiment. We shall explore also the brain's response to video segments that portray the same events described by the (imageable) sentences. Our aim is to understand further the involvement of the human mirror neuron system, and Broca's area, in situated language processing. The data will inform theories of 'embodied cognition' — the grounding of cognition, and in this case, of combinatorial semantics and event representations, in the neural substrates that support interaction with the external world. If there is a common representational substrate that underpins both language and visual scene interpretation, the human mirror neuron system is a prime candidate for being a part of that substrate (at least in respect of the interpretation of particular kinds of event).

APPENDIX *(i.e. further details on Modeling and Imaging ideas)*

IMAGING

If the representations that underpin the interpretation of real-world events also underpin situated language processing, can neuroimaging shed light on the commonalities? Certain actions (e.g. cutting, which involves tools such as scissors) tend to be accompanied by activity in premotor cortex and the middle temporal gyrus, and by gamma-band responses in the 30Hz range over motor cortex (e.g. Pulvermüller et al, 1999). MEG studies of manual manipulation suggest that Brodmann's area 44 is a part of the human 'mirror-neuron' system (Nishitani & Hari, 2000). The fact that the mirror neuron system is implicated in BA44 is important given the psycholinguistic and neuropsychological data that implicates Broca's area (BA44/45) in thematic processing. If parts of Broca's area are implicated in the mirror neuron system, perhaps this same area supports thematic processing precisely because such processing provides a representational 'mirror' of the event being described (cf. 'simulation'; e.g. Barsalou, 2003). Our work on establishing common components in the neural activity evoked by sentences, (static) scenes, and movies, will provide valuable data in respect of this issue. We shall contrast sentence stimuli developed by Glenberg & Kaschak (2002), which exhibit strong motoric components (e.g. '*the man will turn on the tap*') with sentences that do not ('*the man will think about the idea*'). Both involve thematic processing, but only the former entail motor activity, and they do so because of their combinatorial semantics that affords the direction of motion (no one word determines it; we thus intend to contrast with cases where e.g. '*push*' and '*pull*' are used, thereby implicating lexical semantics). We shall thus use MEG to track the time course of the development of the (pre)motor response, and co-register with fMRI to localize the response. One question we shall ask is how the neural activity (both in terms of location and, importantly, time-course) varies as a function of hearing the sentence alone, or of hearing the *same* sentence but after previously viewing a video segment that depicts the action that the sentence describes. In the sentence-only condition, the mental representation will presumably be more abstract than in the prior-video condition, which will entail an episodic component (presumably implicating the hippocampal system). This design allows the *same* stimulus to be observed in two very different conditions, and aside from differences in involvement of the hippocampal system, it will be informative to explore what other differences result, both in terms of location and time-course as a function of condition. By collecting data during the prior video, we can compare activity in BA44 to the language alone, to the video alone, and to their combination. The data will inform theories of how cognition, and in this case, combinatorial semantics and event representation, are grounded in the neural substrates that support interaction with, and memory of, the external world.

MODELING

Computational modeling will enable precise predictions regarding the deployment at different stages of development of different kinds of knowledge, and the timing with which distinct kinds of information are applied as a sentence, or word, unfolds. Our previous research employed a Simple Recurrent Network (SRN) to map information from one modality onto information from another, even when the two modalities were never concurrent (Altmann, 2002; Altmann & Dienes, 1999; Dienes et al., 1999). The network had to predict at its output layers whatever the next input would be. It would learn about structure in one domain, and when given structure in a second domain, would encode the new structure in the same representational substrate that encoded the earlier structure. If structure in one domain correlated with structure in the other, these structures would be encoded together in the representational substrate. We shall investigate how a network can learn to map linguistic input onto a visual world, when that world is abstracted, by the network itself, across 'retinal' input, and how the emergence of abstract representations in one domain is mediated by the emergence of abstract representations in the other. **OBJECT LEARNING** Each object will be represented as a pattern of activity across a set of input units that instantiate a 'retina' (cf. Plunkett et al., 1992; we intend a more simplistic 'course-coding' of the retina). Learning about an object will consist in the network having to predict where the object will be at the next moment in time (a tracking task; cf. Mareschal et al., 1999). After training on individual objects, we shall introduce multiple objects 'interacting' (e.g. one object hitting another, or pushing another, and so on). **VOCABULARY AND SENTENCE LEARNING** Object names will be instantiated across 'language' units. We shall explore different 'linguistic' training regimes (cf. Plunkett et al., 1992): We shall, for example, interleave object learning with sentence-event pairings (noun-verb-noun sequences will be paired with scenes portraying interacting objects; given the limits on the information an SRN can practicably hold, simulations will be limited to 3-word sequences). We shall explore also the consequences of completing object learning to criterion before introducing linguistic input (see Altmann, 2002, for how to prevent unlearning of earlier learned structure). In these simulations, object-label pairings will be interleaved with 'sentence-event' pairings, or sentential training (in the absence of the visual input) will be interleaved with sentence-event pairings. We shall use statistical techniques to explore the nature of the abstractions that the network forms as a function of training regime. **PLAUSIBILITY AND THE LANGUAGE/WORLD MAPPING** We shall run simulations in which the events and corresponding sentences differ in frequency; a square bumping a circle may be more likely than a square bumping a triangle. Following Elman (1990), we anticipate that '*square bump*' will output a composite representation in which '*circle*' is more dominant than '*triangle*'. We aim to model the anticipatory eye movements observed in the human data (outputting a visual representation of what will most likely be bumped, with appropriate positional information). We shall simulate also the 'blank screen' data described earlier (the network should output a visual representation of e.g. the circle in the position in which it had last been 'seen'). **SIMULATING AND PREDICTING PRIMING DATA** We shall use the model to explore predictions regarding the relative efficacy of different kinds of prime (associative vs. category) in the visual priming paradigm. When, in the development of the network, does category priming emerge relative to associative (co-occurrence) priming? What factors might prevent such emergence? Do these predict the child data (SERIES 9)? **ENCODING MULTIPLE LEVELS** Finally, we shall modify the network to enable unfolding sequences of *phonetic input* to be mapped onto the units mediating between the 'linguistic' and 'visual' domains. The aim here will be to explore the influence of object learning, when concurrent with sentence-event pairing, on the emergence of word-like representations, and to enable predictions regarding the timing with which phonetic information is integrated with 'higher-level' information within the system.

The modeling is intended to show how certain behaviours may be founded on certain computational and statistical principles; we make no claim regarding the validity of the training in respect of infant learning, vocabulary acquisition, or grammatical development. SRNs as currently implemented cannot learn large-scale realistic lexicons, or grammars. But they will help us understand better the computational principles that enable integration of information across domains, and will help formulate testable hypotheses. The modeling will have consequences for theories of early acquisition, and how the emergence of structure (and categories) in one domain may influence the emergence of structure (and categories) in another. It will allow us to determine, in theory, whether the linguistic mediation of eye movements (in network terms: the anticipation of an object at a particular location given the linguistic input) might be an emergent consequence of a common representational substrate supporting interpretation in both the linguistic and visual domains.