DRAFT NOTES ON A NEW NEURAL MODEL OF LANGUAGE PROCESSING

Martin I. Sereno (1995, unpublished)

What and Where is Working Memory for Discourse/Scene?

One of the most basic acts of cognition that humans as well as many other animals must execute almost every hour of the day might be called scene assembly—the act of acquiring blocks of sensory information and then combining them in some form to generate a temporarily persisting representation for the purpose of directing behavior. When a baboon comes into a small clearing, for example, it might make 10 or 20 quick saccades around the scene to check for predators, possible food items, the location of particular members of the troop, possible escape routes, and so on. There is a great deal of behavioral evidence that some residue of the having looked around persists for some time and forms the basis for ongoing behavior (feeding, trying the copulate, taking the best escape route when a predator happens by). The capacity for scene assembly in working memory would appear to be a fundamental requirement for survival.

It is an unfortunate fact, rarely explicitly stated, that we currently have only the barest glimmer of insight into how such information is temporarily stored in the brain or even whether it consists of some kind of active recirculating pattern, or whether it exists merely as a pattern of changed weights with the potential to affect neural firing in the future. We have some interesting clues, however, about where these patterns must be.

Working memory, long-term memory, and intermediate-term memory

Perhaps one of the most striking facts about the famous patient H.M, who sustained a bilateral loss of most of the hippocampus, amygdala, as well as the entorhinal, perirhinal, periamygdaloid, and parahippocampal cortex, is that he has retained the ability for scene assembly as well as discourse assembly. He is perfectly capable of understanding and temporarily assembling an interpretation of a long sequence of verbal instructions—for example, an explanation of the rules of the Tower of Hanoi problem—despite not being able to consciously remember any of the explanation an hour or a day later. This strongly suggests that the hippocampus must not be the site of this temporary assembly. Evidence instead points to the perirhinal and parahippocampal cortex (perhaps including the hippocampus proper) as a kind of intermediate-term storage where information is kept for weeks or months before being incorporated back into the permanent long term storage in higher cortical sensory areas.

So we have reasonable evidence that this assembly process must go on in higher neocortical sensory areas. Before trying to use build on thisIn the next section, we argue that much of this process must go on in higher visual areas.

What is language for?

[for modifiying other patterns]

Language as Code-directed Scene Comprehension

Vision as the primary modality in monkeys, apes, and humans

Vision is very important to primates; in fact, over 50% of the cortex in primates probably

including humans, consists of areas devoted to specifically visual processing. This is not to deny that information about an object perceived via another modality—say the somatosensory system—might be able to enter visual areas in the form of a visual copy of the somatosensory areas' activity pattern (see e.g., experiments by Haenny et al. (1988) in macaque visual area V4 using a somatosensory-visual matching task). But it does suggest that we should carefully distinguish a visual copy of a somatosensory stimulus (in a visual area with a visual map) from a somatosensory copy of a visual stimulus (in a somatosensory area with a somatosensory map).

The notion of language as (primarily) vision stems entirely from the fact that monkeys are diurnal animals with particularly well-developed visual systems. It is not hard to find animals whose primary modality is not vision. Bats, for example, perceive the world primarily through their auditory systems. Many grazing animals have very specialized mouths and lips. These specializations are often much more dramatic when one looks inside the brain. The somatosensory cortex of a llama, for example, contains absolutely enormous representations of the lower and upper lips that are orders of magnitude larger than the representation of all the remaining body parts combined (Welker, 1966). Catfish have taste buds covering their entire bodies and even their fins, and have multiple 'gustotopic' maps of these receptors in their brains (Finger, 1978###). Several groups of fish emit weak electric fields (Heiligenberg, 1992###)

Perhaps one of the most striking examples of a primary modality other than vision comes from the duck-billed playtpus. When foraging for food underwater, this peculiar mammal closes both its eyes (eyelids) and ears (with little flaps) and perceives the world almost entirely through its exceedingly sensitive bill, which is densely covered with both somatosensory receptors but also electroreceptors. Maps of the platypus brain have revealed an absolutely enormous representation of the bill with a finely intercalated mosaic of somatosensory and electrosensory patches (Krubitzer et al., Neurosci abstr###). Visual are auditory cortex are very small by comparison.

Underspecification of information in a single glance—visual polysemy

[...]

Tying together information from non-adjacent glances—visual anaphora

[...]

Visual syntax of possible scenes

When looking around a new scene—as one does upon entering a new room, for example—there would seem to be a great deal of flexibility with regard to the order in which fixations are made; to a first approximation, it is difficult at first to see any analogue to ordering constraints in language (e.g., "the dog bit the man" *versus* "the man bit the dog"). Some early studies (Yarbus, 1967; Noton and Stark, 1971) suggested that scan paths (measured during several minutes of fiaxtion of a photograph) reflected to some extent what the subjects were looking for. Later studies, however, produced little evidence for rigid scanpaths during successful recognition performance and the field lost interest. The information arriving in the visual cortex, however, is quite constrained in one respect: adjacent glances are typically made from nearby locations. The head and/or eyes may rotate or translate, but it is rare to move to the opposite side of the room in between glances.

There is no such intrinsic constraint on the sequences of takes in a film or video—one can splice together any two scenes taken from any camera angle. Filmmakers, nevertheless, have learned a series of conventions respecting the constraints of eye and head movement that make the comprehension of sequences of takes as effortless as possible. It is important to remember that the film viewer is denied any efference copy about what direction the eye of the camera has 'looked' for each new take. One such convention concerns how one typically would film a conversation in close-ups. A camera is aimed at each participant's head with the convention that the cameras must be on the same side of the line connecting the two participants. The visual effect of cutting back and forth between these two viewpoints is that a head rotates as it changes identity; this gives the viewer the impression of having looked back and forth between the two. If instead, one person is filmed from one side of the pair and the other person if filmed from the other side, the visual effect of a cut between these two cameras is that one head simply turns into another head without any rotation. This gives the impression that one person has suddenly changed into another without giving the impression of having looked from one to the other.

I do not want to argue that the ordering constraints on glances are as strict as those with respect to words. However, there are some definite constraints on what sequences of glances can successfully convey a particular meaning (a conversation *versus* a rapid metamorphosis).

Impossible objects

So-called impossible objects demonstrate another kind of constraint on the possible legal sequences of glances that bears a certain resemblance to linguistic syntactic constraints. Such an object may make sense locally (cf. have a degree of legal constituent structure) yet as one looks to different parts of it, it becomes impossible to make sense of the object as a coherent whole.

Difficulty of investigating visual syntax

Ordering constraints on glances have not been as well studied as ordering constraints on words. One major reason for this is that it is much harder to manipulate scenes than words. A linguist or psycholinguist can easily experiment with different word orders in the office or lab. There is currently no equivalent way of rapidly generating arbitrary sequences of glances for the purpose of investigating their syntax and semantics. However, this is beginning to change with the availablity of virtual environments.

Another difficulty is that words are much better defined than glances. Particularly upon entering a new place, it seem quite unlikely that the glances one experiences line up nicely with a pre-existing visual lexicon. On the other hand, one can learn many new words (a major part of learning about a new field—say geology—involves learning the meaning of a thousand or more new words), and one can become extremely familiar with a particular visual environment (for example, one's bedroom) to the extent that particular glances might come close to achieving a certain lexical status.

Glances and words (as before)

Some linguists have independently suggested that visual representations may be very important in the semantics of natural language (Jackendoff, 1983; 1987; Fauconnier, 1985; Lakoff, 1987; Langacker, 1987). An idea common to several different approaches is that more concrete visual

meanings may have been extended by analogical processes to deal with more abstract objects and relations. The present proposal goes further in suggesting a particularly direct relationship between the mechanisms of scene and discourse comprehension.

The integration of successive glances in the comprehension of a visual scene requires a kind of serial assembly operation similar in some respects to the integration of word meanings in discourse comprehension. Primates (but also many other animals) make long series of fixations at the rate of several new views per second during scene comprehension. Each fixation brings the retina to a new part of the visual scene and generates a burst of activity in V1, which largely replaces the burst caused by the previous fixation. Higher visual areas with less precise retinotopy somehow integrate information from these disconnected activity sequences to generate an internal representation of the location and identity of the relevant objects in the current scene (e.g., predators, food items, particular conspecifics, escape routes, suitable sleeping trees, etc.) that can serve as a basis for action. Many aspects of this process are redolent of linguistic integration—e.g., the underspecified, context-free information in an isolated glance is sharpened and focused by context (cf. polysemy); information from temporally distant glances must be tied together (cf. anaphora).

One main difference between scene and discourse comprehension is, of course, that scene comprehension is tied closely to the current scene. Discourse comprehension might best be thought of as a kind of fictive visual scene comprehension directed, in the case of spoken language comprehension, by sequences of phoneme representations in secondary auditory cortex. The advantage of linguistic discourse comprehension is that we are no longer tied to the current scene. However, once the appropriate visual word meaning patterns have been called up and bound together, the nature and interactions of the composite pattern may be conditioned mainly by the prelinguistic rules of interaction of scene representations in primate visual areas networks. In this sense, a large part of what has been called linguistic syntax and semantics might *not* be modular with respect to the neurobiology of vision.

*Why code-directed assembly is har*d

[it seems too easy to do much explaining]

[it’s hard because it’s online and word rec is hard and sticking together is hard]

[cf. cells]

Transcortical sensory aphasia (temporal and parietal visual areas)

There is in fact substantial evidence that visual areas in humans are involved in specifically linguistic functions. There is a kind of aphasia confusingly called 'transcortical sensory' aphasia (i.e., 'across-from-the-language-cortex' aphasia!) that is generated by a lesion in left human inferotemporal cortex (Rubens and Kertesz, 1983). Many of these lesions are so posterior and ventral that they are associated with overt visual field defects. Transcortical sensory aphasics have poor, "Wernicke's-like" comprehension, yet paradoxically (at least in the context of traditional models of language comprehension), can repeat words effortlessly. Far from being 'across from the language cortex', the visual areas in posterior inferotemporal cortex damaged in these patients may be the primary site of semantic processing in sighted humans. Transcortical sensory aphasics

recover more quickly than patients with more dorsal lesions; this may only be an indication that the functions performed by visual cortex in language comprehension are less lateralized than those performed by auditory cortex. This is consistent with what has been known for some time with respect to primate visual areas; permanent deficits in visual pattern recognition in monkeys require *bilateral* inferotemporal cortex lesions (Gross, 1973). There is no need to assume that all the cortical areas involved in language comprehension are equally lateralized; for example, the functions performed by the superior temporal gyrus (see below) may be more lateralized than the functions performed by the inferotemporal cortex.

Basal temporal language area (ventral inferotemporal cortex)

In the past few years, there have been a number of reports that language behaviors can be arrested by stimulation of a substantial area of the ventral temporal lobe (the fusiform gyrus—a rather ill-defined area with an anterior-posterior extent of almost 10 cm). The interference in ongoing linguistic behavior produced by stimulating this area is indistinguishable from that produced by stimulating the classical Broca's and Wernicke's areas (Ojemann, ###). This area has only recently been made accessible to stimulation through the use of temporarily implanted subdural grids inserted to aid in planning epilepsy surgery.

Pictures and word studies (Potter,Kutas)

Psycholinguistic experiments using pictures inserted into sentences and picture-word priming (Potter et al, 1986; Vanderwart, 1984) suggest that it is surprisingly easy for visually represented concepts to be integrated into ongoing linguistic discourse comprehension. This may be another indicator of the closeness of visual category representations to linguistic meanings. None of this implies that visual meaning patterns need *look* like pictures; the patterns in question would be distributed across several visual areas that lack any systematic representation of retinal space.