tion and speech recognition. A masking field does this by giving the chunks that represent larger groupings, up to some maximal length, a prewired competitive advantage over those that represent smaller groupings. It was shown how this bias could develop from simple developmental growth laws (Cohen & Grossberg 1986). The network clarifies how the most predictive chunk can be maximally activated, while less predictive chunks are less activated, and chunks with insufficient evidence are merely primed. Such a network naturally explains such data as the Magic Number Seven (Grossberg 1978a; 1986; Miller 1956), and predicted data about the word length effect (Samuel et al. 1982; 1983), which shows that a letter can be progressively better recognized when it is embedded in longer words of lengths from 1 to 4. This is the speech analog of the word superiority effect, which it also explains, unlike the Seidenberg and McClelland (1989) model. Masking fields have recently been used, within an ART framework, to quantitatively explain data about how future word sounds can reorganize conscious percepts of earlier word sounds (Grossberg & Myers 1999; Repp et al. 1978). None of the distributed models mentioned by Page can explain these data. More recent developments of ART continue to analyse how a network can automatically discover, through incremental learning in real time, the optimal level of compression with which to represent different input environments.

Page mentions "binding" as one means of generating distributed representations. One mechanism for this is the horizontal connections that exist in neocortex, notably in layers 2/3. Recent modeling work has clarified how bottom-up, horizontal, and top-down interactions interact within the laminar circuits of neocortex, notably visual cortex, to bind together distributed activations into coherent boundary representations (Grossberg 1999; Grossberg & Raizada 1999). This work opens the way toward the very large task of showing how distributed information may be coherently bound in other parts of sensory and cognitive neocortex.

Page notes that both view-specific and view-invariant representations of familiar objects can be found in IT cortex. Such representations have been shown to self-organize in a number of ART-based models; see Bradski and Grossberg (1995) for one such model and related references. A key issue here is that working memories play a useful role in generating these representations. These working memories are "distributed," yet are also clearly localist.

Page quotes the assertion of McClelland and Rumelhart (1981) and Rumelhart and McClelland (1982) that their Interactive Activation (IA) model is a canonical model "that characterizes the qualitative behavior of other models." Actually, the original IA model had serious defects. These defects illustrate that all localist models are not created equal, and that one must exercise as much caution in choosing among them as one does between localist and nonlocal distributed models. In particular, I early noted that the IA model had unrealistic processing levels (phonemes, letters, words) and bottom-up input pathways (both excitatory and inhibitory). These properties were inconsistent with key data, and prevented the model from being able to stably learn from its inputs-even though the authors did not attempt to make the IA model learn (Grossberg 1984; 1987). Later versions of the model changed these properties to be consistent with previously published ART properties; e.g., those in Grossberg (1978a). In this sense, the IA model is dead, and has been subsumed by ART. Problems within models like IA can lead people who prefer nonlocal distributed models to conclude that their models are better. A more proper conclusion is that IA was not an adequate model, localist or not.

Page provides a useful critique of the McClelland et al. (1995) attempt to explain how interactions between the hippocampus and neocortex may control learning and memory. He leaves out at least one issue that I find devastating to all models of this type. Grossberg and Merrill (1996) provide a critique which builds upon this concern. It involves the issue of representation, which is key to all discussions of localist versus distributed coding. In particular, this model proposes that the hippocampus rapidly encodes information which is then later transferred to neocortex. But there is no evidence of which I am aware that the hippocampus can represent the types of information from vision, audition, and so on, that would need to be represented there for this proposal to be plausible. Saying that the information is represented by hippocampus in compressed form does not help, because then one needs to explain how it gets decompressed in the cortex. I am amazed that authors of such models have not bothered to respond to this critique. I hope that it does not take as long as it took the stability-plasticity issues to get discussed which were introduced with ART in 1976.

The Law of Practice and localist neural network models

Andrew Heathcote and Scott Brown

Department of Psychology, The University of Newcastle, Callaghan, 2308, NSW, Australia. **{heathcote; sbrown}@psychology.newcastle.edu.au psychology.newcastle.edu.au/**

Abstract: An extensive survey by Heathcote et al. (in press) found that the Law of Practice is closer to an exponential than a power form. We show that this result is hard to obtain for models using leaky competitive units when practice affects only the input, but that it can be accommodated when practice affects shunting self-excitation.

In a recent survey, Heathcote et al. (in press) analyzed the form of the Law of Practice in 7,910 practice series from 475 subjects in 24 experiments using a broad range of skill acquisition paradigms. When the practice series were not averaged over subjects or conditions, an exponential function (mean response time, $RT = A + Be^{-\alpha N}$, where A is asymptotic RT, B is the amount that learning decreases RT, and N is practice trials) provided a better fit than a power function ($RT = A + BN^{-\beta}$) for the majority of cases in every paradigm. The defining property of an exponential function is that its relative learning rate, $RLR = -dRT/dN/(RT - A)$ equals a constant (α). In contrast, the power function's RLR decreases hyperbolically to zero, $RLR = \beta/N$. Previous findings in favor of a power function (e.g., Newell & Rosenbloom 1981) used practice series averaged over subjects and/or conditions. When exponential practice series with different rates (α) are averaged, the RLR of the average decreases, because fast learners (with large α) control the rate of change early in practice, while slow learners (with small α) dominate later in practice (see Brown & Heathcote, in preparation, for detailed analyses of averaging effects). As theories of skill acquisition model the behavior of individuals, not averages, Heathcote et al. concluded that the "Law of Practice" is better characterized by an exponential than a power function. Hence, the power function prediction made by Page's model does not accord with recent empirical results.

We believe that an exponential law of practice is extremely difficult to obtain using Page's approach to practice effects in competitive leaky integration networks (Equation 5). To see why, consider the time (t) it takes the activation ($x(t)$) of a leaky integrator ($dx/dt = I - kx$, where I is input and k is leakage rate and $x(0) = 0$) to reach a criterion χ.

$$t = \frac{1}{k}\ln\left(\frac{I}{I - k_\chi}\right) \tag{1}$$

The RLR of (1) with respect to I decreases to zero. If we assume, as Page does, that practice decreases t by increasing I, the RLR of (Eq. 1.) with respect to N will decrease to zero unless $I(N) \geq O(N^2)$ for large N. Such a faster than linear increase in input is difficult to justify. The increase of I with N is slower than linear for Page's "noisy-pick-the-biggest" model. Even if all instances, rather than just the maximally activated instance, were to contribute to

I, the increase would be only linear. Page's simulation results (Fig. 6) indicate that the same power-like effects of increasing *I* apply to the time it takes competing leaky integrators to pass an activation criterion.

However, competitive leaky integrators can account for Heathcote et al.'s (in press) findings if practice alters shunting terms, such as the weights of self-excitatory connections.[1] Consider a two-unit system of the type discussed by Usher and McClelland (1995), with normalized inputs *I* and $(1 - I)$ and linear threshold transfer functions:

$$dx_1/dt = I - (k - \epsilon)x_1 - \delta x_2 \quad (2)$$

$$dx_2/dt = 1 - I - (k - \epsilon)x_2 - \delta x_1 \quad (3)$$

A response is made when the activation of one unit exceeds a criterion, χ. Assume that as practice proceeds, the self-excitatory weight, ϵ, approaches the leakage rate *k*, using a weight-learning rule like Page's Equation 2:

$$d\epsilon/dN = \lambda(k - \epsilon) \quad (4)$$

In simulations with Gaussian noise added (Eq. 2, 3) at each step of the integration (Page's N_1 term in his Eq. 5) and larger values of *I* so errors did not occur, learning series were consistently better fit by an exponential than by a power function. Insight into this result can be gained from the analytic result for the one unit case (i.e., Eq. 2 with competitive weight, $\delta = 0$, which was also better fit by the exponential in simulations):

$$t = \frac{1}{k} e^{\lambda N} \ln\left(\frac{I}{I - k_\chi e^{-\lambda N}}\right) \quad (5)$$

For a linear Taylor approximation to (Eq. 5), *RLR* decreases marginally with *N*, but asymptotically approaches λ rather than zero. Heathcote et al. (in press) found that an APEX function ($RT = A + Be^{-\alpha N}N^{-\beta}$), which has a *RLR* that decreases to an asymptote greater than zero, consistently fit slightly better than an exponential function. We found the same pattern of fit to our simulation results for both the one and two-unit models. The parameter estimates for these fits also concurred with the survey results. Estimates of the power function *A* parameter were implausibly small (as *N* increases *t* approaches χ/I for the linear Taylor approximation to [Eq. 5], whereas most power function *A* estimates were zero). Fits of a power function with an extra parameter (*E*) to account for prior practice ($RT = A + B(N + E)^{-\beta}$) produced implausibly large *B* estimates, mirroring Heathcote et al.'s (in press) findings with the survey data.

Given limited space it is not possible to quantitatively examine this type of model further (see Heathcote 1998, for related findings and Heathcote & Brown, in preparation, for a detailed analysis). However, the findings presented are sufficient to demonstrate that Heathcote et al.'s (in press) results are not incompatible with the overall localist neural network approach. Indeed, learning in shunting connections, both self-excitatory and competitive, provides an adaptive mechanism for consolidating and differentiating local response representations (cf. Usher & McClelland 1995, who note that the "units" in such models may correspond to collections of neurons bound together by mutually excitatory connections). Reduced leakage with practice can also explain Jamieson and Petrusik's (1977) finding (cited in Usher & McClelland 1995) that the difference between error and correct RTs decreased with practice. As leakage approaches zero, a leaky integrator approximates a classical diffusion process, for which error and correct RTs are equivalent.

NOTE

1. We also obtained an exact exponential result for inputs that (1) increase with practice according to a learning rule like Page's Equation 2 ($I = M(1 - e^{-\lambda N})$), (2) are nonstationary (decreasing with presentation time *t*, as $I = 1/(t + (b - cI))$, $b/c >$ M), and (3) have a shunting effect on a single unit's activation ($dx/dt = (U - x)I$). We will not pursue this model here, as it is very different from Page's approach (see Heath 1992, and Smith 1995, for more on nonstationary inputs, and Heathcote 1998 for more on shunting inputs).

Localism as a first step toward symbolic representation

John E. Hummel

Department of Psychology, University of California, Los Angeles, CA 90095.
jhummel@lifesci.ucla.edu **www.bol.ucla.edu/~hummel/**

Abstract: Page argues convincingly for several important properties of localist representations in connectionist models of cognition. I argue that another important property of localist representations is that they serve as the starting point for connectionist representations of symbolic (relational) structures because they express meaningful properties independent of one another and their relations.

Page's arguments and demonstrations make a compelling case for the essential role of localist representations in connectionist models of cognition (and cognition itself). One important property of localist representations that Page does not emphasize (although he mentions it at the end of sect. 7.3), concerns the role of localist nodes in the representation of relational structures. I argue that localist representations share a crucial property with the kinds of representations that are necessary for relational representation in connectionist systems – namely, independent representation of meaningful entities – and that they therefore play an essential role in the ability of connectionist models to account for symbolic aspects of cognition.

The notion of a "localist representation" is subtle because localism is not a property of a representation, but of the relationship between a representation and the entities it represents. To borrow Page's example, the activation pattern *−woman, +politician,* and *−actor* is a distributed representation of Tony Blair, but a local representation of *woman, politician,* and *actor.* Every representation is local at some level. Even a "fully distributed" representation is localist with respect to some entities, in that each node has an equivalence class of entities to which it corresponds. The equivalence class may be difficult or impossible for the modeler to understand (as in the case of the hidden nodes in many BP networks), but unless a node is always active (in which case it carries no information), its activity will correspond to some state of affairs in the network's universe: The node is a localist representation of that state of affairs. As such, the important question is not whether a representation is localist or distributed, but whether it is localist *with respect* to a meaningful state of affairs in the network's universe.

In this sense, the question of localist versus distributed maps onto the question of *independence* (a.k.a., *separability;* Garner 1974) versus *nonindependence* (a.k.a., *integrality*) in mental representation. If meaningful concepts, entities or dimensions map onto individual nodes (or in the case of dimensions, nonoverlapping populations of nodes) – that is, if the system is localist with respect to those entities or dimensions – then the system represents those entities as independent of one another. To the system, the entities or dimensions are separable (cf. Cheng & Pachella 1984). If individual nodes respond to *conjunctions* of entities or properties, then the resulting representation is integral with respect to those properties (e.g., nodes that respond to specific conjunctions of shape and color constitute an integral representation of shape and color). One hidden limitation of many "fully distributed" representations (e.g., those that emerge in the hidden layers of BP networks) is not only that they lack individual nodes to respond to individual entities (the limitation Page emphasizes), but also that they typically constitute integral, rather than separable representations of the important entities or properties in the network's universe.