Age of Acquisition in Connectionist Networks

Karen L. Anderson (kanders@cs.ucsd.edu)
Garrison W. Cottrell (gary@cs.ucsd.edu)
Computer Science and Engineering Department 0114
Institute for Neural Computation
University of California, San Diego
La Jolla, CA 92093-0114 USA

Abstract

Recently, there has been a resurgence of interest in the role of the Age of Acquisition (AoA) of an item in determining subjects' reaction time in naming words, objects, and faces. Using the number of epochs required to learn an item as a direct measure of AoA in connectionist networks, Smith, Cottrell & Anderson (in press) have shown that AoA is a stronger predictor of final Sum Squared Error than frequency. In this paper, we replicate Smith *et al*. using more realistic frequency distributions for the items, and examine why some patterns may be learned earlier than others. First, we have found that the same patterns tend to be learned early and late by networks differing in their initial random weights; hence, the issue is, what property of the patterns determines AoA? We have found that even very weak pattern similarity structure is a strong predictor of AoA when frequency is controlled for. Also, we have found evidence that such a similarity structure may still be an important factor in determining AoA even when pattern frequency is varied.

Introduction

Ever since Carroll & White (1973) reanalyzed Oldfield & Wingfield' s (1965) naming latency data and discovered that frequency was not significant when AoA was considered, controversy has surrounded discussions of the import of the two variables. Technological and methodological refinements have led to agreement that both frequency and AoA play significant roles. Hence, interest has returned to the pursuit of understanding the mechanisms underlying AoA effects.

It had been proposed recently (Morrison & Ellis, 1995; Moore & Valentine, 1998) that connectionist networks would be incapable of exhibiting AoA effects because training on late patterns would cause "catastrophic interference" resulting in the unlearning of early patterns. However, this sort of interference is only found if training on early patterns ceases. Ellis & Lambon Ralph (2000) demonstrated AoA effects in a neural network by training the net on an "early" set of patterns and then simply adding a second set of "late" patterns halfway through training.

Smith et al. (in press) independently demonstrated AoA effects in networks. In contrast to the staging method of Ellis & Lambon Ralph (2000), where AoA is assumed to correspond to the time at which patterns are presented to the network (early or late), all patterns were presented to the model from the outset. AoA can then be *measured* for each pattern individually as the time during training when the pattern is learned. Using this more natural definition, Smith et al. reported significant effects of AoA on naming latency (defined as the residual error on a pattern after training is completed, a measure of the network's *uncertainty*).

What we would like to know is why certain patterns are learned earlier than others, and how early learning of a pattern comes to affect the network' s performance. Ellis & Lambon Ralph's (2000) approach cannot be used to find out why patterns are acquired in a particular order as it *imposes* an order by staging pattern presentation. Instead, we vary properties of the patterns and then measure AoA, as in Smith et al. Ellis & Lambon Ralph (2000), do suggest why early AoA is important for final performance – the network is more "plastic" earlier in training, so items that are learned first have the opportunity to make the biggest impression on the weights.

We also want to know whether our finding that AoA is a stronger predictor of final error than frequency survives a more realistic version of Zipf's (1935) frequency distribution than was used by Smith et al. Here we show that it does.

Methods

Our investigation is organized around a series of experiments in which we replicate and extend network simulations and analyses previously reported by Smith et al. (in press). We begin with one of the simplest connectionist models of lexical access, an autoencoder network. This kind of network simply reproduces its input on its output through a set of hidden units, and has seen surprisingly wide application in cognitive modeling. We then extend our simulations to more complex mappings.

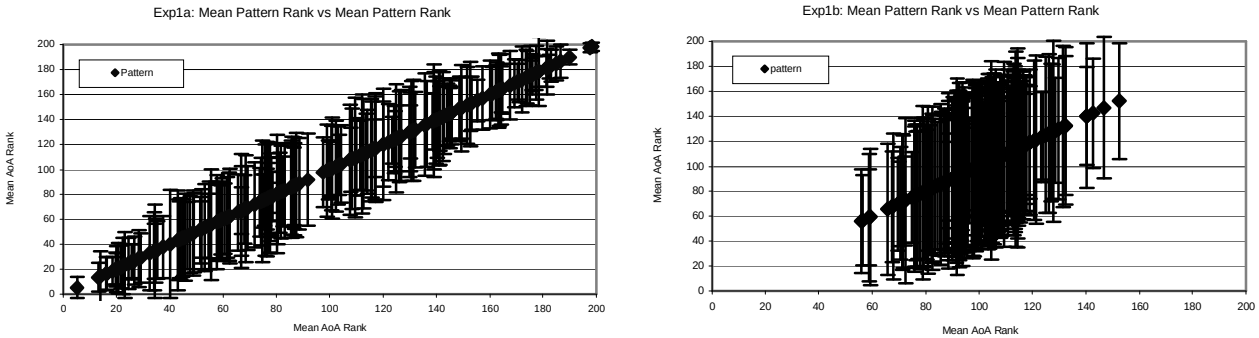

Figure 1: Comparison of pattern AoA variance between experiments using the same pattern set for all simulations (left) and 10 different pattern sets (right).

Table 1: Average correlations between pattern similarity measures and AoA.

	AoA	Mean Cosine	Mean R^2	Density	Mean Distance
AoA	1.0000				
Mean Cosine	0.0487	1.0000			
Mean R^2	-0.4751	-0.0399	1.0000		
Density	0.0806	0.9886	-0.0412	1.0000	
Mean Distance	0.1069	-0.3058	-0.0437	-0.1825	1.0000

Experiment 1

Smith et al. (in press) report finding a strong correlation between AoA and SSE in their first experiment in which they trained ten autoencoders on the same set of equally frequent patterns. Training all networks on the same set of patterns ignores the possibility that the order in which the patterns in the set will be acquired by the network may depend on some property of the training set. To examine this possibility, we replicated Smith et al.'s first experiment in two ways: first, using the same set of randomly generated patterns (an exact replication) and, second, using a different pattern set for each network. The first replication allowed us to perform an analysis of the AoA rank order correlation of patterns between pairs of networks – if networks trained on the same pattern set tend to acquire patterns in the same order then the rank order correlations between pairs of networks should be significant, implicating a property of the training set in driving acquisition order. The second replication allowed us to see whether Smith et al.'s finding concerning the correlation between AoA and SSE maintained across a larger set of patterns.

Methods For the first replication, ten groups of ten networks were trained with all of the networks in a group using the same pattern set. In the second replication, a single group of ten networks were trained with each network using a different pattern set. For both replications, the pattern sets consisted of 200 randomly generated 20-bit patterns in which each bit had a 50% chance of being on. All networks were 20-15-20 autoencoders trained via backpropagation, using learning rates of .001, momentums of .9, and initial random weights between 0.1 and -0.1. All patterns were presented every epoch. Training was continued until 98% of the patterns were acquired (where "acquired" means its SSE went below 2.0). The AoA of a pattern was taken to be the first epoch in which it was acquired.

Results Smith et al. (in press) reported a correlation coefficient of 0.749 between SSE and AoA averaged over all 10 networks. For both replications, we found similar mean correlations: 0.773 (0.038) and 0.756 (0.050), for a randomly chosen group in the first (same pattern) replication and the group in the second (different patterns) replication, respectively. Thus, our replication supports the finding of Smith et al. that AoA and SSE are strongly correlated.

Although we arrived at that same result in the first replication, our second replication does not support the conclusion that AoA is independent of properties of the training set. Our examination of the AoA rank order correlation between groups of networks trained on the same pattern set revealed that networks trained from different initial weights tend to learn the patterns in a set in a similar order. The pair-wise AoA rank order correlations between networks in the same group averaged over all pairs in all groups (N=450) were found to be 0.485 (σ=0.061), using Kendall's τ, and 0.665 (σ=0.071), using Spearman's ρ. Figure 1 illustrates this relationship. Both graphs

in the figure plot the mean AoA values for each pattern on both axes. The graph on the left is for 10 networks in one of the groups in the first replication using the same pattern set for each network. In order to estimate how chance behavior would look, we simply aligned the different pattern sets used in the second simulation based on pattern numbers and, in the graph on the right in Figure 1, we plot the means and standard deviation for all patterns with the same number. Note how the means in the plot on the left do not cluster about the center as do those in the plot on the right, and that those on the left have smaller standard deviations.

Having found that *some* property of the training set contributes to the AoA of the patterns in the set, our next goal was to attempt to identify what that property might be. Note that in choosing random 20 bit patterns, we are selecting vectors randomly from a 20 dimensional space. Since the maximum number of vectors that can be mutually orthogonal in such a space is 20, and we are selecting 200 vectors, each vector in the set will necessarily be closer to some vectors in the set than to others. This unavoidable clustering of patterns in the vector space is what we refer to when we speak of the similarity structure of a randomly chosen set of training patterns. Since the patterns are randomly chosen, the average pair-wise correlation between patterns is small (0.0581, σ = 0.0056 for an exemplary set), but non-zero.

For one of the pattern sets used in the first replication, we computed for each pattern the mean cosine, R^2, and Euclidean distance between the pattern and all others, and the pattern density (% bits "on"). The correlations between these measures and the patterns' AoA values were computed for each network and then averaged together. As Table 1 shows, the negative mean R^2 between a pattern and all others in the network is on average the best predictor of the pattern's AoA. We performed a repeated measures multiple regression analysis (Lorch & Myers, 1990) using mean R^2, mean Euclidean distance and density as predictors of AoA, and found that the null hypotheses that the mean regression coefficients are equal to 0 can be rejected with $p<<0.000001$, $p=0.000013$, and $p=.020976$ for mean R^2, mean Euclidean distance, and density, respectively. Thus, we are led to believe that the small and subtle structure reflected by the inter-pattern correlations among the patterns in even a randomly chosen set has a strong role in determining the order in which those patterns will be learned.

Experiment 2

In this experiment we again replicate and extend Smith et al. (in press). Like Smith et al., we aim to show that AoA effects persist in our model when

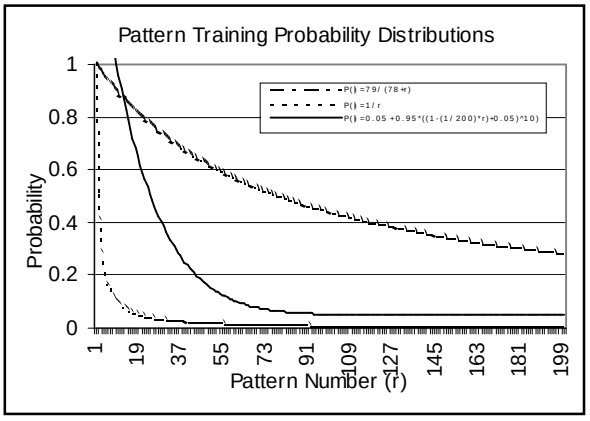

Figure 2: Comparison of training probability distributions.

frequency is added as a variable, and to compare the strengths of these effects to those found in human studies. We improve upon Smith et al. by first, using more realistic frequency distributions and second, by examining the role the shape of the frequency distribution has on the relative contributions of frequency and AoA to naming latency.

Methods We again use ten autoencoders with differing pattern sets, but we vary the frequency of presentation of the patterns within each set. In manipulating pattern frequency, we aim to simulate the well-known Zipf distribution, where a small number of words occur very frequently – that is, the frequency of a word is proportional to the reciprocal of the word's frequency rank. We took two approaches to simulating this distribution. In the first approach, we randomly assign ranks to patterns and train on each pattern with probability 1/rank in each epoch. In the second approach, we take account of the fact that the most frequent words tend to be function words (like "a", "the", "and", etc.) and that human naming studies seldom use such words. Hence, a more accurate model of the frequency distribution of words used for naming stimuli should start lower on the Zipf curve. In order to determine a reasonable starting point, we needed to make an estimate of the frequency ranking of the most frequent word likely to be used in a naming study. To do so, we examined the Celex database (Baayen, Piepenbrock & Gulikers, 1985), and found the rank of the most frequent noun with an imageability rating of 500 or greater in the MRC Psycholinguistic database (Coltheart, 1981). The 500+ imageability criterion was chosen somewhat arbitrarily (the mean rating for words in the MRC database is 450), but was intended to find roughly where concrete nouns show up on the Zipf's curve. The noun selected by this process was "man" with a rank of 78. Hence, our second replication of this experiment randomly

Table 2: Network and human naming study correlation data.

	Networks			Object Naming				Word Naming
	Smith et al.	1/r	79/(78+r)	E&M	S&Y	BM&E	C&W	M&E
r(aoa, sse)	0.749	0.727	0.763	0.626	0.683	0.700	0.77	?
r(logf,sse)	-0.730	-0.462	-0.324	-0.405	-0.456	-0.455	?	-0.388
r(aoa,logf)	-0.283	-0.259	-0.212	-0.377	?	?	?	?
r(log-aoa, sse)	?	0.755	0.826	?	?	?	?	0.244
r(log-aoa, logf)	?	-0.524	-0.273	?	?	?	?	-0.414

(E&M = Ellis & Morrison, 2000; S&Y = Snodgrass & Yuditsky, 1998; BM&E = Barry, Morrison & Ellis, 1997; C&W = Carroll & White, 1973; M&E = Morrison & Ellis, 2000)

assigned a rank, r, between 1 and 200 to each pattern and then presented that pattern with probability (79/78+i) for training on each epoch.[1]

Our central motivation for using more than one frequency distribution in this (and subsequent) experiments is to determine how the shape of the distribution might influence the relative contributions of AoA and frequency to SSE. We hypothesized that training with a frequency distribution from the beginning of Zipf's curve would tend to make the frequency of a word a stronger determinant of its final SSE than would training with a distribution that started after the curve began to flatten. We were also interested in verifying the results obtained by Smith et al., since they used only a single pattern set and just a Zipf-*like* frequency distribution. In particular, they presented pattern r for training each epoch with a probability given by:

$P(r) = 0.05 + 0.95*((1-(1/200)*r)+0.05)^{10})$

Figure 2 shows a graph comparing all three distributions. Note that the Smith et al. distribution has many more "high-frequency" words than does the 1/r distribution, and that it spans a larger range of probabilities than does the 79/(78+r) distribution.

Results Table 2 shows the correlation coefficients obtained from the three network models, as well as regression coefficients obtained from human object and word naming studies. The results show that the network model correlations look much more similar to the object naming data than to the word naming data. This is a bit counter-intuitive given that the networks are being trained to autoencode — word naming is a less arbitrary mapping than object naming and, hence, seems like a better match to the autoencoding task. The results of our next experiment suggest a reason for this discrepancy. We put off further discussion until then.

The main difference between the network models' correlations is that Smith et al.'s r(logf,sse) is much greater than both the other two network models and the human data. Examining Figure 3, we might suppose that the 79/(78+r) distribution has a weaker frequency effect than Smith's distribution since the frequency differences among patterns are not as pronounced. The 1/r distribution may be weaker than Smith's for a similar reason — while it covers a maximal range of frequencies like Smith's, it has relatively few at the high frequencies and, so, little differentiation in terms of frequency for the vast majority of its patterns. As both the models with 1/r and 79/(78+r) are closer to the human data than the model with Smith's curve, though both are at somewhat opposite extremes in terms of frequency distribution, support is lent to the notion of using a true Zipf based distribution. Furthermore, the slight weakening of the effect of frequency on SSE in the 79/(78+r) model compared to the 1/r model suggests that the choice of data set used in human naming experiments (object names will not be at the top of the Zipf curve) could influence the observed strength of correlation between naming latency and frequency and, possibly, explain some of the differences in findings reported in these studies.

Experiment 3

Having demonstrated AoA effects in the presence of frequency in networks trained to perform an autoencoding task, Smith et al. (in press) then examined how different levels of consistency in the mapping task represented by the pattern set influenced AoA and frequency effects. While spelling to sound is reasonably consistent mapping, spelling to meaning or faces to names are not. Again, we were interested in replicating Smith et al. to see whether their results still held when using the more realistic 1/r and 79/(78+r) frequency distributions and unique training pattern sets for each simulation.

Methods The networks were modified compared to the previous experiments in order to make learning the less consistent pattern sets possible: the number of hidden units was increased to 50, and the objective function was changed from SSE to cross-entropy.

[1] The scale factor of 79 is used only to minimize the number of epochs required for learning the set – it simply guarantees that the most frequent pattern is presented exactly once every epoch, while the relative frequencies of the patterns remain unchanged.

Ten pattern sets were randomly created, as before. From this set of ten, eleven sets of ten were created by randomly flipping bits of the target patterns with eleven levels of probability evenly distributed between 0.0 and 0.50 − in 100% consistent pattern sets, the target patterns were exactly the same as the input patterns (autoencoding); in the 0% consistent case, each bit in the target pattern had a 50% chance of being flipped from the input setting (a completely random mapping, like sound to meaning).

Results The graphs in Figure 3 plot for each level of mapping consistency the mean correlation coefficients and the mean coefficient p values of multiple regressions (N=10 for each point) on network SSE with AoA and frequency as the independent variables. The plots reveal that AoA is a stronger and more significant predictor of naming latency than is frequency in our model across all levels of consistency. As we previously noted, these charts may help explain why the data from Experiment 2 look more like object naming than word naming. Even though word naming is a more consistent mapping than object naming, it is still not 100% consistent, as was the task used in Experiment 2. From the graphs of variable significance on the bottom in Figure 3, it is obvious that the case of 100% consistency is somewhat of a discontinuity, resembling 0% consistency more than it does 90% consistency. Autoencoding is not a good model of word naming tasks.

Experiment 4

We view the mean pair-wise AoA rank order correlation between simulations trained using the same pattern set as a measure of the contribution of pattern set similarity structure to determining the order in which words are acquired. In analysis of the AoA effects observed in the networks of experiment 1, we computed this measure for several groups of simulations and found it to be significant. Since experiment 1 was concerned only with autoencoding networks, we wondered whether the effect of pattern structure has as much influence on pattern AoA in networks trained to perform less consistent mapping tasks. We were also curious as to whether the order in which patterns were presented for training would have much effect on the ordering of AoA among the patterns. To answer these questions, we designed our final experiment.

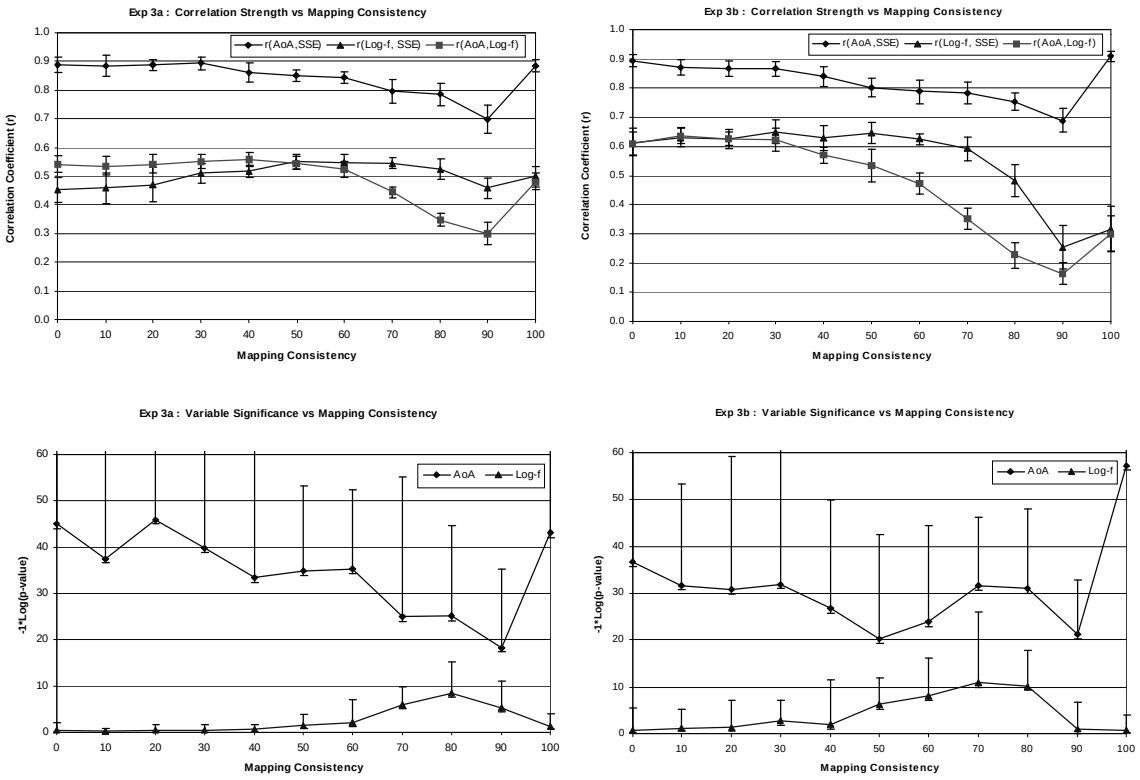

Figure 3: Comparison of the effect of consistency on correlation strength (top) and significance (bottom) between models trained with a 1/r frequency distribution (left) and a 79/(78+r) frequency distribution (right).

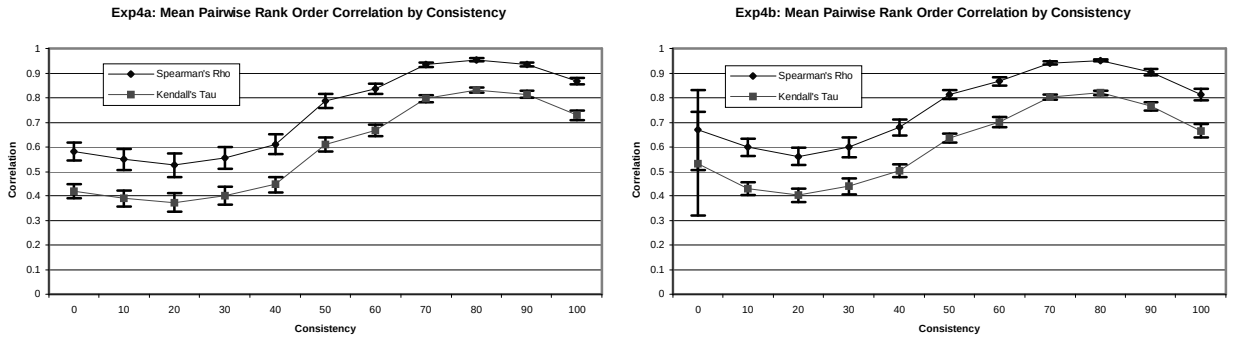

Figure 4: Comparison of pairwise pattern AoA rank order correlations across consistency levels between sets of networks trained with (right) and without (left) randomized pattern presentation order.

Methods One pattern set was arbitrarily selected from each consistency group used in experiment 3, to create a set of eleven training sets with varying levels of consistency ranging between 0% and 100%. For each level of consistency, two sets of ten networks were trained from different random initial weights. The first set was trained with every pattern presented for training in the same order on every epoch. For the second set of ten, all patterns were presented in a new and random order each epoch. Because we were interested only in observing the influence of pattern set similarity structure across training set consistency levels, all patterns in all sets were trained with a uniform frequency distribution. Otherwise, the networks were the same as those in experiment 3.

Results The graphs in Figure 4 plot the mean rank order correlations for each level of consistency. They reveal that, not only is pattern set similarity structure important at all levels of consistency, but that it is also mostly independent of pattern presentation order. The one notable difference between random and non-random presentation ordering occurs at 0 consistency, showing up as a large standard deviation in the plot on the right in Figure 4. This experiment also reveals that similarity structure is generally more influential on AoA at higher levels of consistency.

Conclusion

We have shown that the similarity structure among items is an important determinant of AoA across a variety of mapping tasks. Future work will concentrate on more realistic similarity structures within the domains and ranges of the mappings, such as similarities between words, between faces, and between meanings. On the issue of frequency vs. AoA, the regressions performed in experiment 3 reveal that AoA is a stronger predictor of naming latency in our models than frequency. While AoA and frequency are clearly correlated, there appears to be a fundamental effect of an item becoming encoded in the network weights before other items. Frequency may be the key, but AoA is the door to performance.

Acknowledgments

We wish to thank GURU, Rich Golden, Mark Smith and Dave Noelle for their contributions and support.

References

Barry, C., Morrison, C. M., & Ellis, A. W. (1997). Naming the Snodgrass and Vanderwart pictures: Effects of age of acquisition, frequency and name agreement. *QJEP*, 50A, 560-585.

Baayen, R. H., Piepenbrock R. & Gulikers, L. (1995). *The CELEX Lexical Database (CD-ROM).* Phila., PA: U. of Penn., Linguistic Data Consortium.

Carroll, J. B. & White, M. N. (1973). Word frequency and age of acquisition as determiners of picture-naming latency. *QJEP*, 25, 85-95.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *QJEP*, 33A, 497-505.

Ellis, A. W. & Lambon Ralph, M. A. (2000). Age of Acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *JEP:LMC*, 26(5), 1103-1123.

Lorch, R.F., Jr., & Myers, J.L. (1990). Regression analyses of repeated measures data in cognitive research. *JEP:LMC*, 16, 149-157.

Morrison C. M. & Ellis, A. W. (2000). Real age of acquisition effects in word naming and lexical decision. *British J. of Psychology*, 91(2), 167-180.

Oldfield, R. C. & Wingfield, A. (1965). Response latencies in naming objects. *QJEP*, 17, 273-281.

Smith, M. A, Cottrell, G. W., and K. L. Anderson (in press). The early word catches the weights. To appear in *NIPS* 12. Cambridge, MA: MIT Press.

Snodgrass, J. G., & Yuditsky, T. (1996). Naming times for the Snodgrass and Vanderwart pictures. *Beh. Res. Meth., Instr. & Comp.*, 28, 516-536.

Zipf, G. K. (1935). *The Psycho-Biology of Language.* Boston, MA: Houghton Mifflin.