November 2008

Does Memory Reflect Statistical Regularity in the Environment?

Lael J. Schooler
Carnegie Mellon University

John R. Anderson
Carnegie Mellon University

Published In
.

Does Memory Reflect Statistical Regularity in the Environment?

Lael J. Schooler **John R. Anderson**

Department of Psychology
Carnegie Mellon University
Pittsburgh PA 15213

Abstract

Anderson and Milson (1989) derived optimal performance functions for memory based on assumptions about the goals of memory, the computational costs of achieving those goals, and the statistical structure of the environment. Based on these assumptions, and a good deal of Bayesian analysis, they accounted for a substantial number of empirical findings. Here we started with the same assumptions about the goals of memory, but instead of simulating the statistical structure of the environment, we analyzed it directly. It was found that the factors that govern memory performance also predict the probability with which words are spoken in children's linguistic environments. These factors include frequency, recency, and spacing between exposures. The ability of these factors to predict word use was analyzed in the context of four laboratory memory phenomena: 1) the power law of practice; 2) the power law of forgetting; 3) the interaction between study spacing and retention interval and 4) the combined effects of practice and retention. These factors predict information demand and lend strong support to Anderson and Milson's claim that memory behavior can be understood in terms of the statistical structure of the environment.

Introduction

The environment constantly makes demands on memory. For instance, we may need to remember a particular friend's telephone number 23 times during the past semester and only once in the current semester, while we have had to recall another friend's number three times during the past semester and three times during the current semester. What do these two patterns predict about the probability of having to recall either friend's number today? Anderson & Milson (1989) proposed that the phenomena of human memory could be understood as memory's response to questions like this. They argued that human memory behaves as if it is performing a Bayesian inference to determine the probability that a given memory will be needed now. They contend that this probability is based on 1) the pattern of prior use of the memory (i.e., its history) and 2) the similarity of the current context to the previous contexts in which it had occurred.

Lacking any direct data about patterns of human memory use, Anderson and Milson adapted statistical models developed to describe patterns of use in other information retrieval systems, such as library borrowings (Burrelll, 1985) and computer file access (Stritter, 1977), to create a model of the demands that the environment might make on human memory. By applying Bayesian inference mechanisms to this model of the environment, they accounted for a number of memory phenomena including the effects of practice, retention interval, and study spacing on memory performance.

The model proposed by Anderson and Milson has basically two assumptions:

(1) Memories vary in their rate of use. Thus one of memory's goals is to estimate a particular memory's true rate of use, based on information about how often it has been needed in the past.

(2) Memories undergo fluctuations in their rate of use and some memories are more volatile than others. Thus, memory must infer changes in a particular memory's rate of use.

Assumption (1) helped to account for practice and forgetting effects. Memories which have been used often and recently would be identified as having high rates of use. Assumption (2) helped to account for the forgetting and spacing effects. Memories that have not been used recently would be identified as having undergone a loss in their rate of use and memories that occurred in massed patterns with long periods of disuse would be identified as undergoing rapid fluctuations in their rate of use. It should be emphasized that the Anderson and Milson model did more than just predict these effects; it generated the parametric form of the practice and forgetting functions and the exact interactions that occurred in spacing experiments.

There are, however, a number of problems with the Anderson and Milson analysis. First, one can question whether the statistics of library usage and file access, the basis of their environmental model, are representative of those faced by people. Second, the published analyses of these non-human systems only justify assumption 1. While Anderson and Milson argued that assumption 2 plausibly held, they could not provide hard evidence for it. The raw data from these sources is unavailable and this is needed to determine whether massing and spacing occurred. Third, the predictions were derived from an interaction between Bayesian inference procedures and a mathematical model of the environment. Replacing this model with actual statistics from the environment would help to remove an entire layer of assumptions from the theory.

These problems can be addressed by undertaking an analysis of the statistical properties of the environment; we need to get detailed records of informational demands being placed on human memory. We have analyzed three such sources of informational demand: speech to children, word usage in newspaper headlines, and electronic mail messages. For each of these sources we have conducted analyses that reveal strong similarities between a number of memory phenomena and statistical properties shared by each of these sources. To illustrate our methods, we report the results we have obtained from our analyses of speech to children. Each word a child hears is another demand to retrieve the meaning of that particular word. Hall & Tirre (1979) collected nearly 100 hours of preschool children's verbal interactions. These have been assembled and standardized in a larger database (MacWhinney & Snow, 1985). We analyzed roughly a quarter of Hall & Tirre's data, the

transcripts of 9 white middle-class children. We excluded the children's own utterances from the analysis, because we are interested in the memory demands made by their environments and not on those demands resulting from production.

Building a case for each analysis requires first, characterizing the relevant results from the experimental literature; second, specifying the method for analyzing the environment; and third, comparing the results of the environmental analysis to its experimental counterpart. In particular, we will discuss four laboratory memory phenomena: 1) the power law of practice 2) the power law of forgetting 3) the interaction between study spacing and retention interval and 4) the effects of practice and retention.

Power Law of Practice

Any memory experiment can be thought of as presenting some items to memory some number of times and then testing for recall at some later time. In the Anderson & Milson model, an experiment was encoded as a set of times that described the pattern of presentation and test. Presentation of information in the environment can be characterized similarly. A single exposure to a piece of information in the lab maps onto the occurrence of an item in the environment. The probability of recall or recognition in the lab maps onto the probability that an item in the environment will be used in the next time unit. This is what Anderson & Milson call "need probability".

Within this framework, the mappings between need probability and performance measures, such as probability of recall and recognition, is somewhat more complex then they might initially appear. One might expect that need probability and performance would simply be proportional to each other, but this is not so. Take for example the relationship between need probability and latency of recall. The model assumes that memories are considered in the order of their need probabilities. Thus, estimating the time it would take to retrieve a given memory requires an estimate of the number of memories with need probabilities greater than the given memory. Anderson and Milson argue that this implies that latency of recall should be a power function of need probability. They argue for a similar power function relating practice and probability of recall. Such power functions will preserve ordinal relationships between need probability and some parametric relationships.

Results from the empirical literature further constrain our expectations about the form of the relationship between frequency and need probability. The relevant experiments are those that examine the relationship between practice and memory performance. This relationship is often characterized as taking a power or exponential form. A power relationship, in its simplest form, means that performance, P, equals the amount of practice, L, raised to an exponent r, the learning rate:

$$P = k L^{r}$$

This can be conveniently transformed into a linear relationship by taking the natural logs of each side:

$$\log P = \log k + r \log L$$

In contrast, an exponential function takes the form:

$$P = k e^{bL}$$

Taking the natural log of this yields:

$$\ln P = \ln K + b L$$

These transformations allow the use of linear regression to determine whether power or exponential relationships best capture the relationship between practice and performance. The transformed power equation implies that if performance and practice are in a power relationship with each other, then when log performance is plotted against log practice the result should be a straight line. In contrast, the transformed exponential equation will be curved when plotted in log log, but straight when log performance is plotted against practice.

Using linear regression in combination with these transformations, we have analyzed the results of several learning experiments and in general the power relationship accounts for more variance than does the exponential form. These results are also consistent with Newel and Rosenbloom's (1981) detailed analyses of many other learning experiments. Based purely on these experimental results we can expect a power law relationship between word frequency, the environmental analog of practice, and the analog of performance, need probability.

We are not too concerned with whether the memory or environmental functions are truly power functions or just good approximations. Rather, the goal is to establish that the functional forms of statistical regularities in the environment are the same as the functional forms of human memory performance.

Method

The analog of the practice curve amounts to calculating the conditional probability that a word will occur in an utterance given that it has occurred *n* times in the previous *k* utterances. We calculate this probability by looking at examples of words that occur *n* times in some period, say, the first 100 utterances and determining how often they occur in the 101'st utterance. Our analysis program first made a list of the words that occurred exactly once in the first 100 utterances. To store the partial results of the analysis the program created a variable, or bin, labelled *occurred-1-time*. Each time a word was found that occurred just once the *instances* field of the bin labeled *occurred-1-time* was incremented by 1. If the word occurred in the critical (101'st) utterance, the *uses* field of the *occurred-1-time* bin was incremented by 1. Dividing the number of uses by the number of instances, yields the need probability for items used exactly once. The analogous process was repeated to calculate the probabilities for words occurring exactly 2,3,4..100 times.

The focus of the analysis then shifted onto utterances 2 through 102, with utterance 102 becoming the critical utterance. Again, the program made a list of all the words occurring exactly once between utterance 2 and utterance 101, incrementing the appropriate fields of the *occurred-1-time* bin. As before, the process was repeated for those words that occurred exactly 2,3,4..100 times. The focus, or window, then covered utterances 3 through 103, with utterance 103 acting as the critical utterance. The process repeated until all utterances between utterance 101 and the last had played the role of the critical utterance.

Results

The Hall corpus contained over 18500 utterances that were on average 7.88 words long. The size of the corpus resulted in many instances of infrequently occurring words; there were well over 1.6 million instances of words

occurring 1 time in a 100 utterance window, and 525,000 words occurring 2 times. In figure 1 we have plotted the probability that a word will be used in the critical utterance against the number of times it has been mentioned in the previous 100 utterances. As the sample was insufficient to generate adequate numbers of high frequency instances, some cells corresponding to frequencies greater than 10 were collapsed together in blocks of five to achieve reasonable levels of accuracy. For example, one cell corresponds to words that were used between 11 and 15 times, another to words used between 16 and 20 times, etc.

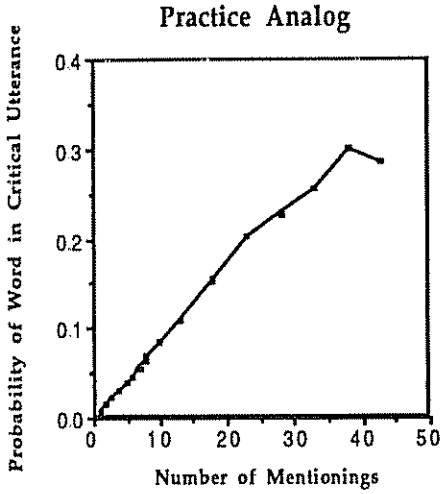

Figure 1. Need Probability (i.e., probability of a word occuring in the critical utterance) as a function of "practice" (i.e., frequency).

Consistent with the empirical and theoretical results the exponential (R^2 = .78) form faired poorly. However, there is nearly a perfect linear relationship between frequency and need probability, (R^2 = .98). A linear function is a special case of a power function, and so is consistent with the theory. It is worth noting that the Anderson & Milson model predicts exactly such a linear relationship between frequency and need probability.

We have identified three power relationships relating the Anderson & Milson model, the practice effect and children's linguistic environment.

1) Theoretically, Anderson & Milson predict that performance, P, should be related to need probability, N, by a power function: $P = a N^b$

2) Empirically, there is a power function relating practice, or frequency, L, to need probability in children's linguistic environments: $N = c L^d$

3) Empirically, there is a power function relating memory performance to practice: $P = f L^h$

Composing the power functions in (1) and (2) leads to a prediction of the power function in 3: $P = a(cL^d)^b = a\,c^d L^{db}$

In contrast, if we had found that need probability was in an exponential relationship with frequency in the environment, $N = ke^{mL}$, then this would imply that memory performance should be in an exponential relationship with practice: $P = a(ke^{mL})^b = a\,k^b e^{mbL}$

Power Law of Forgetting

Another classic memory phenomenon is the retention function-the effect of the interval since the last exposure to a memory item on memory performance. Typically, this is studied by giving subjects a constant experience with one or more studies of an item, and then varying the retention interval. We fit a number of experiments that manipulated retention. Again it turns out that there is a power relationship between the independent measure of retention interval and a number of performance measures. The same form holds in time scales from seconds all the way up to years. This is again despite the frequent claim (e.g., Loftus 1975) that the forgetting function is exponential in nature.

As we have observed in the analysis of the practice function, a power function with respect to need probability implies a power function with respect to performance. Thus, the critical question is whether there is a power function in the environment. That is, if we look at the probability of an item being used as a function of how long it has been since it was last used, do we get a power function?

Method

We are interested in calculating the conditional probability that a word will occur in the critical utterance given that it has not been mentioned in the last k utterances. Each time a word was found that had last occurred in the first utterance, the instances field of the bin labeled *last-seen-in-utterance-1* was incremented by 1. If the word occurred in the critical utterance, the *uses* field of the *last-seen-in-utterance-1* bin was incremented by 1. This process was repeated, in turn, for all items that had not been mentioned since the 2, 3,4...100 utterance. As in the practice analysis, the window shifted to cover utterances 2 through 102.

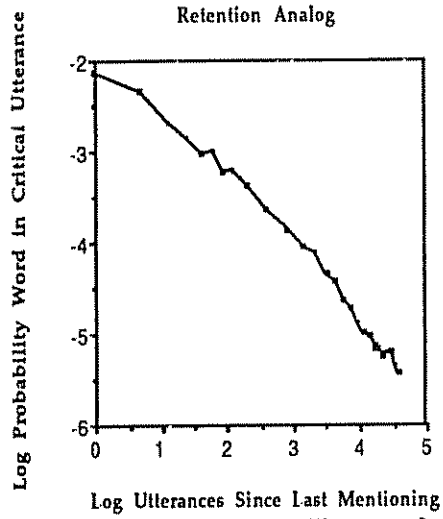

Figure 2. Log Need Probability as a function of log "retention" interval (i.e., the number of utterances since the word was last mentioned). Straight line indicates power function.

Results

Figure 2 plots the power form of the data. Here the data is best captured by a power (r^2 =.98) relationship as compared to the exponential (r^2 = .88) and linear (r^2 =.55) forms. Note that here, unlike the practice function, the data is only fit by a power function and not by a linear function.

Spacing Effect

Up to this point we have considered independently the contribution that retention and practice make on performance. Now we will look at how these interact in the spacing effect. The spacing effect results from manipulating the spacing between study presentations. Glenberg (1976) illustrated the nature of this interaction by orthogonally manipulating the lag between two study presentations and the lag between the second exposure and test (Figure 3). The size of the retention interval produces the largest effects; short test lags are always better than long ones, indicated by the separation between the various curves. It is apparent, however, that the lag between the two studies interacts with retention interval. If the test is going to be given after a long delay it is best to have a long lag between studies. In contrast, when testing occurs after short delays it is best to have short lags between studies. Apparently as the length of the retention interval increases better performance is obtained with longer study lags.

Method

The parallel to the effect of study spacing is the probability that an item will occur in the critical utterance, given a particular combination of spacings between previous occurrences. To start, the program collected all the items used exactly twice in the first 100 utterances. The interval between the first occurrence and second occurrence corresponds to the "study" lag and the interval between the second occurrence and the critical (101 st) utterance the "test" lag. For example an item that first occurred in utterance 30 and then again in utterance 60, would have a study lag of 30 and a test lag of 41. As this is an example of a *study-30-test-41*, the instances field would be incremented by 1. In addition, if the word occurs in the critical utterance, the uses field would be incremented by 1. This process was repeated for each word that occurred twice. As in the previous analyses the focus next shifts to utterances 2 through 102.

Results

Compared to the retention and practice analyses there were relatively few instances for each combination of study and test lag. For example, in the practice analysis there were over 1.6 million instances of items used exactly once, and in the retention analysis there were over 26 thousand instances of words that last occurred in the previous utterance. In contrast, there were only 607 instances of items with study and tests lags respectively of 1 and 5, the most common combination in the spacing analysis. To gain adequate statistical power we had to aggregate the data, collapsing it into a three by three matrix. This matrix consists of short (1-9 utterances), medium (10-30), and long (31-99) test and study lags.

The interaction between study and test lag that was apparent in the Glenberg experiment is evident in the child language data (Figure 4). The shorter the interval between the second occurence and the target utterance the higher the probability that the word will be mentioned. This corresponds to Glenberg's general result that the shorter tests lag lead to better performance. What is striking is the interaction between the spacing between the occurence of the words and the length of the interval between the second exposure and the critical utterance. When the interval between the second exposure and the critical utterance is long, the probability that the word will be included in the critical utterance increases with the spacing between occurences. In contrast, when the interval between the second exposure and the critical utterance is short, the probability that the word will be included in the critical utterance decreases with the spacing between occurences.

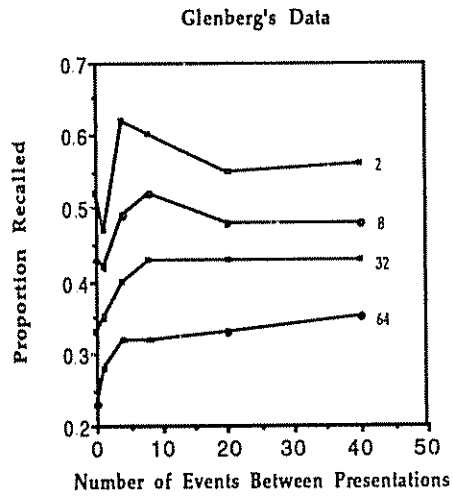

Figure 3. Proportion recalled as a function of the number of events between presentations and the number of events (2,8,32,64) between the second presentation and test.

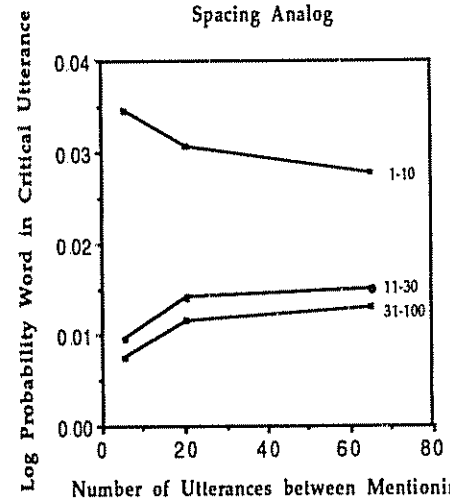

Figure 4. Need Prob. as a function of the num. of utterances between occurrences and the num. of utterances (1-10,11-30,31-100) between the 2nd occurrence and the critical utterance.

Combinded Effects of Practice & Retention

Recently, there has been considerable interest in how different degrees of initial study affect the form of the retention function(Loftus, 1985). If a retention curve takes the form of a power relationship in standard coordinates, then there will be a linear relationship between log performance and log time. Further, we can interpret the slope of this line as a decay rate. Parallel slopes indicate that the retention functions associated with various levels of practice share the same decay rate.

A number of experiments bear directly on this question. For example, Krueger (1929) had subjects learn word lists to various degrees of over learning. Subjects were then asked to recall the lists after retention intervals that lasted between 1 and 28 days. More recently, Hellyer (1962) looked at similar issues but on a much shorter time scale. In this experiment subjects were presented with "syllables" composed of 3 consonants. These were presented from 1 to 8 times at a rate of 1 per second. Subjects were then required to recall the consonants after retention intervals ranging from between 3 and 27 seconds.

Subjects in the Hellyer experiment demonstrated nearly perfect recall at the three second retention interval. When measured in terms of the probability of recall, such near perfect performance approaches the limits of the measure (i.e., approaches 1). In contrast, a power function is unbounded above. Therefore a bounded performance measure, such as probability of recall, can obscure what could best be described in terms of a power relationship. This problem can be easily overcome by using an unbounded measure. Here, we used the odds of recall, another standard performance measure. If p is probability of recall, then o = p/ (1-p) will be the odds of recall.

The transformed results of both experiments are plotted in log log coordinates to test the degree to which they exhibit power law forgetting (Figures 5 & 6). In addition, we have estimated the decay parameters for the various retention curves. In these experiments the interaction between the degree of initial learning and the rate of forgetting is not statistically significant (Tables 1 & 2). Though perhaps a hint of an interaction can be seen in the Krueger data. In short, the effects of practice and retention appear to be essentially additive.

Degree of Learning	Decay Parameter	95% Confidence Interval	R^2
100%	-1.37	(-1.88,-.87)	.96
150%	-1.02	(-1.20,-.88)	.98
200%	-1.02	(-1.22,-.82)	.98

Table 1. Decay parameter estimates (i.e., regression coefficients) for the various levels of over learning in Krueger (1929).

Degree of Learning	Decay Parameter	95% Confidence Interval	R^2
1	-1.78	(-2.72,-.85)	.99
2	-1.70	(-2.35,-1.05)	.98
4	-1.33	(-1.78,-.88)	.99
8	-1.82	(-2.66,-.98)	.98

Table 2. Decay parameter estimates (i.e., regression coefficients) for the various numbers of stimulus presentations in Hellyer (1962)

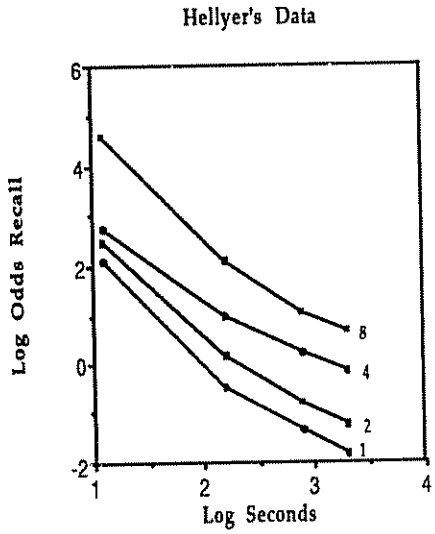

Figure 5. Retention curves at 4 practice levels.

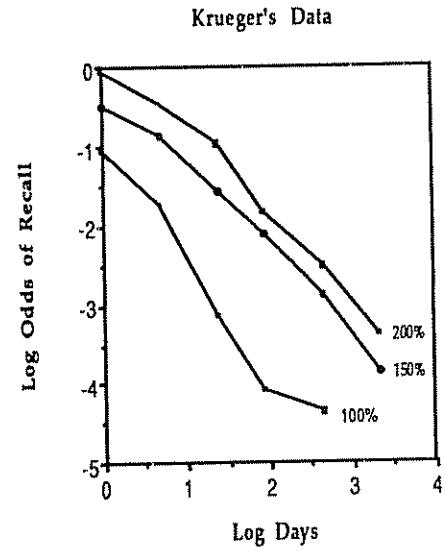

Figure 6. Retention curves at 3 practice levels.

Method

This analysis involves calculating different retention functions for various levels of practice. The program gathered all the words that occurred exactly once in the first 100 utterances. From these it selected those words that last occurred in the first utterance. For each word in this list the *instances* field of the bin labeled *occurred-1-last-seen-1* was incremented by 1. In addition, if the word occurred in the critical (101'st) utterance, the *uses* field was incremented by 1. This process was repeated for all those words that had occurred 1 time and had not been seen since the 2, 3,4...100 utterance. The analogous process was repeated to calculate similar probabilities for words occurring exactly 2,3,4...100 times. As in all the analyses the window shifted to utterances 2 through 102 and so on.

Results

As with the previous analyses, the data was aggregated to obtain reliable statistics. The data was collapsed into cells by the degree of practice and the length of the retention interval. Each cell had a range of 5 mentionings and 5 utterances. For example, one cell represented all instances of words that had been mentioned from 6 to 10 times and had not been mentioned for from 21 to 25 utterances, whereas another cell consisted of examples of words that had been mentioned from 16 to 20 times and were unmentioned for from 1 to 5 utterances. Each curve plots successively higher ranges of practice (Figure 7). The bottom curve represents words used from 1 to 5 times, the second from 6 to 10, etc. The plots use log retention for the abscissa and log odds of need probability as the ordinate. Certain combinations of practice and retention ranges were rare. For instance, there were only 5 instances of words that were used 21 to 25 times and had last been mentioned for between 46 and 50 utterances. So the graph contains only those points for which over 500 instances have been recorded.

Plotting the data in log log coordinates enables us to use regression to estimate the decay rates. These results indicate that the decay parameters associated with low frequency words are approximately the same as those associated with the high frequency words (Table 3). But, as with the Krueger results, there is a hint of an interaction between frequency and retention interval, with the lower frequency words associated with slightly faster decay rates.

Degree of Learning	Decay Parameter	95% Confidence Interval	R2
1-5	-.63	(-.67,-.59)	.98
6-10	-.57	(-.70,-.45)	.87
11-15	-.39	(-.49,-.28)	.92
16-20	-.41	(-.52,-.30)	.98
21-25	-.46	(-.67,-.25)	.98

Table 3. Decay parameter estimates (i.e., regression coefficients) for the various practice/retention curves.

Conclusion

Anderson and Milson derived optimal, rational, behavior functions based on 1) assumptions about the goals of memory, 2) the computational costs of achieving those goals, and 3) the structure of the environment. Based on these three sets of assumptions, and a good deal of Bayesian analysis, they accounted for many empirical findings. Here we started with the same assumptions about the goals of memory, but instead of simulating the statistical structure of the environment based on plausible assumptions, we analyzed the patterns in the environment directly. In this analysis we found statistical analogs to some of the most robust memory phenomena. Not only did we find analogs to single variable functions, such as practice and retention curves, we found an analog to the interaction that underlies the spacing effect. That these analogs can be found in two distinct linguistic domains and one interpersonal one lends sound support to the framework that Anderson and Milson developed. In short, memory appears to be exquisitely tuned to regularities in the statistical structure of the environment.

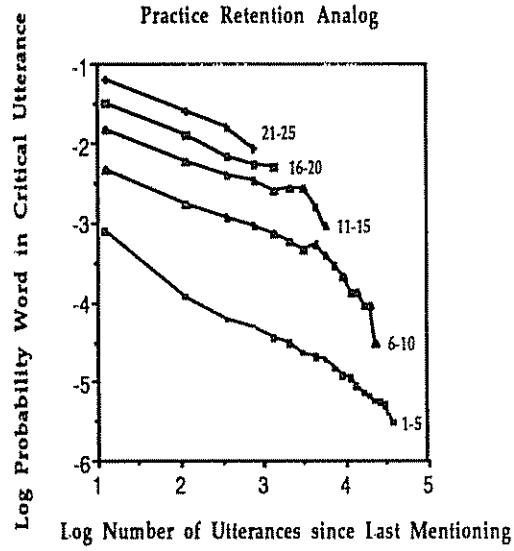

Figure 7. Retention curves at 4 frequency levels.

References

Anderson, J.R. & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review, 96 (4),* 703-709.

Burrell, Q.L. (1985). A note on aging on a library circulation model. *Journal of Documentation, 41,* 100-115.

Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *JVLVB, 15,* 1-16.

Hall, W. S., & Tirre, W.C. (1979). *The communicative environment of young children: Social class, ethnic and situation differences.* University of Illinois, Center for the study of Reading.

Hellyer, S. (1962). Frequency of stimulus presentation and short-term decrement in recall. *JEP* 64, 650.

Krueger, G.R. (1985). The effects of overlearning on retention. *JEP, 12,71-78.*

Loftus, G.R. (1985). Evaluating forgetting curves. *JEP: LMC,* 11, 397-406.

MacWhinney, B., & Snow, C. (1990). The child language data exchange system: an update. *Journal of Child Language,* 17, 457-472.

Newell, A. & Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In J.R. Anderson (Ed.), *Cognitive skills and their acquisition* (1-55). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stritter, E. P. (1977). *File Migration.* Unpublished doctoral dissertation, Stanford University, Stanford, CA.