A COMPLEMENTARY-SYSTEMS APPROACH TO ABSTRACT AND EPISODIC SPEECH PERCEPTION

Stephen D. Goldinger

Arizona State University
Goldinger@asu.edu

ABSTRACT

Over the past two decades, numerous experiments have shown that spoken word perception creates detailed memory traces, containing not only word meanings, but also extraneous, perceptual or contextual details. This is shown, for example, by voice-specific priming effects. Based on such results, exemplar theories suggest the mental lexicon may consist of accumulated episodic traces. Although an episodic approach is well-suited to explain priming, ample evidence suggests that language also entails abstract representations. Certainly at the segmental level, there are logical constraints that require unitization. An optimal theory may include stable abstract representations, combined with context-sensitive episodic traces. This paper summarizes new tests examining word perception from a *complementary-systems* perspective, wherein reciprocal neural networks represent hippocampal and cortical memory systems [13]. In this approach, detailed episodic traces and holographic, abstract traces combine to create behavior in real-time, allowing perceptual or memorial data to appear more or less "episodic," depending on myriad factors. I summarize new results and simulations on perceptual priming, and discuss the model with respect to perceptual learning in speech.

Keywords: Exemplar Models, Neural Networks, Priming, Perceptual Learning.

1. INTRODUCTION

In cognitive science, two fundamental ideas guide theory and research [15]. First, we assume that perception, categorization, etc. entail stable mental representations, ranging from idiosyncratic personal memories to shared, general knowledge. Second, these representations are activated and manipulated by transient operations, leading to behavior. Although this generic "*structures and processes"* framework is widely accepted, long-standing debates have concerned the precise nature of mental representations, whether the behavioral domain is perception, long-term memory, or language [e.g., 8, 12, 24]. In general terms, the question regards *degrees of abstraction*: Memory theorists have long recognized that people may retain episodic details of specific experiences [25], but that they more generally interact with perceptual objects in a categorical manner, failing to encode specific details, or perhaps failing to encode anything, once momentary cognition is past [10]. This seems especially true in language, as people may retain broad, narrative themes, but often forget precise wording. Indeed, the power of language derives from its generative nature [14], as a small set of elements (e.g., segments) can be arranged into a large set of words, which can, in turn, be arranged into an infinite set of messages.

Central goals in speech research are to elucidate the nature of segmental and lexical representations in long-term memory and the processes that access those representations. Most theories posit that speech input is recoded into strings of units, which are then matched to entries in the mental lexicon. This standard view emphasizes abstract properties of lexical entries: Speech signals with idiosyncratic information (e.g., voice, accent, speaking rate) are reduced to sequences of ideal, abstract phonemes, which activate ideal, abstract words. This view is undeniably correct at many levels of analysis; tremendous evidence supports abstraction in speech perception, production, and memory. However, growing evidence suggests there is more to the story: *Phonetically relevant surface details* of spoken words have wide-ranging effects in experiments. For example, repetition priming is more robust when presentation voices are constant across word repetitions [2, 5]. In normal listening, people acquire considerable knowledge about the speech of their interlocutors. This goes beyond spectral voice characteristics; people are closely attuned to the temporal and articulatory styles of their interlocutors, as shown in both perception and production measures [4]. Indeed, people recognize familiar voices even if spectral characteristics are

distorted by sinewave synthesis or reversal [19, 20]. Notably, speech perception is essentially unaffected by random variation in *non-phonetic* acoustic attributes, such as loudness or fundamental frequency [22]. Together, the findings suggest that, in natural conversation, people learn voice qualities, but are mainly sensitive to different speakers' unique manners of achieving phonetic contrasts, their articulatory habits. This has profound theoretical implications, as it places rational limits on the phonetically (or in some cases, conceptually) relevant aspects of episodes that are encoded and used by perceptual and memorial processes.

2. TOWARD A RATIONAL THEORY: ABSTRACT AND EPISODIC SYSTEMS

Given findings such as voice-specific priming, it becomes clear that *purely* abstract models of word perception are incomplete – some mechanisms are required to allow perceptual learning, and to apply prior episodes to new acts of perception. In prior work, an approach was taken directly from the categorization literature, starting with a strong hypothesis that the mental lexicon holds detailed episodic traces, rather than abstract units. For example, Goldinger [5, 6] applied an exemplar model, MINERVA 2 [8, 9], to examine data in word perception, production, and memory. In this model, separate memory traces are created for each perceptual experience (in this case, recognizing isolated words). Aggregates of memory traces are activated to create experience; abstraction occurs during retrieval. MINERVA 2 thus accounts for specificity and generality of memory with one set of exemplar traces. In word perception, it can exhibit behaviors that reflect either singular traces or aggregates, depending on the manner of testing.

Although exemplar models provide a simple solution to the abstract-episodic dilemma, they are not wholly satisfying, for several reasons. First, exemplar models seemingly demand vast memory resources. To avoid this, many assume partially overlapping (holographic) traces, with abstractions naturally forming over time [23]. This solution is elegant, but blurs the line between episodic and abstract representations. (In MINERVA 2, traces are not combined in storage, but they decay over time, effectively reducing their individuality.)

Second, any reasonable theory must postulate that memories are created by reference to *psychological experience*. Episodes cannot store "raw data," but must reflect attended cognitive events. Complex waveforms strike the ears, but words and ideas are retained, perhaps with some surface and/or contextual information. More generally, upon hearing speech waveforms, listeners generate meanings, perhaps in ways that reflect personal biases, affect, etc. Although this is clearly the correct theoretical posture, we again see the conceptual conflict with exemplar models: To adequately explain behavior, they must presume that abstract knowledge is imposed upon each encountered stimulus. As a result, each stored "exemplar" is actually a product of perceptual input combined with prior knowledge, the precise balance likely affected by many factors.

Third, but related to the prior point, perceptual processing does not only elaborate upon stimuli at the "global" level, such as endowing meaning to an utterance. Because language relies upon discrete segments (i.e., phonetic contrasts), perception also imposes unitization, although the units' exact form is subject to debate [7]. Models include various means to discover internal structures, such as segments within words, with varying degrees of success. In many cases, featural or segmental units are simply provided as input. This is certainly true for MINERVA 2, which includes no perceptual processes; vectors are simply stored in its memory, with elements corresponding to units such as letters or phonemes. In an elaborated "intertrace resonance" version, reciprocal loops among traces can reinforce the perception and differentiation of such internal structures. Although this elaboration allows the model to better reflect discrete linguistic behavior (e.g., reading nonwords), the previous concern arises again, as the nature of "exemplars" becomes unclear. Perceptual unitization prevents episodic models from devolving into *template models*, storing perceptual objects without internal structure, but it also complicates all encoding and matching procedures.

2.1. The Complementary-Systems Approach

Conceptual elaboration and segmental analysis are fundamental to language perception, and both carry similar implications for exemplar theories: Encoded traces must reflect perceptual operations, along with considerations of selective attention, encoding biases, etc. The exemplar substrate must contain *cognitive objects*, which begs many deeper questions (e.g., what are the temporal boundaries of an episode?). Even if we set aside such issues,

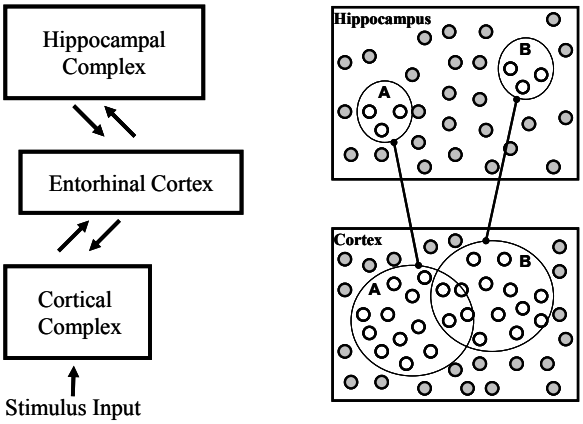
focusing on perception of isolated words, we can now appreciate a core theoretical challenge: Data and logic suggest that perception entails abstract analysis, but data and common experience suggest that perception also creates individuated memories. The challenge arises because, if a memory model brings prior experience to bear on a stimulus, it will naturally distort perception toward its stored prototypes. Conversely, if a model encodes all stimulus details, its core representations will be distorted, as in *catastrophic forgetting* [13]. Both of these processes occur in real memory, but on far "smaller scales" than models would suggest.

Consider a generic example of *perceptual learning* in speech [loosely based on 3, 17]. In a lexical decision task, a participant is repeatedly exposed to items with a mispronunciation of the same segment (e.g., /s/ produced as /ʃ/), but the remaining segments induce "word" decisions. In a relatively short time, people come to understand that /ʃ/ "stands for" /s/, and they will transfer that segment-level knowledge to novel stimuli, but this generalization is limited: It applies mostly to the speaker heard in training [3, 11], and it does not readily spread to other segments. Stated in simple terms, listeners learn that "this person says /s/ strangely," and adjust behavior accordingly. They do not change their segmental interpretation for other speakers (akin to catastrophic forgetting), and they are readily able to "unlearn" the pattern for the test speaker. Thus, people use abstract, segmental knowledge to "correct" the input, and use speaker- specific knowledge to guide generalization. This balance is difficult to achieve in a single memory system, such as a Hebbian neural network [23].

Like studies of repetition priming, studies of perceptual learning suggest that behavior arises from the combined action of abstract and episodic knowledge sources. In the literature, almost all theories, whether abstract or episodic, propose monolithic solutions to questions of lexical (or other) representation. Although simple models are desirable, there are many sources of evidence that mammalian brains do not rely on unitary learning systems. As reviewed in depth by McClelland, McNaughton and O'Reilly [13, also 16], the brain has evolved cortical systems that may subserve slow learning of general patterns, and hippocampal systems for fast learning of stimulus complexes.

In the *complementary learning systems* (CLS) approach, reciprocal neural networks are proposed, with a fast-learning "hippocampal" network and a more stable "cortical" network. The hippocampus is specialized to rapidly memorize specific events. Implemented in CLS, it is a sparse network that assigns pattern-separated representations to stimuli, allowing fast learning without catastrophic interference [16, 18]. In contrast, the cortex is specialized to slowly learn statistical regularities of the environment. In CLS, it assigns overlapping representations to similar stimuli, so it respects shared structures, as in words. The complementary systems are schematically illustrated in Figure 1.

Figure 1: General structure of the CLS model (left); schematic illustration of pattern overlap in cortex, and separation in the hippocampus (right, adapted from [16]).



With respect to word perception, the CLS approach has several attractive features. It solves the abstract-episodic dilemma by positing an openly hybrid memory system, but this is not an ad-hoc or simplistic solution: Although the model includes episodic and abstract representations, the systems are completely inter-dependent. Traces in the hippocampus are formed by streams of input from different parts of cortex. In the present case, these would include input from the cortical system that segments spoken words and assigns meanings, but also might include streams representing visual input, emotional responses, etc. Thus, although the hippocampus is designed to learn unique traces, its input has already experienced some degree of abstraction; this occurs in the earliest moments of word perception. In complementary fashion, cortical representations are created by accumulated episodes, stored in the hippocampus and slowly consolidated back into cortex. This cycle creates a means for gradual learning in the cortical system, without catastrophic interference. For example, extensive exposure to unique traces (e.g., regional accent) will slowly affect abstract knowledge (the

"Madonna effect"). CLS also offers an elegant solution to the *plasticity-stability* dilemma [1], preventing the either memory system from unduly overwriting the other. By aligning its mechanisms with brain anatomy, the CLS is compatible with hemispheric differences in font-specific priming [12], and is compatible with a far broader literature in neuroscience.

3. AN APPLICATION TO PRIMING

The CLS model has been extensively developed [16, 18], including an anatomically motivated model of the hippocampus that exceeds the present needs. For the present example, two sets of simulations were conducted. The first used the full model from Norman & O'Reilly (available online: *http://compmem.princeton.edu*), which is designed to provide estimates of recognition memory. Although it qualitatively fit the results reported below, similar results were also possible with a simpler three-part network, as shown in the left side of Figure 1: In this model, the "cortical network" was a classic PDP model [13], with three layers and delta-rule learning. This interactive-activation network resembled TRACE, trained to convert feature vectors into lexical outputs, with two hidden layers for internal units. The outputs were structured as strings of segments, with predetermined "semantic" units added; these corresponded to lexical access.

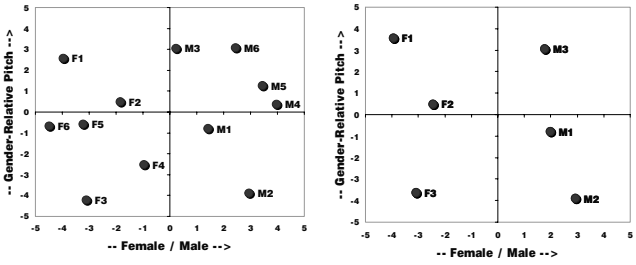
Output from the cortical network was passed through a simple "entorhinal cortex" network, a single-layer self-organizing map. In addition to the lexical output, this network received a random set of the original "acoustic" features, some "voice" features, and some "context" features. These were fed into the entorhinal network as separate inputs, but the self-organizing map functioned to blend them into a single distributed representation. In essence, this network creates an elaborated version of the original input – it preserves some perceptual and contextual features, but also contains each word's "name" and "meaning." When simulating an experiment on voice-specific priming, I held the contextual cues constant, varying the voice cues to match inter-voice similarities.

Once the input streams were combined in the entorhinal network, its output was fed into the hippocampal network. This network was another self-organizing map, based on a Hebbian learning model from Stark and McClelland [23]. This was an attractive choice because it produces "settling times" that reflect word perception efficiency, an index of familiarity. Finally, as shown in Figure 1, the hippocampal network connects back to cortex, which critically allows recent experiences to help influence early perceptual processes. Such mutual connectivity is required for long-term learning, as in learning unfamiliar phonemic contrasts, and is also naturally related to the guidance of selective attention, as prior traces can optimize performance.

3.1. Voice-Sensitive Priming

To evaluate the CLS model, I simulated a new experiment that assessed voice-specific priming, as a function of inter-voice distances in psychological space. This was an extensive experiment, but only a few aspects are relevant here. Following a prior study [5], the first step was to select a set of voices that covered a broad, continuous range of distances from one another, determined by multidimensional scaling. Twelve paid volunteers each recorded three tokens of 1000 bisyllabic words, and also initial syllables for each. A subset of tokens was used to collect direct voice similarity ratings from 102 volunteers, leading to the solution in Figure 2, left panel. From this space, 6 voices were chosen to maximize dispersal; another 66 people provided new similarity estimates to the subset, leading to the solution shown in Figure 2, right panel.

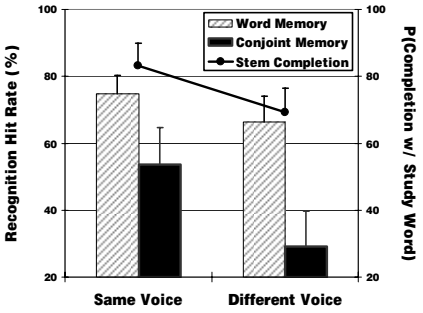
Figure 2: Two-dimensional MDS space for 12 voices (left) and rescaled space for final 6 voices (right).



The 6 voices chosen for the experiments had a nice distribution of inter-voice distances. In one experiment, 216 people initially heard 600 two-syllable words, with 100 per voice. After a 2-hour break, they heard 900 words in a three-way recognition task, wherein participants classified spoken words as *old/same-voice*, *old/different-voice*, or *new*, Test items included 300 old/SV words (50 per voice). Note, however, that "same-voice" items were not identical to study items; they were different tokens. There were 300 old/DV words, with all study-test voice combinations used equally (15 combinations, 20 times each). There were also 300 new words, with 50 per voice. In a

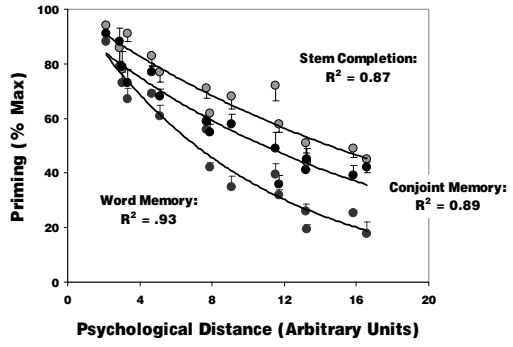
separate experiment (N = 144), participants in the test phase heard initial syllables of old and new words, completing them with the first syllable that came to mind [2]. By counterbalancing the old and new stems (new syllables, not truncated words) across groups, priming scores were derived within and across voices. The overall results are shown in Figure 3; in the figure, "word memory" denotes hit rates to words, irrespective of voice memory, and "conjoint memory" denotes word and voice hits. Voice effects were robust in all tests.

Figure 3: Results, shown as word and contingent recognition hit rates, and study stem-completion rates.



Although the observed voice effects replicated and extended prior work [5], the main goal of this experiment was to evaluate the relation of voice similarities to priming, at a finer grain of analysis. To achieve this, a series of tests were conducted, combined here for brevity. For each voice, SV hit rates were scaled to equal 100%, then hit rates for all DV trials were converted to proportions of that SV value, separately for each voice combination. Thus, SV priming is set to a common scale, and we can evaluate degrees of DV priming, contingent upon psychological similarity for each voice pair. The results are shown in Figure 4.

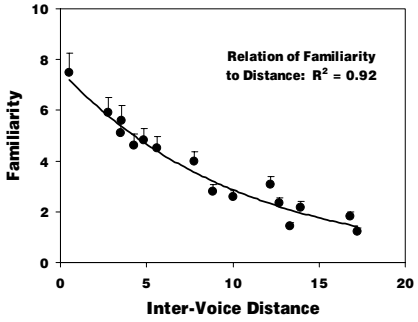
Figure 4: The relation of priming to inter-voice distances, with best-fitting exponential functions.



In every case, there was a consistent decline in priming as psychological distance between voices increased. In fact, the results closely approximate *Shepard's Law* [21], with excellent fits of exponential decay functions to the data. To test the CLS model, a simulation was conducted, with words coded using an input "alphabet," based on features. Of primary importance, different "voices" were created, vectors structured such that their dot products recreated the distance matrix for the six-voice scaling solution. In the simulation, words in all voices were presented, as in the experiment, with constant context. In the self-organizing map, this causes an interesting pattern: Although similar words should generally overlap (and they do in the cortical network), the hippocampal network is less constrained. In this case, the shared context of all study words creates a large region, defining the experimental setting. In this region, words are next organized by voices, as they are the second best source of constant input. There is considerable overlap, however, across voices. Finally, words within and across voice regions naturally organize themselves by segmental similarity. (Recall that "semantic" features were random, not expected to affect representation.) In short, after the "study session," the hippocampal network shows a rough approximation of the input space.

In the "test session," old words were presented to the model, with original and changed voices, as in the experiment. Following Stark & McClelland [23], presentation of each word leads to a series of cycles in the model, until it settles to a solution. In the present case, settling times closely approximate stimulus familiarity. The simulation results are shown in Figure 5. For consistency with Figure 4, voice-similarities have been multiplied by a constant; *familiarity* denotes inverse settling times, multiplied by 10. As shown, same-voice trials lead to fastest settling; performance declines steadily with larger voice changes.

Figure 5: Simulation of voice familiarity across magnitudes of subjective voice changes.



4. DISCUSSION

The simulation of voice-sensitive priming was useful, as it shows that activity in the hippocampal

network reflects changes in the surface features of words. One last test deserves brief mention. In the CLS framework [16, 18], the hippocampal network is not required to fully recreate perceptual input, such as generating fully segmented words. Because hippocampus is reciprocally connected to cortex, it can help reinstate lexical activity, given partial input. In another simulation, I consistently paired one word with one voice, repeated 20 times in a list of 400 random words in five other voices. As the model is devised, voice features are entered directly into the entorhinal network. Nevertheless, when the unique voice was presented to the model with a random string of features, the previously overlearned word was most strongly activated (or perhaps hallucinated) in the cortical network. This is a crude example, but it suggests that high-level representations in the hippocampus, such as task-specific strategies, can help guide or misguide early perceptual processes.

In testing the CLS model, O'Reilly & Rudy [18], found that it is capable of transitive learning. As mentioned earlier, perceptual learning in speech is transitive, as a phonetic principle discovered in a sample of words is easily extended to new words. Indeed, the empirical profile of perceptual learning seems to demand an account such as CLS: Given a few tokens of strangely pronounced words, the listener quickly adjusts perception of other words sharing the odd segment. But this change is mainly limited to the particular acoustic token, is mainly speaker-specific, and does not otherwise adversely affect phonetic perception. The abstract lexicon is required to interpret an odd segment; episodic memory is required to both generalize and delimit the effect. The CLS offers a rapprochement for abstract and episodic theories of language; both forms of representation are mutually created in a reciprocal loop, uniting long-term memory with real-time perception.

5. REFERENCES

[1] Carpenter, G., Grossberg, S. 1991. *Pattern recognition by self-organizing neural networks.* Cambridge: MIT Press.

[2] Church, B., Schacter, D. 1994. Perceptual specificity of auditory priming: Memory for voice intonation and fundamental frequency. *J. Exp. Psych: Learn., Mem., & Cog.*, 20, 521-533.

[3] Eisner, F., McQueen, J. 2005. The specificity of perceptual learning in speech processing. *Perception & Psychophysics,* 67, 224-238.

[4] Fowler, C., Brown, J., Sabadini, L., Weihing, J. 2003. Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *J. Mem. and Lang.,* 49, 396-413.

[5] Goldinger, S. 1996. Words and voices: Episodic traces in spoken word identification and recognition memory. *J. Exp. Psych: Learn., Mem., & Cog.*, 22, 1166-1183.

[6] Goldinger, S. 1998. Echoes of echoes? An episodic theory of lexical access. *Psych. Review*, 105, 251-279.

[7] Goldinger, S., Azuma, T. 2003. Puzzle-solving science: The quixotic quest for units in speech perception. *J. Phonetics*, 31, 305-320.

[8] Hintzman, D. 1986. "Schema abstraction" in a multiple-trace memory model. *Psych. Review,* 93, 411-428.

[9] Hintzman, D. 1988. Judgments of frequency and recognition memory in a multiple-trace memory model. *Psych. Review,* 95, 528-551.

[10] Keenan, J., MacWhinney, B., Mayhew, D. 1977. Pragmatics in memory: A study in natural conversation. *J. Verbal Learning and Verbal Behavior,* 16*,* 549-560.

[11] Kraljik, T., Samuel, A. 2005. Perceptual learning in speech: Is there a return to normal? *Cog. Psych.*, 51, 141–178.

[12] Marsolek, C. 2004. Abstractionist versus exemplar-based theories of visual word priming: A subsystems resolution. *Quarterly J. Exp. Psych.,* 57A, 1233-1259.

[13] McClelland, J., McNaughton, B., O'Reilly, R. 1995. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psych.. Review,* 102, 419-437.

[14] Miller, G. 1965. Some preliminaries to psycholinguistics. *American Psychologist,* 20, 15-20.

[15] Neisser, U. 1967. *Cognitive Psychology*. New York: Appleton-Century-Crofts.

[16] Norman, K., O'Reilly, R. 2003. Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning-systems approach. *Psych. Rev.,* 110, 611–646.

[17] Norris, D., McQueen, J., Cutler, A. 2003. Perceptual learning in speech. *Cog. Psych.,* 47, 204-238.

[18] O'Reilly, R., Rudy, J. 2001. Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psych. Review, 108*, 311-345.

[19] Remez, R., Fellowes, J., Rubin, P. 1997. Talker identification based on phonetic information. *J. Exp. Psych: Human Perc. & Perf.*, 23, 651-666.

[20] Sheffert, S., Pisoni, D., Fellowes, J., Remez, R. 2002. Learning to recognize talkers from natural, sinewave, and reversed speech samples. *J. Exp. Psych: Human Perc. & Perf.*, 28, 1447-1469.

[21] Shepard, R. 1987. Toward a universal law of generalization for psychological science. *Science*, 237, 1317-1323.

[22] Sommers, M., Barcroft, J. 2006. Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *J. Acoust. Soc. Am.,* 119, 2406-2416.

[23] Stark, C., McClelland, J. 2000. Repetition priming of words, pseudowords, and nonwords. *J. Exp. Psych: Learn., Mem., & Cog.,* 26, 945-972.

[24] Tenpenny, P. 1995. Abstractionist vs episodic theories of repetition priming and word identification. *Psychonomic Bulletin & Review*, 2, 339-363.

[25] Underwood, B. 1969. Attributes of memory. *Psych. Review,* 76*,* 559-573.