Chunking Processes and Context Effects in Letter Perception

Emile Servan-Schreiber
Department of Psychology, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213
Schreiber@lentil.psy.cmu.edu

Abstract

Chunking is formalized as the dual process of building percepts by recognizing in stimuli chunks stored in memory, and creating new chunks by welding together those found in the percepts. As such it is a very attractive process with which to account for phenomena of perception and learning. Servan-Schreiber and Anderson (1990) demonstrated that chunking is at the root of the "implicit learning" phenomenon, and Servan-Schreiber (1990; 1991) extended that analysis to cover category learning as well. This paper aims to demonstrate the potential of chunking as a theory of perception by presenting a model of context effects in letter perception. Starting from a set of letter segments the model creates from experience chunks that encode partial letters, then letters, then partial words, and finally words. The model's ability to recognize letters alone, or in words, pseudo-words, or strings of unrelated letters is then tested using a backward masking task. The model reproduces the word and pseudoword superiority effects.

To overcome the limited capacity of its short term memory a mind organizes its input into familiar chunks (Miller, 1956). From this fact we can directly derive two more: First, when confronted with a set of input features, a mind will seek to recognize configurations of features, or chunks, that it has stored in its long term memory, and the resulting percept in short term memory will consist of those recognized chunks. Second, additional chunks will be created, and stored in long term memory, by welding together some of the chunks that make up the percept. We have here a general description of an adaptive recognition machine that learns continuously in order to perceive better: A chunking machine.

The facts that minds perceive and learn by chunking have been heavily documented (e.g., Bartram, 1978; Buschke, 1976; Chase & Simon, 1973; Johnson, 1970; Newell & Rosenbloom, 1981), yet most current models of perception, and letter perception in particular, overlook those facts (e.g., McClelland & Rumelhart, 1981; Oden, 1979; Massaro & Sanocki, in press). In this paper I demonstrate that the perceptual advantage of letters in words and pseudowords over letters in unrelated letter strings is a natural characteristic of a chunking machine that has learned, from scratch, to recognize letters and words.

The Chunking Model

Chunks. A chunk is a long term memory hierarchical structure whose constituents are chunks also. There are two kinds of chunks: Elementary chunks are those that the cognitive system never had to create. They are assumed to be the output of an elementary perceptual system. All other chunks are created by welding together lower level chunks. (Any theory of chunking must assume a limit on chunk size, the number of chunks that can be welded together into a new chunk. For simplicity, this theory assumes that it is 2, the lowest number that still enables chunk creation.)

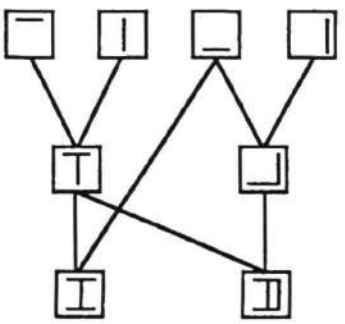

Figure 1. A potential hierarchical network of chunks that encodes the letters T, I, and D in terms of simple line segments. Structurally similar letters may share subchunks, and entire letters may be subchunks of more complex letters.

The letter perception model assumes that simple straight line segments are elementary chunks. As Figure 1 illustrates by showing a potential hierarchical structure of three chunks for the letters T, I, and D, the letters that were used as stimuli in the simulation were made of such simple segments.. Note that, as in this case, structurally similar letters may share subchunks.

Perception. Chunks are used to perceive stimuli, and given the recursive and hierarchical nature of chunks the perception process is necessarily cyclical and bottom-up. Starting with an elementary percept that contains all the elementary chunks present in the stimulus (e.g., letter segments), each cycle of perception seeks to reduce the number of chunks currently in the percept by replacing pairs of chunks in the percept by a chunk that encodes their co-occurrence. This operation is called *encoding*. For example, if the chunks w, x, y, and z are in the percept and there are chunks (w x) and (y z) in memory, then the next cycle of perception puts (w x) and (y z) in the percept in place of their constituent chunks. And if the chunk ((w x) (y z)) is also in memory, then the percept can be encoded further on the following cycle. The process can continue to cycle until the percept cannot be encoded further.

Encoding occurs in parallel on each cycle. As the example above illustrates, two or more chunks can encode the percept simultaneously in one cycle. But there are potential conflict situations: For example, if the percept contains the chunks [w,x,y,z], and there are chunks (w x), (x y), and (y z) in memory, then (w x) and (y z) are compatible encoders while (x y) is incompatible with both of them. To resolve such conflicts, the model first collects the set of all the candidate encoders, then randomly selects a subset of those that are all compatible. Thus the percept could be encoded either as [w,(x y) z] or as [(w x),(y z)].

Note that the choice that is made is not without consequence for the next cycle of perception. If there is a chunk ((w x) (y z)) in memory then it has a chance to encode the percept on the following cycle if it is [(w x),(y z)], but not if it is [w,(x y),z]. As is common with simple hill-climbing procedures, this bottom-up perception process can easily get stuck in a non-optimal encoding of the stimulus.

A simple way to avoid getting stuck in non-optimal encodings is to allow the process to backtrack through a *decoding* operation. To decode a chunk that is in the percept is to remove it and replace it by its subchunks. For example, if the chunk (x y) is decoded in the percept [w,(x y),z] then the resulting percept is [w,x,y,z].

In the model, every perception cycle consists of an encoding stage followed by a decoding stage. Every chunk that is in the percept at the end of an encoding stage has a probability, dp, of being decoded before the onset of the next cycle, and if a chunk is decoded at the end of a cycle, it is forbidden to be an encoder on the immediately following cycle. A chunk's dp is determined throughout the perception process in the following way: Elementary chunks come into the process with an initial probability of being decoded. Then every encoder chunk comes into the percept with a dp that is equal to the average of its subchunks' dps. Finally, and most importantly, whenever an encoding stage has failed to retrieve any encoder the dp of every chunk in the percept is decreased by a small amount (e.g., .01). Thus, the perception process will oscillate between different percepts, but oscillations will become more and more unlikely as the dps of the chunks in the percept decrease. Eventually, the process settles on a stable percept when the dp of each chunk is zero.

This is a straightforward application of simulated annealing to bottom-up encoding. It has the nice property of being likely to settle in one of the more, and often the most, encoded interpretation of the input, as the following example illustrates: Consider Figure 2. It represents the different percepts that the process oscillates between when presented with a D, and given a hypothetical network of chunks. In each network in the figure the chunks that make up the percept are enclosed in bold squares. Thus percept P1 is the elementary percept that consists of the elementary segments of a D, while percept P6 contains only a D chunk. Encoding proceeds from top to bottom on the page, while decoding proceeds from bottom to top following the arrows. There are two possible minima: P5 which represents an "I" and an isolated segment, and P6 which represents a "D". To move from P5 to P6 requires at a minimum 1 decoding followed by 2 encodings, but to move from P6 to P5 requires at a minimum 2 consecutive decodings followed by 1 encoding. Because encoding is guaranteed at every cycle (provided that there exists a pertinent encoder), while decoding is probabilistic, it is easier to encode than to decode. Therefore, given any probability of decoding it is easier to move from P5 to P6 than to move from P6 to P5, and that difference increases as the probability of decoding chunks decreases. So the process will tend to settle on P6 (D) much more often than on P5 (I+ |).

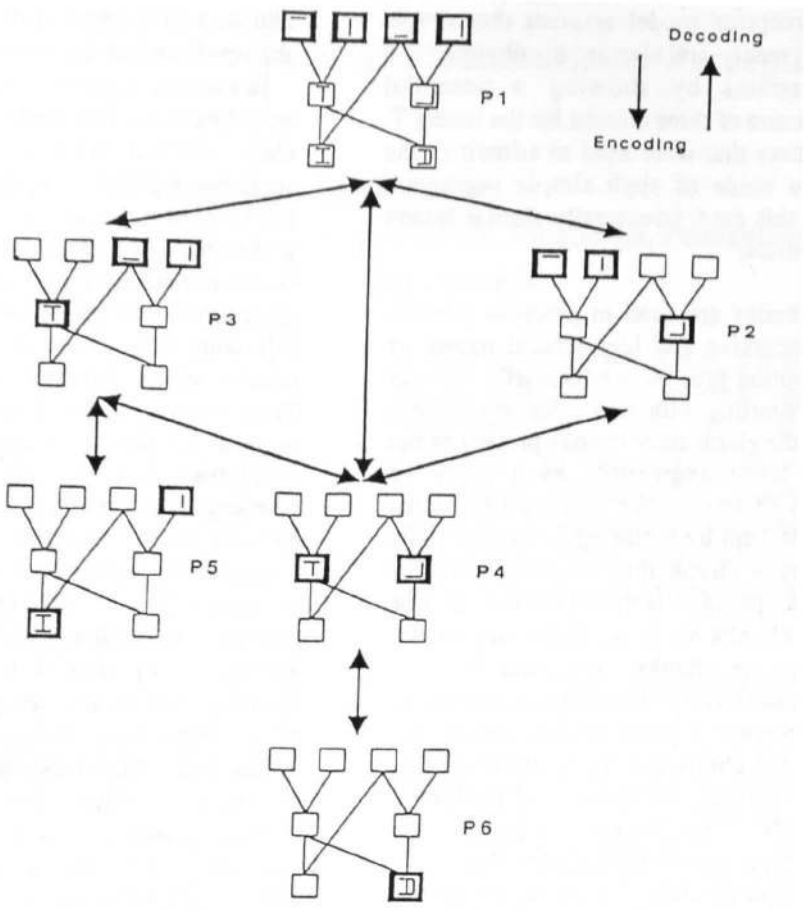

Figure 2. Starting with percept P1 (the segments of the letter "D") the perception process will tend to oscillate between the two possible minima P5 ("I" + "|") and P6 ("D") before settling preferably on P6, the most encoded interpretation of the input. The chunks of each percept are enclosed in bold squares, and arrows indicate how encoding and decoding transform one percept into another.

Learning. Once the perception process has settled on a final percept, a collection of chunks, a new chunk is created by selecting a pair and welding it into a new chunk (if it does not already exist in memory). In cases where the final percept consists of a single chunk, the creation process is not engaged. The selection of the pair of chunks that will be welded into a new chunk is essentially random but may be constrained. For example, when letter segments are welded together into a letter or a partial letter, a constraint may be that the two segments selected must be connected or parallel.

The combination of the perception and chunk creation processes allows the model to continuously grow a network of chunks from its experience with successive stimulus exposures, given only a minimum set of elementary chunks to start with.

Test of the Model

The model was tested with respect to its ability to reproduce several important results in the letter perception literature. They are: (1) The perceptual advantage of letters in words over letters in pronounceable nonwords (also called pseudowords), letters in strings of unrelated letters, and letters presented alone. (2) The perceptual advantage of letters in pseudowords over letters in strings of unrelated letters. (3) The reversal of the advantage of letters in words over letters presented alone at long exposure durations. For a review of those results see McClelland & Rumelhart (1981), or Massaro & Sanocki (in press).

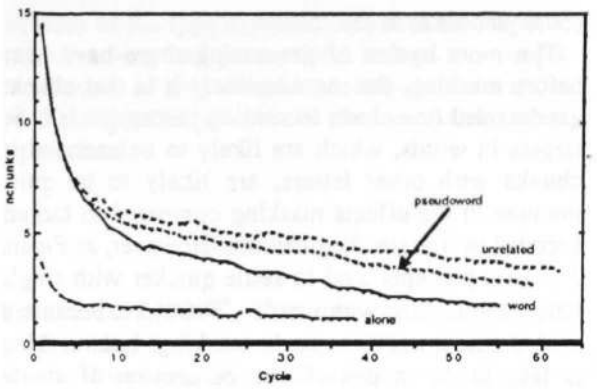

Figure 3. Evolution of the number of chunks in a percept, *nchunks*, with increasing number of perception cycles. There is one curve for each type of stimulus: a single letter (alone), a four letter word (word), a four letter pseudoword (pseudoword), and a string of four unrelated letters (unrelated). Individual curves are plotted up to the cycle where, on average, perception has settled on a stable percept.

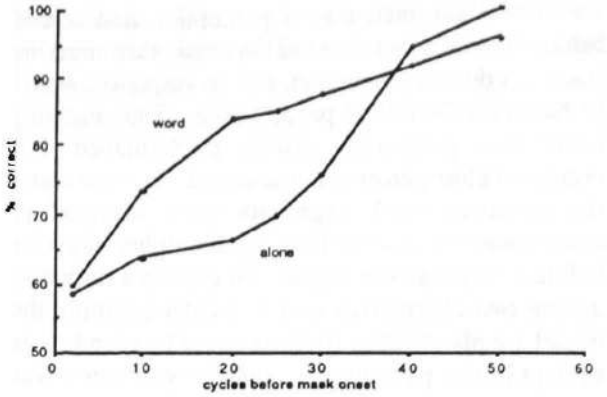

Figure 4. Evolution of the percentage of correct forced-choice recognition of target letters, presented in words or alone, with increasing number of processing cycles before masking.

All these results were obtained in experiments where a high contrast stimulus letter is presented for a short duration in a particular context, followed immediately a high contrast masking stimulus. The subject is then asked to choose among two possible letters which one was in a particular position. For example the subject might see "WORK" quickly followed by a mask "####" and be asked whether "K" or "D" was in the fourth position. As in this example, when the context forms a word with the target, the foil does also, therefore guessing is controlled for.

Before any simulation of those results could be attempted, the model first had to grow a network of chunks to represent letters and words. To start it was given a small set of elementary chunks to represent letter segments of different lengths and orientations. It then learned chunks to recognize the 26 individual letters of the alphabet. Once it could perfectly recognize any letter presented alone, it learned chunks to recognize each word in a sample of 288 four letter words. In the end, from the original 8 elementary segment chunks the model grew a network containing 18 partial-letter chunks, 51 letter chunks, 1665 partial-word chunks, and 1339 word chunks. These numbers indicate a large amount of redundancy in the representation of letters and words.

Time course of encoding. In a first experiment the model was tested for its ability to encode words, pseudowords, single letters, and strings of unrelated letters. Figure 3 plots for 144 stimuli of each type the average number of chunks, *nchunks*, in a percept after a given number of perception cycles. Individual curves are plotted up to the cycle where, on average, perception has settled on a stable percept. For this experiment, and the two that follow, the initial decode-probability was set at .30. Note that the process settles faster when a single letter is presented that when a multi-letter stimulus is presented. For instance, even though the letters in a word are processed in parallel, recognizing a whole word takes longer than recognizing a single letter. Note also that the less related the letters in a multi-letter stimulus are (as evidenced by the final nchunks), the longer it takes to settle on a final percept. These two results can only be attributed to a kind of lateral interference that is an emergent property of the annealing process. To put it simply, the less related the letters are, the more chunks there are in the percept at any cycle, so the more chance of decoding there is, and therefore the longer it takes to encode.

Words vs. single letters. In a second experiment, the recognition of letters in words and letters presented alone was compared given different numbers of cycles before masking. The mask used by the model was the union of the letter segments in "O", "X", and "+", and masking was simulated by adding to a percept those spurious segments of the mask that were not already present in the stimulus.

The model assumed that if perception had settled before the scheduled onset of the mask, then masking could not disrupt perception, that is, responses would be based on the settled percept only. Thus masking could only potentially disrupt performance if it occurred before perception had settled. In those cases, the spurious mask segments were allowed to contaminate the percept for a small number of cycles before a response was made. To choose a response, among two alternatives in a particular position, the model simply checked if either could be found in its percept in that particular position. If yes then it was chosen, else, if neither or both were perceived in that position, then the model chose randomly. (Two or more letters could be recognized in a position because of the spurious segment introduced by the mask.)

Figure 4 plots the results of that experiment involving 144 targets in words and 144 equivalent targets presented alone. This simulation assumed 5 cycles of masking before a response was made. Like the human subjects of Massaro and Klitzke (1979) the model produced an advantage for letters in words that was eventually reversed at late masking onsets.

There are essentially three possible encoding states for the target letter in a percept before masking introduces spurious segments: (1) The target may not be encoded as a single chunk, possibly due to decoding or, more simply, encoding failure. (2) The target may be encoded as a chunk that is not part of a larger chunk, possibly due to the decoding of a larger chunk, or the fact that it was presented alone. (3) The target may be encoded as a chunk that is further encoded in a larger chunk, or a hierarchy of larger chunks. Masking has potentially different effects in any of those three states: (1) If the target is not even encoded as a chunk then the spurious segments of the mask can prevent future encoding of the target by being encoded together with target segments into chunks incompatible with the target's structure. (2) If the target is encoded as a free standing chunk then on each masking cycle there is a probability that this chunk gets decoded with dire consequences as in (1), or that a spurious mask segment gets encoded with the target chunk into a chunk that represents another letter (for instance, the addition of a single letter segment to the chunk for "P" can transform it into an "R"). (3) If the target chunk is well hidden within a further encoded hierarchy, for example in a chunk for a complete word, or part of a word, and these larger chunks resist decoding during the masking cycles, then masking has no harmful effect. But if those larger chunks get decoded, then there may be dire consequences as in (2).

The more cycles of processing there have been before masking, the more unlikely it is that chunks get decoded (the closer to settling perception is). So targets in words, which are likely to be encoded in chunks with other letters, are likely to be quite immune to the effects masking compared to targets encoded in free standing chunks. However, as Figure 3 shows, percepts tend to settle quicker with single letter stimuli than with words. Therefore, because a settled percept is immune to masking, late masking is less likely to disturb the perception of single targets than that of targets presented in words.

Words vs. pseudowords vs. strings of unrelated letters. A final experiment compared the forced-choice recognition of targets in words, pseudowords (e.g., MIPE), and unrelated letter strings (e.g., TCKU). 30 cycles of perception were allowed before 10 cycles of masking. There were 144 stimuli in each condition. The results are in Table 1. Like human subjects the model produced a large advantage for words and a smaller advantage for pseudoword over strings of unrelated letters (e.g., McClelland & Rumelhart, 1981).

The time course of encoding of the different stimulus types plotted in Figure 3 indicated that words are encoded as fewer chunks than pseudowords, themselves encoded as fewer chunks than strings of unrelated letters. This is simply a reflection of the different amounts of relatedness between the letters in the three stimulus types. And as the analysis of the previous experiment demonstrated, more compact encoding directly translates into less adverse effect of masking.

Table 1

Correct forced-choice recognitions of letters in words, pseudowords, and strings of unrelated letters.

Word	Pseudoword	Unrelated
84.4 %	78.6 %	70.3 %

To conclude, briefly, this limited testing of the model demonstrated its potential as a theory of letter perception. Further testing is certainly warranted, but considering that the same chunking analysis was successfully applied elsewhere to "implicit learning" and to category learning (Servan-Schreiber & Anderson, 1990; Servan-Schreiber, 1990; 1991), there is some reason to be confident that chunking

processes of the type explored here underly much of human learning and perception. Indeed, one major contribution of this chunking analysis is to show how these apparently unrelated phenomena are in fact deeply related.

References

Bartram, D. J. 1978. Post-iconic visual storage: Chunking in the reproduction of briefly displayed visual patterns. *Cognitive Psychology* 10:324-355.

Buschke, H. 1976. Learning is organized by chunking. *Journal of verbal Learning and Verbal Behavior* 15:313-324.

Chase, W. G., & Simon, H. A. 1973. Perception in chess. *Cognitive Psychology* 4: 55-81.

Johnson, N. F. 1970. The role of chunking and organization in the process of recall. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* 4:171-247. NY: Academic Press.

Massaro, D. W., & Sanocki, T. (in press). Visual information processing in reading. In D. Willows, R. Kruck, & E. Corcos (Eds.), *Visual processes in reading and reading disabilities.* Hillsdale, NJ: Lawrence Erlbaum.

Miller, G. A. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review* 63:81-97.

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

Servan-Schreiber, E. 1991. The Competitive Chunking Theory: Models of Perception, Learning, and Memory. Ph.D. diss., Dept. of Psychology, Carnegie-Mellon University.

Servan-Schreiber, E. 1990. Classification of dot-patterns with competitive chunking. In Proceedings of the Twelfth Annual Conference of the Cognitive Science Society, 182-189. Boston, Mass.

Servan-Schreiber, E., & Anderson, J. R. 1990. Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16:592-608.

McClelland, J. L., & Rumelhart, D. E. 1981. An interactive activation model of context effects in letter perception, part 1: An account of basic findings. *Psychological Review* 88:375-407.

Oden, G. C. 1979. A fuzzy logical model of letter identification. *Journal of Experimental Psychology: Human Perception and Performance* 5:336-352.