Exploring the Continuum of Unit Size in Word Identification

Catherine L. Harris
64 Cummington St., Department of Psychology
Boston University, Boston, MA 02215
charris@bu.edu (617) 353-2956

Abstract

Connectionist approaches to word recognition suggest that the units of word identification are not part of a fixed architecture, but emerge through extracting co-occurrence regularities. One implication of this idea is that unit-status, and the size of units, may be a matter of degree. This paper investigates the possible unit status of common word collocations, such as adjective-noun pairs (*next step, large part*) and verb-preposition combinations (*look out, appear in*). On analogy to the pseudo-words used in word superiority experiments, I contrasted letter detection in near-collocations (*next stem, barge part*) and random pairs (*next role, power part*) with performance on collocations (which had been defined as frequent combinations in a printed corpus). Although letter detection for collocations was not better than single words, detection was impaired for random pairs relative to single words and collocations. Near-collocations had a paradoxical effect that was only partially anticipated: an enhancing effect when letter targets were in the first word, and an inhibiting effect when targets were in the second word. Because reaction times were 400msec slower in the latter case, it was inferred that the near-collocations have a time-dependent effect, one of initial activation of neighbors, followed by inhibition.

Introduction

The word superiority effect (WSE) refers to the finding that laboratory subjects are more accurate at detecting a letter in a word than in a non-word or a letter alone (Reicher, 1969). Letter detection is also enhanced in pseudowords (strings which embody the orthographic regularities of English), even for pseudowords which aren't pronounceable (McClelland & Johnston, 1977). For example, *N* can be detected more easily in *SLNT* than in *SDNR*. These and other effects were explained in McClelland & Rumelhart's (1981) interactive-activation model of word recognition, a model which illustrated how elements in an interacting system can mutually constrain each other, and how rule-governed behavior can emerge in the absence of explicit rules.

McClelland & Rumelhart's IA model explained the enhancing effect of pseudowords by proposing that the familiar letter clusters (such as *SL* and *NT* in *SLNT*) activate the many words of which they are a member, and these words feed-back activation to their component letters, allowing, for example, S and *N* to receive more activation when viewed as part of SLNT than when viewed in the context of *SDNR*.

IA achieves these results by representing words as units which receive and send inhibitory and excitatory signals. Non-words do not have these characteristics. For humans, an open question is which mental entities have unit status, and why. Presumably the letter strings we call words come to have unit status via readers' frequent (and usually early) exposure to these letter combinations. Elman (1991) has shown that words as perceptual units can emerge from a back-propagation network trained to predict the next letter in a letter sequence composed of words strung together without separations. In a discussion of "subsymbolic psycholinguistics" Van Order, Pennington & Stone (1990) describe how the units of word identification emerge through extracting co-occurrence regularities, which they call "covariant learning". As they point out, "...any relatively invariant correspondence, at any grain size equal to or larger than the grain size of our subsymbols, may emerge as a rulelike force..." (Van Order et al, p. 504).

If unit status is a matter of degree, then units smaller than words, and units larger than words, could come to have a degree of unit status. The hypothesis that units smaller than words may have a type of unit status has been well researched, although usually in the context of determining the representational status of morphemes (e.g., are morphemes stored separately from the words of which they are a part; can readers search for morphemes in text; do morphologically related words prime each other).

In the current paper, I investigate the possible unit status of common word combinations. For simplicity, I look at two types of combinations: noun combinations (noun+adjective, adjective+noun and noun-noun: *next step, night club*) and verb-preposition combinations (*look out, appear in*). Since we recognize the cohesive quality of these pairs, it may seem obvious that they must have a type of unit status. But the nature of this cohesive quality has important implications for theories of lexical representation and language processing. Current theories propose that words are stored in a separate data structure (the lexicon) and are individually accessed and assembled into larger units (Forster, 1979; for a review, Emmorey & Fromkin, 1988). Although proponents of a unitary lexicon acknowledge that semantically cohesive word combinations (such as word compounds, cliches and idioms) may have their own "entries" in the lexicon, these com-

pound items are considered the exception to the intrinsic nature of the mental lexicon, which is a compendium of individual words and their meanings.

If common word combinations have unit status, then we may want to view the mental lexicon as being composed of items of varying size. The modal unit size may correspond to the word, but this would result from the statistical attributes of words, which itself may be a result of "functional unitization" (Van Orden et al., 1990) and the usefulness of this size for human language processing (Harris, 1994).

One method of investigating whether collocations like the noun compound *"night club"* have unit status is to see if these two words prime each other. Experiments have shown priming for noun compounds in lexical decision tasks (Hodgson, 1991; Harris, unpublished data). But enhanced lexical decision at most shows an associative link between two words. Hodgson (1991) has argued that semantic priming reflects an attempt at semantic integration initiated by the language comprehension system after lexical representations are accessed.

A more stringent test of the unit-status of a word combination would be if the hypothesized unit was able to feed activation down to the level of letters. I thus chose a forced-choice letter detection task as my method of exploring activation between units at one granularity level and units at another level.

WSE experiments typically contrast detection of a letter alone, a letter in a word, and letters in various types of non-words. One type of non-word is pronounceable and differs from a word by a single letter. A typical finding is that letter detection is facilitated in a pseudo-word relative to a random letter string.

One set of items was constructed to be analogous to a WSE experiment (see Table 1). These were the nouns materials. In these materials, subjects detected letters under 5 conditions: letter alone, target word alone, collocation, near-collocation (adjacent word is one letter removed from a collocation) and random word-pair (adjacent word is frequency-matched to the adjacent word in the collocation but is not associatively related to it).

These materials allowed me to answer the following questions:

- Does a collocation enhance detection, compared to a single word? A positive finding would be strong support for the unit-status of collocations. A negative finding could simply mean that letter detection is too fast for the enhancing effect of the collocation to be observed.
- Does a near-collocation enhance or inhibit detection, relative to a non-collocation? One might expect a near-collocation to enhance letter detection relative to a non-collocation, on analogy to how a pseudo-word enhances detection relative to a non-word. This possibility is diagrammed in Figure 1a. On the other hand, a near-collocation could inhibit detection, if the near-collocation competes with the collocation (Figure 1b). Support for this view comes from the "Word Inferiority Effect" (Chastain, 1986). Decreased letter detection occurs for letter

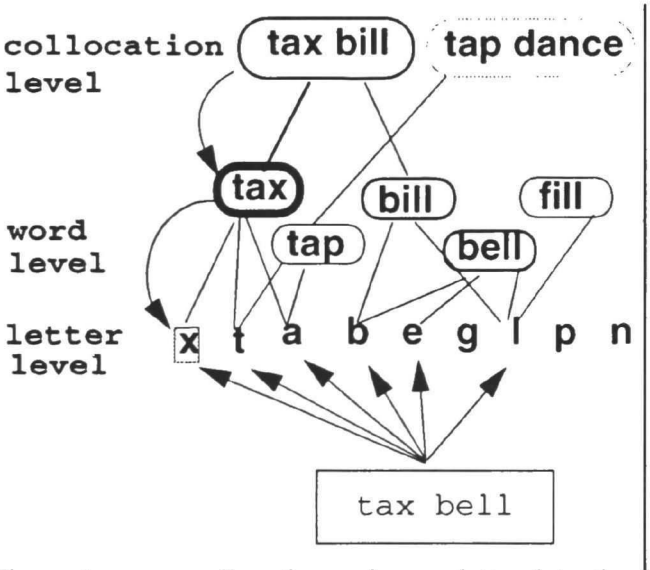

Figure 1a. near-collocation enhances letter detection relative to a non-collocation: The unit for bill is activated by letters in the input. This spreads activation to the collocation tax bill, which then feedback activation down to tax and the target letter x.

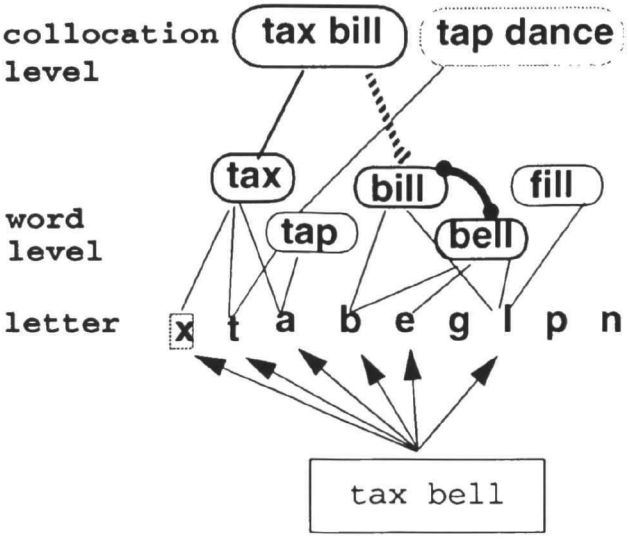

Figure 1b. near-collocation inhibits letter detection: Units for bill and bell are both activated by the input and compete, inhibiting each other. This prevents bill from activating tax bill, or may even lead to inhibition of tax bill (dotted line).

Table 1: Table 1: Example of Stimulus Materials

Sample Noun Materials

Collocation	Near-Colloc	Non-Collocation	Forced Choice Letters
(Letter target in first word)			
tax bill	tax bell	tax deep	x g
night club	night clue	night wall	n e
(Letter target in second word)			
focal point	vocal point	cargo point	o a
free world	tree world	open world	r u

Sample Verb Materials

High Freq.	Low Freq.	Anomalous	Forced Choice Letters
find out	find off	find for	d e
keep in	keep over	keep if	k d
give up	give on	give oy	u a
show up	show under	show ip	h l

strings in which an additional letter was interposed part-way through string exposure, if both strings made a word; e.g., *cat*, with interposed *s* to make *cast*). Chastain interprets this to mean that competition between words is inhibiting both words' activation levels, thereby decreasing letter-level activations.

- Does the effect of context differ depend on whether the target is the first or second word in a collocation? One reason *not* to expect enhanced letter detection in collocations compared to single words is that excitatory feedback from collocations to words, and from words to letters, may take too much time; by the time the feedback reaches the letter level, the word (and its letters) may already have been recognized. I hoped to be able to distinguish this possibility by comparing the patterns of letter detection when the target is in the second word, rather than the first, on the assumption that more time is required to read the second item in a two-word pair than the first.

The verb materials were used to explore the effects of frequency of a collocation, by contrasting high- and low-frequency verb-prepositions pairs.

Method

Materials

The experimental stimuli were 55 noun pairs and 27 verb-preposition pairs, constructed using the criteria described below. Subjects also saw 55 single words and 75 single letters.

Verbs and nouns used in the study were selected by filtering the Brown Corpus (Francis & Kucera, 1982) frequency listings for words that were less than 8 characters long and appeared more than 100 times in the million-word corpus. The frequency of each words' left and right neighbors (in the Brown Corpus) was then tabulated.

Noun materials

Nouns were selected for inclusion in the study if they followed or proceeded another content word such that the pair had a frequency of at least 2 (mean for the final set was 6) and the resulting string was 11 characters or less (counting the blank space as a character). 55 nouns met this criteria. 28 were the first member of the pair, and 27 were the second member of the pair. Two types of control items were also selected (see Table 1). The near-collocations are items in which one letter has been changed (preserving word-status) in the word which is *not* targeted for letter detection. The random-pair control items keep constant the word targeted for letter detection, and pair it with a word which is length- and frequency-matched to the analogous item in the collocation.

Verb materials

Verbs were most frequently followed by prepositions (*e.g., find out* and *live in* had counts of 34 and 30, respectively). Because there were theoretical reasons to believe that verb+preposition may have unit status (Harris, 1990; 1994) (and to increase stimulus homogeneity), verb items were restricted to verb+preposition (or particle) pairs. Ten prepositions with clear semantic content were selected. Frequencies were obtained for all verb+preposition combi-

nations. 27 verbs were found which met the following criteria: Each verb occurred in a high-frequency verb-preposition pair (mean frequency of 10 counts per million), a low-frequency pair (occurred only 1 time per million), and an illegal pair (never occurred in the Brown corpus, and was judged anomalous by two independent raters). In order to keep repetition of the prepositions to a minimum (to avoid repetition priming), in the "illegal" condition, 18 of the 27 verbs were paired with either a non-preposition or a nonword (which was one letter removed from the preposition used in the high-frequency condition).

Subjects

Subjects were 28 Boston University undergraduates who participated for course credit. All subjects were native speakers of English.

Procedure

Subjects focused on a fixation window and pressed a button to initiate each trial. The target letter, word or word pair appeared for 30 msec and was followed by a pattern mask for 250 msec. The letters for the forced choice task then appeared above the masked stimulus and remained visible until the subject pressed a top or bottom button to signal which letter had been in the corresponding position. The 212 experimental trials followed 27 practice trials.

```
    o
    a
top tan
```

Results

Noun Combinations

The dotted line in figure 2 allows comparison of the mean

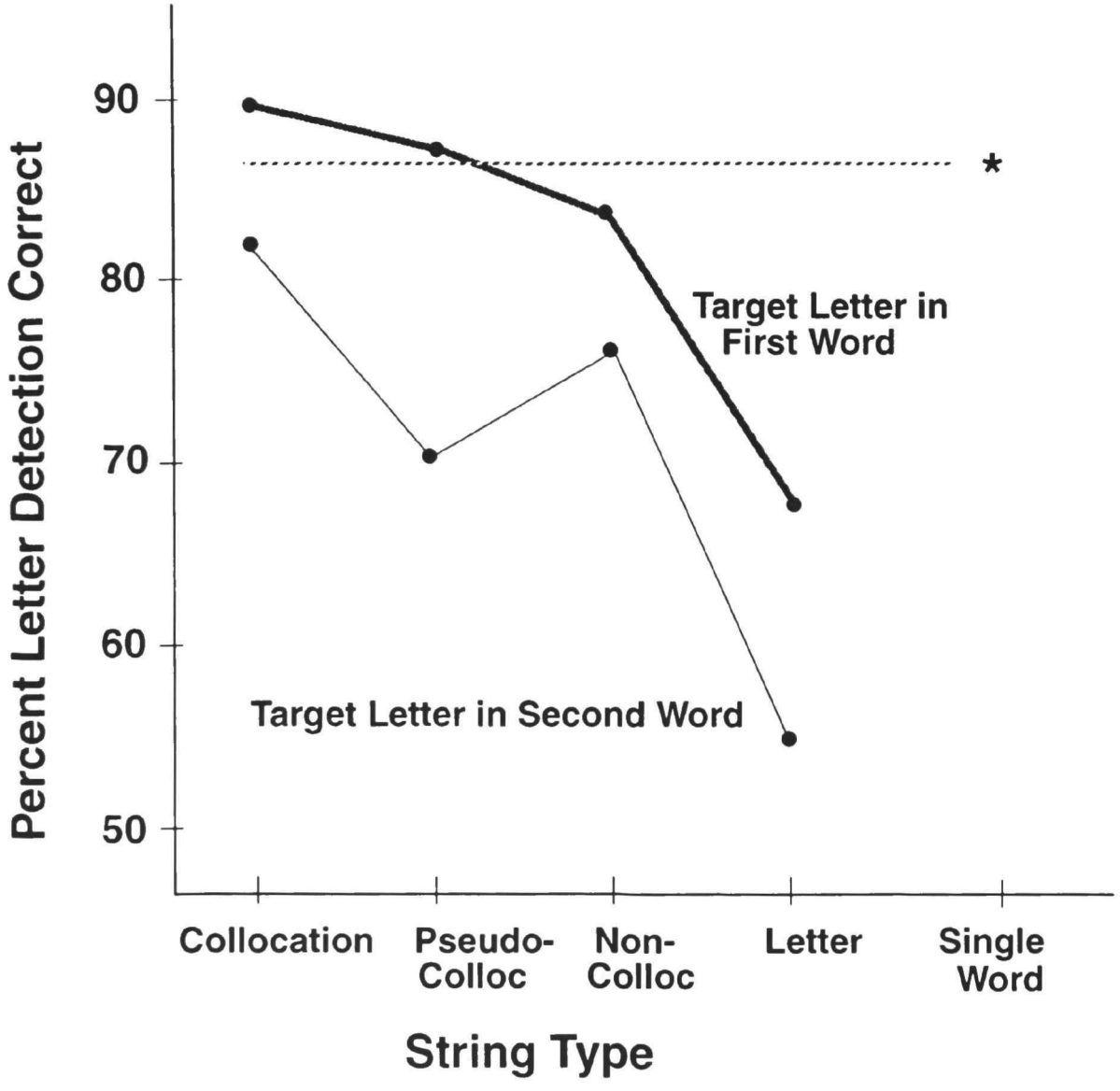

Figure2: Percent of correct letter detection for noun pairs.

detection rate for single words compared to other conditions. Letter detection was worse when letter targets were in the second word, for all string conditions; $F(1,53)=22$; $p < .001$. The condition X target word interaction shown in Figure 2 is significant; $F(3,159)=2.8$; $p < .05$, as are the 2 X 2 anovas on target word and pseudo-word vs. non-colloc; $F(1,53)=5.7$, $p < .02$ and colloc and non-colloc; $F(1,53)=7.8$; $p < 0.005$. (The only means in Figure 2 which aren't significantly different from each other are colloc and Near-Colloc, and Near-Colloc and non-colloc, when the target letter is in the first word. Because all stimuli were left-adjusted on the screen, collocations in which the second-word contained the letter-detection target didn't have a matching "single word" condition.[1]

Figure 3 shows response time for each of the string types Button-pressing times were an average 426 msec slower when the target occurred in the second word; $F(1,287)=94$; $p < .0001$.

Verb Combinations

Table 2 shows that percent correct was better for all conditions containing a word than the single letter condition; $F(4,104)=10$. However, there was no advantage the high frequency verb combinations over the low frequency or anomalous word pairs. The difference between response time means for the high frequency and the anomalous conditions was statistically reliable; $F(1,26)=6$; $p< .03$. implying that the anomalous word pairs incur a reaction time cost.

Data Summary

The current study showed the following:

- The basic word superiority effect (better detection of letters in words than in single letters) holds when words are part of word pairs, with a total string length of up to 11 characters. The target letter may appear in any of the 11 positions.
- Detection of letters in collocations is equivalent to detection in single words. Relative to single words and collocations, detection in random pairs is impaired.
- Letter detection in a near-collocations (*tax bell*) is inhibited relative to a collocation (*tax bill*) or an unassociated word pair (*tax deep*), but only if the target letter is in the second word of the word pair.
- Frequency and legality of verb+preposition pairs did not modulate letter detection.

Conclusions

My method of establishing the reality of units larger than words was to show that letter detection in collocations is better than detection in single words. A positive finding would have supported the proposal that activation can accrue to word combinations and enhance letter detection via top-down activation. However, detection in collocations and single words was found to be equivalent.

1. *All analyses are within-item anovas, averaging over subjects.* Results are comparable for anovas with subjects as the random factor, although some *F* values are smaller.

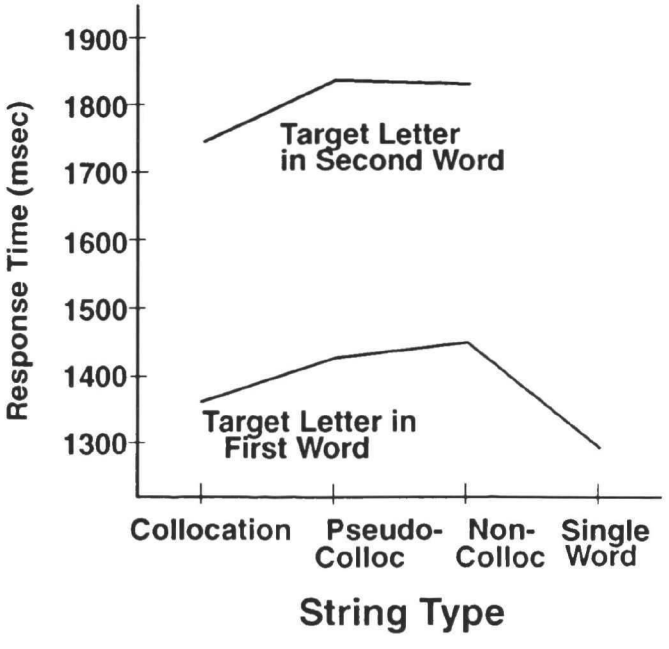

Figure 3: Response latencies for noun pairs

Table 2: Verb Combinations

Condition	% correct	RT
letter alone	.68	1174
single word	.83	1230
high freq word pair	.81	1282
low freq word pair	.85	1330
anomalous pair	.84	1368

Some support for the unit-status of common word combinations was provided by the finding that detection was impaired in random pairs compared to collocations, at least for the noun materials. But how could a non-collocation impair letter detection, if random pairs aren't units, and don't send inhibition and excitation? One possibility is that this is an effect of automatic semantic integrative processes. To make sense of the non-collocation, the processor initiates a search, which activates many candidate word-units. These interfere with units which are legitimately activated by the perceptual display, thus impairing letter detection. Support for this scenario is that response times for the random pairs were greater than for the collocations, for both noun and verb materials.

The current experiment contained materials analogous to those in WSE experiments, by comparing collocations and near-collocations. In WSE experiments, pseudo-words are sometimes as good at enhancing letter detection as real words. This study found results comparable to that of WSE studies, but with a twist. When the target letter was in the first word, near-collocs showed letter detection performance similar to that of collocations, and better than that of random pairs, suggesting they enhanced the activation of word units (as depicted in Figure 1a). When the target letter was in the second word, detection of letters in near-collocs was worse than in non-collocs, suggesting the near-collocs inhibited word units, as depicted in Figure 1b. A plausible reason for the difference in these conditions is time: detection was over 400 msec slower when the target letter was in the second word. Thus, near-collocs play an initially enhancing role, followed by an inhibiting role.

Future Work

One reason the collocations did not lead to higher detection rates than the single words may have been because the collocations are much less frequent as units than single words: they had an average frequency of 6 per million (range: 2 to 43), while the single words had an average frequency per million of 316 (range: 13 to 807). An experiment which includes a single-word, low frequency control condition is currently underway.[2]

A second method of demonstrating the beneficial effects of context is to investigate whether being one-letter away from a collocation facilitates letter detection in unpronounceable (orthographically illegal) non-words.

Subjects could be asked to detect the letters in bold:

---

let **d**own, let **d**owx, act **d**owx (choose **d** or **g**)
come fro**m**, come frx**m**, fact frx**m** (choose **m** or **g**)

According to standard WSE findings, **d** in *down* should show superior detection rates to *dowx*, but there is no prediction regarding how detection of **d** in *dowx* varies if preceded by either *let* or *act*. If units larger than words can facilitate letter detection, then we predict superior facilitation in *let dowx*, since it is a neighbor of *let down*.

2. I think one of the anonymous reviewers for suggesting this comparison.

Acknowledgments

This work was supported by a grant to the author from the McDonnel-Pew Program in Cognitive Neuroscience. I thank Brendan Kitts and two anonymous reviewers for helpful comments on an earlier version of the paper.

References

Chastain, G. (1986). Word-to-letter inhibition: Word-inferiority and other interference effects. *Memory & Cognition, 14*, 361-368.

Elman, J.L. (1990) Finding structure in time. *Cognitive Science, 14*, 179-211.

Emmorey, K.D., & Fromkin, V.A. (1988). The mental lexicon. In F.J. Newmeyer (Ed.), *Language: Psychological and biological aspects.* New York: Cambridge University Press.

Forster, K. (1979) Accessing the mental lexicon. In R.J. Wales & E. Walker (eds.) *New approaches to language mechanisms.* Amsterdam: North-Holland.

Harris, C.L. (1994) Coarse coding and the lexicon. In C. Fuchs and B. Victorri, (Eds.), *Continuity in linguistic semantics.* Amsterdam: John Benjamins.

Harris, C.L. (1990) Connectionism and cognitive linguistics. *Connection Science* 2, 7-34.

McClelland, J.L. & Johnston, J.C. (1977) The role of familiar units in perception of words and nonwords. *Perception & Psychophysics, 22*, 249-261.

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.

Reicher, G.M. (1969) Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology, 81*, 274-280.

Van Orden, G.C., Pennington, B.F. & Stone, G.O. (1990) Word identification in reading and the promise of subsymbolic psycholinguistics. *Psychological Review, 97*, 488-522.