Towards a Scalable Holographic Word-form Representation

Gregory E. Cox and George Kachergis
Department of Psychological and Brain Sciences
Cognitive Science Program
Indiana University

Gabriel Recchia
Cognitive Science Program
Indiana University

Michael N. Jones
Department of Psychological and Brain Sciences
Cognitive Science Program
Indiana University

Corresponding author: Greg Cox (grcox@indiana.edu)
240-426-1573
Department of Psychological and Brain Sciences
1101 E. 10th St.
Bloomington, IN 47405

Abstract

Phenomena in a variety of verbal tasks, e.g., masked priming, lexical decision, and word naming, are typically explained in terms of similarity between word-forms. Despite the apparent commonalities between these sets of phenomena, the representations and similarity measures used to account for them are not often related. To show how this gap might be bridged, we build on the work of Hannagan, Dupoux, and Christophe (2011) to explore several methods of representing visual word-forms using holographic reduced representations and evaluate them on their ability to account for a wide range of effects in masked form priming, as well as data from lexical decision and word naming. A representation that assumes that word-internal letter groups are encoded relative to word-terminal letter groups is found to predict qualitative patterns in masked priming as well as lexical decision and naming latencies. We then show how this representation can be integrated with the BEAGLE model of lexical semantics (Jones & Mewhort, 2007) to enable the model to encompass a wider range of verbal tasks.

Towards a Scalable Holographic Word-form Representation

Introduction

Verbal stimuli have a long history in experimental psychology, whether as the primary object of study (e.g., masked priming), as a window onto memory processes (e.g., episodic recognition), or as part of larger linguistic phenomena (e.g., sentence comprehension). They are also, of course, tremendously important in everyday experience. Given their ubiquity, it is reasonable to believe that words—at least by adulthood—have highly developed representations that enable them to be used in such myriad tasks. To date, however, studies using verbal stimuli have tended to focus exclusively on only a few dimensions of word representations, for example, their orthographic or phonological features, relative frequency, semantic features, or part-of-speech. Likewise, models of phenomena in various verbal tasks are not usually formulated in a way that makes it easy to accommodate findings from other verbal tasks; rather, separate mechanisms and/or representations must be posited in order to broaden a model's applicability.

A major stumbling block on the road to a unified account of lexical processing has been the lack of a unified representation for words that incorporates information about orthography, phonology, semantics, and syntax. The search for such a representation has motivated the MROM-p model that integrates orthography and phonology to account for naming, identification, and lexical decision data (Jacobs, Rey, Ziegler, & Grainger, 1998). More recently, Davis (2010) has made further headway in the direction of a unified word-form representation with the SOLAR model of word-form learning and encoding (Davis, 1999), which accounts for a wide variety of priming effects in word recognition as well as lexical decision latencies. Such accounts, though impressive, have not yet begun to address how perceptual properties of words might be integrated with their semantic or syntactic properties. A promising route toward such a unification is via Holographic Reduced Representations (HRRs; Plate, 2003), which make use of the mathematical tools of holography to encode a wealth of structured information within a single distributed

representation. HRRs have had considerable success in psychology as models of memory (Borsellino & Poggio, 1972; Murdock, 1982) and vision (Le Cun & Bengio, 1994). A notable application of HRRs to lexical representations is the BEAGLE model (Jones & Mewhort, 2007), which uses HRRs for words that capture information about their semantic content as well as word-order and other local syntactic constraints. BEAGLE's lexical representations have also been extended to accommodate perceptual information about word referents (Jones & Recchia, 2010).

In this paper, we explore methods of further enriching lexical representations by encoding word-form information using HRRs, with the aim of identifying word-form encoding schemes that are consistent with extant data on word-form priming, relatively simple to implement, and that are capable of scaling up to models of the entire lexicon. We first review some theoretical proposals for different word-form encodings as well as relevant empirical results. We then provide a brief overview of the mathematics of HRRs before evaluating a number of holographic word-form encodings. We find that a HRR that encodes both local letter information (bigrams) and the position of letters relative to both terminal letters produces the best fit to qualitative trends in masked priming data. We then compare various word-form encodings, implemented as HRRs, to response latencies in lexical decision and word naming and show how HRRs for word-form can be integrated into the BEAGLE model of lexical semantics. Finally, we suggest other ways in which HRRs can lead us toward a unified account of various verbal phenomena.

Theories of Word-Form Encoding

A variety of word-form representations have been proposed to account for phenomena arising from orthographic similarity, many of which are reviewed and compared in Davis and Bowers (2006) and Hannagan, Dupoux, and Christophe (2011). Our primary concern in this paper is not to develop a complete theory of visual word processing, but rather to demonstrate how the form of the "output" of visual word recognition—a word-form representation—makes predictions about the perceived similarity between verbal stimuli. The following overview, in

which we touch upon various theories for word-form encoding and relevant empirical findings, is quite cursory, and the reader is directed to the resources cited above for a more complete review of theories and results in visual word processing.

Evidence from Masked Priming

The majority of the empirical evidence for different theories of word-form encoding comes from studies of masked priming in lexical decision (Forster & Davis, 1984). In this task, a pattern mask is presented for a moderate amount of time, usually 500 ms, followed by a brief flash (around 50 ms) of a string of lowercase letters (which may or may not be a word) which is then replaced by an uppercase letter string, which remains for a fixed time or until the participant makes a response. Participants must decide as quickly and accurately as possible whether the second (uppercase) string is a word or nonword, while the first (lowercase) string serves as an unconscious prime. Certain kinds of primes facilitate processing of the target string (i.e., produce a decrease in the latency to produce a correct response, relative to a neutral prime). The relative amount of facilitation across different types of primes can be thought to index the relative similarity of each prime to the target, in terms of whatever word-form encoding is used by the human visual word recognition system. A theory of word-form representation should, then, strive to result in representations that capture the pattern of similarity implied by the masked priming data.

Hannagan, Dupoux, and Christophe (2011) define four types of effects from the masked priming literature, based on the relationship of the prime to the target and the impact of that relationship on response latency:

Stability. These effects reflect the fact that a word is maximally similar to itself, and that minor changes like single insertions, deletions, and repetitions should produce less facilitation—and therefore be less similar—than when the prime is identical to the target.

Edge effects. The importance of the initial and final letters for word recognition has been affirmed by a number of studies using a variety of paradigms. The spaces that border the edges of

a word mean that there is less lateral masking for the first and last letters, thus enabling them to be accurately perceived, even far from the point of eye fixation (Townsend, Taylor, & Brown, 1971). In terms of the importance of outer letters in assessing word similarity, primes that overlap with target words in their outer letters, and especially in their initial letters, produce more interference effects in lexical decision than do primes that overlap in interior letter positions (Davis, Perea, & Acha, 2009), and degradation of exterior letters slows reading to a greater extent than degradation in other positions (Jordan, Thomas, Patching, & Scott-Brown, 2003). Further, in masked priming, shared terminal letters produce much more facilitation than do letters shared in other positions (G. W. Humphreys, Evett, & Quinlan, 1990). Thus, one should expect that a reasonable word-form encoding should produce representations that are more similar when they differ in internal letters than when they differ in external letters.

Transposed letter effects. In masked priming, a prime formed by transposing letters in the target word produces more facilitation than does a prime that replaces them entirely, and consistent with the edge effects described above, transpositions that retain the exterior letters produce more facilitation than do other transpositions (Perea & Lupker, 2003). An extreme case of this advantage for transpositions that preserve exterior letters was investigated by Guerrera and Forster (2008), who found that a prime created by transposing every letter of a word with its neighbor to the right (e.g., 21436587 from 12345678) produced no facilitation whatever[1]. A more subtle effect is that a transposition of two non-adjacent letters produces more facilitation than does replacing the transposed letters, but less than replacing just one letter (Perea & Lupker, 2003). Finally, a "neighbor once removed" (an internal transposition followed by a substitution; Davis & Bowers, 2006) produces less facilitation than a single internal-letter substitution.

Relative position effects. Finally, facilitation has been found to occur when the prime preserves the relative position of the letters of the target, but disturbs their absolute position within the word (Grainger, Granier, Farioli, Van Assche, & van Heuven, 2006). Such primes may

consist of the target minus some letters at the beginning or end, or they may simply retain the external letters and the majority of the internal letters. A related effect is that a target that contains a repeated letter receives just as much facilitation if the prime is missing that letter as it does when it is missing a non-repeated internal letter.

Hannagan, Dupoux, and Christophe (2011) codified the above findings into 20 criteria for assessing word-form encoding schemes, which is given in Table 1.

Slot Coding

Perhaps the simplest way to represent a visual word-form is with a code that associates each letter with its exact position in the word. That is, a word-form can be considered as a series of slots which are filled by letters. For example, *word* would be represented as $\{w_1, o_2, r_3, d_4\}$, where each letter-slot l_n is a unique code for letter l in position n. This is essentially the approach taken by the interactive activation model of McClelland and Rumelhart (1981), in which each possible combination of a letter and its position is represented by a single unit. An obvious drawback of this approach is its inability to account for transposition effects, since according to this encoding theory, *wrod* is just as similar to *word* as it is to *weld*, since an *o* in position 2 bears no similarity to an *o* in position 3. Further, it is not clear how one could compare words of different lengths by this encoding, like *word* and *sword* or *world*, since one must first decide how the two words should be aligned. Moreover, the general idea that single letters are the sole building blocks of word-form representations contradicts findings that extra-letter information (e.g., relative letter position, spacing) is important in word identification (Mewhort & Johns, 1988). The Overlap model (Gomez, Ratcliff, & Perea, 2008) overcomes many of the deficiencies of slot coding by allowing for uncertainty about letter location, but at the cost of free parameters specifying the amount of uncertainty for each position (when fitted to data, uncertainty about the initial position is less than for internal positions, consistent with edge effects).

N-Gram Coding

Rather than encoding the absolute positions of each single letter, it is also possible to encode a word-form as a collection of substrings of the word, and thereby capture information about the relative positions of letters in the word. If those substrings are all of size 2, then the word-form is encoded as a set of bigrams, although the substrings could, in principle, be of any size, hence "n-grams". Further, the n-grams may be "closed"—consisting of only contiguous substrings—or "open"—allowing for substrings that contain letters that are not contiguous in the original word-form. Thus, a closed bigram code for *word* would be $\{wo, or, rd\}$, while an open bigram code for *word* would be $\{wo, wr, wd, or, od, rd\}$. Because the effect of n-gram coding is to capture relative position, rather than absolute position, it is able to account for similarity by transposition. For example, using an open bigram coding, *wrod* ($\{wr, wo, wd, ro, rd, od\}$) shares all but one n-gram with *word*, but only one with *weld* ($\{we, wl, wd, el, ed, ld\}$). It is also possible to use n-grams of multiple sizes: Using an open n-gram code with $1 \leq n \leq 3$, *word* would consist of $\{w, o, r, d, wo, wr, wd, or, od, rd, wor, wod, wrd, ord\}$. The use of n-grams at varying scales can allow for more graded similarity measures.

However, as can be seen from these examples, n-gram encoding alone cannot account for end-effects: An open n-gram code represents all letters in the word equally (i.e., each letter appears in the same number of n-grams), and so similarity is equally affected by replacements or transpositions in any part of the word, in contrast to data showing that replacements or transpositions at the ends of words produce a greater disruption to perceived similarity. A closed n-gram code over-represents internal letters relative to terminal ones, thus predicting that replacements or transpositions of internal letters have a greater effect than replacements or transpositions of terminal letters, again contrary to the data.

An extension to the basic n-gram approach is offered by the SERIOL model (Whitney, 2001). In SERIOL, words are represented as sets of open bigrams, but the bigrams are allowed to take on continuous activation values, rather than simply being "present" or "absent". Bigrams

that begin closer to the beginning of the word receive greater activation. Bigrams are also activated inversely with the size of the gap between their component letters, with the exception of the bigram containing the initial and final letters, which receives a boost in activation, thus allowing SERIOL to account for end effects.

Spatial Coding

Although this paper will not directly address it, we briefly describe a third class of models of visual word-form representations. This is the "spatial coding" approach exemplified by SOLAR (Davis, 1999; Davis & Bowers, 2006; Davis, 2010). In this approach, word-forms are encoded as a pattern of activity over an abstract space—perhaps implemented as a neural field (Davis, 2010)—describing letters and positions, with word-form similarity corresponding to the degree of overlap between these activity patterns. Explicit coding of initial and final letters allows the model to account for edge effects, while the nature of spatial coding—in particular, the similarity in a letter-node's level of activation when it lies in a similar position in a word—allows SOLAR to easily capture similarity by transposition, relative position, and substitution (Hannagan, Dupoux, & Christophe, 2011). While the full SOLAR model involves the need to set parameters that we wish to avoid here, spatial coding in general is a power technique for word-form encoding, especially for its ability to provide more graded similarity measures than are possible with most slot-based or n-gram-based encodings. The virtue of graded similarity measures is, fortunately, shared by the word-form encoding techniques we explore in this paper, holographic reduced representations.

Holographic Reduced Representations

Consider the problem of encoding a set of items, for example, the set of words $\{\text{cat}, \text{catch}, \text{cut}, \text{catcher}\}$. One approach to storing this information would be to store each word separately, in a list with its own label: 1) cat, 2) catch, 3) cut, 4) catcher. This is called *localist* encoding, because each word is stored "locally" in its own region of memory that doesn't overlap

with any other regions. Localist encoding has the advantage that there is no chance of confusion between the items that are stored, but fails to represent similarity between items and has difficulty dealing with noise (e.g., there is no obvious way to match "catch" with "cafch"). In contrast to localist representations, distributed representations store an item as a pattern of activation over many units (for example, neurons), with each unit playing a role in representing many items (Hinton, McClelland, & Rumelhart, 1986). Similar items can then be represented in a way that reflects their similarity, e.g., "cat" and "cut" would represented with similar patterns of unit activity. Distributed representations thus allow for confusability between similar items ("cat" and "cut"), but are also robust against noise ("cafch" is more similar to "catch" than anything else).

Holographic Reduced Representations (HRRs; Plate, 2003) are a type of distributed representation that allows for the encoding of hierarchically structured information as a pattern of activity over a set of units. The advantage of HRRs is that adding more information does not entail adding more units to the representations. HRRs have also been shown to provide a solution to the problems of variable binding (tracking which features belong to which items) and representation of hierarchical structure that have plagued many distributed representation frameworks (Plate, 2003). At the lowest level, a HRR is composed of several random vectors, each representing, say, a letter. These "atomic" vectors bear no similarity to one another, but they can be compositionally combined via two operations: binding ($\circledast$) and superposition ($+$). Binding takes two HRRs and produces a third HRR that is independent of (not similar to) the two HRRs that were used to construct it. Superposition takes two HRRs and creates a composite HRR which is partially similar to the two original HRRs. Superposing HRRs thus allows recognition, but not recall. Binding is invertible, and thus allows recall of one bound HRR by probing with the HRR with which it was bound. This kind of invertible binding operator is reminiscent of light holography, hence the term (see Plate, 2003).

In combination, binding and superposition can be used to implement a variety of word-form encoding schemes that simultaneously represent structure at multiple levels. For example, the

word *cat* may be represented as the superposition of bound substrings of the word, e.g.: $c + a + t + c \circledast a + a \circledast t$, where each letter is represented by a unique random vector. This strategy of chunking long sequences (e.g., letters in words, words in sentences) allows the representation to capture similarity at many resolutions: *cat* will be similar to *catch*, but *catcher* will be more similar to *catch* by virtue of more shared substrings. Hence, HRRs explicitly indicate which associations are being stored and engender similar representations of similar objects. The similarity of two HRRs relies both on the contents of the representations (e.g., *cat* and *catch* both have the letters c, a, and t) and on the structure of the stored associations (e.g., *cat* can be made more similar to *cut* if the association $c \circledast t$ is included in their HRRs). A quantitative description of the vectors and operators in HRRs is given in the next section.

HRRs can also represent predicates and other structures by representing a role as a random vector and binding it to the role-filler. Thus, HRRs are a reduced description (Hinton, 1990; Plate, 2003), and may be used to encode hierarchical and even recursive structures in a representation that does not grow in dimensionality. In the context of visual word-form processing, the similarity patterns produced by a particular HRR encoding have been found to be equivalent to the similarities produced by a back-propagation network trained for location-invariant visual word recognition (Hannagan, Dandurand, & Grainger, 2011). However, the explicit construction of HRRs—which require the modeler to specify the basic vectors and ways to combine them—allows for easier interpretation and comparison between encodings than is possible with learned representations in standard neural network models.

Circular Convolution

Circular convolution is one candidate binding operator for HRRs, as it is neurally plausible (Eliasmith, 2004) and approximately invertible via a correlation operation (Plate, 2003). To minimize noise, the random n-dimensional vectors representing the "atoms" (letters, in our case) from which more complex representations (words) are constructed have each of their n elements drawn independently from a normal distribution $\mathcal{N}\left(0, \frac{1}{n}\right)$. The circular convolution of two vectors

A and B is

$$C = A \circledast B$$

where each element c_j of C is

$$c_j = \sum_{k=0}^{n-1} a_k b_{j-k \bmod n}.$$

Circular convolution is depicted schematically in Figure 1, which shows how circular convolution can be thought of as a compressed version of the outer product of two vectors. For example, let $A = [1, 2, 3]$ and $B = [4, 5, 6]$. Then, $C = A \circledast B$ can be computed:

$$c_0 = a_0 b_0 + a_1 b_2 + a_2 b_1 = 1 \times 4 + 2 \times 6 + 3 \times 5 = 31$$

$$c_1 = a_0 b_1 + a_1 b_0 + a_2 b_2 = 1 \times 5 + 2 \times 4 + 3 \times 6 = 31$$

$$c_2 = a_0 b_2 + a_1 b_1 + a_2 b_0 = 1 \times 6 + 2 \times 5 + 3 \times 4 = 28$$

Note that the output vector of a circular convolution is the same dimensionality as each input vector, unlike techniques in other models that produce outputs with greater dimensionality (e.g., Murdock, 1982; M. S. Humphreys, Bain, & Pike, 1989).

Circular convolution is commutative, associative, and distributes over addition. In our implementation, we use $n = 1024$-dimensional vectors unless otherwise specified, and rather than $O(n^2)$ time circular convolution, we make use of the fast Fourier transform to compute convolutions in $O(n \log n)$ time[2]. When using circular convolution as a binding operator, an appropriate superposition operator is vector addition. Thus, using A and B as given in the above example, they can be superposed $A + B = [1 + 4, 2 + 5, 3 + 6] = [5, 7, 9]$.

Computing Similarity

Throughout this paper, we will compute the similarity of two HRRs via the normalized dot-product, otherwise known as the cosine of the angle between the two HRR vectors or simply

their "cosine similarity". If A and B are vectors, their cosine similarity is:

$$\text{sim}\,(A, B) = \frac{A \bullet B}{\|A\|\|B\|} = \frac{\sum_{i=0}^{n-1} a_i b_i}{\sqrt{\sum_{i=0}^{n-1} a_i^2}\sqrt{\sum_{i=0}^{n-1} b_i^2}}.$$

This similarity measure is always in the range $[-1, 1]$. The expected cosine similarity of two random vectors (e.g., letters c and a) is 0—that is, they are orthogonal. HRRs representing bound items (e.g., $c \circledast a$) are independent of (orthogonal to) their components (c or a). Thus, the expected values of $\text{sim}\,(c, c \circledast a)$ and $\text{sim}\,(a, c \circledast a)$ are 0. HRRs comprised of superposed vectors (e.g., $c + a$) will, on average, have a positive similarity to each component, such that $\text{sim}\,(c, c + a) > 0$. Identical vectors have maximal similarity (1).

Holographic Word-Form Representations

We now use convolution-based HRRs to implement a variety of word-form encoding schemes. This is much in the spirit of Hannagan, Dupoux, and Christophe (2011), who compared localist implementations of word-form encodings with ones implemented as binary spatter codes (BSC, a different type of HRR; Kanerva, 1994), and evaluated the benefits of the various theories and their implementations with respect to the similarity constraints imposed by masked priming studies (summarized above and in Table 1). Our primary focus will be on the construction of HRRs via circular convolution, rather than BSCs. Further, because we are interested in constructing HRRs for word-forms that can scale up to models of the entire lexicon, we also investigate the ability of holographic word-form encodings to account for latencies in standard lexical decision (LD) and word naming tasks.

Slot Coding

Slot coding can be implemented as an HRR in a fairly straightforward way. First, we create a random "atomic" vector of length $n = 1024$ for each of the 26 letters of the alphabet, where the elements of the vectors are sampled independently from a normal distribution with mean zero and

variance $\frac{1}{n}$. These letter vectors have an expected similarity of zero. In addition, we create another random vector p (also with an expected zero similarity to the letter vectors) that will be used to encode individual letter positions. Recall that circular convolution takes two HRR vectors and produces a third that is not similar to the first two. Thus, by convolving p with itself ($p \circledast p$), we create a new vector that is not similar to p. Likewise, convolving p with itself three times produces yet another vector that is not similar to either p or $p \circledast p$. We denote the convolution of p with itself k times as p^k.

Because each of the vectors $p, p^2, p^3, \ldots$ are not similar to one another, we can use them to represent independent slots in a word-form. Individual letters can be bound to their appropriate slot and the entire word form is the superposition of those letter-slot bindings. Thus, *word* can be represented:

$$word = w \circledast p + o \circledast p^2 + r \circledast p^3 + d \circledast p^4,$$

where w, o, r, and d are the 1024-dimensional vectors used to represent each letter. When evaluating slot coding against the constraints from Hannagan, Dupoux, and Christophe (2011) (Table 2), we find that it fails in the ways discussed above, namely, to capture the fact that transposed letter primes produce more facilitation than replaced letters, that replacement of terminal letters produces less facilitation than replacement of internal letters, and that preservation of relative (but not absolute) position of letters in a prime still produces facilitatory priming. More in-depth discussion of the merits and demerits of slot coding may be found in Davis and Bowers (2006) and Hannagan, Dupoux, and Christophe (2011).

Open N-gram Encoding

We next evaluate two open n-gram schemes, one that uses unigrams (i.e., single letters) and bigrams and another that uses all n-grams for $1 \leq n \leq 4$. As above, each of the 26 letters of the alphabet (i.e., a unigram) is assigned a random vector of length n. To create an n-gram larger than a unigram, we use a non-commutative version of circular convolution, in which each operand

is randomly permuted prior to being convolved (Plate, 1995). For example, the bigram *wo* would be represented as $L(w) \circledast R(o)$, where w and o are random vectors as just described and L and R are (invertible) functions which randomly permute the entries in a vector. The quadrigram *word* would be represented as $L(L(L(w) \circledast R(o)) \circledast R(r)) \circledast R(d)$, which is a simple iterative application of the operation used for bigrams. Although we will henceforth omit the L and R permutation functions for clarity, all n-grams (including those created with terminal-relative encoding, below) are created with this non-commutative variant of circular convolution.

Thus, a complete open n-gram representation of *sword*, with $1 \leq n \leq 2$, would be:

$$\begin{aligned} sword = {} & s + w + o + r + d \\ & + s \circledast w + s \circledast o + s \circledast r + s \circledast d \\ & + w \circledast o + w \circledast r + w \circledast d \\ & + o \circledast r + o \circledast d + r \circledast d \end{aligned}$$

Similarly, we can construct an open n-gram representation for *word* like so:

$$\begin{aligned} word = {} & w + o + r + d \\ & + w \circledast o + w \circledast r + w \circledast d \\ & + o \circledast r + o \circledast d + r \circledast d \end{aligned}$$

By this encoding, we can see that $sword = word + s + s \circledast w + s \circledast o + s \circledast r + s \circledast d$, that is, "sword" contains all the unigrams and bigrams of "word", plus some others. Recalling that superposition $(a + b)$ results in HRRs that have non-zero similarity to their components (a and b), the shared n-grams between *sword* and *word* mean that $\text{sim}\,(sword, word) > 0$.

A space character may also be introduced to differentiate contiguous and non-contiguous n-grams. We will represent this character with an underscore ("_") and treat it like any single

character, i.e., it is represented as a random vector. Treating non-contiguous n-grams this way, the n-gram representation of *sword* (with $1 \leq n \leq 2$) would become:

$$\begin{aligned} sword = {} & s + w + o + r + d \\ & + s \circledast w + (s \circledast \_) \circledast o + (s \circledast \_) \circledast r + (s \circledast \_) \circledast d \\ & + w \circledast o + (w \circledast \_) \circledast r + (w \circledast \_) \circledast d \\ & + o \circledast r + (o \circledast \_) \circledast d + r \circledast d \end{aligned}$$

We evaluated four open n-gram schemes against the criteria in Table 1: One with $1 \leq n \leq 2$ (the "bigram" scheme) and another with $1 \leq n \leq 4$ (the "quadrigram" scheme), both with and without the use of a space character in non-contiguous n-grams. Results are given in Table 3. Overall, the effect of adding longer n-grams is to reduce the similarity values, since any change to the word-form also changes more n-grams. Representing non-contiguous n-grams with internal space-markers allowed for the satisfaction of condition 15 and decreased the similarity resulting from transpositions, particularly long-range transpositions.

The most obvious flaw with all of these schemes is their inability to account for edge effects (conditions 9 and 11), which arises from the fact that all letters are equally represented in a word's representation (i.e., they each appear in the same number of n-grams). In addition, they do not assign enough of a penalty to transpositions, leading them to predict more positive priming in condition 13 than is actually observed (although the open quadrigram coding with spaces is very close to satisfying that constraint). Although n-gram codes without space markers assign too high a similarity value to "neighbors once removed" (condition 15; Davis & Bowers, 2006), introducing space markers into non-contiguous n-grams solves this problem. Finally, all of these schemes assign a higher similarity value to condition 20 (*123456* priming 1232456) than to condition 19 (*123256* priming 1232456), indicating that the absence of a unique letter in the prime (condition 19) disturbs similarity more than the absence of a repeated letter (condition 20),

since the repeated letter leads to the coding of some redundant n-grams, while a unique letter naturally contributes unique n-grams to the representation.

Terminal-Relative Encoding

To account for the end effects that open n-gram coding cannot reproduce, many models assign extra weight to terminal letters (i.e., the initial and final letters). One could simply assign these weights either arbitrarily or by estimating weight values from data (similar to the approach taken by Gomez et al., 2008). Neither of these solutions would work well for our purposes, since we are looking for a holographic word-form encoding that can scale to the entire lexicon and that can be used in various capacities to model different phenomena. Adding parameters (like relative weights on terminal letters) would reduce the generality of our approach.

The solution we pursue to the problem of giving more emphasis to terminal letters is inspired by two models of word perception. Clark and O'Regan (1998) introduced a simple model of word recognition to explain the fact that the optimal fixation point for an English word is slightly to the left of center. Their model assumed that both terminal letters could be perceived correctly, regardless of the location of fixation, while only two letters (i.e., a contiguous bigram) could be perceived at the fixation point[3]. Whitney (2001), meanwhile, uses principles of neural information processing to motivate a model (SERIOL) that results in relatively strong encoding of contiguous bigrams, as well as bigrams containing terminal letters. Thus, both these models imply that contiguous bigrams are important to word recognition. These can be considered the local structure of the word-form. These models also imply that global word-form structure is available by virtue of the enhanced perceptibility (and persistence, in the case of SERIOL) of the terminal letters.

The holographic encoding scheme we propose—which we refer to as terminal-relative (TR) encoding—is closely related to the simplified model of Clark and O'Regan (1998), but may also be considered a "discretized" relative of SERIOL (because, unlike in SERIOL, n-grams in our encoding are either present or absent, and are not continuously weighted). To encode a word-form

using TR encoding, we first encode all unigrams and contiguous bigrams:

$$sword = s + w + o + r + d + s \circledast w + w \circledast o + o \circledast r + r \circledast d$$

Next, for any unigram or bigram that does not contain a terminal letter, we encode a chunk representing that n-gram's position relative to each terminal letter:

$$\begin{aligned} sword = {} & s + w + o + r + d + s \circledast w + w \circledast o + o \circledast r + r \circledast d \\ & + s \circledast d + s \circledast w + w \circledast d + s \circledast o + o \circledast d + s \circledast r + r \circledast d \\ & + (s \circledast w) \circledast d + s \circledast (w \circledast o) + (w \circledast o) \circledast d \\ & + s \circledast (o \circledast r) + (o \circledast r) \circledast d + s \circledast (r \circledast d) \end{aligned}$$

Note that because of the iterative application of this second step, the first and last bigrams are included twice in the representation, thus increasing their "strength" in a manner similar to the weighting that arises in the SERIOL model. Thus, the set of n-grams that comprise the word *sword* are

$$\{s, w, o, r, d, 2 \times sw, wo, or, 2 \times rd, swd, swo, wod, sor, ord, srd\} \ .$$

Again, the effect of this encoding scheme is to represent the position of letters relative to both their local context—contiguous bigrams—and their global context—relative to the beginning and end of the word, hence "terminal-relative" encoding. As a result, it should be expected to capture effects at both low and high levels of word-form structure. The non-contiguous bigrams and trigrams that result from this encoding (e.g., *sor* in *sword*) can be represented either with (*s_or*) or without (*sor*) a space character, as described above. Evaluating the TR scheme, both with and without spaces, on the criteria in Table 1 produces the results in Table 4.

TR encoding is able to capture all of the qualitative masked priming criteria whether spaces are used or not[4]. Once again, the effect of including spaces is primarily to increase the penalty for

transpositions. The emphasis placed on local contiguity, in combination with the global influence of terminal letters, means that TR correctly assigns condition 13 the lowest similarity value, because this condition not only involves transposition, but also a change in the terminal letters. Further, the influence of the missing unique letter in condition 19 is now mitigated, allowing TR encoding to correctly predict equal priming effects in conditions 19 and 20.

Comparison with Lexical Decision and Word Naming Data

We now investigate how well certain of the above holographic word-form encoding schemes can account for reaction times in lexical decision (LD) and word naming. Given that the amount of facilitation observed in masked priming is indicative of the similarity between the prime and target letter strings—as measured by whatever encoding is used by the visual word recognition system—we can say that masked priming is a measure of "local" similarity. Latency in lexical decision and word naming can then be thought of as relating to the "global" similarity between the target word and the lexicon, e.g., the number of other words in the lexicon that are similar to the target word, which has been known to affect RT in LD (Andrews, 1997). Here, we assume that the same word-form representation is used in masked priming, LD, and naming; it is just a matter of what similarities are being computed.

To assess each encoding's ability to relate to LD and naming RT, we rely on the lexical decision and word naming latencies recorded by the English Lexicon Project (ELP; Balota et al., 2007). We first generate representations for all words in the ELP database for which mean LD and naming latencies have been collected[5]. For each word in the lexicon, we create a HRR for its word-form. First, we compute the cosine similarity of the HRR for each pair of words in the lexicon. Then, for each word in the lexicon, we take the mean of all its pairwise similarities that are greater than 0.4. This has the effect of estimating roughly how many, and to what degree, other words are "sufficiently" similar (by some encoding) to a target word across the lexicon. Further, although a process model of LD and naming is beyond the scope of this paper, the use of a threshold seems more cognitively plausible, since it restricts to a manageable size the subset of

lexical memory against which a probe must be compared (indeed, such "activation" thresholds are often employed in memory models for that reason, e.g., Shiffrin & Steyvers, 1997; Kwantes, 2005). While the choice of a particular threshold of 0.4 is somewhat arbitrary, any value greater than about 0.25 preserves the qualitative trends in the following analyses.

After generating the mean above-threshold similarity described above for each word in the lexicon using slot coding, open bigrams (with space markers), open quadrigrams (with space markers), and terminal-relative encoding (also with space markers), we computed the Pearson's correlation between mean similarity and LD and naming latencies for each word, as well as several relevant lexical variables: word length, log-frequency[6] in the HAL corpus (Burgess & Livesay, 1998), and orthographic neighborhood size (the number of other words in the lexicon that differ by exactly one letter substitution, often called Coltheart's N; Coltheart, Davelaar, Jonasson, & Benner, 1977). These correlations are given in Table 5. Consistent with empirical findings that neighborhood density facilitates LD and naming (Andrews, 1997), mean similarity for all HRR word-form encodings is negatively correlated with LD and naming RT. Slot coding produces the strongest negative correlations, followed by TR encoding, bigrams, and quadrigrams. The parity between slot coding similarity and orthographic neighborhood size in terms of theirs correlations with LD and naming RT is sensible when one considers that orthographic neighborhood counts the number of single-letter substitution neighbors a word has; single-letter substitutions result in relatively small decrements in slot-coding similarity (see Table 2, conditions 6, 7, 9, 10, and 11), while other orthographic transformations (insertions, deletions, transpositions) are more disruptive to slot-coding similarity and thus may not even reach the 0.4 threshold. The low correlations between LD and naming RT and quadrigram similarity arise from overall lower similarities by quadrigram coding, which makes it hard to discriminate between more and less similar words across the entire lexicon. Bigram and TR encoding are better able to separate similar and dissimilar words, and so mean above-threshold similarity by those encodings better captures differences between words that are similar to many or fewer other words in the lexicon.

It is also worth noting in Table 5 the correlations between similarities—both by the holographic encodings and by orthographic neighborhood size—and word length and log-frequency. The current investigation does not attempt to model "pure" effects of length and frequency, which are inhibitory and facilitative, respectively. Thus, to better assess the effect of orthographic similarity on LD and naming RT, we should partial out the effects of length and frequency. We did this by computing linear regressions with LD and naming RT as outcome variables and length and log-frequency as predictors. The residuals from these regression models constitute the variation in LD and naming latencies that are not accounted for by linear effects of length or log-frequency. Pearson's correlations between these residuals and the various similarity measures are given in Table 6. Accounting for length and frequency, the correlations between LD and naming RT and various similarity measures are reduced, particularly for orthographic neighborhood size. Among the holographic word-form encodings, slot coding still has the strongest correlations with LD and naming RT, followed by TR encoding, then bigram and quadrigram encoding.

This indicates that, unlike the discrete measure of orthographic neighborhood size, the more graded measures of "global" similarity afforded by holographic word-form encoding tend to account for more variability in LD and naming latency beyond that accounted for by word length and frequency. We caution that these results should not be interpreted as strong support for slot coding of word-form, *per se*, given the deficiencies of slot coding to account for masked priming effects, as shown above. Rather, this supports the idea that, in the absence of a process model of LD or naming, similarity metrics that result in greater discrimination between similar and dissimilar word-forms across the lexicon do a better job of predicting LD and naming latencies. Slot coding presents a particularly stringent criterion for similarity, while open n-gram schemes allow for a broader range of partial similarity. TR encoding lies in between these two extremes, and as such can account for slightly less variability in LD and naming RT than slot coding, but more than either n-gram scheme and certainly more than orthographic neighborhood size (when

controlling for length and frequency). More generally, these results emphasize the utility of using theory-driven holographic word-form encodings to define graded similarity metrics across the corpus, that can account for variability in LD and naming RT beyond that from length and frequency.

Incorporating Word-form into a Model of Lexical Semantics

Again, a major advantage of the HRR approach is its ability to encode a variety of information in the same format. By encoding word-form information holographically, we can integrate it with holographic representations of semantic and syntactic information.

The BEAGLE Model

BEAGLE (Bound Encoding of the AGgregate Language Environment) is a convolution-based HRR model that learns both word meaning and word order information from natural language text corpora (Jones & Mewhort, 2007). BEAGLE uses a unique high-dimensional (e.g., 1024) vector—the *environmental* vector—to represent each word's physical form, as well as a *context* vector to store word co-occurrence information, and an *order* vector to encode which words appear before and after the given word.

A corpus is first broken up into individual sentences. As BEAGLE reads each sentence, the environmental vectors of the other words in the sentence (with the exception of function words, "the", "a", "to", etc.) are superposed on each word's context vector. Take, for example, the sentence "the cat played catch with the dog." The context vector for "cat", c_{cat}, would be updated

$$c'_{cat} = c_{cat} + e_{played} + e_{catch} + e_{dog},$$

where e_{played}, e_{catch}, and e_{dog} are the environmental vectors for "played", "catch", and "dog" respectively. The other words in the sentence are updated in a similar fashion. Words that have similar contexts grow more similar to one another, since their context vectors tend to hold the

same set of superposed vectors. Thus, BEAGLE learns a semantic space; the semantic similarity of any two words can be found by taking the cosine between the two words' context vectors, exactly as was done above to determine the similarity of two word-form HRRs.

BEAGLE learns word order by binding n-grams containing a placeholder vector for the current word and superposing the bound n-grams in the current word's order vector. For example, in the sentence "the cat played catch with the dog", the word-level n-grams in which "cat" occurs are "the cat", "cat played", "the cat played", "cat played catch", "the cat played catch", etc. The parameter λ specifies the maximum size of these word-level n-grams; here, we set $\lambda = 3$. To encode these n-grams in the order vector for "cat", o_{cat}, we replace each instance of "cat" with the placeholder vector (called Φ) that is randomly generated for each BEAGLE simulation. Then, o_{cat} is updated

$$\begin{aligned} o'_{cat} =& o_{cat} + e_{the} \circledast \Phi + \Phi \circledast e_{played} + e_{the} \circledast \Phi \circledast e_{played} \\ &+ \Phi \circledast e_{played} \circledast e_{catch} + e_{the} \circledast \Phi \circledast e_{played} \circledast e_{catch} + \ldots \end{aligned}$$

where the non-commutative variant of circular convolution (described above) is used. Words that appear not just with the same words, but *with the words in the same order* thus grow more similar to one another, capturing aspects of a word's usage and syntactic function. It is a simple matter to compare two words on the basis of *both* their semantic and syntactic contents: add together the order and context vectors for each word, and then compute the cosine between these summed vectors, e.g., $\mathrm{sim}\,(o_{cat} + c_{cat}, o_{dog} + c_{dog})$. For more details on the BEAGLE model, see Jones and Mewhort (2007).

BEAGLE nicely captures many aspects of syntactic and semantic similarity between words. However, in previous versions of BEAGLE, each word's environmental vector was generated randomly and thus were approximately orthonormal. We replaced BEAGLE's random environmental vectors with the TR HRR word-form encoding defined above. In this way, we may

capture similarity on the basis of orthography (e.g., *cat* and *catch*), and perhaps additional semantic relationships (e.g., *catch* and *catcher*). The use of HRRs to model the lexicon allows us to incorporate orthographic structure into a model of semantics and word order within a unified framework.

Memory Blocking

By using structured representations for words, we can also apply BEAGLE to additional empirical findings. For example, memory blocking experiments use orthographically-similar word forms (e.g., HOLSTER) as primes to block the retrieval of the target word (HISTORY) in a fragment completion task (H_ST_R_) (e.g., Smith & Tindell, 1997). In this paradigm, subjects tend to perseverate on the prime, likely due to its orthographic similarity to the fragment. However, perseveration must also be a function of semantics: primes will bring to mind a specific meaning, and fragments will activate a number of contexts of varying similarity to the block (and the unobserved target). When the orthographic similarity of a block strongly matches the target, they will have nearly the same context, and perseveration may be greater than when a fragment has a more diffuse semantic activation. Thus, using structured word-form representations in BEAGLE allows us to make item-level performance predictions based on both semantics and orthography of the primes, the targets, and the fragments.

Using TR HRRs in BEAGLE with a window size of three, trained on the TASA corpus, we examine the twelve targets, blocking primes, and fragments used by Smith and Tindell (1997). Specifically, we look at the cosine similarity of each fragment's TR HRR to its target's TR HRR, and to its blocking prime's TR HRR. As shown in Table 7 (columns four and five), these similarities are all quite high, and the target is not consistently more similar to the fragment than the blocking prime. That is, according to TR-encoding, the blocking prime has equal or greater similarity to the fragment than the target, even though the blocking prime cannot actually fit in the spaces of the fragment. From a purely orthographic standpoint, blocking in most of the Smith and Tindell (1997) stimuli is unsurprising. However, as mentioned above, high semantic similarity

of a blocking prime to a target may diminish a large orthography-based blocking effect.

To explore this possibility, semantics are added to the orthographic comparisons in the final two columns. BEAGLE's context vectors for the targets and the blocking primes are superposed on their respective word-form HRRs. The fragments never occur in TASA, and thus have no context. Instead, we construct a context for each fragment by taking the weighted sum of context vectors for words that have orthographic cosine similarity to the prime of greater than 0.4 (the same threshold used in the LD and naming comparison, above). Table 7 shows the similarities of each fragment and its average context to the target and its context (column 6), and to the blocking prime and its context (column 7). While the added noise of the context vectors moderately decreases the overall similarity values, the magnitude of similarities between targets and blocking primes changes drastically, and even reverses in one case. For example, *crumpet* was much more similar to *cu_p__t* than *culprit* when only word form was considered, but with BEAGLE's semantic information, this relationship was strongly reversed. Thus, BEAGLE can make predictions based on orthographic and semantic similarities of each blocking prime, target, and fragment about the size of blocking effects. It would be informative to see whether high word-form similarity between fragments and targets result in less of a memory blocking effect than in items with higher similarity between fragment word-forms and blocking prime word-forms. Moreover, it is possible in principle to estimate the relative weighting of semantic and orthographic contributions by looking at item-level human data. Unfortunately, we do not have item-level human data for comparison at present.

General Discussion

In this paper, we have demonstrated how various proposed word-form encodings can be implemented as holographic reduced representations, and how the resulting representations may be used to make predictions about performance in masking priming tasks, and in unprimed lexical decision and word naming tasks. We have also introduced a novel holographic

representation for word-forms that is relatively simple to compute, satisfies a variety of empirical constraints on word similarity, and shown how this orthographic representation (and, in principle, others) can be integrated with semantic and syntactic information in a unified model of the lexicon. By using HRRs, we can be explicit about the information included in our lexical representations, but retain the flexibility and neural plausibility of parallel distributed models.

To support further investigation into holographic word-form representations, at `http://www.indiana.edu/~clcl/holoword/`, we provide code in Python as well as a Windows graphical interface (along with tutorials) that allow users to generate HRRs for word-forms using the encoding schemes described in this paper. Slot coding, n-grams, and TR encoding are supported using either circular convolution or binary spatter coding, and a variety of parameters (e.g., maximum n-gram size, vector dimensionality) may be manipulated. Empiricists can investigate similarity between visual verbal stimuli while varying holographic encoding parameters, while modelers may wish to incorporate HRRs for word-forms into their own work.

As we have emphasized, a major advantage of HRRs are their extensibility. The BEAGLE model, augmented to include principled environmental vectors, can be applied to a variety of additional tasks beyond the fragment-completion task. Perceptual-level uncertainty about printed text—even in the absence of stimulus degradation—has been shown to affect sentence comprehension (Levy, Bicknell, Slattery, & Rayner, 2009), an effect to which an extended BEAGLE model could be applied. The inclusion of orthographic information also enables inferences about semantics based on shared word-form properties: For instance, a model could come to represent the meanings of prefixes and suffixes, or could generalize information across words (or to new words) on the basis of shared phonaesthemes, sub-lexical units that indicate semantic similarity (Otis & Sagi, 2008). A more practical concern is that the inclusion of orthographic information in models like BEAGLE may obviate the need to lemmatize the corpora on which they are trained.

The techniques presented in this paper could, in principle, enrich lexical representations

even further. Phonological information could be encoded in much the same manner as we have dealt with orthographic structure (for an example of HRRs applied to phonology, see Harris, 2002). In terms of orthography, our representations could be extended to an even lower level by imposing similarity constraints on the letter representations themselves, derived from empirical studies of confusability (e.g. Bouma, 1971). Here, we have descended only one rung on the ladder of representations, but it is a ladder that extends both downward and up.

References

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, *4*(4), 439–461.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., et al. (2007). The english lexicon project. *Behavior Research Methods*, *39*(3), 445–459.

Borsellino, A., & Poggio, T. (1972). Holographic aspects of temporal memory and optomotor responses. *Biological Cybernetics*, *10*(1), 58–60.

Bouma, H. (1971). Visual recognition of isolated lower-case letters. *Vision Research*, *11*, 459–474.

Burgess, C., & Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kucera and Francis. *Behavior Research Methods, Instruments, & Computers*, *30*(2), 272–277.

Clark, J. J., & O'Regan, J. K. (1998). Word ambiguity and the optimal viewing position in reading. *Vision Research*, *39*, 843–857.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Benner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555). Hillsdale, NJ: Erlbaum.

Davis, C. J. (1999). *The self-organising lexical acquisition and recognition (SOLAR) model of visual word recognition*. Unpublished doctoral dissertation, University of New South Wales, Sydney, Australia.

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychological Review*, *117*(3), 713–758.

Davis, C. J., & Bowers, J. S. (2006). Contrasting five different theories of letter position coding: Evidence from orthographic similarity effects. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(3), 535–557.

Davis, C. J., Perea, M., & Acha, J. (2009). Re(de)fining the orthographic neighborhood: The role of addition and deletion neighbors in lexical decision and reading. *Journal of Experimental*

Psychology: Human Perception and Performance, *35*(5), 1550–1570.

Eliasmith, C. (2004). Learning context sensitive logical inference in a neurobiological simulation. In S. Levy & R. Gayler (Eds.), *Compositional connectionism in cognitive science* (pp. 17–20). Menlo Park, CA: AAAI Press.

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 680–698.

Gomez, P., Ratcliff, R., & Perea, M. (2008). The overlap model: A model of letter position coding. *Psychological Review*, *115*(3), 577–601.

Grainger, J., Granier, J., Farioli, F., Van Assche, E., & van Heuven, W. J. B. (2006). Letter position information and printed word perception: The relative-position priming constraint. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(4), 865–884.

Guerrera, C., & Forster, K. (2008). Masked form priming with extreme transposition. *Language and Cognitive Processes*, *23*(1), 117–142.

Hannagan, T., Dandurand, F., & Grainger, J. (2011). Broken symmetries in a location-invariant word recognition network. *Neural Computation*, *23*(1), 251–283.

Hannagan, T., Dupoux, E., & Christophe, A. (2011). Holographic string encoding. *Cognitive Science*, *35*(1), 79–118.

Harris, H. D. (2002). Holographic reduced representations for oscillator recall: A model of phonological production. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the 24th Annual Meeting of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum Associates.

Hinton, G. E. (1990). Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence*, *46*, 47–76.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 77–109). Cambridge, MA: MIT Press.

Humphreys, G. W., Evett, L. J., & Quinlan, P. T. (1990). Orthographic processing in visual word identification. *Cognitive Psychology*, *22*, 517–560.

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review*, *96*(2), 208–233.

Jacobs, A. M., Rey, A., Ziegler, J. C., & Grainger, J. (1998). MROM-P: An interactive activation, multiple readout model of orthographic and phonological processes in visual word recognition. In J. Grainger & A. M. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (pp. 147–188). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1–37.

Jones, M. N., & Recchia, G. (2010). You can't wear a coat rack: A binding framework to avoid illusory feature migrations in perceptually grounded semantic models. In S. Ohisson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Jordan, T. R., Thomas, S. M., Patching, G. R., & Scott-Brown, K. C. (2003). Assessing the importance of letter pairs in initial, exterior, and interior positions in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 883–893.

Kanerva, P. (1994). The spatter code for encoding concepts at many levels. In *Proceedings of international conference on artificial neural networks* (Vol. 1, pp. 226–229). London: Springer-Verlag.

Kwantes, P. J. (2005). Using context to build semantics. *Psychonomic Bulletin & Review*, *12*(4), 703–710.

Le Cun, Y., & Bengio, Y. (1994, Oct.). Word-level training of a handwritten word recognizer based on convolutional neural networks. In *Pattern recognition* (Vol. 2, pp. 88–92).

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers

maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, *106*(50), 21086–21090.

Lupker, S. J., & Davis, C. J. (2009). Sandwich priming: A method for overcoming the limitations of masked priming by reducing lexical competitor effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 618–639.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. an account of basic findings. *Psychological Review*, *88*(5), 375–407.

Mewhort, D. J. K., & Johns, E. E. (1988). Some tests of the interactive-activation model for word identification. *Psychological Research*, *50*, 135–147.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*(3), 609–626.

Otis, K., & Sagi, E. (2008). Phonaesthemes: A corpus-based analysis. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 65–70). Austin, TX: Cognitive Science Society.

Perea, M., & Lupker, S. J. (2003). Transposed-letter confusability effects in masked form priming. In S. Kinoshita & S. J. Lupker (Eds.), *Masked priming: The state of the art* (pp. 97–120). New York, NY: Psychology Press.

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, *6*, 623–641.

Plate, T. A. (2003). *Holographic reduced representations*. Stanford, CA: CSLI Publications.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166.

Smith, S. M., & Tindell, D. R. (1997). Memory blocks in word fragment completion caused by involuntary retrieval of orthographically related primes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 355–370.

Townsend, J. T., Taylor, S. G., & Brown, D. R. (1971). Lateral masking for letters with

unlimited viewing time. *Perception & Psychophysics*, *10*(5), 375–378.

Whitney, C. (2001). How the brain encodes the order of letters in a printed word: The SERIOL model and selective literature review. *Psychonomic Bulletin & Review*, *8*(2), 221–243.

Footnotes

[1]The lack of any priming effect in this condition may be due not just to reduced orthographic similarity, but also to lexical competition, because interchanging each pair of letters in a word makes it more similar to other words than to the target (Lupker & Davis, 2009). Using a modified priming procedure that reduces the possibility of lexical competition, Lupker and Davis (2009) were able to find positive priming in this condition. Indeed, although we do not model lexical competition here, many encoding schemes—including all of the ones that we investigate—predict positive priming in this condition. However, because such priming requires different procedures to be apparent, we retain the constraint from Hannagan, Dupoux, and Christophe (2011) that this condition simply produces the weakest facilitation in the cases considered, rather than no facilitation at all.

[2]This follows from the fact that convolution of the raw vectors (in the "time domain") is equivalent to elementwise multiplication of the discrete Fourier transform of the vectors (in the "frequency domain").

[3]The idea that terminal letters can be correctly perceived regardless of fixation location is bolstered by work showing that the absence of lateral masking at word edges facilitates letter perception (Townsend et al., 1971).

[4]For comparison, Hannagan, Dupoux, and Christophe (2011) find that neither SOLAR nor SERIOL can account for all of these constraints.

[5]In addition to this constraint, we excluded words that had a frequency of zero in the HAL corpus (Burgess & Livesay, 1998), as well as words that contained punctuation (e.g., apostrophes). In total, $38,876$ words were analyzed.

[6]Because word frequency is extremely positively skewed, it is common to take the logarithm of word frequency to make the scale more sensible.

Constraint Family	Condition	Prime	Target	Criterion
Stability	(1)	*12345*	12345	> .95
	(2)	*1245*	12345	< (1)
	(3)	*123345*	12345	< (1)
	(4)	*123d45*	12345	< (1)
	(5)	*12dd4*	12345	< (1)
	(6)	*1d345*	12345	< (1)
	(7)	*12d456*	123456	< (1)
	(8)	*12d4d6*	123456	< (7)
Edge effects	(9)	*d2345*	12345	< (10)
	(10)	*12d45*	12345	< (1)
	(11)	*1234d*	12345	< (10)
Transposed letter effects	(12)	*12435*	12345	> (5)
	(13)	*21436587*	12345678	= Min
	(14)	*125436*	123456	< (7) and > (8)
	(15)	*13d45*	12345	< (6)
Relative position effects	(16)	*12345*	1234567	> Min
	(17)	*34567*	1234567	> Min
	(18)	*13457*	1234567	> Min
	(19)	*123256*	1232456	> Min
	(20)	*123456*	1232456	= (19)

Table 1

Word-form similarity constraints from masked priming. Digits refer to unique letters in the target word, while "d" indicates a different unique letter. "Min" refers to the minimum similarity value across all conditions. Reproduced, with permission, from Hannagan, Dupoux, & Christophe (2011).

Constraint Family	Condition	Slot coding similarity
Stability	(1)	1.0
	(2)	.45
	(3)	.55
	(4)	.55
	(5)	.60
	(6)	.80
	(7)	.83
	(8)	.67
Edge effects	(9)	**.80**
	(10)	.80
	(11)	**.80**
Transposed letter effects	(12)	**.60**
	(13)	0.0
	(14)	**.66**
	(15)	.60
Relative position effects	(16)	.85
	(17)	0.0
	(18)	**.17**
	(19)	.62
	(20)	**.47**

Table 2
Similarity values for prime/target pairs given in Table 1, derived from slot encoding. Values are averaged over 1000 Monte Carlo simulations. Violated constraints are indicated in bold.

		No spaces		With spaces	
Constraint Family	Condition	Bigram	Quadrigram	Bigram	Quadrigram
Stability	(1)	1.0	1.0	1.0	1.0
	(2)	.82	.72	.73	.52
	(3)	.97	.86	.91	.71
	(4)	.84	.71	.79	.58
	(5)	.40	.24	.40	.24
	(6)	.67	.48	.67	.48
	(7)	.71	.55	.71	.55
	(8)	.47	.24	.48	.24
Edge effects	(9)	**.67**	**.52**	**.67**	**.52**
	(10)	.67	.52	.67	.52
	(11)	**.67**	**.52**	**.67**	**.52**
Transposed letter effects	(12)	.93	.72	.66	.38
	(13)	**.87**	**.34**	**.72**	**.24**
	(14)	.86	.53	.67	.33
	(15)	**.67**	**.48**	.53	.31
Relative position effects	(16)	.73	.62	.73	.62
	(17)	.73	.62	.73	.62
	(18)	.73	.47	.63	.34
	(19)	.86	.73	.83	.62
	(20)	**.96**	**.80**	**.87**	**.66**

Table 3
Similarity values for prime/target pairs given in Table 1, derived from open n-gram encoding schemes. Values are averaged over 1000 Monte Carlo simulations. Violated constraints are indicated in bold.

Constraint Family	Condition	No spaces	With spaces
Stability	(1)	1.0	1.0
	(2)	.75	.75
	(3)	.92	.92
	(4)	.80	.80
	(5)	.38	.38
	(6)	.62	.62
	(7)	.72	.72
	(8)	.41	.40
Edge effects	(9)	.54	.54
	(10)	.65	.65
	(11)	.54	.54
Transposed letter effects	(12)	.69	.54
	(13)	.32	.23
	(14)	.66	.53
	(15)	.54	.46
Relative position effects	(16)	.60	.60
	(17)	.61	.60
	(18)	.63	.45
	(19)	.86	.85
	(20)	.86	.85

Table 4
Similarity values for prime/target pairs given in Table 1, derived from TR encoding schemes. Values are averaged over 1000 Monte Carlo simulations.

	LD RT	*Naming RT*	*Slot*	*Bigram*	*Quadrigram*	*TR*	*Length*	*Frequency*
Naming RT	0.73	—	—	—	—	—	—	—
Slot	-0.35	-0.36	—	—	—	—	—	—
Bigram	-0.15	-0.15	0.43	—	—	—	—	—
Quadrigram	-0.04	-0.04	-0.10	-0.18	—	—	—	—
TR	-0.16	-0.15	0.29	0.58	0.01	—	—	—
Length	0.55	0.54	-0.37	-0.07	0.05	-0.10	—	—
Frequency	-0.61	-0.53	0.27	0.21	0.05	0.17	-0.35	—
ON	-0.34	-0.37	0.49	0.29	-0.07	0.23	-0.56	0.29

Table 5

Pearson's correlations between a number of lexical variables and mean RTs in LD and naming, based on a subset of the ELP database (Balota et al., 2007). "ON" is orthographic neighborhood size. All correlations significant at $p < 0.001$*, with the exception of the correlation between quadrigram and TR similarity (*$p = 0.02$*).*

	Residual LD RT	*Residual Naming RT*
Residual Naming RT	0.49	—
Slot	-0.11	-0.15
Bigram	-0.04	-0.06
Quadrigram	-0.06	-0.05
TR	-0.07	-0.08
ON	-0.01	-0.08

Table 6

Correlations between across-lexicon similarity measures and LD and naming RT, with the linear effects of length and log-frequency partialled out. All correlations significant at $p < 0.001$*, with the exception of the correlation between ON and residual LD RT (*$p = 0.03$*).*

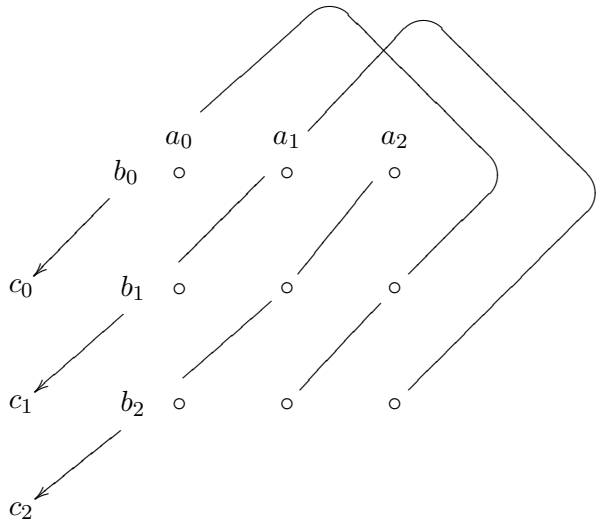
Blocking Prime	Target	Fragment	sim(Frag., Targ.)	sim(Frag., Block.)	sim(Frag., Targ.+con.)	sim(Frag., Block.+con.)
analogy	allergy	a_l__gy	0.64	0.64	0.63	0.46
brigade	baggage	b_g_a_e	0.48	0.70	0.35	0.53
cottage	catalog	c_ta__g	0.47	0.62	0.36	0.50
charter	charity	char_t_	0.69	0.65	0.59	0.53
cluster	country	c_u_tr_	0.38	0.49	0.45	0.57
crumpet	culprit	cu_p__t	0.36	0.58	0.44	0.45
density	dignity	d__nity	0.66	0.79	0.57	0.68
fixture	failure	f_i_ure	0.67	0.71	0.55	0.59
holster	history	h_st_r_	0.49	0.49	0.30	0.45
tonight	tangent	t_ng__t	0.46	0.64	0.34	0.43
trilogy	tragedy	tr_g__y	0.54	0.59	0.26	0.57
voyager	voltage	vo__age	0.57	0.78	0.55	0.69

Table 7

Blocking primes, targets and fragments from Smith & Tindell (1997) and the cosine similarities of their TR HRRs with and without BEAGLE's context information superposed.

Figure Captions

Figure 1. A schematic depiction of circular convolution. See the text for further detail.


a_0
a_1
a_2
b_0
b_1
b_2
c_0
c_1
c_2