Reading the written language environment: Learning orthographic structure from statistical regularities

Teresa Marie Schubert (corresponding author)[a], Trevor Cohen[b], Simon Fischer-Baum[c]

[a] Department of Psychology, Harvard University

33 Kirkland St

William James Hall 918

Cambridge, MA 02138

Teresa_Schubert@fas.harvard.edu

[b] Department of Biomedical Informatics and Medical Education, University of Washington

Box 358047

Seattle, WA 98195

cohenta@uw.edu

[c] Department of Psychological Sciences, Rice University

Sewall Hall

P.O. Box 1892

6100 Main Street

Houston, TX 77005

sjf2@rice.edu

ABSTRACT

Statistical regularities in the environment impact cognition across domains. In semantics, distributional approaches posit that similarity between words can be derived from regularities of the contexts in which they appear. Here, we study how regularities in written text impacts readers' knowledge about orthography: Can similarity between characters be learned from the written environment? Adapting methods from distributional semantics, we model the contextual similarity among alphanumeric characters in a large text corpus. We find modest correlations between model-derived similarities with similarity derived from a behavioral experiment. Beyond this result, model-derived similarity from neural embedding models captures key aspects of orthographic knowledge, like case, letter identity and consonant-vowel status. We conclude that the text environment contains regularities that are relevant to readers and that statistical learning from is a promising way for this information to be acquired. More broadly, our results imply that the statistical regularities are relevant not only at the level of word semantics but also individual written characters.

Keywords: distributional semantics, orthography, statistical learning, reading, consonants and vowels, abstract letter identities

This work was supported by the National Institutes of Health [R01 LM011563] and the National Science Foundation [CAREER 1752751].

Readers know a great deal about alphanumeric characters, not only how they can be legally combined but structured similarity between them. For example, we can easily distinguish between BLORK and BLO&K or B1ORK, knowing that only the former consists of letters exclusively. We agree that BLORK could be a word of English but BLRK could not, in part because it does not contain any vowel letters - A, E, I, O, or U. We know that PET and pet consist of the same letter identities, yet we use their case difference to distinguish between Positron Emission Tomography and the family dog. Aspects of this information are taught to some extent explicitly in school, but the current study investigates whether this knowledge is also available in a bottom-up manner from the written language environment.

A number of distinctions within orthography are relevant to accessing the correct phonological and semantic information for a given orthographic string. At a broad level, the distinction between digit and letter characters is relevant for determining text type (e.g., a database or a novel) and these characters require different types of semantic access (quantity information for digits, lexical semantics for words). There is also some evidence that processing speed and facility differs between these character types (Schubert, 2017; Starrfelt & Behrmann, 2011). Within the letter domain, multiple key distinctions are encoded by the English writing system. These include case, case-invariance, and consonant/vowel status. Whether a letter is in upper or lower case informs a reader about its location within a sentence, and some homonyms are distinguished only by case, including some that are acronyms or proper names (e.g., ‘Jack’ vs. ‘jack’). Access to the appropriate word semantics in these instances requires encoding of the letter case (for evidence that orthographic processing is affected by expectations that certain words are capitalized, see e.g., Perea, Jiménez, Talero, & López-Cañada, 2015; Peressotti, Cubelli, & Job, 2003; Kinoshita & Norris, 2018). Aside from these particular situations, case

invariance is an important property for fluent readers: the knowledge that ‘a’ and ‘A’ represent the same letter identity regardless of their differing appearance. The existence of abstract letter identities is a defining characteristic of the Roman alphabet and allows readers to understand that ‘FRIDGE’ and ‘fridge’ are the same word (Besner, Coltheart, & Davelaar, 1984; Bowers, Vigliocco, & Haan, 1998; Kinoshita & Kaplan, 2008; Polk & Farah, 1997, 2002; Rothlein & Rapp, 2014). The final key distinction is between consonant and vowel letters. Recent evidence has suggested that these two sub-categories of letters affect processing by structuring the input representations (Chetail & Content, 2012, 2014; Chetail, Drabs, & Content, 2014; Chetail, Ranzini, De Tiège, Wens, & Content, 2018; Schubert, Kinoshita, & Norris, 2017) and/or affecting the speed of lexical access by constraining the matching entries (e.g., Carreiras, Vergara, & Perea, 2009; Duñabeitia & Carreiras, 2011; New, Araujo, & Nazzi, 2008; New & Nazzi, 2014; Vergara-Martínez, Perea, Marín, & Carreiras, 2011). Furthermore, in the domain of spelling, acquired deficits can differentially affect consonant and vowel letters (Buchwald & Rapp, 2006; Caramazza & Miceli, 1990; McCloskey, Badecker, Goodman-Schulman, & Aliminosa, 1994; Miceli, Capasso, Benvegnù, & Caramazza, 2004). These distinctions between alphanumeric characters are vital to accurate reading, yet the degree to which this knowledge can be acquired implicitly remains unknown.

Statistical learning about orthography

How do literate adults come to have knowledge about their written language? Statistical learning, or the ability to acquire knowledge about patterns in the input, has long been proposed as a mechanism by which children and adults learn language implicitly (for a review, see: Romberg & Saffran, 2010). Much of this work has focused on learning of transitional probabilities, such as knowing that the sound /t/ is frequently followed by /i/ (as in TEA /ti/,

TEEN /tin/, and TEAM /tim/), and never by /l/ (no English words begin with TL… */tl/). This allows a learner to acquire knowledge of the legal sequences in a language: to know that /ti/ but not */tl/ is allowable. Famously, even infants are sensitive to transitional probabilities of phonemes in a speech stream (Saffran, Aslin, & Newport, 1996) and show sensitivity to the frequency with which phonemes are presented (Maye, Werker, & Gerken, 2002). These early findings and many that followed suggest that some type of learning mechanism allows children to pick up on statistical regularities in the environment to acquire knowledge about their spoken language.

Could similar statistical learning mechanisms apply to written language? The acquisition of written language is distinct from the acquisition of spoken language because spoken language abilities can be acquired through exposure alone while written language must be taught explicitly. However, there is a growing body of evidence suggesting that both children and adult readers are capable of learning implicit statistical patterns from written language. In this context, transitional probabilities capture the fact that the letter Q is often followed by U but never by K (QUEEN but not *QK).

In children, sensitivity to orthographic statistical regularities can be observed as soon as their exposure to written language begins in earnest. In a 2008 review, Deacon and colleagues (Deacon, Conrad, & Pacton, 2008) concluded that from the earliest years of schooling, children's reading and spelling behaviors are affected by statistic regularities of letter strings (words) they have seen. For example, after just a few months of reading instruction, children have picked up on some regularities such as allowable letter doublets (e.g., 'LL' and 'EE' but not 'WW'). By first grade, children prefer pseudowords that conform to the statistics of their text environment over those that do not, on the dimensions of letter frequency and letter co-occurrences (Cassar &

Treiman, 1997; Pacton, Perruchet, & Fayol, 2001), and tend to reproduce these statistics in spelling tasks (Pollo, Kessler, & Treiman, 2009). Adults' spelling behaviors also reflect the regularities of letter distribution. For example, when asked to spell pseudowords, English-reading adults mimic the statistical patterns of English letter doubling (Treiman & Boland, 2017). This evidence suggests a sensitivity to the statistical patterns of the text environment, specifically the transitional probabilities between letters. Furthermore, these and similar studies reveal that readers are sensitive to and reproduce the most frequent ways in which particular phonemes are realized in written words (for a review see: Kessler, 2009).

The majority of work in orthographic statistical learning considers sensitivity to transitional probabilities or conditional relationships between orthography and phonology, accounting for effects of frequent bigrams (e.g., QU and LL vs. QK and II) and letter-sound relationships (e.g., spell /l/ with LL but do not spell /k/ with KK). In addition to these types of regularities, we can also consider what the text environment reveals about the elements of the orthography themselves. That is, what is an 'a', and how does it relate to other elements in the orthography, like '7', 'A' or 'U'? Our approach to investigating how this type of orthographic knowledge could be acquired through statistical learning is the distributional hypothesis—which has been a fruitful approach to modelling the semantic relationships between words on the basis of their text environment—applied to the single character level.

The distributional hypothesis states that we learn about a given element of a set on the basis of the other elements with which it is likely to co-occur (see discussion in, e.g., Harris, 1954). In semantics, the hypothesis is that words that are similar in meaning often occur in similar contexts. That is, they are likely to appear with a similar set of words, or to quote Firth (1957), "you shall know a word by the company it keeps." Furthermore, a contextual

representation of which words a given word is likely to co-occur with forms part of the knowledge of what that word means and how it should be used (e.g., Firth, 1957; Miller & Charles, 1991). Evidence for the distributional hypothesis as it applies to semantic knowledge has been demonstrated empirically by the successes of techniques for text analysis including Hyperspace Analog of Language (HAL, Lund & Burgess, 1996) and Latent Semantic Analysis (LSA, Landauer & Dumais, 1997). LSA in particular served as a primer to the power of using a general learning mechanism to extract rich similarity information from a large distributed corpus of language, as semantic representations generated by LSA are highly related to many aspects of semantic processing in experimental data.

Both HAL and LSA have been described as 'count' models that tally word co-occurrences within a set window size or document (Mandera, Keuleers, & Brysbaert, 2017). From these counts, a measure of similarity between any two given words can be derived. More recent models with more complex architectures, including the class of neural embedding models (e.g., *word2vec*) developed by researchers at Google (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), instead attempt to predict a word or its context, and have been shown in many situations to out-perform count style models (Baroni, Dinu, & Kruszewski, 2014; Mandera et al., 2017; though this success seems to be due to properties other than the predictive nature: Johns, Mewhort, & Jones, 2019; Levy, Goldberg, & Dagan, 2015).

Leaving aside model type, successes of the distributional semantics approach include the ability to model judgements of semantic similarity (Landauer & Dumais, 1997), the ability to fill in the final word in a sentence (cloze task, Snyder & Munakata, 2008), the magnitude of semantic priming in lexical decision tasks (e.g., Günther, Dudschig, & Kaup, 2016; Hollis &

Westbury, 2016; Jones, Kintsch, & Mewhort, 2006; Mandera et al., 2017), and the distributed patterns of brain activity in semantic processing regions in response to both words (Fischer-Baum, Bruggemann, Gallego, Li, & Tamez, 2017) and pictures (Carlson, Simmons, Kriegeskorte, & Slevc, 2014). In our work, we extend the spirit of this approach, considering whether contextual similarity is also relevant to our knowledge of orthography, and the extent to which "you shall know a *letter* by the company it keeps" (with apologies to Firth).

The current study

In this study, we consider how much readers can learn about the characters that comprise their orthography purely on the basis of statistics of the text environment. We employ two classes of models, a class of LSA-inspired vector accumulation models (Kanerva, Kristofersson, & Holst, 2000) and a class of more-complex neural embedding models (Mikolov, Chen, et al., 2013). We trained models of each class on the same text corpus, resulting in a measure of similarity of characters in the environment as learned by each model. We submitted these model-derived similarities to two tests. In the first test, we correlated the model-derived similarity with behaviorally-derived character similarity, both solely and in combination (using multiple regression) with other measures of empirically-derived similarity. The behaviorally-derived character similarity takes the form of reaction times from a task in which participants had to decide whether two characters are the same or different; the use of such a data set as a measure of similarity has been well established (e.g., Podgorny & Garner, 1979 [visual similarity]; Rothlein & Rapp, 2017; Wiley, Wilson, & Rapp, 2016 [more abstract types of similarity]). This test allows us to evaluate the extent to which the model-derived similarity is related to a broad measure of character similarity displayed by human readers. In the second test, we directly evaluated whether the models that are most highly-correlated with the behavioral data learned

appropriate structured similarity among orthographic categories. Here we tested for sensitivity to letter/digit status, case invariance, upper/lowercase status, and consonant/vowel status. As reviewed above, these categories have been shown to have psychological reality and affect reading behavior; testing whether the model-derived similarity reflects these categories serves as a proof of concept for a distributional source of this knowledge.

MATERIALS AND METHODS

Text Modeling

The models we used fall into two classes: vector accumulation and neural embedding. The vector accumulation models are based on random indexing models by Kanerva and colleagues (Kanerva et al., 2000), adapted for the first time to the single character rather than word level. The neural embedding models are also adaptations of two prior algorithms: skipgram-with-negative-sampling (SGNS, Mikolov, Chen, et al., 2013) and Embeddings Augmented by Random Permutations (EARP, Cohen & Widdows, 2018), also newly adapted to the single-character level. To distinguish them from their word-based counterparts, we refer to the models in this paper as RandInd-char, SGNS-char, and EARP-char. The primary difference between the classes is that the vector accumulation models simply tally character occurrences while the neural embedding models are neural networks that progressively learn about each character's context. Both learn incrementally from a sliding window that moves across the entire text corpus. After a training period, in which each document of the corpus is presented in a randomly-determined order, the output of interest reflects what each model learned about each character in the corpus. (See below for details of the model architectures and learning procedures.) This output takes the form of a similarity matrix for every pairwise combination of

uppercase letters, lowercase letters, and digits (62x62 similarity values). For Figure 3, the model-derived similarity matrices were visualized using the Uniform Manifold Approximation and Projection (McInnes, Healy, Saul, & Großberger, 2018, https://github.com/lmcinnes/umap).

For both classes, we generated 10 runs of each model specification, each with different random initializations of the vectors at the onset of training. We analyzed the central tendency (median) of each model across the 10 runs to avoid drawing conclusions based on incidental properties of the random starting point.

Model architectures

The vector accumulation models (RandInd-char) have relatively few trainable parameters. A vector for each character (this is analogous to the 'semantic' vectors in LSA) is initialized as a zero vector, and each character is also assigned an initial random vector of 100 dimensions. During training, for each window, the random vector for every surrounding character is added to the semantic vector for the central character. Due to this simple addition procedure, the shuffling of document order across training runs does not have any impact on this model class. The outputs of interest after training are the vectors for each character, which are compared using a cosine metric to derive pairwise similarities between characters.

The neural embedding models (SGNS-char and EARP-char) are more complex, consisting of shallow neural networks with a single hidden layer (100 hidden units). The network is initialized with random weights for the hidden units (generating variability not present in the random indexing models). During training, for each window the network is trained to predict the surrounding characters with a high probability of being present in that context (within the window). Learning in these models can be framed as back-propagation via stochastic gradient descent, with a linearly-decreasing learning rate across training epochs. In addition to predicting

positive examples (observed context characters in the window), these models implement 'negative sampling' (Goldberg & Levy, 2014; Mikolov, Sutskever, et al., 2013): they are also trained to correctly reject negative examples that do not appear in the window. The outputs of interest from the fully-trained model are the (100-dimensional) vectors of input weights to the hidden units for each character. As with the vector accumulation models, a cosine metric is used to produce a similarity value for each pairwise combination of characters.

Comparing the model classes, the neural embedding models have twice as many trainable parameters as the random indexing models due to the randomly-initialized hidden weights that change as documents are presented during training. This class also has additional variability from the document-order shuffling that occurs for each run: the decreasing learning rate means that documents presented earlier in training have a larger effect on the vectors than ones presented later. (This shuffling does not affect the random indexing models because they use simple addition and A+B = B+A). A third difference is that only the neural embedding models include negative sampling- learning of characters that are *not* present in the context, in addition to ones that are. Thus, the generalization across the 10 runs is of particular relevance to the neural embedding models which have more non-zero vectors at training onset and random variability in the negative sampling procedure. See Supplementary Methods and Cohen & Widdows, (2018) for further details of the algorithm implementations for both model classes.

Training corpus

For training we used the popular TASA (Touchstone Applied Science Associates, Inc.) corpus, which includes a variety of document types (including novels, educational texts, and newspaper articles) with a total of approximately 73 million characters (including spaces), comprising 12 million words and 44,486 documents.

Testing model-window properties

We manipulated three properties across both classes of models. The first is a property of the corpus, the second a property of the moving window, and the third a property of how characters are considered within the window (see schematic depiction in Figure 1). The first property, *Word Boundaries*, controls whether the model represents spaces between words or whether the spaces are removed to create a single continuous 'word' in each document. We tested models with and without word boundaries. The second property, *Window Size*, controls the radius of the moving window of characters considered in each iteration. We tested models with radii of 2, 3, 4, 5, and 6 characters. Finally, we tested models with four different ways of representing the *Positions* of the characters in the window relative to the central character. The first version is a model that does not differentiate between the order of characters within the context window. (For the vector accumulation model this is akin to a 'bag of letters' model, for the neural embedding it is a SGNS model.) The remaining three versions encode character position in one of three ways: Directional, Positional, and Proximal. The Directional models represent separately the preceding (left) and following (right) characters in the window, with no distinction within these sets ("eat" and "tea" would result in different encodings for the character "a", but "beats" and "beast" would not). The Positional models separately represent each character within the window, without a similarity structure among them. The Proximity model is similar to the Positional model, but with higher similarity between characters that are similarly situated in relation to the central character ("abets" and "baest" would result in orthogonal encodings for the character "e" with the positional model, and similar but not identical encodings with the proximity model). (Refer to further examples in Figure 1B.) In total, we generated and tested 80 models (2 *Model Classes* x 2 *Word Boundaries* x 5 *Window Size* x 4 *Position*). The

resulting contextual similarity matrices are available for download at the following DOI: 10.17605/OSF.IO/P4QU9.

Behavioral Experiment

There is a range of tasks that could be used to elicit character similarity, just as a range of tasks has been used to determine semantic similarity and benchmark algorithms for distributional semantics. The task we employed here is not designed to provide explicit judgments about character context, but rather an indirect measure of similarity. In that way it is more similar to an analogy task (e.g., the Test Of English as a Foreign Language [TOEFL] Mandera et al., 2017) than a cloze or semantic rating tasks (e.g., Landauer & Dumais, 1997; Snyder & Munakata, 2008) in the context of testing models of text-derived semantic similarity.

To this end, we elicited a data set of behavioral character similarity in a same/different task with digits and uppercase letters. In this task, pairs which require a 'Same' response are physically identical ('D D' or '4 4') and those that require a 'different' response are physically non-identical ('D C' or 'D 4'). In a same/different task such as this, reaction time on the 'Different' trials reflects the similarity between each letter pair: More similar stimuli in a pair require longer response times (RT) to decide they differ (Courrieu, Farioli, & Grainger, 2004; Podgorny & Garner, 1979; Rothlein & Rapp, 2017; Wiley et al., 2016; Zhai & Fischer-Baum, 2019).

Forty-seven undergraduate students (34 women, mean age = 19.8 years) provided informed consent to participate in the experiment, receiving credit in a psychology or cognitive science course. The stimuli were 25 uppercase letters and 9 digits; O and 0 were excluded. The characters were combined pairwise into 561 trials requiring a Different response. The ordering of the stimuli in the Different pairs (e.g., '4 D' and 'D 4') was counterbalanced across participants

so that each participant saw each pair in only one order. The 578 Same trials consisted of 17 presentations of each of the 34 Same pairs.

On each trial, the two characters in a pair were presented in the center of the screen, in 18-pt. fixed-width Consolas font, separated by approximately two character spaces. Stimulus presentation ended upon response or after 1500 ms elapsed; a fixation cross intervened between trials for 500 ms (±100 ms randomly-determined jitter on half of the trials). Participants responded by pressing a 'same' or 'different' key as quickly as possible, one key with each hand. The hands used for the response keys were counterbalanced across participants. A 40-trial practice block using symbol stimuli (e.g., %, $) preceded the main experiment. E-Prime 2.0 (Psychology Software Tools, Pittsburgh, PA) was used for stimulus presentation and response collection. The entire experiment, including practice and a break, took about 45 minutes.

Only data from the Different trials were analyzed for this study. Data were cleaned by removing error trials (6.7% of trials) and outliers (trials with RT more than 2.5SD below or above each participant's mean RT). The resulting RTs were then normalized by dividing each trial's RT by that participant's mean RT. These data are available in the Supplementary Materials.

Correlations between model-derived similarity and behavioral similarity

For each model class (vector accumulation, neural embedding), we correlated the model-derived character similarity space from each model with the character similarity space of the behavioral data. We classified as the best model the one with the highest median correlation (Spearman's rho) across the ten runs. We report *p*-values that meet the Bonferroni-corrected threshold ($\alpha = .05/40 = .00125$) for multiple comparisons, corrected for the number of models in each class.

Reliability check and noise ceiling

As an additional check on the robustness of the results, we conducted a sampling procedure from the adult data. Each sample consisted of data from a randomly-selected subset of 30 participants, then tested against the medians (across the 10 runs) for each of the models. The highest-correlated model for each sample was recorded and the sampling procedure was repeated 10,000 times. This procedure allowed us to determine the extent to which the winning model fits some peculiarity of the full behavioral data set rather than a common pattern across samples of participants' data.

Additionally, we borrowed a method from representational similarity analysis (RSA) to calculate a noise ceiling for our behavioral data (Nili et al., 2014). The noise ceiling indicates the range in which a model is accounting for more than the minimum information present in the data (lower bound) and gives a maximum possible fit (upper bound), both based on the inherent noise in the data. The lower bound is calculated by averaging the correlations between each participant and the averaged group. The upper bound is calculated by averaging, in a leave-one-out manner, the correlations between each participant and the remaining participants as an averaged group.

Multiple regression

We conducted multiple regression to investigate the contributions of other types of similarity to the empirical data and to quantify the unique contribution of contextual similarity. On the basis of prior work, we tested the contributions of visual similarity, name similarity, and character frequency (e.g., Rothlein & Rapp, 2017; Wiley et al., 2016).

Two measures of visual similarity were used which differ in visual abstractness. *Pixel overlap* measures the number of overlapping pixels between any two characters, normalized by the total number of pixels present (Fischer-Baum et al., 2017; Kriegeskorte, Mur, & Bandettini, 2008; Marinus et al., 2016; Schubert, 2017; Starrfelt & Behrmann, 2011). This is a stimulus-driven measure of similarity, meant to roughly approximate low-level visual representations of each character as they were presented in the experiment. We computed the pixel overlap in MATLAB using black characters centered within a white background. *Feature overlap*, unlike pixel overlap, is a font-independent measure of visual similarity. The features used in this set are quite generic, including features such as “horizontal bar” and “line slanted 45-degrees left”; abstract enough to describe letters and digits regardless of the specific font in which they are presented. This measure is meant to approximate readers’ knowledge of the typical shape of letters and digits, rather than the specific visual details of the stimuli. (See further detail in Schubert, 2017.) *Name similarity* indexes the overlap between the spoken names (e.g., 3 as ‘three’ /θɹi/, D as ‘dee’ /diː/) of the two characters of a pair. Each name was decomposed into its constituent phonemes and phonological features (e.g., /diː/ as consonantal, coronal, anterior, voiced for /d/ and high, front, tense for /iː/), and overlap was computed as in Rothlein and Rapp (2014). We also included a predictor that was the difference in first-order character frequency between the two characters of each pair. Character frequency has been shown to affect character

recognition speed (Jones & Mewhort, 2004; New & Grainger, 2011; Walker & Hinkley, 2003); in the current context this predictor also allows us to test whether the model-derived similarity extends beyond capturing first-order statistics.

Sensitivity to orthographic categories

In a second analysis, we explored the extent to which the model that best-correlated with the empirical character similarity also captured categorical distinctions in alphanumeric characters: between letters and digits, between upper and lowercase letters, between orthographic consonants and vowels, and knowing that different characters (g, G) correspond to the same abstract letter identity. This analysis asks whether these types of categorical distinctions about alphanumeric characters can be learned from the distributional properties of the bottom-up input alone. For each comparison, we computed the average similarity within each category (e.g., letters, digits separately) and the average similarity between members of the category (e.g., letters and digits together). We then conducted one-tailed t-tests to evaluate the significance of the category difference. It is worth noting that significance in this analysis does not provide a measure of how likely that any specific vector could be classified as belonging to one orthographic category or another. Instead, this analysis allows us to test whether or not the distribution of vector values differ by category membership, which indicates that this distinction has at least, in part, been learned.

These analyses were run separately for each the 10 runs of the winning models; we report the most conservative results. While the correlation and multiple regression analyses are conducted on the subset of the contextual similarity that pertains to digits and uppercase letters only (minus 0 and O), because these are the characters used in the behavioral experiment, this analysis also considers all digit and letter characters in both cases with two exceptions. For the

consonant/vowel distinction, the letter Y was excluded due to its uncertain category membership. Furthermore, same-identity pairs were excluded from the consonant/vowel comparison because they could inflate the within-category similarity (same-identity pairs also by definition have the same consonant/vowel category).

RESULTS

Does model-derived similarity relate to behavioral similarity?

Simple correlations

In the behavioral task, average response time on Different trials was 507 ms and 493 ms on Same trials, and participants were 93.3% correct across the experiment. We computed correlations between the model-derived contextual similarity and the behaviorally-derived similarity. Figure 2 shows all of the correlations by model class and property. As can be appreciated in Figure 2A, the vector accumulation models have small variability across the 10 runs of each model, due to fewer values randomly initialized across each run as the neural embedding models (see Methods). Overall, for this class there is a general pattern that larger radius sizes lead to higher correlations with behaviorally-derived similarity. The different positional schemes do not appear to have a systematic relationship with the correlation, as can be seen by the intermixing of the colors for each radius size. In Figure 2B, for the neural embedding models, there is higher variability across runs, and a more subtle but still present trend for larger radius sizes to result in higher correlations.[1] Here, models without position coding (basic, in orange) perform more poorly than the models with permutation or proximity coding.

[1] In fact, a previous set of model iterations for the neural embedding models resulted in a winning model with no word boundaries and positional encoding (as here) but radius 6 outperformed radius 5.

To determine the best-fitting model, we took the median of the correlations across the ten model runs. Among the vector accumulation models, we found the highest correlation with the model with a radius of 5 characters, no word boundaries, and no position encoding ($\rho = 0.17$, $p < .001$). The highest correlation for the neural embedding models was with the radius 5, no boundary, positional model, with a slightly smaller but still significant $\rho = 0.15$, $p < .001$.

Both of these winning correlations fall within the noise ceiling (lower bound: $\rho = 0.09$; upper bound: $\rho = .20$), suggesting that both models relate to a non-trivial proportion of the variance present in the behavioral data. To check the reliability of our results, we repeated the correlations between model-derived and behaviorally-derived similarities on 10,000 random sub-samples of the behavioral data. For the vector accumulation models, the radius 5, no word boundary, no-position encoding model was most successful across 63% (6311/10000) of the samples. Likewise for the neural embedding model, the same model we found in the analysis with the entire behavioral data set (radius 5, no word boundaries, positional encoding) was the most highly-correlated model for approximately 37% of the samples (3744 times), the highest number of samples across the models. These sampling outcomes suggest that the success of the winning models is not due to a peculiarity of the full set of behavioral data but is consistent across smaller and repeated samples of the behavioral data.

Multiple regression

The correlations do not reach the upper bound of the noise ceiling, suggesting that model-derived contextual similarity is not the only source of behavioral similarity in this task. We conducted multiple regression to test how other sources of similarity combine with contextual similarity to predict the behavioral data. Four predictors were entered into the base model: visual

pixel similarity, visual feature similarity, name similarity, and frequency.[2] Without contextual similarity included, this base model fits the data, accounting for 14% of the variance $F(4,556) = 24.73$, $p < .001$, adjusted $R^2 = 0.1449$. Adding contextual similarity from the highest-correlated vector accumulation model to the base model, we find that it is a significant predictor (b= .016, se = .0063, $t = 2.47$, $p = 0.014$), and this full model is also significant and accounts for 15% of the variance ($F(5,555) = 21.18$, $p < .001$, adjusted $R^2 = 0.1527$). If we instead add contextual similarity from the highest-correlated neural embedding model to the base model, it is also a significant predictor (b= .020, se = .0076, $t = 2.59$, $p = 0.010$) and contributes to a significant full model accounting for 15% of the variance ($F(5,555) = 21.32$, $p < .001$, adjusted $R^2 = 0.1536$).

Directly comparing the regression models including the vector accumulation or neural embedding-based contextual similarity predictors indicates a very slightly higher variance accounted for with the neural embedding predictor ($R^2_{\text{EARP-char}} = .1536$, $R^2_{\text{RandInd-char}} = 0.1527$) but no statistical advantage when either predictor is added to a regression model already including the other ($p > .27$).

Do the models learn the structure of English orthography?

We next explored the structure in the model-derived contextual similarity of the highest-correlated model from each class. The similarity space is depicted (projected into two dimensions) in Figure 3, where it is possible to visually appreciate the dissimilarity (as distance) between different characters. From this figure, which displays similarity from one run of the winning vector accumulation (Figure 3A) and neural embedding (Figure 3B) models, stark

[2] Correlations between these four predictors can be found in Table S1. Briefly, feature and pixel overlap are correlated with each other and also with frequency (ps < .01). Despite this, the variance inflation factors (VIF) for all predictors are less than 1.08, indicating that it is possible to examine their unique contributions to the regression model.

differences between the model classes can be seen in the degree of structure among the letter character categories (e.g., letter/digit, upper/lowercase). To quantify this structure, we carried out t-tests to compare within- and across-category similarity.

First, we explored whether the model learned a distinction between letters and numbers. For each run, we calculated the average similarity of all digits to every other digits (e.g., 1-2, 3-6, 8-9), of all letters to every other letter (e.g., A-B, D-q, f-p) and of all digits to all letters (e.g., 1-A, 3-y). For the winning vector accumulation model, the average similarity of digits to each other was .90, the average similarity of letters to each other was .84, and the average similarity of digits to letters was .40. For the winning neural embedding model, the average similarity of digits to each other was .89, the average similarity of letters to each other was .26 and the average similarity of digits to letters was .20. For both model classes, the within category similarities (digits to other digits, letters to other letters) were significantly larger than the digit to letters similarities (RandInd-char: dig-dig vs. dig-let *t*s(563) > 30.08, $p < .0001$, let-let vs. dig-let *t*s(1844) > 69.76, $p < .0001$; EARP-char: dig-dig vs. dig-let *ts*(563) > 26.83, *p*s < .0001, let-let vs. dig-let *t*s(1844) > 2.16, *p*s < .016). Based solely on bottom-up contextual information about the distributional statistics of which other alphanumeric symbols are likely to appear together, both model classes learned that digits are more similar to other digits than they are to letters, and that letters are more similar to other letters than they are to digits, however it appears that the vector-accumulation model outperforms the neural embedding model in learning the broad class distinction between letters and digits.

Within the category of letters, readers make further important distinctions. We compared the average similarity of letters within the same case (e.g., A-C, d-f) and all letters across case (e.g., a-F, B-v). For the vector accumulation model, we found that within case similarity (.91)

was significantly higher than cross case similarity (.80; *t*s > 11.68, *p*s < . 0001). Likewise, for the neural embedding model, within case similarity (.41) was significantly higher than cross case similarity (.13; ts(1324) > 20.63, *p*s < . 0001). Moving to still more sophisticated orthographic distinctions, we compared letters that map on to the same abstract letter identity across case (e.g., A-a, B-b) with those that do not (e.g., A-b, B-a). For the vector accumulation model, letters that shared the same abstract letter identity had a similarity of .82 and those that did not had a similarity of .80, resulting in a non-significant difference (*t*s < .8, *p*s > .21). However, for the neural embedding model, we found greater similarity when letters shared the same abstract letter identities (.33) than when they did not (.13; *t*s(674) > 5.59, *p*s < .0001).

Finally, we investigated what the models learned about consonant and vowel letters. For the vector accumulation model, we found that consonants were similar to consonants (similarity = .82) and vowels to vowels (similarity = .85), but cross-category similarity was also quite high (.84), leading to no difference by consonant/vowel status (con-con vs. con-vow: *t*s(1158) < -2.73, ps > .99, vow-vow vs. con-vow: *t*s(438) < 1.03, ps > .15). For the neural embedding model, consonants were more similar to other consonants (similarity = .32) and vowels were more similar to other vowels (similarity = .32) than consonants were to vowels (similarity = .13; con-con vs. con-vow: *t*s(1158) > 12.22, *p*s < .0001, vow-vow vs. con-vow: *t*s(438) > 4.53, *p*s < .0001). Both models learned the broad categorical distinctions of letters versus digits and upper versus lowercase letters, likely because these kinds of characters tend to appear in different contexts. However, only the neural embedding model demonstrates fine-grained knowledge of the structure of English orthography that we know that readers are sensitive to: that particular (upper and lower case) characters map onto a single identity, and letters can be divided into sub-classes, specifically vowels and consonants.

Upon discovering the stark difference between the random indexing and neural embeddings models in their ability to learn orthographic categories, we wondered whether it can be attributed to particular features of this model class. One feature present solely in the neural embeddings model class is the ability to learn second-order relationships, which arises due to training of character context vectors (output weights) in the model. Because these vectors also learn, it allows the trained model to encode similarity between characters that occur next to similar but not identical other characters. As a result, neural embeddings models can learn more general similarity such as 'occurs next to digits,' whereas the random indexing models can learn only specific similarities between particular character pairs. We generated a hybrid neural embeddings model without learning of the character vectors to test the impact of this property, and found that these models were still able to distinguish consonants from vowels (con-con sim = .399, vow-vow sim = .449, con-vow sim = .227, con-con vs. con-vow: $t(1158)=14.79$, vow-vow vs. con-vow: $t(438)=8.04$) and were sensitive to abstract letter identities (same = .384, diff = .215 , same vs. diff ID: $t(674) = 6.01$. (See Supplementary Methods for details of these hybrid models.) This result is particularly striking in comparison to the corresponding random indexing model (no boundary, positional, radius 5), which as a property of the class also not have learning of character vectors, and does *not* acquire the consonant-vowel distinction or abstract letter identity (con-con sim = .657, vow-vow sim = .706, con-vow sim = .676, con-con vs. con-vow: $t(1158) = -1.69$, vow-vow vs. con-vow: $t(438) = 1.03$; same ID = .566, diff ID = .547, same vs. diff ID: $t(674) = 0.83$). Thus, it seems that the learning occurring at multiple locations within the neural embeddings model is not responsible for its ability to learn orthographic category structure.

The other feature that differs between random indexing and neural embeddings models and may underlie their difference in performance is the use of negative information. Recall that the neural embeddings models learn not only to predict characters that are present in a given character context, but also characters that are *not* present, via negative sampling where the probability of a character appearing as a negative sample is derived from its frequency in the corpus.[3] Johns and colleagues (2019) suggested that across multiple methods of implementing negative information (with and without explicit prediction), models with this property largely out-perform models with positive information alone. We found a similar pattern with our models: A hybrid EARP-char model without negative sampling had nearly uniform similarity values (close to 1) across all character pairs, and thus did not show any distinctions among character categories. Thus, it seems negative sampling allows the neural embeddings model class to out-perform the vector accumulation class in learning orthographic structure.

DISCUSSION

Prior research has shown that our knowledge about the semantic relationship between words can be captured, in part, by algorithms that consider the similarity of the contexts in which each word appears. We have shown that the same approach to statistical learning applied to single alphanumeric characters can capture some of what readers know about their orthography. In testing two classes of models that learn contextual similarity from text, we found that the two classes had different strengths in learning about the characters in English. The first class, vector accumulation, are very simple models that essentially track contextual character frequency. These models had a numerically higher correlation with human behavioral performance on a

[3] The original *word2vec* assigns the probability of a token being drawn for negative sampling as $F^{0.75}$where F is the frequency of the token of interest (number occurrences / total tokens in corpus). Following *FastText*, and the original EARP experiments, we used the *Semantic Vectors* default of $F^{0.5}$.

same-different task. They also captured the distinction between letter and digit characters, a difference which underlies the alphanumeric category effect, a general tendency for digits to be processed more quickly/easily than letters (Schubert, 2017). Polk and Farah first presented evidence that this distinction could be learned from context in 1994 (Polk & Farah, 1994, 1995), and our work confirms theirs, showing that a simple model can indeed pick up on the differing distributions of letters and digits in the text environment. However, orthographic structure is richer than the distinction between the broad categories of letters and digits. Our findings suggest that vector accumulation models cannot learn about this more detailed structure.

The second class of models we tested, neural embedding models, have more complex architectures, involving learning at multiple locations within the model and learning from both positive and negative context examples. Neural embedding models were better able to capture the orthographic categories among letters, including the distinctions between uppercase and lowercase letters, vowels and consonants, and the case-invariance principle. Curious as to which properties of this model class might underlie the greater ability to these capture fine-grained distinctions, we implemented two modified versions of the winning EARP-char model: one in which character vectors are fixed during training, and one in which negative sampling is not used. We found that only models with negative sampling learn orthographic structure, speaking to the importance of negative sampling for learning appropriate similarity among both words (as in Johns et al., 2019) and single characters. The failure of EARP-char models without negative sampling is consistent with the tendency of random indexing models to converge upon a single vector when trained over repeated iterations (Cohen, Schvaneveldt, & Widdows, 2020). We have demonstrated a strength of the neural embeddings model class in that their use of negative information allows the learning of structured similarity consistent with English orthography. The

ideal statistical learner would extract similarities that correlate highly to the behaviorally-derived similarity and that *also* demonstrate sensitivity to orthographic categories, while the models here differ in their ability to match both types of information.

There are three main implications of our results. The first is that the distributional statistics of text contain information that is relevant to readers' knowledge of their written language. Though the mere presence of distributional information in the text environment is perhaps trivial, our work demonstrates that this information is functionally related to readers' orthographic processing. The second implication is a proof of concept that statistical learning is a mechanism by which readers can learn about their written language at the level of individual characters. An algorithm implementing a particular type of statistical learning extracted character similarity information that is related to the similarity demonstrated by readers. Our testing did not reveal that all model classes and properties are equally successful; many did not correlate with human behavior above the noise inherent in the data. The successful ones may be considered as a simple test that statistical learning broadly can be an effective learning tactic. To be clear, we do not contend that readers have either a vector accumulation or neural embedding model with radius 5 and no sensitivity to word boundaries in their heads (see below for additional discussion of this point). In fact, it is likely that none of the implementational details of the models we tested are analogous to the mind's solution to statistical learning from the text environment. However, statistical algorithms can acquire some aspects of knowledge that are also present in the reader's mind. The third and broadest implication is that the distributional hypothesis, which states that we learn how things relate to each other by comparing the contexts in which they appear, might be a more general property of how humans learn from the environment and not just an account of how we learn word meanings.

The similarity learned by the most successful neural embedding model demonstrated sensitivity to important distinctions within English orthography. On the basis of the distribution of letters in the bottom-up input, letters were differentiated from digits and letter-specific features of orthographic representations like case and abstract letter identity were extracted. Perhaps most striking is the learning of the consonant-vowel sub-classes. The consonant-vowel distinction is most clearly a property of the phonological systems, in which the difference between these phoneme classes is grounded in acoustic and articulatory differences. Yet, the model was able to learn a distinction between orthographic consonants and vowels, based only on the distribution of letters, without any knowledge of how those letters are pronounced. This result could explain why the consonant-vowel status of letters can have an effect of early stages of letters string processing, well prior to the point of activating the phonology (e.g., Chetail et al., 2018), and why deaf readers show sensitivity to the consonant-vowels status of letters (e.g., Olson & Nickerson, 2001). Distributional properties of the written input allow us to learn that the characters of our writing system contain elements and sub-classes with different distributional properties.

We do not postulate that the precise specifications of our computer model describe the specific statistical learning mechanism by which language learners learn about the environment. Instead, our work, like Landauer and Dumais's first introduction of LSA in 1997, serves as an indication of the vast amounts of data available in the environment, and the rich distributional structure that can be extracted from it, even at the level of individual alphanumeric characters. The properties that allowed the best fit to the behavioral data are not predictions about those used in mental computations for tracking letter statistics. Instead our contribution is at Marr's computational level: A description of inputs and computations that arrive at the desired outputs.

The specific implementation, be it as EARP-char or as a set of cognitive processes or as neural wetware, is not our main concern.

Relationship between the winning model and computational models of reading

In addition to evaluating broad classes of models, we investigated specific properties within each model class. For the neural embedding model that was able to learn fine-grained orthographic structure, we found that the best performing model learned over a sliding window with a radius of 5 characters on either side, ignoring word boundaries and representing each character position within a window as totally distinct. First, let us emphasize that this model was not designed to be a mechanistic description of how the human mind extracts statistical regularities from text. Rather than a sliding window centered on each letter in turn, reading depends on fixations on the left portion of a word (in scripts read left-to-right) followed by saccades that typically jump eight or nine letter spaces (Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001; Rayner, Slattery, & Bélanger, 2010). Nor are the models described above designed to be a theory of how we process strings of letters for reading aloud or lexical decision. Instead, these algorithms process the statistics of the text to determine how much information is available in this input, without semantics or phonology. Models of distributional semantics likewise do not respect the specifications of the visual word recognition system when they are run with window sizes of 5 or even 11 words at a time (e.g., Baroni et al., 2014); they are not meant to model the human semantic system but instead to test the extent to which semantic information can be extracted from word context. As Landauer and Dumais (1997) describe about the original LSA model with respect to word semantics, our EARP-char models represent an "abstract computational method" to examine whether the data present in letter context is related to human knowledge about letters.

Still, we found a specific set of properties that maximized how much the model-derived similarity from the text relates to human knowledge about letters. Below, we consider how well these properties map onto the assumptions about how orthographic input is coded in theories of visual word recognition. The space of proposed theories is wide (see Rapp & Fischer-Baum, 2014 for review), including theories that the core units of visual word recognition are ordered letter pairs, or open bigrams (e.g., Grainger, Granier, Farioli, Van Assche, & van Heuven, 2006; Whitney, 2008), theories that letters are coded by flexible or uncertain positions defined relative to word boundaries (e.g., Fischer-Baum, Charny, & McCloskey, 2011; Norris, Kinoshita, & van Casteren, 2010), and theories that words are recognized by maximally aligning input strings to known letter strings based on common letter sequences (e.g., Davis, 2010). While the features of these theories of visual word recognition do not perfectly align with the properties that we compared in the models, we can consider how the winning model relates to the assumptions of these theories.

First, the winning model makes no reference to word boundaries, yet they play a key role in all theories of word recognition. They are a critical component in theories that assume that letter position is defined relative to the boundaries of the word (e.g., Fischer-Baum et al., 2011; McCloskey, Fischer-Baum, & Schubert, 2013), but even in other theories a special status is given to the first letter in the word (e.g., Davis, 2010; Whitney, 2008). That models without word boundaries performed best suggests that statistical regularities both within and across words are relevant sources of information about character context, even though word boundaries are vital to the task of visual word recognition (e.g., Epelboim, Booth, Ashkenazy, Taleghani, & Steinman, 1997; Kohsom & Gobet, 1997; Perea & Acha, 2009; Rayner, Fischer, & Pollatsek, 1998). Second, the window size of 5 letters to the left and right of fixation is somewhat larger than what

has been assumed in word recognition theories. For example, open bigram theories assume that readers track bigrams with at most two intervening letters, a far narrower window than the winning model treats as relevant (Grainger et al., 2006). However, we can tentatively conclude that position coding relative to the center of the sliding window is more compatible with certain theories of visual word recognition than others. Open bigram theories would predict that the winning model should be either a directional model –readers are only sensitive to which letters precede and follow the center letter (Grainger & van Heuven, 2004) – or a proximal model – that readers are sensitive to preceding and following letters, weighted by distance (Grainger et al., 2006). However, the winning model was a positional model, more in line with theories that assume we recognize words by activating sequences of letters represented in different positions, either relative to some anchoring point (e.g., Fischer-Baum et al., 2011; Norris et al., 2010), or coded in a dynamic way to best align with existing mental representations of familiar words (e.g., Davis, 2010). The evidence favoring these theories from our analyses is weak, particularly since the models perform very similarly across these different property settings. But, on balance, we believe that the coding scheme adopted by our winning statistical learning model is slightly more consistent with positional models of letter position encoding than open bigram models.

The gap between model and behavioral similarity

The models we tested do not correlate extremely highly with the behavioral data but they are within the noise ceiling and the results are consistent across sub-samples of the data. We take the modest correlation to indicate both the promise of this method and the distance between this approach and an ideal orthographic statistical learner. We consider two potential reasons for why a maximal correlation was not observed.

Given that our correlations did not reach the upper bound of the noise ceiling, there is some additional variability in the adults' character similarity data that is not captured by contextual similarity. That is, the implicit knowledge revealed by the RT data extends beyond knowledge of the contexts in which characters appear. The multiple regression analyses described above provide some support for this conclusion. In addition to contextual similarity, adults' same-different reaction times were influenced by the visual similarity of characters. This result is not surprising considering the visual presentation; prior work has suggested this task is highly sensitive to visual similarity (Podgorny & Garner, 1979; Rothlein & Rapp, 2017; Wiley et al., 2016; Zhai & Fischer-Baum, 2019). This feature of the behavioral data can also be appreciated in Figure S2, which displays the difference the behaviorally- and model-derived similarities; pairs with high visual similarity (e.g., 8B, YV) have high similarity behaviorally but not in the model (dark pink cells in the figure). However, the significant contribution of contextual similarity above and beyond measures of visual similarity is consistent with the idea that contextual knowledge plays a role in determining broader character similarity. Other factors may also have impacted behavioral performance for particular character pairs, such as number magnitude (e.g., a distance effect may lead participants to be slower to indicate that '2 3' are different than '2 8'), letter-to-phoneme mapping, or bigram frequency (e.g., 'T H' vs. 'T L'). Our experiment was not designed to exclude the influence of these factors and it is unclear how the models would pick up this type of knowledge. In sum, the speed with which people decide whether or not two characters are identical is influenced by many factors, with contextual similarity playing only one role. Future work with other tasks that directly assess character contextual similarity but avoid the influence of factors such as visual similarity may result in higher correlations with contextual similarity as extracted by statistical learning algorithms.

The second potential reason that the correlation between model-derived and behaviorally-derived similarities did not reach the upper bound of the noise ceiling is that the algorithms may be picking up on relationships that adults are not sensitive to. For example, the algorithms extract complex relationships among the characters within the moving window and is equally effective at extracting these relationships regardless of linguistic variables that are known to affect the speed and effectiveness of adults' reading behavior, such as lexical frequency and syntactic complexity. In some ways then, we could consider the algorithms overpowered, picking up on contextual similarity that is perhaps more complex or otherwise beyond readers' capabilities. Adults' ability to track first-order transitional probabilities is well-documented (see discussion in: Siegelman, Bogaerts, Elazar, Arciuli, & Frost, 2018), but the neural embedding algorithms in particular are tracking higher-order relationships beyond the capability of human statistical learning capabilities. In discussing readers' sensitivity to statistical regularities such as how a particular sound is spelled, Kessler (2009) notes that while spelling behavior among multiple alternatives is non-uniform and correlated with patterns in a large corpus, there is not a perfect replication of the text statistics. Thus, while adults are statistical learners in some senses, they are not ideal learners, and powerful statistical learning algorithms are capable of extracting similarities that adults are not. Our conclusion is that literate adults use distributional knowledge about which characters co-occur to learn the relationship between alphanumeric symbols, in a manner that is similar to—but certainly not identical to—the algorithms we tested.

Cross-linguistic considerations

Our study only considered a single language, using an English language corpus and English readers as our study population. We propose that while the specific similarities learned between alphanumeric characters are determined by language, the broader claim that readers

extract statistical regularities from text is universal. Frost (2012) argues that accounting for readers' sensitivity to the statistics of their text environment explains some reading effects which appear to be language-specific, without requiring language-specific mechanisms. Because different languages have different distributional properties, if readers' behavior is affected by these properties, divergent reading behavior across languages is expected (see also: Chetail, Balota, Treiman, & Content, 2015). Additionally, some authors suggest that sensitivity to statistical regularities may be particularly important in readers of inconsistent orthographies, such as English (Samara & Caravolas, 2014). This seems to predict that the extent of fit between model-derived context similarity and behaviorally-derived similarity may differ in interesting and predictable ways depending on language properties. In future work, we hope to extend our modeling efforts to other languages to determine the extent to which statistical learning algorithms with the same properties capture relevant regularities in other languages, or if the best-fitting properties are language-specific.

Conclusions

In summary, we demonstrate that certain distinctions and similarities within orthography can be derived bottom-up from the environment. We show that a statistical learning algorithm with particular properties captures aspects of knowledge that adult readers have about their orthography, suggesting that there is rich contextual information at the level of alphanumeric characters and that readers are sensitive to statistical regularities of their written language environment.

Figure Captions

Figure 1. Schematic depiction of the properties of the statistical learning algorithms employed in this study. Panel A: Demonstration of the context (moving window) for the underlined central character, for each setting of the properties *Window Size* and *Word Boundary*. In the first line, the model attempts to learn that [b], [r], [w], and [n] appear in contexts with [o]. Panel B: Details of the *Position* property, noting which characters are distinguished by a model learning the context of any given letter [x].

Figure 2. Scatter plots of correlation (Spearman's rho) values between the modeled contextual similarity (40 model specifications, 10 runs of each) for each model class and the behavioral similarity. Color corresponds to the position encoding scheme. Gray shading indicates noise ceiling. A. Correlations for the vector accumulation models. (Low variability for "Basic" models is due to the simplicity of this model: Few parameters change between runs leading to highly similar outputs.) B. Correlations for the neural embedding models.

Figure 3. Projection of the model-derived similarity spaces into two dimensions (arbitrary distance units). An example model run is depicted here for each model class. A. Projection of the vector accumulation model that was correlated highest with behavioral similarity (RandInd-char radius 5, no boundary, no position encoding). Note the far separation between letter and digit characters and moderate separation between uppercase and lowercase letters. B. Projection of the neural embedding model that was correlated highest with behavioral similarity (EARP-char radius 5, no boundary, radius 5, positional). Note the clustering of vowel and consonant letters in addition to the separations seen in A.

References

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, *1*, 238–247.

Besner, D., Coltheart, M., & Davelaar, E. (1984). Basic processes in reading: computation of abstract letter identities. *Canadian Journal of Psychology*, *38*(1), 126–34.

Bowers, J. S., Vigliocco, G., & Haan, R. (1998). Orthographic, phonological, and articulatory contributions to masked letter and word priming. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(6), 1705–1719.

Buchwald, A., & Rapp, B. C. (2006). *Consonants and vowels in orthographic representations. Cognitive Neuropsychology* (Vol. 23).

Caramazza, A., & Miceli, G. (1990). The structure of graphemic representations. *Cognition*, *37*(3), 243–297.

Carlson, T. A., Simmons, R. A., Kriegeskorte, N., & Slevc, L. R. (2014). The emergence of semantic meaning in the ventral temporal pathway. *Journal of Cognitive Neuroscience*, *26*, 120–131.

Carreiras, M., Vergara, M., & Perea, M. (2009). ERP correlates of transposed-letter priming effects: the role of vowels versus consonants. *Psychophysiology*, *46*(1), 34–42.

Cassar, M., & Treiman, R. (1997). The beginnings of orthographic knowledge: children's knowledge of double letters in words. *Journal of Educational Psychology*, *89*(4), 631–644.

Chetail, F., Balota, D., Treiman, R., & Content, A. (2015). What can megastudies tell us about the orthographic structure of english words? *Quarterly Journal of Experimental*

Psychology, *68*(8), 1519–1540.

Chetail, F., & Content, A. (2012). The internal structure of chaos: letter category determines visual word perceptual units. *Journal of Memory and Language*, *67*(3), 371–388.

Chetail, F., & Content, A. (2014). What is the difference between oasis and opera? roughly five pixels: orthographic structure biases the perceived length of letter strings. *Psychological Science*, *25*(1), 243–9.

Chetail, F., Drabs, V., & Content, A. (2014). The role of consonant/vowel organization in perceptual discrimination. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 938–961.

Chetail, F., Ranzini, M., De Tiège, X., Wens, V., & Content, A. (2018). The consonant/vowel pattern determines the structure of orthographic representations in the left fusiform gyrus. *Cortex*, *101*, 73–86.

Cohen, T., Schvaneveldt, R. W., & Widdows, D. (2020). Reflective random indexing and indirect inference: a scalable method for discovery of implicit connections. *Journal of Biomedical Informatics*, *43*(2), 240–256.

Cohen, T., & Widdows, D. (2018). Bringing order to neural word embeddings with embeddings augmented by random permutations (earp).

Courrieu, P., Farioli, F., & Grainger, J. (2004). Inverse discrimination time as a perceptual distance for alphabetic characters. *Visual Cognition*, *11*(7), 901–919.

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychological Review*, *117*(3), 713–58.

Deacon, S. H., Conrad, N., & Pacton, S. (2008). A statistical learning perspective on children's learning about graphotactic and morphological regularities in spelling. *Canadian*

Psychology, *49*(2), 118–124.

Duñabeitia, J. A., & Carreiras, M. (2011). The relative position priming effect depends on whether letters are vowels or consonants. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1143–63.

Epelboim, J., Booth, J. R., Ashkenazy, R., Taleghani, A., & Steinman, R. M. (1997). Fillers and spaces in text: the importance of word recognition during reading. *Vision Research*, *37*(20), 2899–2914.

Firth, J. R. (1957). *A synopsis of linguistic theory: Studies in linguistic analysis*. Oxford: Blackwell.

Fischer-Baum, S., Bruggemann, D., Gallego, I. F., Li, D. S., & Tamez, E. R. (2017). Decoding levels of representation in reading: a representational similarity approach. *Cortex*, (March), 1–15.

Fischer-Baum, S., Charny, J., & McCloskey, M. (2011). Both-edges representation of letter position in reading. *Psychonomic Bulletin & Review*, *18*(6), 1083–9.

Frost, R. (2012). Towards a universal model of reading. *The Behavioral and Brain Sciences*, 1–17.

Goldberg, Y., & Levy, O. (2014). Word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *ArXiv*.

Grainger, J., Granier, J.-P., 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), 865–84.

Grainger, J., & van Heuven, W. J. B. (2004). Modeling letter position coding in printed word perception. In P. Bonin (Ed.), *Mental Lexicon: Some Words to Talk about Words* (pp. 1–

23). Nova Science Publishers.

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: evidence from priming studies. *Quarterly Journal of Experimental Psychology*, *69*(4), 626–653.

Harris, Z. (1954). Distributional structure. *Word*, 146–162.

Hollis, G., & Westbury, C. (2016). The principals of meaning: extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin and Review*, *23*(6), 1744–1756.

Johns, B. T., Mewhort, D. J. K., & Jones, M. N. (2019). The role of negative information in distributional semantic learning. *Cognitive Science*, *43*(5).

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*(4), 534–552.

Jones, M. N., & Mewhort, D. J. K. (2004). Case-sensitive letter and bigram frequency counts from large-scale english corpora. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 388–96.

Kanerva, P., Kristofersson, J., & Holst, A. (2000). Random indexing of text samples for latent semantic analysis. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, 1036.

Kessler, B. (2009). Statistical learning of conditional orthographic correspondences. *Writing Systems Research*, *1*(1), 19–34.

Kinoshita, S., & Kaplan, L. (2008). Priming of abstract letter identities in the letter match task. *Quarterly Journal of Experimental Psychology*, *61*(12), 1873–85.

Kohsom, C., & Gobet, F. (1997). Adding spaces to thai and english: effects on reading.

Proceedings of the 19th Annual Meeting of the Cognitive Science Society, 388–393.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(November), 4.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: the latent semantic analysis theory of acquisition , induction, and representation of knowledge. *Psychological Review*, *1*(2), 1–72.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211–225.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*, 203–8.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: a review and empirical validation. *Journal of Memory and Language*, *92*, 57–78.

Marinus, E., Mostard, M., Segers, E., Schubert, T. M., Madelaine, A., & Wheldall, K. (2016). A special font for people with dyslexia: does it work and, if so, why? *Dyslexia*, *22*(3), 233–244.

Maye, J., Werker, J. F., & Gerken, L. A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), 101–111.

McCloskey, M., Badecker, W., Goodman-Schulman, R. A., & Aliminosa, D. (1994). The structure of graphemic representations in spelling: evidence from a case of acquired dysgraphia. *Cognitive Neuropsychology*, *11*(3), 341–392.

McCloskey, M., Fischer-Baum, S., & Schubert, T. M. (2013). Representation of letter position in single-word reading: evidence from acquired dyslexia. *Cognitive Neuropsychology*, *30*(6), 396–428.

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: uniform manifold approximation and projection. *Journal of Open Source Software*, *3*(29), 861.

Miceli, G., Capasso, R., Benvegnù, B., & Caramazza, A. (2004). The categorical distinction of vowel and consonant representations: evidence from dysgraphia. *Neurocase*, *10*(2), 109–121.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv:1301.3781v3*, 1–12.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26.* (pp. 3111–3119). Curran Associates, Inc.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic synonymy. *Languages and Cognitive Processes*, *6*(1), 1–28.

New, B., Araujo, V., & Nazzi, T. (2008). Differential processing of consonants and vowels in lexical access through reading. *Psychological Science*, *19*(12), 1223–1227.

New, B., & Grainger, J. (2011). On letter frequency effects. *Acta Psychologica*, *138*(2), 322–328.

New, B., & Nazzi, T. (2014). The time course of consonant and vowel processing during word recognition. *Language, Cognition and Neuroscience*, *29*(2), 147–157.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A

toolbox for representational similarity analysis. *PLoS Computational Biology*, *10*(4).

Norris, D., Kinoshita, S., & van Casteren, M. (2010). A stimulus sampling theory of letter identity and order. *Journal of Memory and Language*, *62*(3), 254–271.

Olson, a C., & Nickerson, J. F. (2001). Syllabic organization and deafness: orthographic structure or letter frequency in reading? *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *54*(2), 421–38.

Pacton, S., Perruchet, P., & Fayol, M. (2001). Implicit learning out of the lab. *Journal of Experimental Psychology-General*, *130*(3), 401–426.

Perea, M., & Acha, J. (2009). Space information is important for reading. *Vision Research*, *49*(15), 1994–2000.

Perea, M., Jiménez, M., Talero, F., & López-Cañada, S. (2015). Letter-case information and the identification of brand names. *British Journal of Psychology*, *106*(1), 162–173.

Peressotti, F., Cubelli, R., & Job, R. (2003). On recognizing proper names: the orthographic cue hypothesis. *Cognitive Psychology*, *47*(1), 87–116.

Podgorny, P., & Garner, W. R. (1979). Reaction time as a measure of inter- and intraobject visual similarity: letters of the alphabet. *Perception & Psychophysics*, *26*(1), 37–52.

Polk, T. A., & Farah, M. J. (1994). Late experience alters vision. *Nature*, *376*, 648–9. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=40359&tool=pmcentrez&rendertype=abstract

Polk, T. A., & Farah, M. J. (1995). Brain localization for arbitrary stimulus categories: a simple account based on hebbian learning. *Proceedings of the National Academy of Sciences*, *92*(26), 12370–3. Retrieved from

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=40359&tool=pmcentrez&rendertype=abstract

Polk, T. A., & Farah, M. J. (1997). A simple common contexts explanation for the development of abstract letter identities. *Neural Computation*, *9*(6), 1277–89. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9248063

Polk, T. A., & Farah, M. J. (2002). Functional mri evidence for an abstract, not perceptual, word-form area. *Journal of Experimental Psychology: General*, *131*(1), 65–72.

Pollo, T. C., Kessler, B., & Treiman, R. (2009). Statistical patterns in children's early writing. *Journal of Experimental Child Psychology*, *104*(4), 410–426.

Rapp, B. C., & Fischer-Baum, S. (2014). Representation of orthographic knowledge. *The Oxford Handbook of Language Production*, (May 2014).

Rayner, K., Fischer, M. H., & Pollatsek, A. (1998). Unspaced text interferes with both word identification and eye movement control. *Vision Research*, *38*(8), 1129–1144.

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest*, *2*, 31–74.

Rayner, K., Slattery, T. J., & Bélanger, N. N. (2010). Eye movements, the perceptual span, and reading speed. *Psychonomic Bulletin and Review*, *17*(6), 834–839.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews. Cognitive Science*, *1*(6), 906–914.

Rothlein, D., & Rapp, B. C. (2014). The similarity structure of distributed neural responses reveals the multiple representations of letters. *NeuroImage*, *89*, 331–44.

Rothlein, D., & Rapp, B. C. (2017). The role of allograph representations in font-invariant letter

identification. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(7), 1411–1429.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*. Retrieved from http://www.bcs.rochester.edu/people/newport/saffran1996.pdf

Samara, A., & Caravolas, M. (2014). Statistical learning of novel graphotactic constraints in children and adults. *Journal of Experimental Child Psychology*, *121*(1), 137–155.

Schubert, T. M. (2017). Why are digits easier to identify than letters? *Neuropsychologia*, *95*, 136–155.

Schubert, T. M., Kinoshita, S., & Norris, D. (2017). What causes the greater perceived similarity of consonant-transposed nonwords? *The Quarterly Journal of Experimental Psychology*, *0*(0), 1–18.

Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment : prior knowledge impacts statistical learning performance. *Cognition*, *177*(April), 198–213.

Snyder, H. R., & Munakata, Y. (2008). So many options, so little time: the roles of association and competition in underdetermined responding. *Psychonomic Bulletin and Review1*, *15*(6), 1083–1088.

Starrfelt, R., & Behrmann, M. (2011). Number reading in pure alexia-a review. *Neuropsychologia*, *49*, 2283–2298.

Treiman, R., & Boland, K. (2017). Graphotactics and spelling: evidence from consonant doubling. *Journal of Memory and Language*, *92*, 254–264.

Vergara-Martínez, M., Perea, M., Marín, A., & Carreiras, M. (2011). The processing of consonants and vowels during letter identity and letter position assignment in visual-word recognition: an erp study. *Brain and Language*, *118*(3), 105–17.

Walker, P., & Hinkley, L. (2003). Visual memory for shape-colour conjunctions utilizes structural descriptions of letter shape. *Visual Cognition*, *10*(8), 987–1000.

Whitney, C. (2008). Supporting the serial in the seriol model. *Language and Cognitive Processes*, *23*(6), 824–865.

Wiley, R. W., Wilson, C., & Rapp, B. C. (2016). The effects of alphabet and expertise on letter perception. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(8), 1186–1203.

Zhai, M., & Fischer-Baum, S. (2019). Exploring the effects of knowledge of writing on reading chinese characters in skilled readers. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *45*(4), 724–731.