Emergence of a compositional neural code for written words: Recycling of a convolutional neural network for reading

T. Hannagan[a,b,1], A. Agrawal[a,b,1], L. Cohen[c,d], and S. Dehaene[a,b,2]

[a]Cognitive Neuroimaging Unit, Commissariat à l'Énergie Atomique et aux Énergies Alternatives, INSERM, Université Paris-Saclay, NeuroSpin, Gif-Sur-Yvette 91191, France; [b]Collège de France, Université Paris Sciences Lettres 75005 Paris, France; [c]Sorbonne Université, INSERM U1127, CNRS UMR 7225, Institut du Cerveau et de la Moelle épinièr, Hôpital de la Pitié-Salpêtrière, Paris 75013, France; and [d]Assistance Publique-Hôpitaux de Paris, Hôpital de la Pitié Salpêtrière, Fédération de Neurologie, Paris F-75013, France

Contributed by S. Dehaene, September 3, 2021 (sent for review March 28, 2021; reviewed by Marco Zorzi and Kenneth Pugh)

The visual word form area (VWFA) is a region of human inferotemporal cortex that emerges at a fixed location in the occipitotemporal cortex during reading acquisition and systematically responds to written words in literate individuals. According to the neuronal recycling hypothesis, this region arises through the repurposing, for letter recognition, of a subpart of the ventral visual pathway initially involved in face and object recognition. Furthermore, according to the biased connectivity hypothesis, its reproducible localization is due to preexisting connections from this subregion to areas involved in spoken-language processing. Here, we evaluate those hypotheses in an explicit computational model. We trained a deep convolutional neural network of the ventral visual pathway, first to categorize pictures and then to recognize written words invariantly for case, font, and size. We show that the model can account for many properties of the VWFA, particularly when a subset of units possesses a biased connectivity to word output units. The network develops a sparse, invariant representation of written words, based on a restricted set of reading-selective units. Their activation mimics several properties of the VWFA, and their lesioning causes a reading-specific deficit. The model predicts that, in literate brains, written words are encoded by a compositional neural code with neurons tuned either to individual letters and their ordinal position relative to word start or word ending or to pairs of letters (bigrams).

reading | VWFA | neural network | literacy | compositionality

Reading acquisition relies on the development of a novel interface between vision and language, in charge of efficiently identifying letters and their ordering (1, 2). This orthographic analysis then feeds the language systems supporting semantics and phonology. Over the past 20 y, some basic features of this interface have been put to light. A specific region of the left ventral occipitotemporal (VOT) cortex, which was labeled the visual word form area (VWFA), is present at a similar location in the brain of every literate subject and is thought to underlie orthographic coding (3). Functional brain imaging has uncovered a host of functional features of the VWFA—for example, tuning to familiar vs. unknown alphabets (4, 5), partial invariance for retinal location (3, 6), invariance for uppercase/lowercase (7, 8), or sensitivity to the frequency of word occurrence (9, 10). Nevertheless, how this region becomes specialized for written words, or even whether it does so, remains a highly controversial issue (11–14).

Current evidence suggests that the VWFA site owes its functional specialization to a combination of two factors. First, according to the neuronal recycling hypothesis, reading preempts and repurposes part of the large region of ventral visual cortex that participates in visual object recognition (11, 15–17). The specific region involved may not only possess a generic architecture for invariant visual recognition, but also a bottom-up sensitivity to some of the shape features relevant to word recognition, such as a preference for high-resolution foveal inputs (18, 19), line junctions (20), and midlevel geometrical features (21). Because these properties are widespread in both hemispheres, however, a second hypothesis may be needed to explain the narrow, reproducible location of the VWFA in the depth of the left infero-temporal sulcus. According to the biased-connectivity hypothesis, this left-hemispheric site exhibits a preexisting biased connectivity, or "connectivity fingerprints," with distant language areas (22–27). Indeed, in agreement with this idea, the precise location of the VWFA in 8-y-old readers can be predicted from their long-distance anatomical connectivity to other brain areas at 5 y of age, before they learned to read (27).

In the present work, we assess to what extent a minimal computational model of those two hypotheses may suffice to account for the emergence of the VWFA during reading acquisition. This study complements a recent work (28) that investigates the emergence of letter representation using unsupervised learning. Here, we focus on the learning of words and how their combinations of letters are represented. Specifically, we simulate a deep neural network whose architecture was not designed for reading, but is inspired from that of the ventral visual cortex and which was shown to provide a good fit to both behavioral and electrophysiological observations on face and object recognition (29). We examine what happens when this

Significance

Learning to read results in the formation of a specialized region in the human ventral visual cortex. This region, the visual word form area (VWFA), responds selectively to written words more than to other visual stimuli. However, how neural circuits at this site implement an invariant recognition of written words remains unknown. Here, we show how an artificial neural network initially designed for object recognition can be retrained to recognize words. Once literate, the network develops a sparse neuronal representation of words that replicates several known aspects of the cognitive neuroscience of reading and leads to precise predictions concerning how a small set of neurons implement the orthographic stage of reading acquisition using a compositional neural code.

Author contributions: T.H., A.A., L.C., and S.D. designed research; T.H. and A.A. performed research; T.H., A.A., and S.D. analyzed data; and T.H., A.A., L.C., and S.D. wrote the paper.

Reviewers: M.Z., Universita degli Studi di Padova; and K.P., Haskins Laboratories.

The authors declare no competing interest.

[1]T.H. and A.A. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: stanislas.dehaene@cea.fr.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2104779118/-/DCSupplemental.

Published November 8, 2021.

NEUROSCIENCE

PSYCHOLOGICAL AND COGNITIVE SCIENCES

network, after being trained to identify pictures of generic object categories, is further taught to identify written words, with and without biased connections to output lexical units.

Aims of the Present Study

Our work had two aims. First, we wanted to see if we could reproduce a list of experimentally observed properties of the VWFA, namely:

- Emergence, after training, of a localized patch of neurons specialized for words, as opposed to other stimuli such as faces or objects (11);
- Recycling of units with modest prior involvement for objects and faces prior to reading acquisition (11);
- Invariance for word size, case, and font (3, 6–8) and to rotation up to ~40° and letter spacing up to ~1.5 letter width (30);
- Monotonically increasing response to a hierarchy of letter strings that increasingly approximate the statistics of words in the learned script (31–33);
- Resistance of word recognition to letter transpositions (also known as the Cmabridge University Effect) (34, 35);
- Sudden loss of reading abilities (pure alexia) when this patch of cortex is lesioned (36–38), with preserved recognition of other visual categories.

A failure to capture some of these properties with a simple feed-forward convolutional neural network (CNN) would be interesting, inasmuch as it may point to the need for additional properties—for instance, recurrent and/or top-down connections (12, 14, 33, 39).

Our second goal was to see if we could predict, in anticipation of future experiments, some of the properties of the neural code for written words. It is currently controversial whether neurons in the reading pathway are specialized for whole words (40), frequent pairs of letters ("bigrams") (41, 42), graphemes that map onto phonemes (43), or individual letters at a specific location (44, 45). Indeed, these possibilities are not mutually exclusive, and multiple codes may coexist, perhaps in different pathways, to support different tasks such as comprehension vs. reading aloud (43, 46). While it is currently nearly impossible to visualize single neurons in the reading pathway in humans, this has been achieved in a nonhuman primate monkey model (47), and advances in intracranial recordings may soon make it possible in the human brain (33, 48). Using as few prior hypotheses as possible, we describe how an artificial neural network encodes visual words, in the hope that its predicted tuning curves may soon become empirically testable.

The Model

Architecture. Our model is based on CORnet-Z, the simplest network in the CORnet family (29). These CNNs all share a common design of four spatially organized modules meant to represent the visual areas V1, V2, V4, and inferior temporal cortex (IT) in the ventral visual pathway (topological modules hereafter), capped by a nontopological decoding structure (dense layer in Fig. 1). We chose the simplest among CORnet models because high performance was not the focus of this study and because it achieved a good balance between training time and fit of visual system data (29). We built three variants of CORnet-Z networks. The illiterate network was trained only on the visual images in ImageNet (ImageNet Large Scale Visual Recognition Challenge 2012 dataset; about 1,300,000 image exemplars distributed over 1,000 classes such as dogs, cars, etc.). Two literate networks were additionally trained to recognize words. To this end, we generated 1,300,000 word exemplars distributed over 1,000 "word classes," where each word class actually corresponded to a single word, while exemplars varied in location, size, font, and scale. The literate networks thus had 1,000 additional output units (Fig. 1). In the unbiased literate network, their afferent connections were distributed across all units in the dense layer. In the biased literate network, to simulate the hypothesis of a biased connectivity from a subregion of visual cortex to language areas, we restricted the output units' afferent connections to a subset of 49 units in the dense layer (this number, corresponding to ~10% of dense-layer units, was guided by performance on preliminary simulations with a simplified model). These biased units will be dubbed "Dense+" hereafter (violet units in Fig. 1*B*), in opposition to unbiased units ("Dense−"; white units in Fig. 1*B*). The code for training the networks is available online at https://github.com/THANNAGA/Origins-of-VWFA along with pretrained models for all conditions.

Stimuli. We used the training and test sets from ImageNet, amounting to 1,000 image classes, each with about 1,300 training exemplars and 50 test exemplars. Images were rescaled to 228 × 228 pixels before presentation. In addition to ImageNet, we also generated written word stimuli. All word exemplars were black-on-white images of dimension 228 × 228 pixels. The word set was selected from 1,750 high-frequency French words known by 4-y-olds, as listed by the French Academy of Amiens. Within this list, we randomly selected 1,000 words, whose length ranged between three and eight letters, without accents. Each word was considered as a class on its own, and in order to match the ImageNet dataset, we generated 1,300 training exemplars of each word, varying in size, font, location, and case, and 50 test exemplars. In the training set, character size (scale) varied randomly between 40 and 80 pixels (with steps of 10 pixels, bounds included); fonts were either "Arial" or "Times New Roman," and case was either upper or lower. Words were randomly shifted away from a central presentation, with shifts ranging uniformly from −50 to +50 pixels along the x axis and −30 to +30 pixels along the y axis. Exemplars were generated with a uniform probability for all of these variables. Though single-word fixations are not uniformly distributed during reading, our choice of a small and uniform positional variability is adequate for a CNN with built-in location invariance and without a fovea. On the other hand, the large variability in font, size, and case is consistent with the rich diversity of reading material to which children are exposed in modern societies. In the test set, the 50 exemplars for each word were generated in the same way, except that the fonts were randomly chosen between "Calibri," "Courier," and "Comic Sans." A few example stimuli are shown in Fig. 1*A*.

Training. Deep-learning models typically involve a single training phase on a single large dataset, whereas children learn to identify faces and objects long before they acquire reading. To simulate reading acquisition as the partial recycling of prior visual recognition abilities, we trained the networks in two stages: an "image" phase (phase 1) followed by an "image + word" phase (phase 2). Word units were only added to the output layer of the network during the latter phase. Throughout phases 1 and 2, the networks were trained with Stochastic Gradient Descent on a categorical cross-entropy loss, using a linear learning rate scheduling (*SI Appendix*).

Neural networks are notoriously prone to the phenomenon of catastrophic interference, whereby learning of a new task, or even on a new dataset, can spectacularly deteriorate previously stored knowledge. This well-studied effect can be alleviated in a number of ways, included interleaved learning and the use of dropout (49, 50). We used interleaved learning: During phase 2, output word units were added to the network, and the new training set was obtained by concatenating and shuffling the original generic image set and the new word set. The sets were matched for number of classes and exemplars, and the presentation probabilities were 50% images and 50% words. An illiterate control network was trained only on ImageNet, for as many epochs as the literate networks.

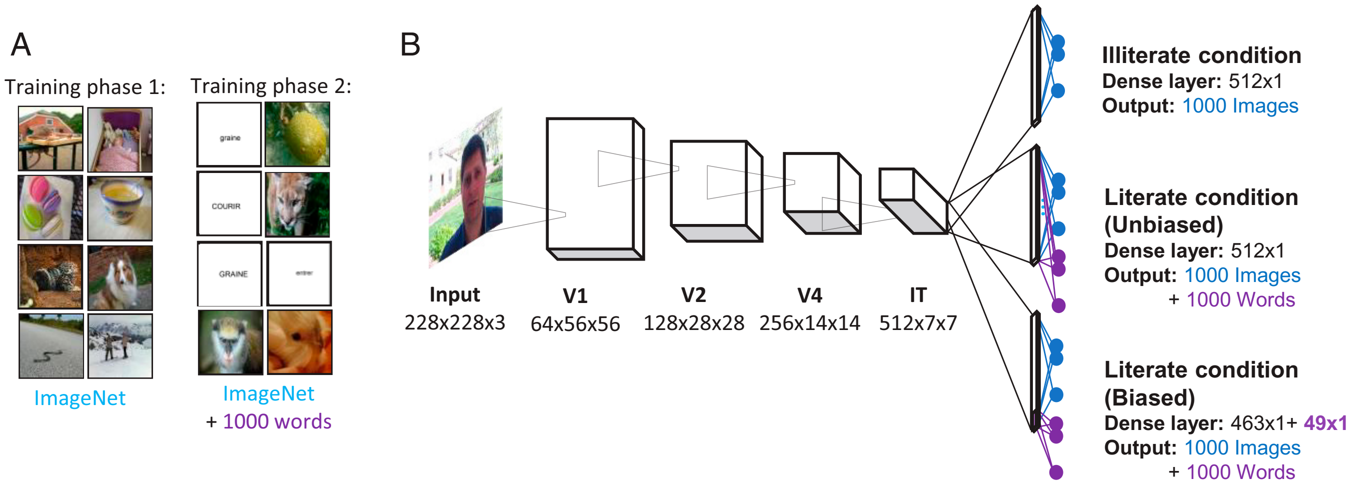

Fig. 1. Network architectures and training schemes. (*A*) Examples of ImageNet and word stimuli illustrating the image sets used in two training phases. (*B*) Testing the biased-connectivity hypothesis by training the CORnet CNN model of the ventral visual pathway under three conditions; 1) illiterate condition: network trained only on ImageNet; 2) unbiased literate condition: network first trained on ImageNet, then on ImageNet and words, with a new set of fully connected output word units; and 3) biased literate condition: network trained on ImageNet, then on ImageNet and words, with biased connectivity between a subset of units in the dense layer, and the output word units.

While it is common to freeze a network's lower-level weights during transfer of learning to new datasets or tasks, here, we trained the full network during the first and second stages, mindful of the fact that an effect of literacy has been reported as early as V1 in the visual pathway (17). Finally, deep-learning models trained on classification usually also augment the training set by performing various visual transforms that do not have an impact on the semantics of the image, such as cropping (which for words had to contain at least 90% of the original input, so as to avoid generating nonwords), resizing, and mirroring. We included such transformations in our training set, but because words are not normally seen in mirror form, and for the sake of a fair comparison of network performance between words and pictures, we excluded mirror transforms for all classes during training.

Results

Training and Recognition Performance. For each model condition and at each epoch of training, we computed the average top-1 accuracy on the test set over five trained networks, i.e., the proportion of correctly classified exemplars in the test set. Fig. 2*A* shows that performance for all networks on ImageNet was on par with earlier reports for CORnet-Z. On the word dataset, the unbiased network converged slightly faster than the biased network. The higher accuracy of the literate network on words than on pictures, despite significant variability in case, font, and size, may be due to the simpler black-on-white layout of the word stimuli. Note that only a small drop of performance on images was seen upon the introduction of words (Fig. 2*A*). Randomly interleaving words with generic image categories during the second phase of training was probably sufficient to avoid interference (even in the absence of dropout, which was not used in our simulations). We observed similar accuracy profiles in the other four runs with new random seeds (*SI Appendix*, Fig. S1).

After training, most of the trained words continued to be recognized when two of their letters were transposed (e.g., "BAGDE"; *SI Appendix*, Fig. S2). The accuracy profile followed an "inverted U" profile as a function of the position of the transposed letters, indicating that transposing the first or the last letter had a greater impact on readability than transposing the inner letters (*SI Appendix*, Fig. S2). Performance was higher for letter transpositions than for substitutions of the same letters. These results were in agreement with experimental observations on letter transpositions (34, 35). Interestingly, similar findings were obtained in baboons trained to perform lexical decision on English words (51), confirming that they arise at a purely orthographic level, as in the present CNN, without requiring additional semantic or phonological influences. The decoding analysis across all layers further suggests that the transposed-letter effect arises in the initial layers of the network and is independent of literacy (*SI Appendix*, Fig. S5).

Changes in Representational Spaces. To measure whether and how literacy changed the networks' representations for images and words, we calculated the mean vector of activation evoked by 80 randomly sampled pictures and words within each layer of the models. We then computed their Representation Dissimilarity Matrix (RDM), which characterizes the structure of internal representations, independently of the specific units activated (52, 53). Finally, we examined how the RDMs resembled each other across the three different networks, an approach known as "second-order isomorphism" (53). As can be seen in Fig. 2*B*, for objects, the RDMs were extremely intercorrelated, with r approaching 1 for the V1, V2, and V4 layers and only a slight drop for the IT, dense, and output layers. Thus, the image-representational space was very similar in all networks and remained largely preserved, even after a network was trained for words (mimicking functional MRI [fMRI] representational similarity results from ref. 11). For words, however, the representational changes were dramatic in the late layers of the literate networks. Both biased and unbiased literate networks converged to highly similar representation space (r approaching 0.85), and this representational space was significantly different from that of the illiterate network ($r = 0.16$ and 0.15, respectively, with unbiased and biased networks in the output layer; resampling test across 10,000 bootstraps, $P < 0.0005$). The change in representational similarity induced by literacy was already detectable in the V4 layer ($r = 0.99$, 0.91, and 0.91 for correlation between unbiased–biased, biased–illiterate, and unbiased–illiterate, respectively), but not V1 or V2—in partial disagreement with fMRI findings that V1–V2, rather than V4, hosts detectable literacy-induced changes in readers of alphabetic languages (refs. 5, 17, and 54; although see also ref. 55).

To visualize those internal changes, we plotted the activations evoked by various word stimuli in the subspace of the first three principal components of word-evoked activity, within each of the three top layers (Fig. 3; although this figure shows only one literate network, i.e., unbiased, very similar results were found for the

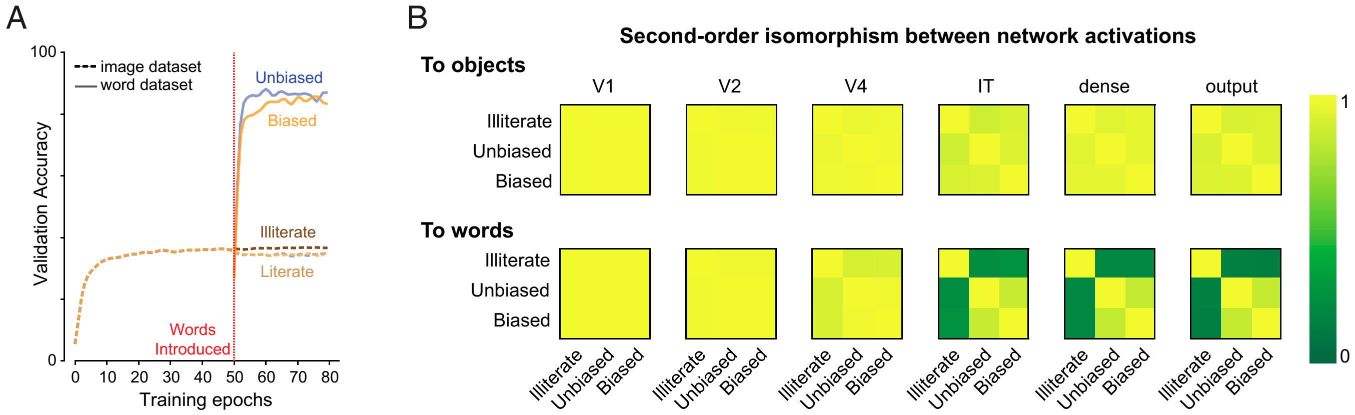

Fig. 2. Changes in performance and internal representations during training. (*A*) Changes in performance: Average top-1 accuracy in the illiterate, unbiased literate, and biased literate networks on the test set, separately for images and words. The introduction of words at epoch 50 (marked by a dashed red vertical line) leads to a sudden increase in word-recognition performance, with only a very slight decrease for images. (*B*) Evidence for changes in word, but not image, representational spaces. For each layer (V1, V2, V4, IT, dense, and output), the figure shows the pairwise correlation [also known as second-order isomorphism (53)] of the RDMs, which characterize the representational space for images (*B*, *Upper*) and words (*B*, *Lower*).

biased network). This plot revealed how the acquisition of literacy led to the emergence of an invariant neural representation of words, with a decrease in the impact of physical parameters (location and size) and an increase in the impact of reading-relevant parameters (word length and word identity). In the illiterate network, activity was primarily segregated by the physical parameters of the words. Retinal location (four quadrants) was strongly segregated in the IT and remained separated into top and bottom stimuli in the dense and output layers. In the literate networks, word location had a much smaller impact in the IT and became indistinguishable in the dense and output layers. A similar change was seen for physical size, which had a massive impact on all layers of the illiterate network, but not the literate ones. Conversely, with literacy, activation became systematically organized according to word length in all three layers. Most importantly, the activation vectors became increasingly segregated by word identity, irrespective of their location and size (Fig. 3). Only in the literate networks, the responses to a given word were all clustered together, irrespective of those variations, with this invariance increasing from the IT to the dense and output layers. The decoding accuracy estimated from each layer confirmed these observations (*SI Appendix*, Fig. S5).

Single-Unit Analysis of Invariance for Size, Font, and Case. We next asked whether we could trace those changes down to the single-unit level. Skilled readers can recognize words effortlessly across changes in size, font, case, and location. At the single-unit level, could we see a trace of those invariances and how they build up across the successive layers? For each unit in the network, and each of three transforms—size, font, and case—we collected two vectors of activations. These vectors were 1,000-dimensional and recorded the activation evoked in the same unit by each of the 1,000 words in the training set, with each word being presented in two distinct, randomly chosen values of the transform under consideration (e.g., two different sizes). We discarded changes in location, as location invariance is essentially imposed in convolutional networks by the hard-wired mechanism of weight sharing, but each word location was randomly sampled. We then computed

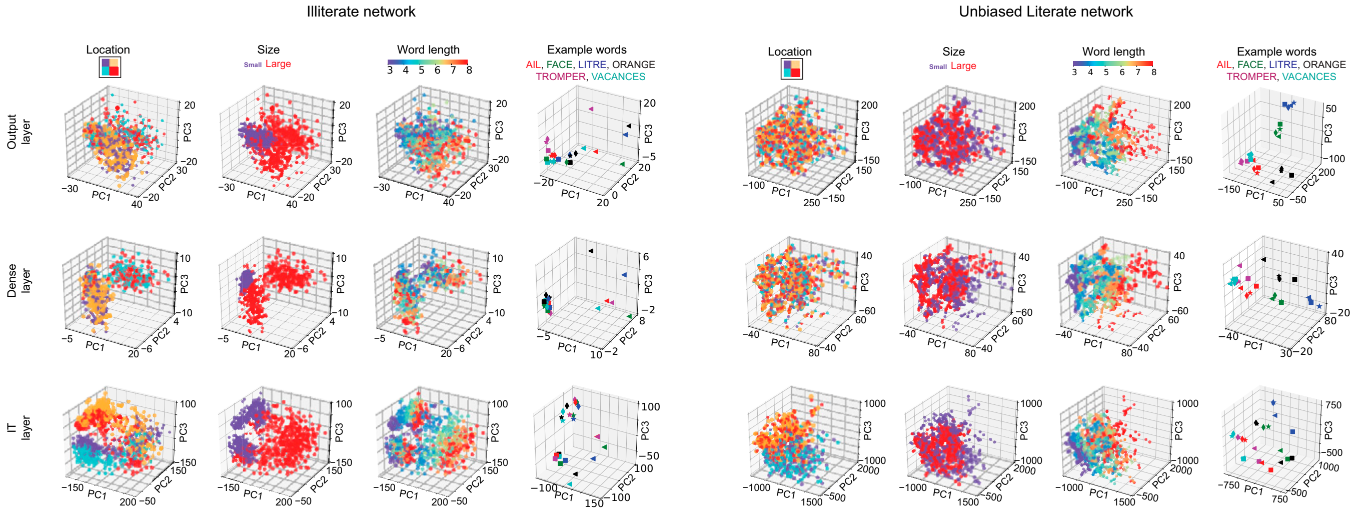

Fig. 3. Principal component (PC) visualization of word representations. The figure shows the activity evoked by various word stimuli in the top three layers of the illiterate and unbiased literate networks (the biased literate results were similar). Dimension reduction was achieved by extracting the first three principal components in a Principal Component Analysis of the activation evoked by 960 stimuli (120 words × 4 positions × 2 sizes). The location of the words varied from −30 to +30 pixels along the horizontal axis and −15 to +15 pixels along the vertical axis, thereby spanning all the four quadrants (first column). Their size varied by a factor of two (Arial font at size 40 and 80 points; second column). Their length varied from three to eight characters (third column). The fourth column shows the activity evoked by six example words, which could be moved rightward (◆), moved downward (◀), or scaled in size (★) from the reference image (■).

the Pearson correlation coefficient between the two vectors for that transform (see ref. 56 for a related invariance index). The correlation should be close to one if and only if the unit exhibits a selectivity profile, across the 1,000 words, which does not vary across changes in size, font, or case. Repeating this operation for all units within a given network layer yielded a distribution of correlation coefficients (Fig. 4*A*). As a global invariance index for an entire layer, we used the medians of those distribution (Fig. 4*B*).

For simplicity, we refer to "pre-literate" as the network prior to epoch 50, before the possible introduction of words, and reserve the term "illiterate" to the network solely trained on images for all 80 epochs. In the preliterate ("pre") and illiterate ("illit") networks (blue and orange curves in Fig. 4), i.e., in the absence of any training with words, the correlation distributions revealed only a modest invariance to word size and font, but not to case. This can be seen in Fig. 4*A* (top layer), where for both networks, distributions were already significantly shifted toward positive values for size (one-sampled, two-sided t tests performed against zero; $mean_{pre}^{size} = 0.258$, $t_{pre}^{size} = 51.289$, $p_{pre}^{size} \ll 0.01$; $mean_{illit}^{size} = 0.197$, $t_{illit}^{size} = 36.776$, $p_{illit}^{size} \ll 0.01$) and for font ($mean_{pre}^{font} = 0.388$, $t_{pre}^{font} = 84.289$, $p_{pre}^{font} \ll 0.01$; $mean_{illit}^{font} = 0.363$, $t_{illit}^{font} = 85.762$, $p_{illit}^{font} \ll 0.01$), but not different from zero or even shifted toward negative values for case ($mean_{pre}^{case} = 0.009$, $t_{pre}^{case} = 2.083$, $p_{pre}^{case} = 0.038$; $mean_{illit}^{case} = -0.058$, $t_{illit}^{case} = -11.971$, $p_{illit}^{case} \ll 0.01$). This finding shows that the tolerance to size and to systematic changes in shape that were acquired for the purpose of image classification generalized in part to the novel task of processing words. Contrariwise, letter mappings across uppercases and lowercases are largely arbitrary (e.g., A and a, E and e) and therefore have to be explicitly learned during reading acquisition (8).

This initial invariance, however, was dramatically enhanced once the networks were trained to recognize words. In both biased and unbiased literate networks (red and green curves in Fig. 4), invariance for size and font started to rise in the IT, while invariance for case only appeared in the dense layer. As could be expected, for all networks and transforms, invariance always culminated at the output level, where unit responses for a word and its transform became highly invariant (correlation coefficient r close to ∼0.9). Strikingly, however, that invariance for all transforms was much stronger in the Dense+ units of the biased network compared to the unbiased network (Fig. 4*B*). This observation shows that biased connectivity promotes the emergence of a restricted cluster of invariant units, mimicking the VWFA (3, 6–8), before the final output stage of the network. While this was clearest in the dense layer, the long tail of the distribution for IT units in Fig. 4*B* indicates that many IT units also acquired some degree of invariance, particularly for size and font, in the course of word-recognition training.

In *SI Appendix*, Figs. S3 and S4, we show how this invariance partially extends to variations in word rotation and letter spacing, which were not present in the training dataset. Briefly, word recognition remained successful up to a rotation of ∼10° and a spacing of ∼0.75 letter width, which is nonnegligible, yet lower than the experimental results (30).

Characterization of Word-Selective Units in the Dense Layer. A typical feature of the VWFA is its activation preference for printed words over other classes of images. Therefore, we next asked whether some units in the dense layer of each network exhibited such word selectivity. A unit was deemed word-selective when its average response to word stimuli was 3 SDs above its responses to faces, bodies, houses, and tools (see *SI Appendix* for details). As a replication set, we also probed these units' responses to a hierarchy of increasingly word-like stimuli, similar to Vinckier et al. (31) (*SI Appendix*, Fig. S6).

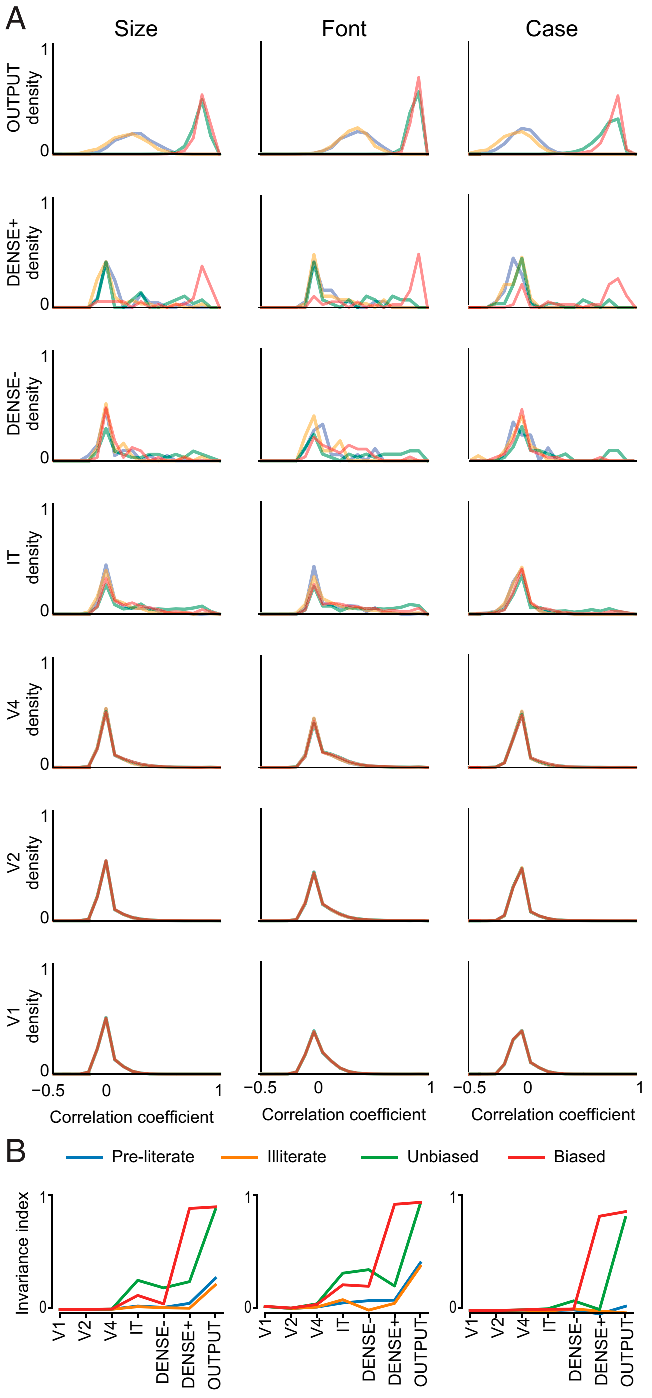

Fig. 4. Invariance for size, font, and case at the single-unit level. (*A*) Distribution, over all the units of a given layer, of the invariance indices for changes in size, font, and case (columns) and for different networks (preliterate in blue, illiterate in orange, unbiased in green, and biased in red). Each curve shows the distribution, over units in a given layer and network, of the mean correlation coefficient between the activities evoked by two different versions of the word (e.g., font A vs. font B). A correlation of zero indicates no invariance. The emergence of a bump near one in the top layers of the literate networks (green and red curves) indicates that many units became both word-selective and invariant over irrelevant stimulus variations. Silent units, i.e., units that did not respond to any of the presented stimuli, were excluded from the distribution. (*B*) Median invariance index (median of the distribution of correlation coefficients) across the hierarchy of layers for the four different networks.

In the preliterate and illiterate networks, we found only a few word-selective units (respectively, five and six units; Fig. 5, *Upper*). Even though those units responded much more strongly to words than to other pictures, their average profile of selectivity did not discriminate between false fonts, infrequent letters, frequent letters, frequent bigrams, frequent quadrigrams, and words. We thus interpret these units as being sensitive to the horizontal shape of words and/or to high-contrast line intersections, either as a result of training or by mere chance initialization of network weights.

Word-selective units were much more numerous in the literate networks (33 units in the biased network, including 29 among the Dense+ units; and 68 units in the unbiased network; Fig. 5, *Lower*). Those units responded to false fonts and infrequent letters, but showed markedly higher responses to all stimuli that contained frequent letters. This indicates that word-selective units in the dense layer became partially selective for the learned letters, rather than for general word shape or subletter features. We observed similar results upon estimating the mean response of all the units in a given network and layer (*SI Appendix*, Fig. S7). While the networks that were trained to recognize words showed increased activations to written words and to word-like stimuli in most layers, the preliterate and illiterate networks failed to respond strongly to strings starting from layer IT (*SI Appendix*, Fig. S7).

We next examined what was the function of those units prior to the acquisition of visual word recognition. Prior to the introduction of words in the training set, word-selective units were overwhelmingly uncommitted. In the two literate networks, when returning to the preliterate stage (where the network was solely trained to classify pictures), these units showed no selectivity toward any of the tested categories and exhibited an overall low response (*SI Appendix*, Fig. S8). This finding mirrors the longitudinal fMRI findings of Dehaene-Lambertz et al. (11), although in this study, the word-selective units did tend to be initially weakly responsive to tools.

We next wondered if these word-selective units exhibit a length effect, i.e., greater average activity for longer words. To address this question in our networks, we computed the activation evoked in word-selective units of the dense layer, for words ranging from three to eight letters. We then evaluated the Spearman rank correlation between length and mean activation. *SI Appendix*, Fig. S9 shows that, on average, the activation of word-selective units in the dense layer of the network increased monotonically with word length. This was true for most units: Only a few units showed no variation or, very rarely, a negative trend (*SI Appendix*, Fig. S9).

Finally, we investigated the causal role of word-selective units in word recognition through focused lesioning i.e., simulating alexia. Interestingly, we found that removing 20% of units in literate networks sufficed to produce a complete impairment on words (*SI Appendix*, Fig. S10). This effect was specific to the word-selective units, and not the lesions of a subset of units such as Dense+ units (or "VWFA units").

A Sparse Neural Code for Words. We next analyzed the nature of the neural code that allowed those selective units to encode 1,000 words. The first question we asked was to what extent the neural code for words is effectively sparse, using only a subset of active units for a given word, or is distributed across all word-selective units. To this aim, we performed selective silencing experiments.

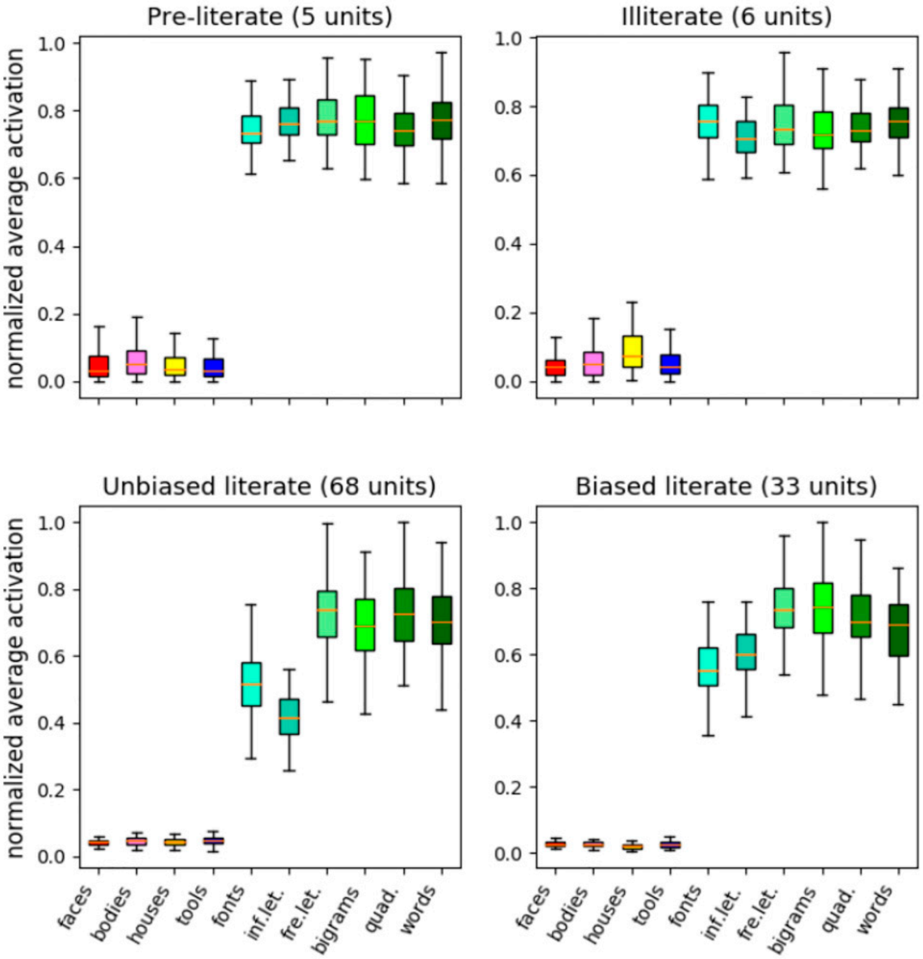

Fig. 5. Emerging selectivity for words and word-like stimuli at the single-unit level. Each panel shows the mean activation profile of word-selective units in the four networks, in response to pictures (faces, bodies, houses, and tools) and to the Vinckier et al. hierarchy of word and word-like stimuli. Preliterate and illiterate networks contained very few string-selective units, and those did not exhibit any differential sensitivity to letter status or to frequency. Word-selective units were more numerous in the two literate networks and showed markedly higher responses to stimuli that contain frequent letters.

We started by computing, for each unit, its distribution of activity across the 1,000 words. Considering the activity pattern produced in all word-selective units by a given word, we then set a unit to zero when its activity for this word fell below a threshold percentage of the unit's global activity distribution and examined the effect on word-recognition performance (solid curve marked "Lowest" in Fig. 6*A*). The results showed that it was possible to silence the word-selective units whose response fell in the bottom half of their global distribution without compromising at all the-word recognition performance of the networks. Silencing in the opposite order, however, starting with the most responsive units (dashed curve marked "Highest" in Fig. 6*A*), had a dramatically different impact. Silencing just the single unit whose response were the strongest resulted in a dramatic drop in performance. This analysis shows that the neural code for words is sparse: Not only is it based on a small number of selective units (68 for the unbiased network and 29 for the biased one), but the strongest responses of these units are absolutely necessary for word recognition, while only the top ∼40% strongest responses are sufficient.

In order to examine the patterns of activity for each word, we silenced the activity of each unit below the median of its activation distribution across all words. This procedure resulted in a sparse "bar code" purified of its unnecessary activity, and yet sufficient to recognize each word (see examples in Fig. 6*B*). Sparsity, defined as the average proportion of units with null activity, was very high (86.7%) in the unbiased literate network, with an average of 9 out of 68 units activated by a given word. Thus, even in the absence of dropout, which is known to produce sparser codes (50), word representations in the networks were effectively sparse. We observed similar effects in the biased literate networks (*SI Appendix*, Fig. S11).

Modeling the Receptive Fields of Word-Selective Units. A simple hypothesis for the effect of word length is that most units respond to critical features, such as letters at a certain location [a bank of letter detectors, which is the front end in most models of word recognition (57–60)] or bigrams and other letter combinations (41, 42, 46, 61). Longer words would then have a greater likelihood of activating a greater number of such detectors.

To test the letter-coding hypothesis, we attempted to model the response of each unit across words as a linear combination of 26 × 8 features, i.e., its constituent letters (26 possibilities) at each of a maximum of 8 possible ordinal positions. To capture the importance of edge letters in reading, our model adopted an end-coding approach, where ordinal letter positions were assigned relative to the exterior positions, with a leftward bias (e.g., in a five-letter word, the initial three letters are assigned positions 1 to 3, while the last two letters are assigned positions 7 and 8) (Fig. 7*D*). Given the large number of features, we performed cross-validated regularized linear regression (least absolute shrinkage and selection operator [LASSO]) to generate a sparse estimate of features that activate a given neuron. The response of each unit was measured along 8,000 stimuli, i.e., each of the 1,000 words in the training set (3 to 8 characters long) presented at 4 different locations and 2 different sizes, while the model matrix comprised $26 \times 8 = 208$ features. While this number may seem large, note that 1) it is relatively small compared to the 8,000 data points of each unit; and 2) the conservative cross-validated LASSO regression attributed null weights to many of those regressors. The LASSO regularization constant was estimated by using fivefold cross-validation.

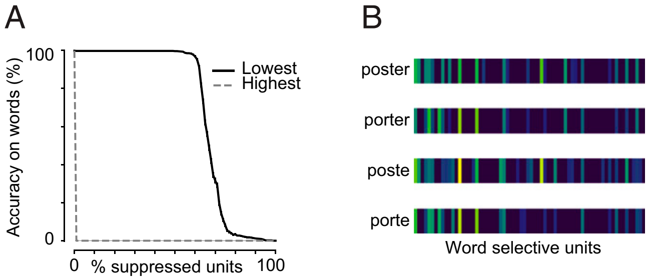

Fig. 6. Characterizing the neural code for words for unbiased network. (*A*) Word-classification performance as a function of activity cutoffs in word-selective units (normalized by performance score without cutoff). Cancelling the bottom-half responses of each unit leaves classification performance intact (black solid curve), while cancelling the top 1% responses removes all word-classification ability (dashed gray curve). (*B*) Example of the sparse neural bar codes sufficient to encode a given word, obtained by keeping only the activity above the median of each unit's activation distribution.

The letter X position model was able to explain a large and significant proportion of the response variability in word-selective units (Fig. 7*A*). Interestingly, the fits for the end-coding model were significantly better than for a purely sequential (i.e., left-to-right), a word-centered (i.e., where letter position is coded relative to word center and only longer strings occupy exterior positions), or an overlapping end-coding model (see *SI Appendix*, Fig. S12 for details). Fig. 7*D* shows some of the reconstructed receptive fields. The units with highest letter-model fits were either selective to a single letter irrespective of its position (e.g., unit no. 282 responded to S at all positions) or were selective to a single or a few letters at a certain position (e.g., unit no. 23 responded primarily to curved letters at the first position and unit no. 53 only to the last position). Some units were ultraselective, responding exclusively to a single letter at a specific location (e.g., unit no. 142 responded to the letter M at the word beginning; *SI Appendix*, Fig. S13). Such narrow selectivity was not seen inside words, however, where most units were selective for certain letters over a range of nearby positions (e.g., unit no. 146 responded to letter O anywhere in the middle of the word). Most units were positively activated by certain letters and negatively activated (i.e., inhibited) by others, in what could be termed a "letter-dipole" configuration (e.g., unit no. 38 encoded "E but not R towards the end of the word"). In some cases, the preferred letters occupied distinct locations (e.g., unit no. 21 seemed to care for letter E at a location left of letter R). In most cases, however, a unit was sensitive (positively or negatively) to letters at roughly the same location, suggesting a factorial code where a product of independent preferences for location and letter determined the unit's response, as proposed in some theories (62, 63) and as observed in neural recordings of monkey inferotemporal cortex (47). Overall, more than half of word-selective units had a broad position-tuning profile, while others exhibited a sharp selectivity toward a specific ordinal position relative to the left or right side of the word (*SI Appendix*, Fig. S14).

Evaluating Letter vs. Bigram Codes. We verified that this letter X position model could predict the neural responses to novel stimuli forming a hierarchical stimulus set, i.e., 300 six-letter strings with infrequent letters, frequent letters, frequent bigrams, frequent quadrigrams, and words (Fig. 7*B*; false fonts were also included as a control for overfitting, since each letter was substituted one-by-one by a roughly similar nonletter shape). Interestingly, we found that the model fit gradually increased from false fonts to word-like stimuli. The increase from high-frequency letters to bigrams, quadrigrams, and words contrasts with, but is not incompatible with, our previous observation that the mean unit activity failed to increase across the hierarchy. It merely suggests that the units became tuned to stimuli that were finely adapted to the statistics of real words. Indeed, Fig. 7*A* is incompatible with the hypothesis that the model units are solely responsive to individual letters: they must also be sensitive to letter combinations.

Indeed, while the proposed letter model was a good overall predictor of the response to words (mean $R^2 = 0.66$), the model fits were poor for a few word-selective units, which exhibited a mixed and complex letter X position selectivity (Fig. 7*D*). We

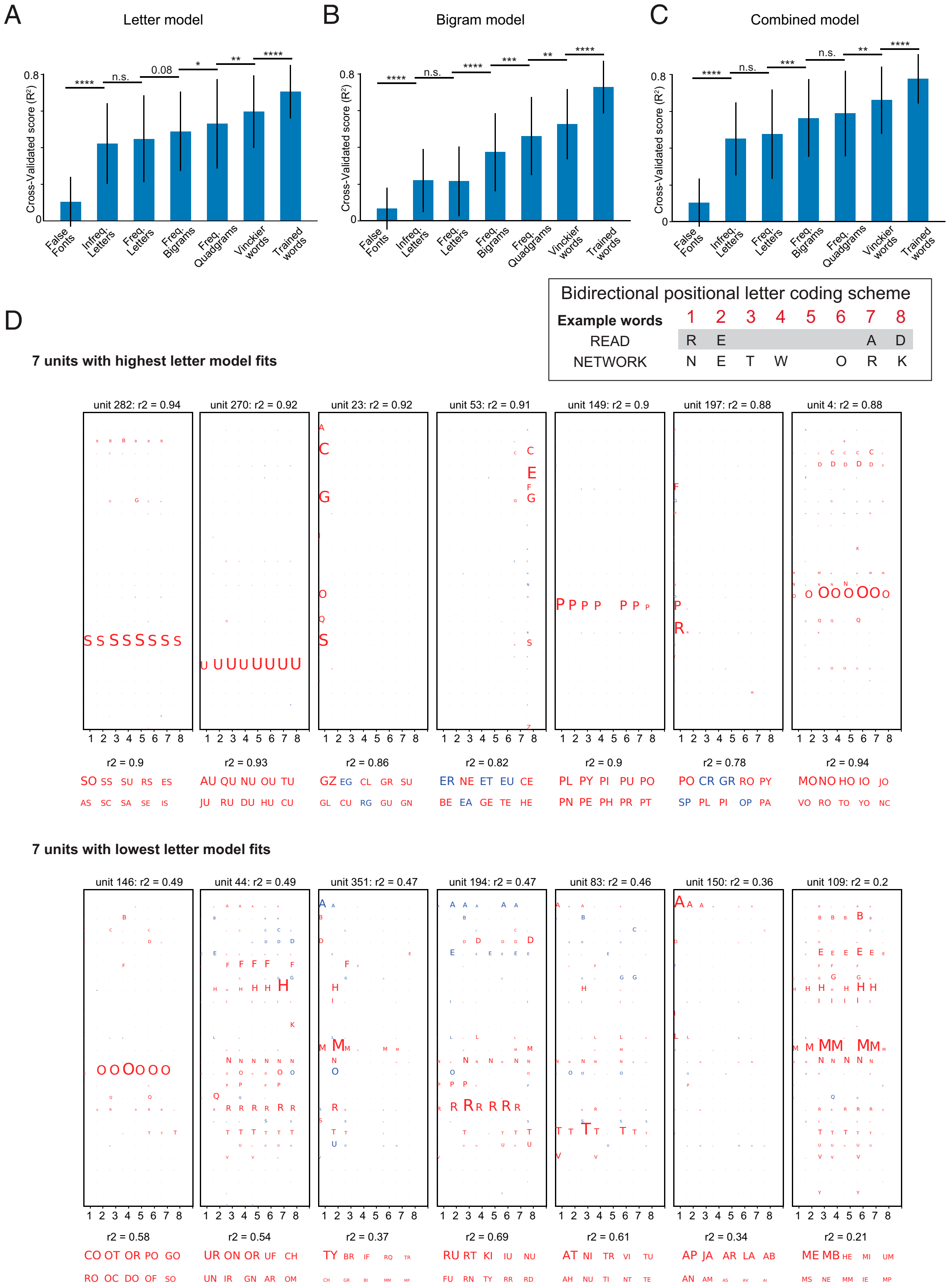

Fig. 7. Modeling the receptive fields of each unit. (*A*) Average fits of the ordinal letter-position model across different categories of stimuli (false fonts, infrequent [infreq.] letters, frequent [freq.] letters, frequent bigrams, frequent quadrigrams, and Vinckier words) when trained on words. Asterisks indicate statistical significance. $^{*}P < 0.05$; $^{**}P < 0.005$; $^{***}P < 0.0005$; $^{****}P < 0.00005$. N.s., not significant. (*B*) Average fits for the bigram model. (*C*) Average fits for the combined model with letter and bigram features. (*D*, *Upper*) Examples of reconstructed letter-based receptive fields of the seven word-selective units with the highest letter-model fits, from the nonbiased literate network. Each matrix shows, for a given unit, which combinations of letters (vertical dimension, 26 levels) x position (horizontal dimension, 8 levels) were significant predictors of the unit's response to words. LASSO regression set many coefficients to zero. Nonnull coefficients are indicated by the size of the corresponding letter. Positive coefficients are indicated in red, and negative ones in blue. Below each panel are bigram model fits for the same units, along with the 10 bigrams with the highest regression weights. Weight magnitude is indicated by letter size, and sign is indicated by colors (red = positive, blue = negative). (*D*, *Lower*) Examples of reconstructed letter-based receptive fields for the seven word-selective units with the lowest letter-model fits.

wondered if their responses would be better predicted using letter bigrams as regressors. To test this idea, we first modeled the word responses as a linear combination of responses to frequent bigrams. The 1,000-word training set comprised 271 unique bigrams (pairs of side-by-side letters, regardless of their position), which were used as features in a second linear regression. Interestingly, the bigram model fits (mean $R^2 = 0.73$) were significantly larger than those of the letter model ($P < 0.00005$, paired t test). However, the better model fits could be a consequence of the greater number of degrees of freedom of the bigram model. Indeed, according to the corrected Akaike information criterion (AICc), both model fits were comparable, and their AICc values did not differ significantly ($P = 0.26$, paired t test). For each unit, Fig. 7*D* shows the bigram that had the highest model coefficients. This analysis often confirmed the letter-based model's conclusions. For instance, unit no. 282, which responded to letter S at any position, was fitted in the bigram model by approximately equal weights for all frequent bigrams containing that letter (SO, SU, SE, ES, etc.). Thus, the most economical description is that this unit simply responds to letter S. However, for units with poor letter-model fits (Fig. 7*D*), the bigram model appeared much more satisfactory, with some units responding strongly to a single bigram (e.g., unit no. 83, bigram AT) or to two of them (e.g., unit no. 109, bigrams ME and MB). Thus, some units behaved as approximate bigram detectors, as postulated, for instance, in the local combination detector model of word recognition (41).

Next, we asked whether these two models explained distinct variance in the word responses. If true, then a combination of the features from both models ($n = 26 \times 8 + 271 = 479$) should yield higher fits compared to either of the individual models. This was indeed true: The observed average model fit ($R^2 = 0.77$) was significantly higher than either of the two models, and, crucially, its average AICc value was also significantly lower ($P < 0.00005$, paired t test).

In summary, although the most parsimonious letter-based model explained a large amount of variance across many word-selective units, the bigram model led to equivalent results, and, most crucially, the best fit was obtained by combining both regressors. Those findings suggest that, with literacy, the networks became attuned to letters and their ordinal position, but also the statistics of their cooccurrence in the trained orthographic system.

Discussion

Our goal was to develop and assess a minimalist model of how the VOT cortex changes during reading acquisition, through the mere recycling of a biologically plausible convolutional network model of object recognition, without introducing ad hoc reading-specific constraints. As occurs in children, a standard CNN was first trained to identify pictures of various objects and scenes and then a set of 1,000 words of different lengths across variations in location, size, font, and case. Furthermore, we tested the biased-connectivity hypothesis, according to which, in literate humans, only part of the VOT cortex becomes specialized for orthographic coding because of its privileged output connections to language areas. To this end, we compared networks whose dense layer was either fully connected to all output units or in which only a subset of dense units were connected to the output layer, simulating a putative VWFA (Fig. 1).

Behaviorally, we found that a network designed for image recognition could easily learn to recognize 1,000 written words. Both biased and unbiased networks reached an accuracy of 80% or more after very few training epochs (Fig. 2*A*). Critically, accurate recognition of abstract word identity occurred across large variations in the physical features of stimuli, particularly across case, thus approximating the perceptual invariance developed by expert human readers. However, invariance for rotation and letter spacing (which were absent in the training set) remained smaller than in humans, suggesting that its simulation may require exposure to a more extended training set. Importantly, in all cases, the performance of biased and unbiased networks was similar, suggesting that the development of an anatomically focal VWFA does not present a functional advantage per se.

The acquisition of reading induced only a barely perceptible deterioration in the previously trained object recognition [note that catastrophic interference (64) was avoided by mixing pictures and words in the training set]. This result is compatible with the fact that general object-perception abilities differ only minimally between illiterate and literate adults (17, 65) and that deficits in object recognition are not typically observed during the acquisition of reading in humans. At the neural level, we found that reading acquisition encroached upon initially uncommitted units that exhibited an overall low response to pictures (*SI Appendix*, Fig. S8). This finding is compatible with longitudinal fMRI data showing that, in the first year of reading acquisition, the VWFA emerges at a site with little or no response to objects or faces (11). It nuances the "direct competition" or "pruning" view of neuronal recycling (66): The acquisition of reading probably does not compete with other visual recognition abilities, such as face recognition, by dislodging them from their locations, but by occupying nearby neural sites that cease to be available for the further growth of face- and object-selective regions, thus forcing them to develop elsewhere [for instance, in the right hemisphere (11, 17)]. Here, we concentrated our analyses on the growth of word responses, but, in the future, it would be interesting to study whether this growth did indeed compete with the network's abilities, either to dedicate units to additional pictures such as faces (11, 17) or to accurately recognize them across mirror inversions (55, 67, 68), thus mimicking the small negative downsides of literacy that have been reported in the literature.

At the neural level, we found that reading acquisition led to the structuring of a distinct neural space for words (Fig. 3). In its top three layers, corresponding to the mid to anterior IT cortex, the network developed critical features of human reading, including invariance for physical location and size, sensitivity to word length, and a segregation of responses according to abstract word identity. Small representational changes were also seen in area V4, but the simulations did not reproduce the reading-related changes in the early visual cortex that have been observed in human fMRI when contrasting literate and illiterate subjects (17) or known letters vs. control stimuli (5, 20, 54). It is possible that those responses reflect top-down inputs from higher-level areas (33, 55), which were not simulated in the present, purely feedforward network. However, it is also possible that the convolutional structure of the network, with its automatic duplication of weights across image locations, prevented the emergence of a reading-related retinotopic specialization, as observed in refs. 17 and 54. The convolution hypothesis is a massive simplification adopted for computational efficiency, and computational resources prevented us from simulating a nonconvolutional network, which might be necessary to capture such fine-grained effects.

At the single-unit level, perceptual invariance emerged from the development, in the top layers of literate networks, of a restricted set of word-selective units whose activation profiles were strongly invariant across changes in size, font, and case (Fig. 4). The most remarkable finding is how compact this representation was. Only 68 word-selective units emerged in the unbiased network, and only 33 in the biased network—and in the latter, only 29 belonged to the Dense+ units and could therefore play a causal role in word recognition. Of course, in real life, each of these dimensions would likely be associated with an entire column of neurons, rather a single unit. Still, the results show that a 29-dimensional space (or even less, assuming that further dimension reduction could be achieved) suffices to recognize 1,000 words. While surprising, this number is on a par with the 50-dimensional space that demonstrably suffices to encode faces at the neural level (69). Intuitively, the statistics of letters are highly redundant (70), and, conversely, a distributed code with 29 dimensions has a large combinatorial capacity (since $2^{29} \sim 536$ million), even when taking into consideration the need for robustness. In fact, we found that, for any given word, the number of required units was even smaller. This is because each word did not evoke a fully distributed code over all word-specific units, but a sparser code: For any given word, silencing the ~60% of word-selective units most weakly activated by this word did not impair identification. Thus, on average, less than 30 active units (for the unbiased network) or 15 units (for the biased network) sufficed to uniquely specify any of the 1,000 trained words.

Fig. 7 represents our best attempt to characterize these dimensions of word encoding, i.e., the receptive fields of the word-selective units. Comparison of various letter-coding models revealed that the neural code for the ordinal position of letters in a word was best described as originating from edge positions. We found that most units are sensitive to the presence of one or a small set of letters at a given position, sometimes with a contrast or letter dipole (e.g., E but not R). However, we also found that the units' receptive field could not be solely described by a sum of responses to individual letters: A better fit was achieved by assuming a sensitivity to neighboring letters, i.e., bigrams. Such a bigram code may be tied to the use of convolutional architecture and/or supervised training, as one previous computational study failed to find coding for letter combinations in a deep generative model trained without supervision (71). The empirical data are conflicting: There is behavioral and brain-imaging evidence that bigrams are a crucial cue to word identity (31, 32, 41, 42, 46, 72), but also recent data suggesting that the bulk of bottom-up orthographic coding may be based on a conjunction of single letters and their positions (33, 44, 45). The present results reconcile both, as they suggest that, in the course of learning, a neural network will make use of all available statistical cues and will develop both letter-by-position codes and bigram-sensitive units.

With respect to the biased connectivity hypothesis, we found that biased and unbiased networks developed similar representations (e.g., Fig. 2). The main difference was that the biased network developed a more compact representation, with twice fewer word-specific units than in the unbiased network, despite equivalent overall word-recognition performance. This compacity came at the expense of a greater sensitivity to focal lesions. By silencing about 20% of the word-specific units, as defined by their category specialization, we made literate networks completely unable to read, thus simulating the main features of pure alexia. The unbiased network was more resilient to small lesions, thanks to more diffuse coding over a larger number of units. Indeed, only the biased network could explain how a focal lesion (restricted to the Dense+ units, corresponding to the VWFA) could yield a complete loss of reading abilities. In the unbiased network, the word-coding units were dispersed haphazardly, such that it is hard to see how they could be targeted by a single lesion. Nevertheless, we acknowledge that our model does not have a notion of cortical topography. If additional assumptions were added, such that neighboring units tended to respond to similar features, as in Kohonen networks and more recent work (73), then even the unbiased network may end up acquiring a functionally localized VWFA. Thus, the biased connectivity hypothesis is neither falsified nor settled by this study, and adding topography to CNNs (73) would be an interesting future project.

Limits of the Present Model

We now address some of the limits of our modeling approach. A first concern is the extreme neurobiological oversimplification of the network architecture we used. Although CNNs may be construed as simplified models of the visual cortex, they lack many properties such as spiking dynamics, neuron subtypes, cortical layers, intricate local connectivity within or between layers and areas, receptor types and densities, temporal delays, realistic learning rules, etc. By construction, the absence of topology in their upper layers also means that none of these models can reproduce the regular topographical organization of category-specific areas in the VOT cortex and the emergence of the VWFA at a fixed location relative to those functional landmarks (11, 17, 19). For the same reason, the left-lateralization of the VWFA falls outside the scope of the model, as it does not distinguish between left and right hemispheres.

These limitations of CNNs as models of the visual cortex are severe, and yet, astonishingly, it is now well documented that when trained on a large number of categories, not only can CNNs reach human-level performance on visual classification tasks, but their single-unit responses can also predict neural, fMRI, and magnetoencephalography activation patterns in the visual cortex of human and nonhuman primates (39, 73–78). Our choice of CNN architectures for modeling VOT cortex is thus conservative, and since the present models already mimic several known properties of the reading system (e.g., invariance for case, word -length effect, pure alexia, etc), we may hope that our networks' predictions may fit the activity evoked by written words as well as others did for faces or objects (39, 73–78).

Another critique concerns the training set. The ImageNet picture set that we used was never intended to mimic the statistics of the images that a young child is confronted with. Similarly, the printed words we used as stimuli may not mimic the progressive introduction of digits, letters, and words, possibly handwritten, that a child receives in the family and the school. In the future, it may be important to include more realistic training sets (e.g., ref. 75) or to allow for unsupervised learning episodes. While introducing more realistic developmental stages, such as an initial letter-by-letter reading stage, may allow us to better capture the interaction of word length and age, it is unclear how this would change the representations of these categories and whether the selectivity analyses that we carried out in the present study would come out differently.

The CORnet-Z network that we used is known to provide a good fit to neural activity despite being relatively shallow, entirely feed-forward, and trained only on ImageNet (29). Nevertheless, a missing ingredient is lateral and recurrent connections (79, 80), which play a critical role in other connectionist models of visual word recognition, as well as in more recent generative (71) or predictive coding accounts (81), and are also omnipresent in the primate visual system. Several simulations have shown that recurrent neural networks provide better predictors of activity in the human visual system than purely feed-forward networks (39, 75, 82). Intracranial recordings suggest that recurrent connections may be particularly important in order to simulate the late part of the neural response to words (33), which is likely to contribute massively to fMRI blood-oxygen-level-dependent signals. In this respect, it is worth mentioning that the present model failed to fully capture an important fMRI dataset, by Vinckier et al. (31), who found a progressive and continuous increase in fMRI activity for stimuli approaching the statistics of real words. Our simulations only captured the difference between stimuli with or without frequent letters (*SI Appendix*, Fig. S7), which fits the initial, but not the late, part of the human intracranial signals evoked by letter strings (33). In the future, it will be interesting to see if the late appearance of the full Vinckier gradient can be reproduced by using a recurrent network architecture. The specialization of the VWFA for reading prevails across all writing systems, ranging from highly regular orthographic scripts to ideographic systems (83). Still, it is possible to decode between two scripts in bilingual subjects (84), and subtle differences exist between the topography of VOT activations during reading different scripts (5), probably due to both bottom-up and top-down specificities. It is thus likely that some fine features of the current results reflect the regularities of French orthography and may differ in languages with different orthographic statistics or using ideographic symbols.

Finally, the present work was limited to modeling the orthographic component of reading, putatively attached to the ventral visual pathway. Yet, decades of research have shown that reading acquisition involves multiple routes and a complex set of visual, phonological, morphological, and lexico-semantic representations (16, 57, 85–87). Other modelers have made the choice of capturing these multiple routes in neural network models, often with great success (e.g., refs. 57–61). Such global modeling of reading, however, usually comes at the expense of not accurately modeling the early visual stages of word recognition; indeed, in those models, the input is usually a bank of abstract letter identities, rather than an actual image of the word (for exceptions, see refs. 71 and 88). Our networks face the converse limitation: They only process visual inputs up to the classification level, akin to lexical access, without any influence from downstream phonological, morphological, or semantic systems. Fortunately, evidence from baboons and monkeys, which have none of those late stages, suggests that visual demands alone may suffice to drive the emergence of orthographic abilities and a VWFA somewhat analogous to humans (47, 89, 90). Nevertheless, focusing on the input orthographic system is a design choice for a first modeling step, and not a theoretical statement, as it is known that phonological representations, for instance, do impact on ventral visual representations in humans (43). Again, a more complex, recurrent architecture, combining visual and phonological inputs, would be needed to accurately capture those observations.

Conclusion and Summary of Main Predictions

Beyond reproducing existing data, the main purpose of our approach was to develop predictions concerning the fine-grained neural code for words, which may soon become testable either with very-high-resolution, high-field fMRI or with high-density intracranial recordings. We therefore end by summarizing the predictions that arise from our simulations.

First, we predict that the readers' inferotemporal cortex contains a sparse neural code for written words, with only ∼30 dimensions of vector coding sufficing to encode the 1,000 most frequent words. Second, we predict that those vector codes are encoded by word-specific neurons that do not respond to other pictures (faces, bodies, houses, or tools). Third, these neurons should be initially uncommitted in the young preliterate brain and only become attuned to letters and their combinations in the course of reading acquisition. Fourth, their overall activation should increase as a function of word length. Fifth, the receptive field of those units should be characterized by a high selectivity for one or a few letters at one or several consecutive locations in the string, with occasional sensitivity to order (a neuron may be selective to letter x around position p1 and to letter y around position p2) and/or an additional inhibition to several other letters (thus forming a letter dipole responding to letter x, but not letter y, around position p).

Finally, according to the biased-connectivity hypothesis, the location of those units should be predictable from their preexisting pattern of projective connections to downstream language areas. And, since we found that even in the biased literate condition, not all Dense+ units ended up being selective to words, we predict that within the human VWFA, unlike monkey face patches (91), some neurons may not respond to words and remain committed to other visual categories. The latter predictions, however, may have to be revised once more realistic, topographic, and recurrent models of reading acquisition are developed.

Data Availability. Neural networks data have been deposited in GitHub (https://github.com/THANNAGA/Origins-of-VWFA).

ACKNOWLEDGMENTS. This work was granted access to the High Performance Computing resources of the Institute for Development and Resources in Intensive Scientific Computing under Allocation 2021-AD011012288 made by the Grand Équipement National de Calcul Intensif. We also gratefully acknowledge the financial support of Collège de France, INSERM, Commissariat à l'Énergie Atomique et aux Énergies Alternatives, Fondation Bettencourt-Schueller, Région Ile-de-France, and an Agence Nationale de la Recherche grant "Toplex" to L.C. and S.D.

1. S. Dehaene, L. Cohen, J. Morais, R. Kolinsky, Illiterate to literate: Behavioural and cerebral changes induced by reading acquisition. *Nat. Rev. Neurosci.* **16**, 234–244 (2015).
2. B. A. Wandell, J. D. Yeatman, Biological development of reading circuits. *Curr. Opin. Neurobiol.* **23**, 261–268 (2013).
3. L. Cohen *et al.*, The visual word form area: Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain* **123**, 291–307 (2000).
4. C. I. Baker *et al.*, Visual word processing and experiential origins of functional selectivity in human extrastriate cortex. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 9087–9092 (2007).
5. M. Szwed, E. Qiao, A. Jobert, S. Dehaene, L. Cohen, Effects of literacy in early visual and occipitotemporal areas of Chinese and French readers. *J. Cogn. Neurosci.* **26**, 459–475 (2014).
6. A. M. Rauschecker, R. F. Bowen, J. Parvizi, B. A. Wandell, Position sensitivity in the visual word form area. *Proc. Natl. Acad. Sci. U.S.A.* **109**, E1568–E1577 (2012).
7. S. Dehaene *et al.*, Cerebral mechanisms of word masking and unconscious repetition priming. *Nat. Neurosci.* **4**, 752–758 (2001).
8. S. Dehaene *et al.*, Letter binding and invariant recognition of masked words: Behavioral and neuroimaging evidence. *Psychol. Sci.* **15**, 307–313 (2004).
9. J. L. Bruno, A. Zumberge, F. R. Manis, Z. L. Lu, J. G. Goldman, Sensitivity to orthographic familiarity in the occipito-temporal region. *Neuroimage* **39**, 1988–2001 (2008).
10. M. Kronbichler *et al.*, The visual word form area and the frequency with which words are encountered: Evidence from a parametric fMRI study. *Neuroimage* **21**, 946–953 (2004).
11. G. Dehaene-Lambertz, K. Monzalvo, S. Dehaene, The emergence of the visual word form: Longitudinal evolution of category-specific ventral visual areas during reading acquisition. *PLoS Biol.* **16**, e2004103 (2018).
12. C. J. Price, J. T. Devlin, The interactive account of ventral occipitotemporal contributions to reading. *Trends Cogn. Sci.* **15**, 246–253 (2011).
13. S. Dehaene, L. Cohen, The unique role of the visual word form area in reading. *Trends Cogn. Sci.* **15**, 254–262 (2011).
14. T. Twomey, K. J. Kawabata Duncan, C. J. Price, J. T. Devlin, Top-down modulation of ventral occipito-temporal responses during visual word recognition. *Neuroimage* **55**, 1242–1251 (2011).
15. S. Dehaene, L. Cohen, Cultural recycling of cortical maps. *Neuron* **56**, 384–398 (2007).
16. S. Dehaene, *Reading in the Brain* (Penguin Viking, New York, NY, 2009).
17. S. Dehaene *et al.*, How learning to read changes the cortical networks for vision and language. *Science* **330**, 1359–1364 (2010).
18. U. Hasson, I. Levy, M. Behrmann, T. Hendler, R. Malach, Eccentricity bias as an organizing principle for human high-order object areas. *Neuron* **34**, 479–490 (2002).
19. R. Malach, I. Levy, U. Hasson, The topography of high-order human object areas. *Trends Cogn. Sci.* **6**, 176–184 (2002).
20. M. Szwed, L. Cohen, E. Qiao, S. Dehaene, The role of invariant line junctions in object and visual word recognition. *Vision Res.* **49**, 718–725 (2009).
21. B. Long, C.-P. Yu, T. Konkle, Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E9015–E9024 (2018).
22. P. Barttfeld *et al.*, A lateral-to-mesial organization of human ventral visual cortex at birth. *Brain Struct. Funct.* **223**, 3107–3119 (2018).
23. F. Bouhali *et al.*, Anatomical connections of the visual word form area. *J. Neurosci.* **34**, 15402–15414 (2014).
24. T. Hannagan, A. Amedi, L. Cohen, G. Dehaene-Lambertz, S. Dehaene, Origins of the specialization for letters and numbers in ventral occipitotemporal cortex. *Trends Cogn. Sci.* **19**, 374–382 (2015).
25. B. Z. Mahon, A. Caramazza, What drives the organization of object knowledge in the brain? *Trends Cogn. Sci.* **15**, 97–103 (2011).
26. R. B. Mars *et al.*, Whole brain comparative anatomy using connectivity blueprints. *eLife* **7**, e35237 (2018).
27. Z. M. Saygin *et al.*, Connectivity precedes function in the development of the visual word form area. *Nat. Neurosci.* **19**, 1250–1255 (2016).
28. A. Testolin, I. Stoianov, M. Zorzi, Letter perception emerges from unsupervised deep learning and recycling of natural image features. *Nat. Hum. Behav.* **1**, 657–664 (2017).
29. J. Kubilius *et al.*, CORnet: Modeling the neural mechanisms of core object recognition. bioRxiv [Preprint] (2018). https://doi.org/10.1101/408385 (Accessed 1 July 2020).
30. L. Cohen, S. Dehaene, F. Vinckier, A. Jobert, A. Montavont, Reading normal and degraded words: Contribution of the dorsal and ventral visual pathways. *Neuroimage* **40**, 353–366 (2008).

31. F. Vinckier *et al.*, Hierarchical coding of letter strings in the ventral stream: Dissecting the inner organization of the visual word-form system. *Neuron* **55**, 143–156 (2007).
32. J. R. Binder, D. A. Medler, C. F. Westbury, E. Liebenthal, L. Buchanan, Tuning of the human left fusiform gyrus to sublexical orthographic structure. *Neuroimage* **33**, 739–748 (2006).
33. O. Woolnough *et al.*, Spatiotemporal dynamics of orthographic and lexical processing in the ventral visual pathway. *Nat. Hum. Behav.* **5**, 389–398 (2021).
34. S. Andrews, Lexical retrieval and selection processes: Effects of transposed-letter confusability. *J. Mem. Lang.* **35**, 775–800 (1996).
35. M. Perea, S. J. Lupker, "Transposed-letter confusability effects in masked form priming" *Masked Priming: State of the Art*, S. Kinoshita, S. J. Lupker, Eds. (Psychology Press, New York, NY, 2003), pp. 97–120.
36. L. Cohen, S. Dehaene, S. McCormick, S. Durant, J. M. Zanker, Brain mechanisms of recovery from pure alexia: A single case study with multiple longitudinal scans. *Neuropsychologia* **91**, 36–49 (2016).
37. J. Dejerine, Contribution à l'étude anatomo-pathologique et clinique des différentes variétés de cécité verbale. *Mem. Soc. Biol.* **4**, 61–90 (1892).
38. R. Gaillard *et al.*, Direct intracranial, FMRI, and lesion evidence for the causal role of left inferotemporal cortex in reading. *Neuron* **50**, 191–204 (2006).
39. C. J. Spoerer, T. C. Kietzmann, J. Mehrer, I. Charest, N. Kriegeskorte, Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLOS Comput. Biol.* **16**, e1008215 (2020).
40. L. S. Glezer, X. Jiang, M. Riesenhuber, Evidence for highly selective neuronal tuning to whole words in the "visual word form area". *Neuron* **62**, 199–204 (2009).
41. S. Dehaene, L. Cohen, M. Sigman, F. Vinckier, The neural code for written words: A proposal. *Trends Cogn. Sci.* **9**, 335–341 (2005).
42. J. Grainger, J. P. Granier, F. Farioli, E. Van Assche, W. J. van Heuven, Letter position information and printed word perception: The relative-position priming constraint. *J. Exp. Psychol. Hum. Percept. Perform.* **32**, 865–884 (2006).
43. F. Bouhali, Z. Bézagu, S. Dehaene, L. Cohen, A mesial-to-lateral dissociation for orthographic processing in the visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 21936–21946 (2019).
44. A. Agrawal, K. V. S. Hari, S. P. Arun, Reading increases the compositionality of visual word representations. *Psychol. Sci.* **30**, 1707–1723 (2019).
45. A. Agrawal, K. Hari, S. Arun, A compositional neural code in high-level visual cortex can explain jumbled word reading. *eLife* **9**, e54846 (2020).
46. J. Grainger, S. Dufau, J. C. Ziegler, A vision of reading. *Trends Cogn. Sci.* **20**, 171–179 (2016).
47. R. Rajalingham, K. Kar, S. Sanghavi, S. Dehaene, J. J. DiCarlo, The inferior temporal cortex is a potential cortical precursor of orthographic processing in untrained monkeys. *Nat. Commun.* **11**, 3886 (2020).
48. M. W. Self *et al.*, The effects of context and attention on spiking activity in human early visual cortex. *PLoS Biol.* **14**, e1002420 (2016).
49. G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors. arXiv [Preprint] (2012). https://arxiv.org/abs/1207.0580 (Accessed 21 January 2021).
50. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
51. J. C. Ziegler *et al.*, Transposed-letter effects reveal orthographic processing in baboons. *Psychol. Sci.* **24**, 1609–1611 (2013).
52. N. Kriegeskorte, M. Mur, P. Bandettini, Representational similarity analysis—connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).
53. R. N. Shepard, S. Chipman, Second-order isomorphism of internal representations: Shapes of states. *Cognit. Psychol.* **1**, 1–17 (1970).
54. C. H. C. Chang *et al.*, Adaptation of the human visual system to the statistics of letters and line configurations. *Neuroimage* **120**, 428–440 (2015).
55. F. Pegado *et al.*, Timing the impact of literacy on visual processing. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E5233–E5242 (2014).
56. I. Goodfellow, H. Lee, Q. V. Le, A. Saxe, A. Y. Ng, "Measuring invariances in deep networks" in *Advances in Neural Information Processing Systems*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, A. Culotta, Eds. (Curran Associates, Inc., Red Hook, NY, 2009), **vol. 22**, pp. 646–654.
57. M. Coltheart, K. Rastle, C. Perry, R. Langdon, J. Ziegler, DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychol. Rev.* **108**, 204–256 (2001).
58. M. W. Harm, M. S. Seidenberg, Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychol. Rev.* **106**, 491–528 (1999).
59. M. S. Seidenberg, J. L. McClelland, A distributed, developmental model of word recognition and naming. *Psychol. Rev.* **96**, 523–568 (1989).
60. M. Zorzi, G. Houghton, B. Butterworth, Two routes or one in reading aloud? A connectionist dual-process model. *J. Exp. Psychol. Hum. Percept. Perform.* **24**, 1131–1161 (1998).
61. D. C. Plaut, J. L. McClelland, M. S. Seidenberg, K. Patterson, Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychol. Rev.* **103**, 56–115 (1996).
62. R. T. McCoy, T. Linzen, E. Dunbar, P. Smolensky, RNNs implicitly implement tensor product representations. arXiv [Preprint] (2019). https://arxiv.org/abs/1812.08718 (Accessed 28 March 2020).
63. P. Smolensky, Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artif. Intell.* **46**, 159–216 (1990).
64. M. McCloskey, N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem" in *Psychology of Learning and Motivation*, G. H. Bower, Ed. (Academic Press, New York, NY, 1989), pp. 109–165.
65. R. Kolinsky, "How learning to read influences language and cognition" in *The Oxford Handbook of Reading*, A. Pollatsek, R. Treiman, Eds. (Oxford University Press, Oxford, UK, 2014).
66. E. C. Kubota, S. J. Joo, E. Huber, J. D. Yeatman, Word selectivity in high-level visual cortex and reading skill. *Dev. Cogn. Neurosci.* **36**, 100593 (2019).
67. F. Pegado, K. Nakamura, L. Cohen, S. Dehaene, Breaking the symmetry: Mirror discrimination for single letters but not for pictures in the Visual Word Form Area. *Neuroimage* **55**, 742–749 (2011).
68. F. Pegado *et al.*, Literacy breaks mirror invariance for visual stimuli: A behavioral study with adult illiterates. *J. Exp. Psychol. Gen.* **143**, 887–894 (2014).
69. L. Chang, D. Y. Tsao, The code for facial identity in the primate brain. *Cell* **169**, 1013–1028.e14 (2017).
70. G. A. Miller, N. Chomsky, "Finitary models of language users" in *Handbook of Mathematical Psychology*, R. D. Luce, R. R. Bush, E. Galanter, Eds. (Wiley, New York, NY, 1963), vol. **1**, p. 419–491.
71. M. G. Di Bono, M. Zorzi, Deep generative learning of location-invariant visual word recognition. *Front. Psychol.* **4**, 635 (2013).
72. F. Vinckier, E. Qiao, C. Pallier, S. Dehaene, L. Cohen, The impact of letter spacing on reading: A test of the bigram coding hypothesis. *J. Vis.* **11**, 1–21 (2011).
73. H. Lee *et al.*, Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. bioRxiv [Preprint] (2020). https://doi.org/10.1101/2020.07.09.185116 (Accessed 1 June 2021).
74. M. Schrimpf *et al.*, Brain-score: Which artificial neural network for object recognition is most brain-like? bioRxiv [Preprint] (2018). https://doi.org/10.1101/407007 (Accessed 1 June 2021).
75. T. C. Kietzmann *et al.*, Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 21854–21863 (2019).
76. D. L. Yamins *et al.*, Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014).
77. M. Eickenberg, A. Gramfort, G. Varoquaux, B. Thirion, Seeing it all: Convolutional network layers map the function of the human visual system. *Neuroimage* **152**, 184–194 (2017).
78. P. Bao, L. She, M. McGill, D. Y. Tsao, A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).
79. J. L. McClelland, D. E. Rumelhart, An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychol. Rev.* **88**, 375–407 (1981).
80. C. Perry, J. C. Ziegler, M. Zorzi, Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychol. Rev.* **114**, 273–315 (2007).
81. M. Heilbron, D. Richter, M. Ekman, P. Hagoort, F. P. de Lange, Word contexts enhance the neural representation of individual letters in early visual cortex. *Nat. Commun.* **11**, 321 (2020).
82. A. Nayebi *et al.*, Task-driven convolutional recurrent models of the visual system. arXiv [Preprint] (2018). https://arxiv.org/abs/1807.00053 (Accessed 10 February 2021).
83. J. G. Rueckl *et al.*, Universal brain signature of proficient reading: Evidence from four contrasting languages. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 15510–15515 (2015).
84. M. Xu, D. Baldauf, C. Q. Chang, R. Desimone, L. H. Tan, Distinct distributed patterns of neural activity are associated with two languages in the bilingual brain. *Sci. Adv.* **3**, e1603309 (2017).
85. M. Melby-Lervåg, S.-A. H. Lyster, C. Hulme, Phonological skills and their role in learning to read: A meta-analytic review. *Psychol. Bull.* **138**, 322–352 (2012).
86. A. Castles, K. Rastle, K. Nation, Ending the reading wars: Reading acquisition from novice to expert. *Psychol. Sci. Public Interest* **19**, 5–51 (2018).
87. P. E. Turkeltaub, L. Gareau, D. L. Flowers, T. A. Zeffiro, G. F. Eden, Development of neural mechanisms for reading. *Nat. Neurosci.* **6**, 767–773 (2003).
88. M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, Reading text in the wild with convolutional neural networks. *Int. J. Comput. Vis.* **116**, 1–20 (2016).
89. J. Grainger, S. Dufau, M. Montant, J. C. Ziegler, J. Fagot, Orthographic processing in baboons (*Papio papio*). *Science* **336**, 245–248 (2012).
90. K. Srihasam, J. B. Mandeville, I. A. Morocz, K. J. Sullivan, M. S. Livingstone, Behavioral and anatomical consequences of early versus late symbol training in macaques. *Neuron* **73**, 608–619 (2012).
91. D. Y. Tsao, W. A. Freiwald, R. B. Tootell, M. S. Livingstone, A cortical region consisting entirely of face-selective cells. *Science* **311**, 670–674 (2006).