Title: Demystifying the Visual Word Form Area: Precision fMRI of Visual, Linguistic, and Attentional Properties of Ventral Temporal Cortex

Authors

Jin Li[1,2], Kelly J. Hiersche[1,2], Zeynep M. Saygin[1,2*]

Affiliations

[1]Department of Psychology, The Ohio State University, Columbus, OH, USA.

[2]Center for Cognitive and Behavioral Brain Imaging, The Ohio State University, Columbus, OH, USA.

*Correspondence to: li.9361@osu.edu and saygin.3@osu.edu

Abstract

After more than three decades of investigation, the nature and even the existence of the visual word form area (VWFA), a region in left ventrotemporal cortex (VTC), is still under debate. Here we provide novel insight on this debate by using precision fMRI to compare the VWFA to adjacent VTC regions, amodal language, and multiple-demand networks. We find that the VWFA is unique in its word-selectivity within VTC. The VWFA also uniquely shows auditory language-selectivity, but this language responsiveness is dwarfed by i) its visual-selectivity even to non-preferred stimuli and ii) language-selectivity of the language network. We additionally identify a region at the tip of the VTC which prefers auditory language to visual stimuli. Finally, almost all VTC fROIs are modulated by attention, not just the VWFA. We propose that the VWFA primarily operates as a visual look-up dictionary of orthographic information, presumably facilitated through connections with canonical language cortex.

Introduction

The ventrotemporal cortex (VTC) consists of numerous regions that each specialize in perceiving abstract visual stimulus categories (e.g., faces, objects, bodies, and places). The visual word form area (VWFA) is perhaps one of the most fascinating of these VTC regions because it seems to be specialized for processing a recent human invention: reading[1,2]. This region responds selectively to orthographic visual stimuli, such as letter strings and words, but only after an individual can read[3–5] and only to languages in which the subject is literate[6]. These orthographic responses are thought to subserve major aspects of typical reading behavior[3,7]. Further, this region is lateralized to the left[8,9] and in approximately the same location across individuals, different languages (English vs. French)[10], and even different writing systems (alphabetic vs. logographic)[11]. Thus, the VWFA is vastly studied as a prime example of how the human brain has the capacity to dedicate cortical tissue for new symbolic representations.

However, the specialization of the VWFA for visual words is not without doubt, and alternative functions of the VWFA have been proposed since its initial discovery. The first main argument supposes that instead of responding specifically to visual words per se, the anatomical location of the VWFA, the posterior fusiform gyrus, is involved in complex shape processing[12] more generally. Using approximate anatomical coordinates from previous reports to define the VWFA, studies have shown that responses in this region to words were not qualitatively higher than responses to line-drawings of objects and false fonts[13]; or that it showed comparable levels of activation for both English letters and characters foreign to the participant (Amharic strings[14]; Korean Hangul characters[15]). Along the same line, some suggest the VWFA is in fact comingled with adjacent high-level regions like the fusiform face area (FFA)[16,17] because of shared visual properties and underlying visual computations between words and faces. The neuronal recycling hypothesis[18] suggests the VWFA was 'recycled' from neural tissue dedicated to another type of visual stimulus category (perhaps faces) (but see[4,19]) and thus retains some of this category-selectivity even after an individual gains literacy.

The second (and related) argument is that the VWFA is not specific for visual processing (let alone word processing) and instead has a 'multiplex' function[20] that is guided by its interactions with upstream and higher-order systems, such as language regions for high-level linguistic processing (e.g., semantics and phonology) and regions for attentional demand[21]). Therefore, the VWFA may respond to words for one type of task but may respond to other stimuli for another task (presumably through coactivation with other brain regions). Consistent with this idea, previous studies have observed intrinsic functional connectivity between the VWFA and the traditional language network[5,22,23] and/or between VWFA and the dorsal attention network[24]. The more extreme version of this theory is that the VWFA not only interacts with the language network when the task conditions call for it, but that it may actually be another core component of the language network, and thus selective to amodal linguistic information. For example, in congenitally blind individuals, the anatomical location of the (putative) VWFA responds to both Braille and to a grammatical complexity manipulation of auditory sentences[25], suggesting that the VWFA may be amodal and involved in linguistic processing more generally (and even in sighted individuals as well). This idea may be further supported by the shared laterality of the VWFA and the language network[26]. Recently, Chen et al.[20] found that the connectivity between the VWFA and the language

or attention networks strengthened during a task to support either linguistic processing or spatial attention, respectively. While connectivity with the amodal language or attention networks does not necessarily imply that the VWFA is part of those networks (and therefore not dedicated to visual word processing but rather to general linguistic or attentional processing), these previous findings do raise important questions about the involvement of the VWFA in language and spatial attention tasks. Only a handful of studies have shown the VWFA can be activated/modulated by auditory speech[27,28]. Is the VWFA responsive and selective for high-level linguistic information and/or modulated by attention? And does the VWFA's response profile (i.e., neural signature) differ from other nodes in the amodal language or attention networks? Or is the VWFA primarily a visual region (and if so, is it unique within the VTC in its processing of visual word information)?

Several methodological limitations in previous studies preclude clear answers to these questions. First, although all referred to their region of interest as the "VWFA", the definition of this region is inconsistent across studies, especially in the choice of the control condition. For example, studies may define the VWFA as a region that responds more to written words than to fixation[29], checkerboards[30], scrambled words[31], photos of objects[23], line-drawings of objects[5], or the average of all other categories[32]. These conditions differ in what they are controlling for (e.g., low- or high-level visual information, semantic content, etc.) and may localize different areas of cortex, thus making it difficult to interpret findings across studies. Second, although located in approximately the same location across individuals, the VWFA is a small region, and the exact location varies from person to person[33]. Relying on group activation or using anatomical coordinates (on a template brain) may not capture word-selective voxels because they do not account for individual variability and/or because they assume a continuous cluster (as with a spherical ROI) which is not a prerequisite for category-selective regions[10,34]. Third, the VWFA is adjacent to many other category-selective regions within the VTC; therefore, to understand the unique function and neural signature of the VWFA, it is also necessary to define as many of these other regions as possible within each individual and test neural preferences of each region across different stimulus conditions. While identifying a region as category-selective does not mean that the region should not respond to other stimuli, it does mean that the region should respond significantly (and robustly, e.g., ~twice as much) higher to the preferred than non-preferred categories[35]. Further, defining other VTC regions allows us to better delineate the functional landscape of the VTC by identifying spatially close and intertwined category-selective regions and comparing their neural signature.

In the present study, we use individual subject precision fMRI with multiple tasks (a total of 13 different experimental conditions) to systematically examine the definition, functional distinctiveness, and spatial specificity of the VWFA, in comparison with other adjacent VTC regions, as well as the language and frontoparietal multiple-demand networks.

Results

VWFA is selective to visual words, showing a distinct neural response signature from adjacent VTC regions

We examined the responses to words as well as other high-level visual stimuli within the VWFA and other spatially adjacent category-selective VTC regions for each individual (see **Figure 1a** for an example subject's fROIs; see **Supplemental Materials, Figure S1** for fROIs for all individuals in sample). We defined fROIs using 3 different methods: selecting the top 150 most significant voxels; the top 10% of voxels within the search space; and using a hard threshold (p<0.005) (see **Methods** for details). In independent runs, we extracted percent signal change (PSC) within the fROI to all conditions across tasks (see **Figure 1b**). Below we report findings for fROIs defined with top 150 voxels, but we observed vastly similar results for the 2 other methods (reported in **Tables S2 & S3**). Additionally, given the left-lateralized nature of the VWFA, that is, the right homologue of the VWFA showed no reliable word selectivity, observed both here and in previous studies, we focus only on word and other category selectivity within the left VTC (see **Figure S3 & S4** for right VTC results). All tests were Bonferroni-Holm corrected for multiple comparisons unless otherwise noted.

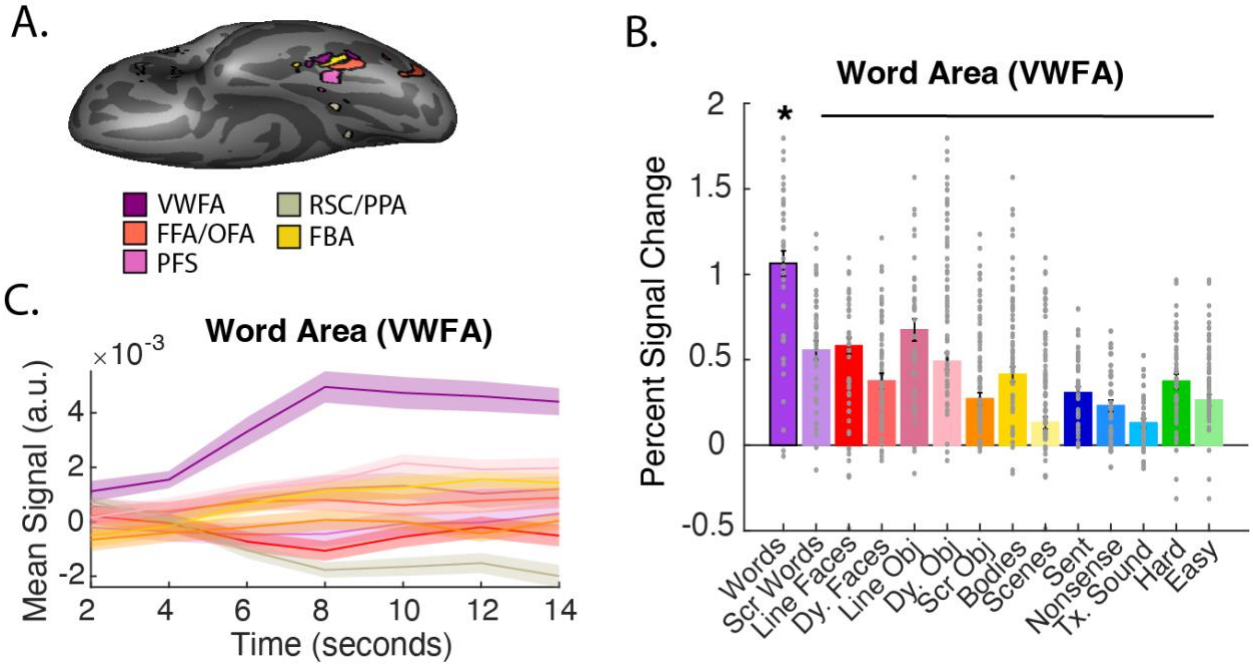

Figure 1. A. VTC fROIs in left hemisphere for an example subject. B. Functional profile of the left VWFA. The mean percent signal change (with standard error bars) to various visual and non-visual conditions are plotted. The preferred category, words, has a thick black outline. Individual subject PSCs are shown with grey hollow circles. Significance is noted (*p<0.05, Bonferroni-holm) for words only, with a black line showing all conditions significantly lower than words. C. Averaged time-course of VWFA's responses to blocks of different experimental conditions. Dark line for mean across all subjects and shading for standard error. VWFA: visual word form area.

Before addressing our main question, we first determined the best contrast to define the VWFA. Specifically, we asked whether using different control conditions (i.e., rapidly presented visual words vs line-drawing objects (W-Obj), words vs. line-drawing of faces (W-F), words vs. scrambled words/parts of words (W-SW), words vs. average of the rest of conditions (W-All), and words vs. fixation (W-Fix)) to define the left VWFA (VWFA) yielded the same, word-responsive region (**Figure S2**). All contrasts resulted in a VWFA significantly more responsive to words than other visual conditions in an independent run, except W-Fix, which does not control for either the abstract concepts inherent in words or the low-level visual features of written words (all $p<2.00x10^{-7}$, except W-Fix $p>0.02$, see **Table S1**). We also examined the overlap between fROIs defined in separate runs for all significant contrasts (i.e., all but W-Fix). A one-way repeated measures ANOVA (rmANOVA) showed a significant effect of contrast ($F(3,165)=3.99$, $p=0.009$), in which the W-All showed significantly higher run-to-run consistency than W-Obj (post-hoc paired test: $t(55)=3.06$, $p=0.003$) and W-SW ($t(55)=2.60$, $p=0.012$, Bonferroni-holm $p=0.073$) but not W-F ($t(55)=0.28$, $p=0.779$). Additionally, the VWFA defined by W-All was the most similar to (i.e. highest overlap with) VWFAs defined using the other contrasts (see probabilistic activation maps across participants for each contrast in **Figure S2** (W-All vs. W-L, $t(55)=12.68$, $p=5.80x10^{-18}$; W-All vs. W-F, $t(55)=13.17$, $p=1.20x10^{-18}$; W-All vs. W-SW, $t(55)=14.26$, $4.01x10^{-20}$) (see **Methods** for details). Thus, we used W-All to define the VWFA in our subsequent analyses, but concluded that W-All, W-F, W-L, and W-SW are all sufficient contrasts for defining the VWFA.

Next, we examine category selectivity in all VTC fROIs by identifying their preferred category (i.e., category with largest effect size) as compared to all other visual categories. **Figure 1B** showed that the VWFA responded highest to visual words and paired t-tests showed that the VWFA's response to words was significantly higher than all other conditions (words vs, line-drawing objects (the second highest), $t(35)=7.03$, $p=3.55x10^{-8}$; see **Table 1** for the statistical results for all comparisons for the VWFA). We observed a similar result when calculating the averaged time-course across the experimental block for all static and dynamic visual conditions for the VWFA (**Figure 1C**): the VWFA showed no preferences to conditions other than words for the entirety of the block. Similarly, adjacent face-, object- and scene-selective regions showed the expected activation profiles, consistent with decades of previous research (e.g., OFA and FFA showed the highest response to faces, PFS to objects, RSC and PPA to scenes; **Figure 2A, Table S1**). Only the FBA lacked clear preferences for its expected category and was therefore excluded from further comparative analyses below. Again, when looking at the time-course, all fROIs (even the FBA), showed increased activity to their preferred conditions but not for other categories (**Figure 2B**). We also explicitly quantified any overlap among the VWFA and other VTC fROIs when selecting the top 150 voxels before assigning them based on their most responsive category (see **Methods**). We found little overlap (if any) within individual: an average of 13, 9, and 3 vertices overlapped between the VWFA and FFA, FBA and PFS, respectively. See **Table S4** for other descriptive information for defined regions when using other two methods to define the fROIs. We also showed the landscape of VTC fROIs across all participants in our study (see **Figure S1**) demonstrating large individual variability and how these regions may blur together (overlap) in group analyses.

VWFA Words vs.	t (df)	p
Scrambled Words	7.03 (35)	3.55×10^{-8} **
Line Faces	7.93 (35)	2.50×10^{-9} **
Dynamic Faces	8.73 (35)	2.60×10^{-10} **
Line Objects	5.57 (35)	2.87×10^{-6} **
Dynamic Objects	7.19 (35)	2.15×10^{-8} **
Scrambled Objects	9.29 (35)	5.65×10^{-11} **
Dynamic Bodies	7.91 (35)	2.70×10^{-9} **
Dynamic Scenes	10.27 (35)	4.20×10^{-12} **

Table 1. Comparing the VWFA's response to words with all other visual conditions (see **Table S2** for other regions). **Bonferroni-Holm $p<0.05$

Critically, the VWFA was unique within the VTC in its robust word-selectivity. Across the VTC, only one other region- the left FFA- showed higher responses to words than to all other visual categories except its preferred category (faces) (one-sample, one-sided t-tests of face selectivity: lFFA, $t(36)=2.64$, $p=6.02 \times 10^{-3}$) but even then, its response to faces was clearly much larger than its responses to words ($t(36)=5.22$, $p=7.66 \times 10^{-6}$). A one-way rmANOVA comparing word selectivity among lVTC fROIs showed a significant main effect for fROI, ($F(5,155)=45.13$, $p=1.40 \times 10^{-28}$). Post-hoc pairwise t-tests indicated that the VWFA is significantly more selective to words than all other fROIs including the left FFA (VWFA vs. all other fROIs, $t(31)>6.59$, $p<2.29 \times 10^{-7}$).

We further examined whether the VWFA or any other region, showed selectivity to a non-expected category (after excluding the response to their preferred category). We found that the VWFA also shows object selectivity ($t(36)=2.72$, $p=5.00 \times 10^{-3}$); however, the VWFA is not unique in showing (low) selectivity to a nonpreferred category (see **Table S3** for details). For example, the left FFA also shows word selectivity (as reported above: $t(36)=2.64$, $p=6.02 \times 10^{-3}$). However, for each fROI, its selectivity to the non-expected categories was much lower than selectivity to its preferred category, and also lower than the selectivity of the fROI typically associated with that category. For example, VWFA's object selectivity was significantly lower than both its word selectivity ($t(35)=6.31$, $p=3.05 \times 10^{-7}$) and the object selectivity of the PFS ($t(36)=7.39$, $p=1.02 \times 10^{-8}$).

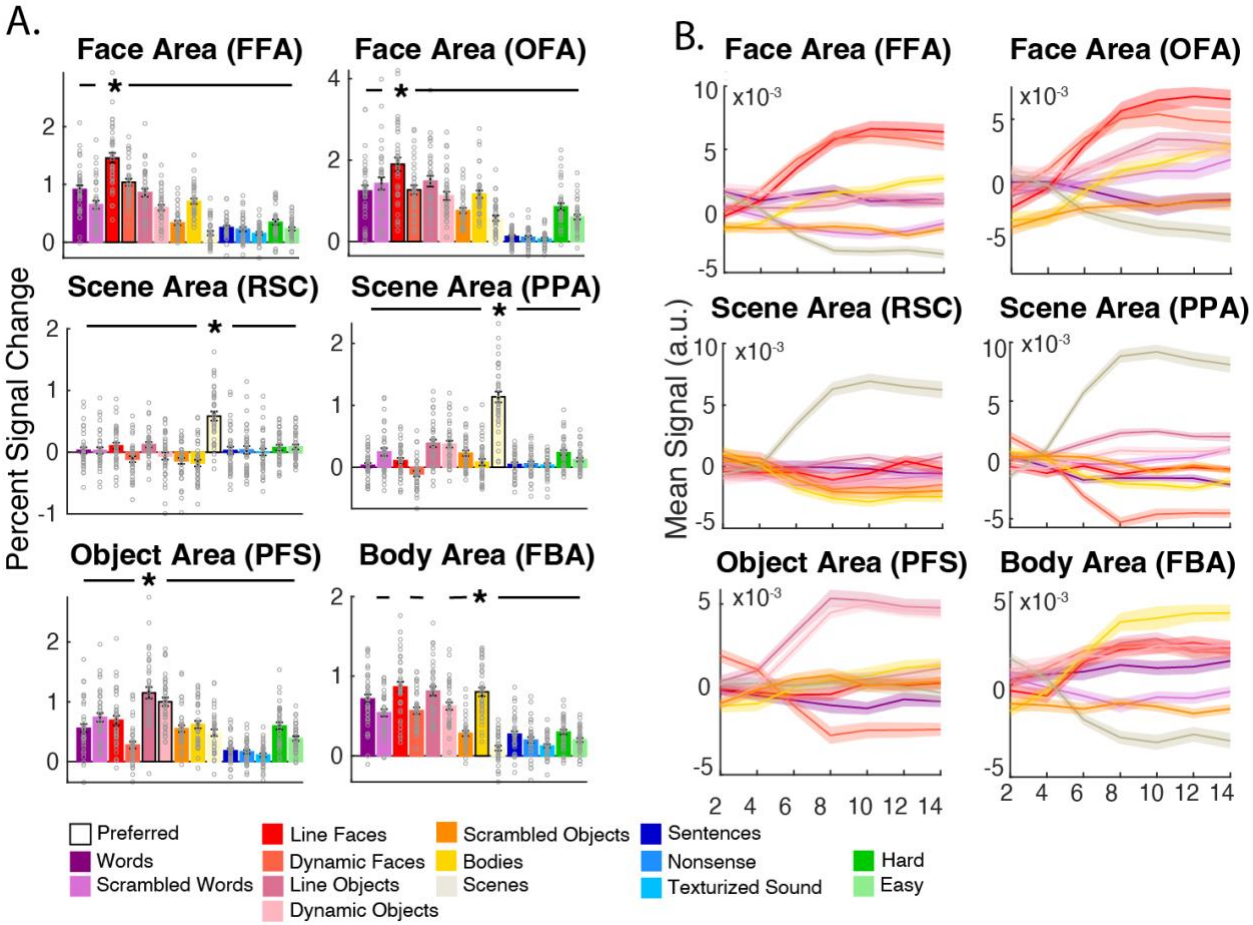

Figure 2. A. Functional profiles for left VTC high-level visual regions. The preferred condition(s) for each fROI is outlined in black. The mean percent signal change (with standard error bars) to various visual and non-visual conditions are plotted. Individual subject PSCs are shown with grey hollow circles. Significance is noted (*p<0.05, Bonferroni-holm) for the preferred category only, with a black lining showing all conditions significantly lower than the preferred category. B. Averaged time-course of each fROI's responses to blocks of different experimental conditions. Dark line for mean across all subjects and shading for standard error. FFA: fusiform face area, OFA: occipital face area, RSC: retrosplenial cortex, PPA: parahippocampal place area, PFS: posterior fusiform sulcus, FBA: fusiform body area.

Very similar results were found when defining the VWFA and other fROIs with two different methods (significance threshold, top 10%, see **Tables S2&S3**), confirming that the functional specificity we observed in the VWFA and other regions is robust and not affected by the method chosen to define the fROIs (as long as the contrast is not words > fixation). We further performed other checks to ensure that these results were not due to high motion or SNR for a particular experiment (see **Supplementary Results 1&2**). We found that while subjects moved less during the static visual experiment than others, this experiment also showed lower SNR than the others (**Figure S8**). Note that even though the static visual (word) localizer has the lowest SNR across the VTC and even within the VWFA (see **Supplementary Results 2**), we still observed that the VWFA showed the highest response to visual words across all conditions, suggesting that this word selectivity is robust and that the effects are unlikely driven by experimental differences.

We replicated the main results (see **Figure S10**) by rescanning a subset of subjects on two additional runs of the static visual localizer. These runs had identical scan parameters to the other three tasks as well as comparable motion and SNR (**Supplementary Results 3**). Altogether, our results showed that by defining each of these VTC regions within individual, and importantly with more stringent control conditions, we find clear evidence of robust word-selectivity for the VWFA.

The VWFA, compared to adjacent VTC regions, selectively responds to auditory language, but is not a core part of amodal high-level language network

Despite the results above showing the VWFA's robust selectivity to visual words, it still remains possible that the region is amodal and selective to auditory linguistic units, and therefore not domain-specific for visual words. Thus, we compared activation to auditorily presented English sentences (Sn) with auditorily presented nonword sequences (nonsense; Ns, controlling for prosody and phonological processing) within the VTC fROIs and frontotemporal language network.

To determine which VTC high-level visual regions respond to auditory linguistic information, we used a two-way repeated measures ANOVA comparing condition (Sn to Ns) in all significant category-selective lVTC regions (VWFA, FFA, PFS, OFA, RSC, PPA). Only the VWFA was sensitive to language (rmANOVA: fROI effect: $F(2.32,76.55)=6.67$, $p=0.0010$; interaction effect: $F(5.00,165.00)=4.40$, $p=8.72\times10^{-4}$; post-hoc t-tests VWFA condition effect $t(33)=2.80$ $p<0.008$; all other regions $p>0.05$). Thus, the VWFA is unique among VTC fROIs in its preference for auditorily presented language (**Table S5**). This result can also be seen by examining the time-course of the VWFA compared to other VTC fROIs during the language task (**Figure 3A**).

But does the VWFA respond to language in the same way as the canonical language network? To answer this, we compared the canonical language fROIs (2 temporal and 3 frontal regions) to the VWFA in their responsiveness to Sn and Ns conditions. Using a two-way (condition by fROI) rmANOVA, we found a significant effect of condition ($F(1,33)=41.75$, $p=2.5\times10^{-7}$), fROI ($F(1.45,48)=88.49$, $p=9.41\times10^{-15}$), and their interaction ($F(2,66)=26.73$, $p=3.13\times10^{-9}$). Unsurprisingly, both the temporal and frontal language fROIs showed language selectivity (Sn > Ns, post-hoc paired samples t-tests: collapsed across temporal: $t(33)=7.43$, $p=1.54\times10^{-8}$; and frontal fROIs: $t(33)=4.59$, $p=6.21\times10^{-5}$). Importantly, the temporal and frontal language fROIs were also significantly more selective to Sn than the VWFA (post-hoc paired samples tests: temporal vs VWFA: $t(33)=11.30$, $p=7.58\times10^{-13}$; frontal vs VWFA: $t(33)=5.21$, $p=9.90\times10^{-6}$) and also to Ns condition than the VWFA (post-hoc paired samples tests: temporal vs VWFA: $t(33)=8.71$, $p=4.53\times10^{-10}$; frontal vs VWFA: $t(33)=4.51$, $p=7.80\times10^{-5}$, see **Figure 3C**). Thus, while the VWFA does show some selectivity to auditory language, its preference for language is dwarfed by responses within the canonical language network.

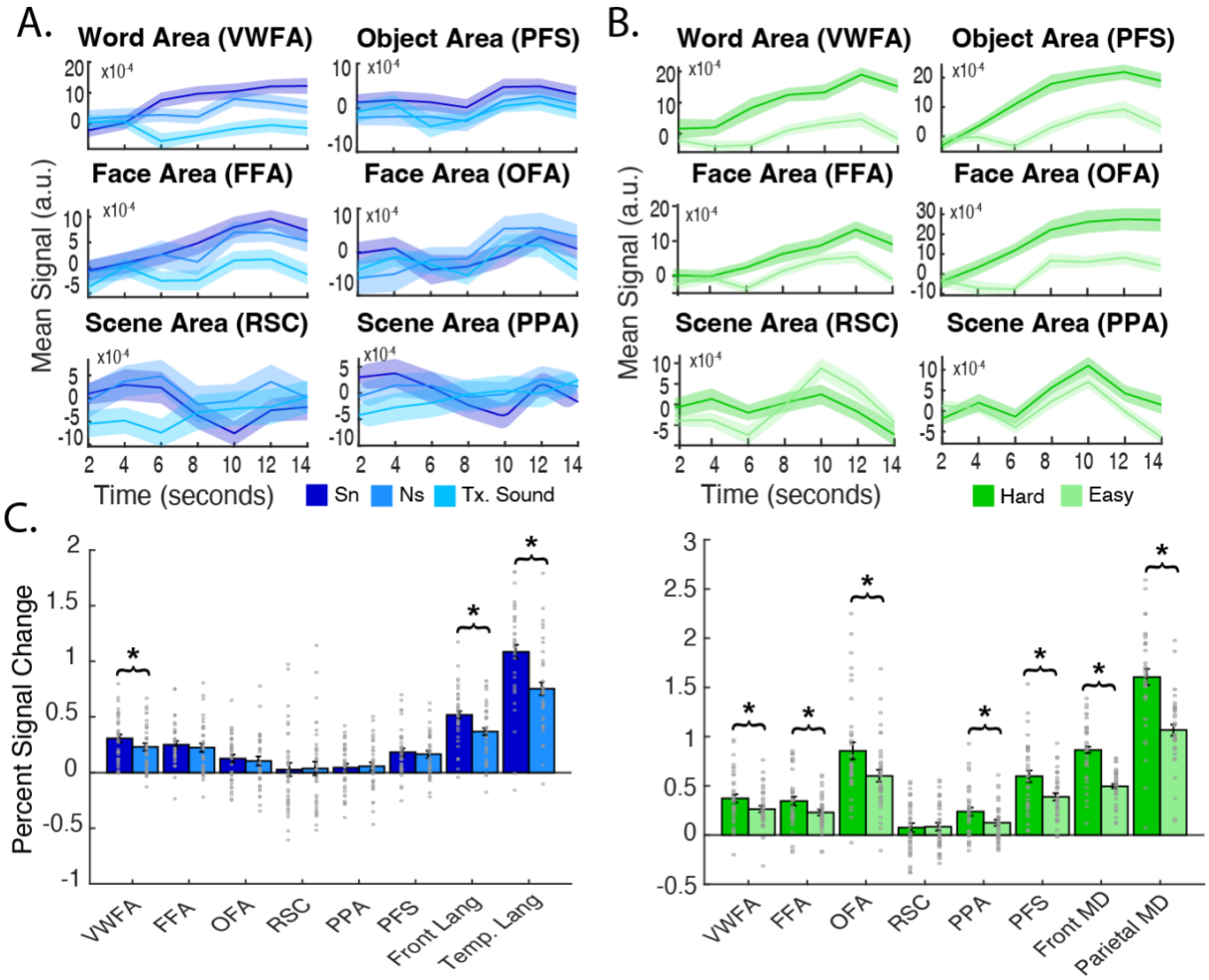

Figure 3. A. The time-course of each VTC fROI during the language task (averaged across all blocks). Dark line for mean across all subjects and shading for standard error. **B.** The time-course of each VTC fROI during the spatial working memory attentional load task (averaged across blocks). Dark line for mean across all subjects and shading for standard error. **C.** Mean percent signal change (with standard error bars) for each VTC fROI and the language fROIs to the language task (left) and the MD fROIs to the MD task (right). Individual subject PSCs are shown with grey hollow circles.

Additionally, not only does the VWFA respond lower to language than the language network, but it also responds significantly more to any visual category (even non-preferred ones) than auditorily presented linguistic stimuli, suggesting that the VWFA is primarily a visual region. If we had only explored language activation within the VTC, we may have concluded that the VWFA was in fact selective to amodal linguistic processing. By comparing the visual vs. language responses of the VWFA, we observe that despite showing language selectivity, the VWFA is decidedly visual (paired t-test between mean response to all non-preferred visual categories and auditory sentences: $t(35)=2.83$, $p=0.0076$).

The VWFA, as well as other VTC regions, are modulated by attentional load

Similarly, we explored the effect of attentional load using a two-way rmANOVA comparing condition (hard vs easy spatial working memory load) by VTC fROIs. The effect of condition ($F(1,35)=38.96$, $p=3.71\times10^{-7}$), fROI ($F(2.89,101.24)=35.47$, $p=2.00\times10^{-15}$), and their interaction ($F(3.71,129.70)=24.93$, $p=1.29\times10^{-14}$) were significant, as was the condition effect in a follow-up one-way rmANOVA for all VTC fROIs ($F(1,35)>17.23$ and $p<0.001$ for all), except the RSC ($F(1,35)=0.147$, $p=0.70$). In other words, almost all VTC fROIs were significantly modulated by attention (see **Table S6** for post-hoc pairwise comparison between Hard and Easy in VTC fROIs). Unlike what was suggested in previous studies, the attentional effect in the VWFA was similar to other fROIs (e.g., FFA) or at times even lower than other fROIs (e.g., PFS and OFA; see full pairwise comparisons in **Table S6**). Time course analyses of VTC fROIs during the MD task also confirmed this ubiquitous attentional effect (**Figure 3B**). Critically, even though the spatial working memory task requires visual processing, we nevertheless found that the magnitude of the attentional load effect in the VTC fROIs was significantly lower than the load effect observed in the frontoparietal multiple demand (MD) fROIs (**Figure 3C**): all VTC regions showed significantly lower activation to the hard condition compared to the frontoparial MD fROIs (Hard: Parietal MD vs VTC fROIs: all $t(32)>7.20$, $p<0.00001$; Frontal: all $t(32)>3.31$, $p<0.002$) (see **Table S6** for post-hoc pairwise comparisons).

We next investigated the resting-state connectivity of the VTC with the language and MD networks. We found that the resting-state connectivity between VTC voxels and functionally defined language fROIs was best explained by each voxel's activation to visual words but not other visual categories (**Figure 4;** paired t-tests comparing the beta values for words vs. the average of other conditions: frontal ($t(32)=3.62$, $p=9.95\times10^{-4}$); temporal ($t(32)=3.02$, $p=0.005$) language network), thus suggesting that language-selectivity within the VTC may be driven by word-responsive voxels' (i.e., VWFA) connectivity to frontotemporal language regions. On the other hand, connectivity between VTC and frontal (or parietal) MD regions was not uniquely explained by VTC voxels' activation to words (words vs. the average of other conditions for the connectivity to frontal ($t(32)=1.16$, $p=0.256$ and parietal ($t(32)=0.23$, $p=0.818$)); rather, all visual selectivities equally contributed to the connectivity with the MD network, in line with our task-based fMRI findings above.

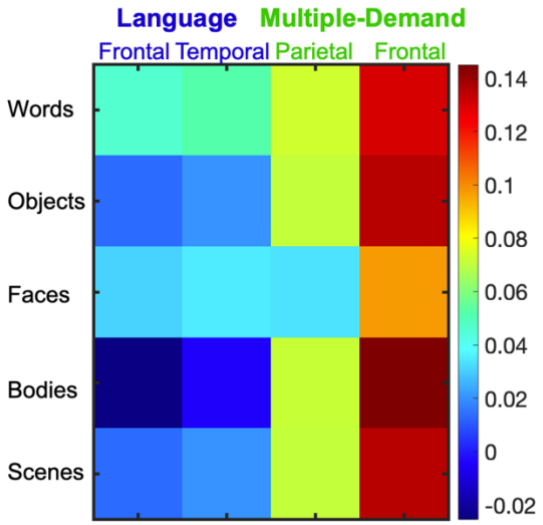

Figure 4: Relationship of voxelwise VTC connectivity to the frontotemporal language fROIs and the frontroparietal MD fROIs by voxelwise PSCs to different visual conditions. Beta values are plotted to characterize how much the connectivity between VTC voxels and language/MD regions was explained by activation to each of the visual categories.

Is there a dedicated language or attention region in the VTC?

Finally, we asked whether we may have missed additional VTC regions that may in fact show language or load-based responses on par with the canonical language or dorsal attention regions respectively. First, we created probabilistic maps to see if any voxels showed language or attentional load responses consistently across subjects. In line with the fROI results above, we did not find a dedicated load-based attention region in either hemisphere. Instead, we saw strong attention modulation throughout the VTC (see **Figure 5B; Figure S6**). In contrast, using a more liberal contrast (sentences vs. texturized sounds (Sn>Tx), controlling for low-level auditory processing, see Methods for more details), we observed two clusters located in the VTC that showed language selectivity: anterior language VTC (aLang-VTC) and medial language VTC (mLang-VTC) (**Figure 6A**, left; similar clusters were observed for Sn vs. Ns, **Figure S7**; see also Figure **S6** for RH hotspots with similar attentional but weaker language activation).

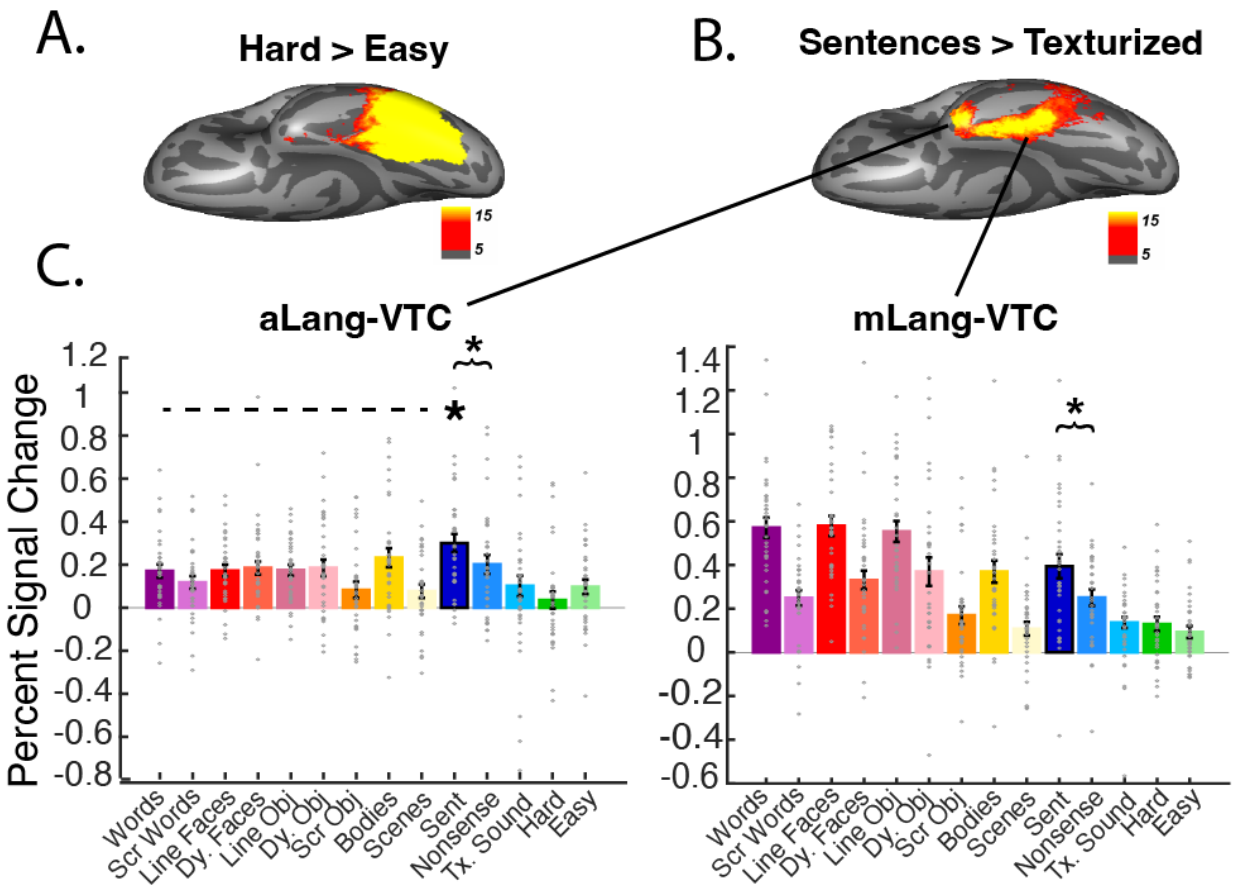

Figure 5: A. Probabilistic map for Hard>Easy showing subjects with overlapping activation during the multiple demand localizer within the VTC. **B.** Probabilistic map for Sn>Tx showing subjects with overlapping activation during the language localizer within the VTC. For A and B, each subject's statistical map was thresholded at p<0.01 and the resulting binarized maps were added together (minimum overlap = 5 subjects) (see **Methods** for details). **C.** Functional profile of the left aLang-VTC and mLang VTC fROIs. The mean percent signal change (with standard error bars) to various visual and non-visual conditions are plotted. The preferred category (i.e., auditory sentences) is outlined in black. Individual subject PSCs are shown with grey hollow circles. Brackets denote a significant difference between sentences and nonsense (paired t-test, *p<0.05). The dashed line denotes a significantly higher response to auditory sentences than average response to visual categories for aLang (paired t-test, *p<0.05).

To quantitively characterize the functionality of these two "language" regions, we defined subject-specific fROIs for mLang-VTC and aLang-VTC within search spaces defined from this probabilistic atlas (see Methods for details) and examined their functional profile (see **Figure 5C)** in comparison to the VWFA and left amodal frontotemporal language network. We examined the language selectivity of the aLang-VTC or mLang-lVTC in comparison to both the VWFA and language network using a condition (Sn, Ns) by fROI rmANOVA. There was a significant effect of condition ($F(1,32)=37.74$, $p=7.18\times10^{-7}$), fROI ($F(2.21,70.57)=62.53$, $p=3.67\times10^{-17}$), and their interaction ($F(2.80,89.45)=14.96$, $p=1.17\times10^{-7}$). As expected, both the aLang-VTC and mLang-VTC were language selective (post-hoc paired-samples t-tests following significant condition effect (Sn>Ns): aLang-VTC, $t(32)=3.80$, $p=0.00061$; mLang-VTC, $t(32)=3.35$, $p=0.002$). However, both

the aLang-VTC and mLang-VTC had language selectivity on par with the VWFA (paired-sample t-tests for Sn responses and Ns responses separately, all $p>0.05$), and significantly lower than the frontal or temporal language fROIs (all $t(32)>2.66$ and all $p<0.05$, see **Table S7**). We also examined the aLang-VTC and mLang-VTC for visual selectivity. The aLang-VTC did not show any visual category preference (one-way rmANOVA of responses to the static localizer: $F(2,66)=0.252$, $p=0.78$; one-way rmANOVA of response to dynamic localizer: $F(2.34,72.4)=11.67$, $p=1.52\times10^{-5}$; post-hoc t-tests show significantly lower response to scenes, but bodies, faces, and objects are not distinguishable). Critically, its response (PSC) to sentences was significantly higher than its average response to all visual categories (paired-samples, two-tailed, t-tests: $t(33)=3.45$, $p=0.0015$). Therefore, aLang-VTC is not a category-selective visual region but rather more of a language region, despite showing weaker language selectivity than the canonical language network. In contrast, the mLang-VTC showed comparable responses to visual conditions as to auditory language (sentences vs. average response to visual stimuli: $t(33)=0.45$, $p=0.66$). Importantly, however, just like aLang-VTC, mLang-VTC responded to different high-level visual stimuli equally and did not show a clear category preference (one-way rmANOVA of responses to the static localizer: $F(2,66)=0.25$, $p=0.78$; one-way rmANOVA of response to dynamic localizer: $F(2.41,79.65)=9.42$, $p=7.92\times10^{-5}$; post-hoc t-tests show significantly lower response to scenes, but bodies, faces, and objects are not distinguishable).

Discussion

Given the continued debates over the existence and functional characteristics of the VWFA, we systematically investigated this region in terms of its preference to different high-level visual stimuli, amodal linguistic information, and attentional demand, aiming to provide a more thorough examination of its neural signature. To our knowledge, this is the first time that the VWFA, along with other closely located category-selective VTC regions were identified in each individual and responses across different visual and nonvisual categories in each region were thoroughly compared to delineate the functional landscape of the VTC. Critically, we found that the VWFA did differ from other regions in its strong preference for visual words. Moreover, the VWFA was selectively modulated by amodal linguistic information but is not a core part of the frontotemporal language or frontoparietal attention networks.

The definition of the VWFA

Methodological discrepancies across previous studies in how they define the VWFA have made it difficult to understand its selectivity. In particular, the VWFA can be defined with different contrasts (i.e., control conditions). When first proposed, the "canonical" location of the VWFA was determined by comparing words vs. rest/fixation[29] or checkerboard[1,36]. These contrasts neither control for visual features nor semantic associations of written words. While later studies continue to update and refine the definition of the VWFA[4,19], the canonical location/coordinates from those earliest studies were often used as a reference as the site of the VWFA (x=-42, y=-57, z=-15)[2], possibly contributing to the confusion as to the role of the VWFA in processing non-orthographic stimuli[37]. Thus, the "myth" of the VWFA may be explained simply by considering the contrasts used to define this region. Indeed, had we only used a contrast that did not control for abstract concepts and low-level visual features, such as fixation tested here, or had we only defined the region by drawing a sphere around the *classic* VWFA coordinates[14] we would be conflating other visual regions with the VWFA, leading to conclusions that the VWFA was tuned for processing general visual shape information (see **Figure S2** (probabilistic map of W-Fix) compared to **Figure S8** probabilistic map of face>objects, objects>scrambled objects). Another potential problem of applying a single coordinate or a group fROI to define the VWFA is the potential of blurring the boundaries between functional regions, either due to individual variability[33] or the closely located mosaic-like functional organization of the VTC. The present study addressed these problems by defining fROIs across the VTC individually. Our results further confirmed the individual variability of the VWFA in terms of its exact location, size, and continuity, as well as the intertwined nature of category-selective VTC regions (see **Figure S1** for depiction of all subjects' fROIs).

We tested multiple and more vigorous contrasts that are also commonly used in the literature: words vs. line-drawings of objects, words vs. line-drawings of faces, words vs. scrambled words and words vs. all other conditions. These experimental control conditions match the low-level visual features and visual similarity (e.g., local contours in scrambled words), complex visual feature preferences (e.g., fovea bias in faces), and abstract semantic content (the identities of the line-drawings of objects were matched

with the visual words), honing in on the high-level perceptual processing of words specifically. All of these contrasts (excluding vs. fixation) resulted in a spatially similar VWFA (see **Figure S2**) with the strongest responses to visual words in an independent run of data (**Table S1**). The VWFA that we identify here likely overlaps with other studies describing this region, sometimes notated as the VWFA-2[38,39], and the mOTS in Lerma-Usabiaga et al.[40], which they proposed to be sensitive to semantic/language unit of visual words. On the other hand, VWFA studies that only use lexical contrasts (e.g., words vs. false fonts[41,42], similar to our scrambled words condition) may identify a separate portion of VTC that is tuned for multiple semantic categories rather than a specific one that distinguishes between orthographic stimuli with semantic meaning vs. non-orthographic stimuli also with semantic meaning, i.e. words and objects[37]. However, in the current study, even when we defined the VWFA using words vs. scrambled words, because we select the most significantly responding voxels and assigned overlapping voxels according to their strongest response (e.g., any voxel selective to words and objects would be assigned to PFS if it was more responsive to objects than words) we still found a VWFA responsive to words more than other categories also representing semantic information (i.e. objects). Consistent with this idea, previous studies have shown that the VWFA is sensitive to orthographic regularity (e.g., frequency of letter bigrams or trigrams) rather than lexical status (distinguishing pseudowords from real words)[42,43]; thus, instead of activating the VWFA which shows strong orthographic selectivity, the lexical contrast may elicit activation in a separate VTC language cluster that is associated with semantic knowledge in general - we identify such a cluster here (mLang-VTC), and discuss it further below.

The VWFA is domain-specific for visual words and is a functionally distinct region within the VTC

To test whether the VWFA is indeed a distinct region from the rest of the VTC, we defined numerous category-selective fROIs along with the VWFA and tested each of their responses to all visual (and other) categories, generating functional profiles for their categorical tuning with high spatial specificity. The VWFA showed minimal overlap with other VTC fROIs. Importantly, we found that the VWFA was functionally different from other VTC fROIs by showing the highest responses to visual words versus other non-word conditions (e.g., almost twice as much to words (average PSC 1.06 +/- 0.07) as to the second highest category (i.e., objects, average PSC of 0.68 +/- 0.07)). So, while the VWFA indeed responds more to objects than to the average of other nonword categories, its responses to words surpass that of other visual categories, which fits with the definition of a category-selective region that is domain-specific. For example, the FFA shows higher activation to objects vs. nonface conditions, but these responses are much lower than its responses to faces (usually twice as low[44], see[45] for discussion).

After showing functional specificity of the VWFA, we then asked: what other categorical selectivity does the VWFA exhibit? Answering this question may give us clues as to how this piece of cortex is able to process words. One line of hypotheses suggests that the region may be (innately) tuned for certain visual features and/or emerge from part of another high-level visual region. For example, the foveal hypothesis proposes that both visual words and faces are processed foveally, i.e., at the center of our visual field[46]. We

did not observe robust face selectivity in the VWFA; however, we found that the VWFA shows object selectivity, which may reflect a general semantic nature of the VTC for high-level visual processing. However, this selectivity is much lower than the VWFA's responses to words, and much lower than the selectivity of face and object-preferring regions like the FFA and PFS. As discussed below, the VWFA might further distinguish itself from adjacent regions due to its privileged connectivity with frontotemporal language cortex and the rich exposure to written scripts, which enables the fine tuning of perceptual processing specifically for orthography.

Amodal linguistic activation in the left VTC

The final question we asked was to what extent the VWFA uniquely responds to high-level linguistic information. Interestingly, we did observe higher activation to auditory sentences vs. nonsense speech within the VWFA, but not for any other VTC regions. This echoes the idea that the VWFA may emerge due to its connectivity with left-lateralized language cortex (i.e., Connectivity Hypothesis[47,48]; with empirical evidence provided in[5,22]). Thus, higher activation to auditory language might be explained by the coactivation between the VWFA and frontotemporal language regions. In line with this idea, we found that the resting-state connectivity between the left VTC frontotemporal language fROIs showed stronger association with VTC voxels' activation to visual words than to other high-level visual categories. However, privileged connectivity does not necessarily mean that the VWFA is an extension of the language network, which is amodal and selective to higher-order linguistic structure[49]. Indeed, the high-level linguistic effect within the VWFA was dwarfed by both the language sensitivity of the frontotemporal language regions and the VWFA responses to even non-preferred visual categories, suggesting that it is visual in nature rather than a core part of the language network.

This tuning for linguistic information due to privileged VWFA-Language connectivity further provides an explanation for why the function of the VWFA reported in previous studies was sensitive to task demand[39,50] and why it might be activated during non-orthographic tasks[51,52]. Specifically, some previous task manipulations explicitly required the association between visual inputs and linguistic and/or semantic information, such as object naming[37,53] and action decision[54]; other task manipulations might implicitly trigger lexico-semantic information of a visual stimulus by using long presentation times. In the current study, by fast presentation of visual stimuli, we minimize top-down effects in favor of exploring the role of the VWFA in rapid visual perception, which is a key function of the VTC[55,56]. Additionally, when we explore the effects of attentional load on the VWFA, we find that the entire VTC is manipulated by load but to a significantly lesser degree than the multiple demand attention network, suggesting that the VWFA is not the only VTC region sensitive to attentional manipulations, nor is it an extension of top-down attention networks. This result is mirrored in our resting-state connectivity findings, where VTC's connectivity with MD was equally associated with VTC voxels' activation to different visual categories. Therefore, our results suggest that the VWFA is a category-selective visual region, enabling efficient perceptual processing of orthographic stimuli, and shows sensitivity to higher-order cognition through interactions with higher-order regions rather than as nodes of these other networks.

A language cluster in VTC that is distinct from VWFA

Interestingly, our exploratory analysis showed that anterior to all of these category-selective regions, there are two language clusters within the left VTC that showed auditory language activation. Importantly, however, by directly comparing the functional response profile of these two regions to the VWFA as well as other VTC regions, we found these two language regions do not distinguish between different visual categories. In fact, the anterior cluster (aLang-VTC) was seated at the tip of the inferior temporal and fusiform gyrus, likely corresponding to the temporal pole region that was previously associated with language comprehension and semantic processing[57,58] regardless of modality. Consistent with this, our results showed that aLang-VTC prefers auditory sentences more than visual stimuli. This region was also the only region within VTC that showed higher preferences for auditory sentences than other stimuli. The more medial and posterior region, the mLang-VTC showed comparable language and visual activation but may be the "basal temporal language area (BTLA)" that is seated between the left temporal pole and the VWFA according to Purcell et al.[58]. Note though, the role and even the anatomical location of the BTLA remains unclear and the term "basal temporal language area" is often used to refer to any or all language areas in the basal temporal lobe. Nevertheless, our results provide some insight about the role of this region: instead of specifically serving as the interface between semantics and orthography per se[58], this multimodal region may play a role in semantic processing of both words and other visual categories[59,60] and does not respond distinctively to visual words in the same way as the closely located VWFA. This is supported by observations that the resection of the BTLA shows the strongest association with deficits in object naming compared to other VTC sites[61].

Limitation and Future directions

The present study systematically examined the role of the precisely defined VWFA and provided a clear characterization of the nature of its orthographic selectivity by looking at its activity to visual words, other non-words stimuli, spoken language and attentional load. However, open questions remain. As mentioned above, the VWFA here corresponded to the mOTS cluster identified with words and word-like stimuli like false fonts, consonant strings and even pseudowords in Lerma-Usabiaga et al.[40]. However, their study averaged the centers of these contrasts with a group analysis. Thus, it is unclear whether the VWFA or another VTC region processes words vs. consonant strings/pseudowords differently, especially considering the hierarchical code for orthographic stimuli present in the VTC[42,62] (although see[6] for a comparably-defined VWFA responding similarly to words and meaningless constant strings). Second, does the VWFA respond differently to visual words compared to other human-invented visual signs (e.g., numbers, traffic signs, music notes)? Changizi et al.[63] noted that there are common line configurations in human-invented visual signs which are not only presented in orthography, but also other visual symbols, therefore the possible sensitivity of the VWFA or other VTC region should be explored with respect to these symbols. Third, is the VWFA sensitive to orthographic regularities like the frequency of different words or other orthographic units? Or does it require constant information transformation between the VWFA and language cortex? For example, if we temporarily

disengaged the language cortex, would the VWFA differentiate stimuli with different levels of orthographic regularities? Finally, what does the VWFA do prior to literacy, and what are the computations are this neural tissue capable of? Our hope is that this paper will clarify the role of the VWFA as a high-level visual, category specific regions for words, allowing research to move forward with understanding how this region develops its specialization as a fascinating example of uniquely human cognition in the brain.

Methods

Participants

Sixty-three healthy adults were recruited from The Ohio State University (OSU) and the local community. As part of an ongoing project exploring the relationship between functional organization of the human brain and connectivity, participants completed a battery of functional tasks in the scanner. In the current study, all 63 individuals completed at least one run of at least one of the functional tasks of interest: a VWFA localizer (N=56), a dynamic visual localizer (N=52), an auditory language task (N=48) and a spatial working memory task (N=57). The full sample of a given task was used to generate probabilistic maps for a given functional contrast. Critically, only individuals who completed two runs of the VWFA localizer and the dynamic localizer as well as at least one run of the other two tasks (N=37, 22 female, mean age: 24.57 age range: 18.01-55.66, 4 left-handed) were included in the main analysis, to ensure independence when defining the VTC category-selective regions and investigating their functional responses. When exploring the activation within the language and MD network, 3 individuals with only one run of the language task were further excluded. Subjects additionally completed a 10-minute resting state scan. All participants had either normal vision or corrected-to-normal vision and reported no neurological, neuropsychological, or developmental conditions. The study was approved by the Institutional Review Board at OSU and all participants provided written consent.

VWFA localizer (visual)

We used a VWFA localizer task[5] to define the VWFA. Briefly, static images of words, scrambled words, objects, and faces were presented in blocks. In our critical visual word condition, a single printed word in black text color was shown and all words were nouns. Each word was presented on a black grid inside a white rectangular box, and for the scrambled words condition, each block of this grid was randomly arranged, to preserve low-level visual features of words (lines, curves intersections). Line drawings of faces and objects (also presented on a grid) were implemented to control for the effects of visual similarity, complexity, and abstract semantic meaning. All stimuli were displayed on a white rectangular box and a gray grid was superimposed on the stimuli to match edges that are present in the scrambled words condition (all stimuli are available for download at https://www.zeynepsaygin.com/ZlabResources.html). For this localizer, participants were asked to perform a one-back task. In each block, 26 stimuli (including 2 repetitions) of the same category were presented one-by-one for 500ms followed with a 193ms ISI. Each run contains 4 experimental blocks for each condition and 3 fixation blocks and 2 runs of the VWFA localizer were collected for each participant.

Dynamic localizer (visual)

To define other VTC category-selective regions that are adjacent to the VWFA, we used a dynamic visual localizer task where participants were asked to passively view video clips (with natural colors) from faces, objects, bodies, natural scenes and scrambled objects[64]. Briefly, the face condition included faces of young children dancing and playing, the object condition included different moving objects (e.g., round block toys swinging), and the body condition showed different body parts (e.g., legs, hands; no faces) naturally moving. These clips were filmed on a black background. Scrambled objects were

generated by scrambling each frame of the object movie clip into a 15-by-15 grid. Scene conditions were recorded from a car window while driving through a suburb. Six 3-s video clips from the same category were shown in each block (i.e., 18s per block) and 2 blocks of each experimental condition were presented in each run (alternating with a palindromic manner) with 3 rest blocks with full-screen colors alternating at the beginning, middle and the end of each run. The order across participants was randomized.

Language task (auditory)

A language task[65] was used to investigate how the VWFA, as well as other VTC regions respond to the lexical and structural properties of auditory language. Participants listened to blocks of meaningful sentences (Sn), nonsense sentences (Ns, controlling for prosody but constructed from phonemically intact nonsense words), and texturized (degraded) speech (Tx, controlling for low-level auditory features). Each run consisted of 4 blocks of each condition and three 14 second fixation blocks. Each block contained three trials (6 seconds each); each trial ended with a visual queue to press a button. Language selective response is usually characterized by Sn>Tx or Sn>Ns, the former targets both linguistic and speech-related processing and the latter specifically targets high-level lexico-semantic information. As a comparison, we also used this task as a localizer to locate traditional frontotemporal language regions (Sn>Tx).

Spatial working memory task

To examine the effect of attentional load on the VTC regions, a spatial working memory task[66] with blocked easy and hard conditions was used. In each trial, participants viewed a grid of six boxes. For the easy condition, three blocks would flash blue sequentially, and participants were expected to remember which of the boxes in the grid had been colored. Two grids would then appear, one with the same pattern of blocks colored as the previous sequence (the match) and the other with a different pattern of blocks highlighted (no-match), and participants used a button press to indicate the match. For the hard condition, 3 pairs of 2 blocks would flash sequentially, increasing the amount of information the participant needed to attend to and remember. Each run consisted of 3 blocks of 2 trials per condition, separated by a 16 second rest blocks with a fixation on screen. The order of conditions within each block was randomized across participants. Hard condition elicits higher activation than the easy condition due to the attentional load, which is the neural signature for the domain-general multiple demand (MD) network[66,67]. Similarly, we define this frontoparietal MD network as a comparison to VTC regions.

Data acquisition and Preprocessing

All structural and functional MRI images were acquired on a Siemens Prisma 3T scanner (at the Center for Cognitive and Behavioral Brain Imaging (CCBBI) at the OSU) with a 32-channel phase array receiver head coil. All participants completed a whole-head, high resolution T1-weighted magnetization-prepared rapid acquisition with gradient echo (MPRAGE) scan (repetition time (TR) = 2300 ms, echo time (TE) = 2.9ms, voxel resolution = 1.00 mm^3). A semi-automated processing stream (recon-all from FreeSurfer) was used for structural MRI data processing. Major preprocessing steps include intensity correction, skull strip, surface co-registration, spatial smoothing, white matter and subcortical segmentation, and cortical parcellation. For functional tasks, the VWFA

localizer task was acquired with echo-planar imaging (EPI) sequence, TR=2000ms, TE=30ms, and 172 TRs. For N=50 participants, 2-mm isotropic voxels were acquired with 100 × 100 base resolution; 25 slices approximately parallel to the base of the temporal lobe to cover the entire ventral temporal cortex. For the rest of participants, slightly larger voxels (2.2-mm isotropic) were acquired for whole brain coverage (54 slices) with the same number of TRs to cover the whole brain. All other tasks were also collected with EPI sequence, but with TR=1000ms, TE=28ms, voxel resolution of 2x2x3 mm^3, 120 × 120 base resolution, 56 slices for the whole-brain coverage, and 244 TRs for the language localizer, 234 TRs for the dynamic localizer, and 400 TR for the spatial working memory localizer. To ensure that our results are not due to differences in scanning acquisition parameters, we collected an additional 2 runs of the VWFA task on 8 participants in our sample using identical parameters to the other 3 tasks, finding similar results (see **Supplementary Results 3**) for all analyses. We additionally used the same scan parameters to collect 1 run of resting-state scans, which lasted approximately 10 minutes (580 volumes).

All tasks were preprocessed in the same manner. Functional data were motion corrected (all timepoints aligned to the first timepoint in the scan) and timepoints with greater than 1mm total vector motion between consecutive timepoints were identified (and later included in first level GLM as nuisance regressors). Data were distortion corrected, detrended, spatially smoothed (3mm FWHM kernel for the static visual and 4mm for the other tasks and 4mm for the static visual task with identical scanning parameters), and then registered from functional space to anatomical space (using bbregister from FreeSurfer). A block design with a standard boxcar function (events on/off) was used to convolve the canonical HRF (standard g function, d=2.25 and t=1.25), and experimental conditions for each task were included as explanatory variables. Six orthogonalized motion measures from the preprocessing stage were included as additional nuisance regressors for each task individually. Resting-state data were motion corrected and smoothed (3mm FWHM kernel). Additionally, we applied linear interpolation for spikes above 0.5 mm and a bandpass filter (0.009-0.08 Hz). Resulting beta estimates, contrast maps and preprocessed resting-state data were registered from functional, volumetric space to anatomical, surface space (using FreeSurfer's mri_vol2surf with trilinear interpolation), and then to FsAverage surface space (using FreeSurfer's mri_surf2surf) for subsequent analyses.

Probabilistic map

Using the contrast maps from the GLM analysis for each subject, we created probabilistic activation maps for different contrasts of interest. Specifically, for each subject, the statistical map based on all available runs of a task from a given contrast was minimally thresholded at $p < 0.01$ (uncorrected) and resulting maps were binarized. The binarized maps of all subjects were then summed. This resulted in a map where the value at each vertex indicates the number of subjects who showed significant activation for that contrast

at that location. We presented any vertex that showed a consistent significant effect of the tested contrast across at least 5 subjects (which is approximately 10% of the participants - the exact percentile varied because the total number of subjects who completed each task was slightly different).

To compare the spatial localization of the VWFAs defined with different control conditions, probabilistic maps of words contrasting line-drawings of faces, objects, scrambled words, average of the three, and fixations were generated respectively with the VWFA localizer. In addition, to explore possible amodal language and attentional load activation within the VTC, we used the auditory language and spatial working memory tasks to generate probabilistic maps within the VTC for language (contrasts: sentences vs. texturized sounds and sentences vs. nonword sounds) and attentional load (contrast: hard vs. easy). Finally, as a sanity check, we also created probabilistic maps for face (faces vs. objects), object (objects vs. scrambled objects), body (bodies vs. objects) and scene (scenes vs. objects) using the dynamic localizer (**Figure S8**). Note that functional parcels from independent studies served as our search spaces to define the functional regions of interests (see **Definition of functional region of interest** below), and our probabilistic heat maps agreed well with most of the reference parcels with a few exceptions: lFFA parcel from Julian et al.[68] was too small to cover the hot spots of face-selective activation around the left fusiform gyrus that we observed in our probabilistic face vs. object map (**Figure S8**). Therefore, we flipped the rFFA parcel to the left hemisphere. Additionally, given the original lOFA and lFBA parcels were missing the hot spots in the face and body probabilistic maps, we further created new lOFA and lFBA parcels based on our thresholded probabilistic maps ($p<0.01$) from the dynamic localizer (face vs. object for lOFA and body vs. object for lFBA) (cluster function from FSL (5.0.10)). We chose vertices that were consistently significant across at least 5 subjects and reported results using updated lFFA, lOFA and lFBA parcels.

Definition of functional region of interest (fROI)

Given the individual variability in the precise location of high-level visual regions across individuals, we used the group-constrained subject-specific (GcSS)[65] method to define subject-specific fROIs. Previously defined atlas (parcels) that show the typical location of the regions across large samples of adults were used as our search spaces: for VTC parcels, VWFA was from Saygin et al. and all other parcels were from Julian et al.[68] (except for lOFA and lFBA, see above); language functional parcels were from Fedorenko et al.[65] (we used a recent updated version based on 220 participants) and MD functional parcels were original parcels from Fedorenko et al.[66] (we used a recent updated version based on 197 participants (see https://tinyurl.com/5e4tp67w for more details for language and MD parcels). All parcels were originally in volumetric spaces and moved to FsAverage surface with the same method mentioned above. For each subject's fROI, we selected the most significant/responsive voxels within a given search space using the statistical maps of the contrast of interest. **Table 3** showed the corresponding localizer tasks and contrasts we used to defined fROIs. Importantly, the most significant 150 voxels were selected and any voxels that responded significantly to multiple conditions were assigned to the contrast that they were most responsive to. After fROIs were defined with one run of data, an independent run of data was used to extract the percent signal change

(PSC) within each individual's fROIs to all conditions of interest (**Table 3**). Any PSC values that exceed +/- 3 standard deviation across subjects were marked as outliers and removed from subsequent analyses.

	fROIs	Task and contrasts to define
	The static VWFA localizer (conditions: words, scr.words, faces, objects)	
VTC fROIs	Visual Word Form Area (VWFA)[a]	W-All[b], words > average of scr. words, objects and faces
		W-Obj, words > objects
		W-F, words > faces
		W-SW, words > scr.words
		W-Fix, words > fixations
	The dynamic localizer (faces, objects, bodies, scenes, scr.objects)	
	Fusiform Face Area (FFA)	faces > average of objects, bodies, scenes, scr.objects
	Occipital Face Area (OFA)	faces > average of objects, bodies, scenes, scr.objects
	Retrosplenial Cortex (RSC)	scenes > average of objects, bodies, faces, scr.objects
	Parahippocampal Place Area (PPA)	scenes > average of objects, bodies, faces, scr.objects
	Posterior Fusiform Sulcus (PFS)	objects > average of scenes, bodies, faces, scr.objects
	Fusiform Body Area (FBA)	bodies > average of objects, scenes, faces, scr.objects
	The language task (auditory sentences, nonsense sounds, texturized sounds)	
	frontotemporal language fROIs[c]	Sentences > Texturized
	The spatial working memory task (Hard, Easy)	
	frontoparietal MD fROIs[d]	Hard > Easy

a. VWFA was defined with the static VWFA localizer. We explored different control conditions to defined the VWFA. The rest of VTC fROIs were defined using dynamic localizer
b. the VWFA defined with W-All was used in the main analysis
c. 6 bilateral language fROIs (3 frontal and 3 temporal)
d. 10 bilateral multiple-demand fROIs (7 frontal and 3 parietal)

Table 3. Tasks and contrasts to define the functional region of interests.

The language probabilistic maps (based on Sn-Tx) revealed that there were two clusters within the lVTC showing language sensitivity consistently across some participants (**Figure 5B**). Thus, to explore the functional properties of these clusters and compare them to the VWFA, we additionally defined two VTC language fROIs (mLang-VTC and aLang-VTC). Specifically, we first created the search spaces for the fROIs based on the Sn>Tx probabilistic map. Just as how we made the lOFA and lFBA parcels, the summed map across individuals was thresholded so that at least 5 subjects showed significant Sn>Tx at a given vertex, which resulted in two continuous parcels at the medial lVTC (mLang-VTC, anterior to the fusiform gyrus) and superior anterior lVTC (aLang-VTC). The same GcSS method was then used to define subject-specific mLang-VTC and aLang-VTC fROIs by intersecting each individual's Sn>Tx statistical map (thresholded at $p<0.005$) with the medial and anterior parcels we created above. Note that we identified these two fROIs along with the other category-selective VTC regions so that we could assign vertices to the condition to which they showed the strongest response (in order to achieve better spatial specificity, e.g. the mLang-VTC parcel overlapped with the VWFA parcel but the fROIs selected within these parcels showed the highest responses to the category of interest).

Connectivity between left VTC voxels and language/MD networks

For each voxel within the left VTC, we calculated its resting-state functional connectivity (FC) with the language and MD networks identified above. Specifically, Pearson's correlation between each voxel's time-course during the resting-state scan and the averaged time-course for each of the language or MD networks were calculated. The resulting values were Fisher z-transformed. We then ran a regression analysis to see how each voxel's activations to different visual categories contribute to their connectivity to language or MD functional regions. Beta values were reported for each condition and paired t-tests were used to compare beta values across visual conditions.

Statistical analysis

Do different control conditions result in functionally and/or spatially consistent VWFAs? We compared VWFAs with different contrasts from three aspects: 1) Paired t-test (two-tailed) was used to compare VWFA's responses to visual words vs. other conditions in the VWFA localizer (**Table S1**). 2) For VWFA defined with each of the contrast, Dice coefficient was used to measure the overlap between the resulting VWFA from two runs. Paired t-tests were used to quantify the reliability of the word-selective responses across time using different control conditions. 3) We further quantified the pair-wise overlap between VWFAs defined with different contrasts using Dice coefficient (**Figure S2B**). Then to examine whether the W-All defined VWFA showed statistically greater overlap with other contrasts, we compared the W-All against each of the other contrast based on their averaged Dice coefficient between each of them with the rest of the three contrasts. For example, when comparing W-All and W-Obj, paired t-test was used to compare their respective overlap with W-F, W-SW and W-Fix.

To thoroughly examine the category-selective responses and functional profile of the VWFA as well as other adjacent VTC regions, the response of each fROI to its preferred condition was compared to all other conditions (paired, two-sided t-tests, corrected using Bonferroni-Holm multiple comparisons for the number of fROIs per hemisphere). All fROIs were also tested for their preferred and any unexpected non-preferred category selectivity. Specifically, for their preferred category, the difference between the response to the preferred condition and the average of all other conditions was calculated; for non-preferred category, we calculated the difference between the condition of interest vs. all other conditions after excluding the response to their preferred category. For example, the VWFA was tested for a preference to objects (average of dynamic and static objects – average of faces (dynamics and static), bodies, scenes, scrambled words, scrambled objects divided by the sum of all conditions). The resulting values were tested against zero using one-tailed one-sample t-test and corrected using Bonferroni-Holm multiple comparisons for the number of fROIs in each hemisphere. One-way rmANOVA and follow-up post-hoc t-test were used to compare word selectivity across fROIs. All ANOVAs were completed in RStudio (V 1.4.1717) using the anova_test function, with subject number and fROI (or condition where applicable) as within-subject factors.

To explore if any significant category-selective lVTC regions also showed language or attentional load effects, we used two-way rmANOVAs to compare conditions (language:

sentences, nonsense; attention: hard, easy) by fROIs. We included any fROI that showed preferences for its preferred category (all visual fROIs except the FBA). Given that some VTC fROIs showed sensitivity to auditory language and attentional load, we wanted to examine if they were as responsive to language and attention as the canonical language and multiple network fROIs would be. We first checked the preference of these language and MD fROIs using paired t-tests (language: sentences vs nonsense sounds; attention: hard vs easy) with Bonferroni-Holm multiple comparisons by hemisphere. Given that all language regions were significantly selective to language, we averaged the PSC values across the frontal fROIs and temporal fROIs separately. We used a two-way ANOVA to examine differences between frontal language, temporal language, and VTC fROIs by conditions (Sn, Ns). Similarly, all MD fROIs showed expected hard > easy attentional load effect and were therefore averaged to create one representative frontal and one representative parietal fROI); then a two-way rmANOVA was conducted to compare frontal and parietal MD fROIs vs. VTC regions by condition (Hard, Easy). Finally, rmANOVA was used to examine the visual category specificity of aLang-VTC and mLang-VTC, as well as to compare their language responsiveness (Sn and Ns) to the lVWFA and the canonical language network.

References

1. Cohen, L. *et al.* Language-specific tuning of visual cortex? Functional properties of the Visual Word Form Area. *Brain* **125**, 1054–1069 (2002).
2. McCandliss, B. D., Cohen, L. & Dehaene, S. The visual word form area: expertise for reading in the fusiform gyrus. *Trends Cogn. Sci.* **7**, 293–299 (2003).
3. Dehaene, S., Cohen, L., Morais, J. & Kolinsky, R. Illiterate to literate: behavioural and cerebral changes induced by reading acquisition. *Nat. Rev. Neurosci.* **16**, 234–244 (2015).
4. Dehaene-Lambertz, G., Monzalvo, K. & Dehaene, S. The emergence of the visual word form: Longitudinal evolution of category-specific ventral visual areas during reading acquisition. *PLOS Biol.* **16**, e2004103 (2018).
5. Saygin, Z. M. *et al.* Connectivity precedes function in the development of the visual word form area. *Nat. Neurosci.* **19**, 1250–1255 (2016).
6. Baker, C. *et al.* Visual word processing and experiential origins of functional selectivity in human extrastriate cortex. *Proc Natl Acad Sci USA* (2007).
7. Kubota, E. C., Joo, S. J., Huber, E. & Yeatman, J. D. Word selectivity in high-level visual cortex and reading skill. *Dev. Cogn. Neurosci.* **36**, 100593 (2019).
8. Cai, Q., Lavidor, M., Brysbaert, M., Paulignan, Y. & Nazir, T. A. Cerebral Lateralization of Frontal Lobe Language Processes and Lateralization of the Posterior Visual Word Processing System. *J. Cogn. Neurosci.* **20**, 672–681 (2008).
9. Dehaene, S. *et al.* How Learning to Read Changes the Cortical Networks for Vision and Language. *Science* **330**, 1359–1364 (2010).

10. Zhan, M., Pallier, C., Dehaene, S. & Cohen, L. Does the visual word form area split in bilingual readers? A millimeter-scale 7T fMRI study. 2022.11.10.515773 Preprint at https://doi.org/10.1101/2022.11.10.515773 (2022).
11. Rueckl, J. G. *et al.* Universal brain signature of proficient reading: Evidence from four contrasting languages. *Proc. Natl. Acad. Sci.* **112**, 15510–15515 (2015).
12. Roberts, D. J. *et al.* Efficient Visual Object and Word Recognition Relies on High Spatial Frequency Coding in the Left Posterior Fusiform Gyrus: Evidence from a Case-Series of Patients with Ventral Occipito-Temporal Cortex Damage. *Cereb. Cortex* **23**, 2568–2580 (2013).
13. Ben-Shachar, M., Dougherty, R. F., Deutsch, G. K. & Wandell, B. A. Differential Sensitivity to Words and Shapes in Ventral Occipito-Temporal Cortex. *Cereb. Cortex* **17**, 1604–1611 (2007).
14. Vogel, A. C., Petersen, S. E. & Schlaggar, B. L. The Left Occipitotemporal Cortex Does Not Show Preferential Activity for Words. *Cereb. Cortex* **22**, 2715–2732 (2012).
15. Xue, G. & Poldrack, R. A. The Neural Substrates of Visual Perceptual Learning of Words: Implications for the Visual Word Form Area Hypothesis. *J. Cogn. Neurosci.* **19**, 1643–1655 (2007).
16. Nestor, A., Behrmann, M. & Plaut, D. C. The Neural Basis of Visual Word Form Processing: A Multivariate Investigation. *Cereb. Cortex* **23**, 1673–1684 (2013).
17. Plaut, D. C. & Behrmann, M. Complementary neural representations for faces and words: A computational exploration. *Cogn. Neuropsychol.* **28**, 251–275 (2011).

18. Dehaene, S. & Cohen, L. Cultural Recycling of Cortical Maps. *Neuron* **56**, 384–398 (2007).
19. Feng, X., Monzalvo, K., Dehaene, S. & Dehaene-Lambertz, G. Evolution of reading and face circuits during the first three years of reading acquisition. *NeuroImage* **259**, 119394 (2022).
20. Chen, L. *et al.* The visual word form area (VWFA) is part of both language and attention circuitry. *Nat. Commun.* **10**, 5601 (2019).
21. Price, C. J. & Devlin, J. T. The Interactive Account of ventral occipitotemporal contributions to reading. *Trends Cogn. Sci.* **15**, 246–253 (2011).
22. Li, J., Osher, D. E., Hansen, H. A. & Saygin, Z. M. Innate connectivity patterns drive the development of the visual word form area. *Sci. Rep.* **10**, 18039 (2020).
23. Stevens, W. D., Kravitz, D. J., Peng, C. S., Tessler, M. H. & Martin, A. Privileged Functional Connectivity between the Visual Word Form Area and the Language System. *J. Neurosci.* **37**, 5288–5297 (2017).
24. Vogel, A. C., Miezin, F. M., Petersen, S. E. & Schlaggar, B. L. The Putative Visual Word Form Area Is Functionally Connected to the Dorsal Attention Network. *Cereb. Cortex* **22**, 537–549 (2012).
25. Kim, J. S., Kanjlia, S., Merabet, L. B. & Bedny, M. Development of the Visual Word Form Area Requires Visual Experience: Evidence from Blind Braille Readers. *J. Neurosci.* **37**, 11495–11504 (2017).
26. Gerrits, R., Van der Haegen, L., Brysbaert, M. & Vingerhoets, G. Laterality for recognizing written words and faces in the fusiform gyrus covaries with language dominance. *Cortex* **117**, 196–204 (2019).

27. Cohen, L., Salondy, P., Pallier, C. & Dehaene, S. How does inattention affect written and spoken language processing? *Cortex* **138**, 212–227 (2021).
28. Yoncheva, Y. N., Zevin, J. D., Maurer, U. & McCandliss, B. D. Auditory Selective Attention to Speech Modulates Activity in the Visual Word Form Area. *Cereb. Cortex* **20**, 622–632 (2010).
29. Cohen, L. *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).
30. Bouhali, F. *et al.* Anatomical Connections of the Visual Word Form Area. *J. Neurosci.* **34**, 15402–15414 (2014).
31. Yeatman, J. D., Rauschecker, A. M. & Wandell, B. A. Anatomy of the visual word form area: Adjacent cortical circuits and long-range white matter connections. *Brain Lang.* **125**, 146–155 (2013).
32. Nordt, M. *et al.* Learning to Read Increases the Informativeness of Distributed Ventral Temporal Responses. *Cereb. Cortex* **29**, 3124–3139 (2019).
33. Glezer, L. S. & Riesenhuber, M. Individual Variability in Location Impacts Orthographic Selectivity in the 'Visual Word Form Area'. *J. Neurosci.* **33**, 11221–11226 (2013).
34. Behrmann, M. & Plaut, D. C. Distributed circuits, not circumscribed centers, mediate visual recognition. *Trends Cogn. Sci.* **17**, 210–219 (2013).
35. Kay, K. N. & Yeatman, J. D. Bottom-up and top-down computations in word- and face-selective cortex. *eLife* **6**, e22341 (2017).

36. Szwed, M. *et al.* Specialization for written words over objects in the visual cortex. *NeuroImage* **56**, 330–344 (2011).
37. Price, C. J. & Devlin, J. T. The myth of the visual word form area. *NeuroImage* **19**, 473–481 (2003).
38. Grotheer, M., Yeatman, J. & Grill-Spector, K. White matter fascicles and cortical microstructure predict reading-related responses in human ventral temporal cortex. *NeuroImage* **227**, 117669 (2021).
39. White, A. L., Kay, K. N., Tang, K. A. & Yeatman, J. D. Engaging in word recognition elicits highly specific modulations in visual cortex. *Curr. Biol.* **33**, 1308-1320.e5 (2023).
40. Lerma-Usabiaga, G., Carreiras, M. & Paz-Alonso, P. M. Converging evidence for functional and structural segregation within the left ventral occipitotemporal cortex in reading. *Proc. Natl. Acad. Sci.* **115**, E9981–E9990 (2018).
41. Brunswick, N., McCrory, E., Price, C. J., Frith, C. D. & Frith, U. Explicit and implicit processing of words and pseudowords by adult developmental dyslexics: A search for Wernicke's Wortschatz? *Brain* **122**, 1901–1917 (1999).
42. Vinckier, F. *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).
43. Binder, J. R., Medler, D. A., Westbury, C. F., Liebenthal, E. & Buchanan, L. Tuning of the human left fusiform gyrus to sublexical orthographic structure. *NeuroImage* **33**, 739–748 (2006).

44. Kanwisher, N., McDermott, J. & Chun, M. M. The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *J. Neurosci.* **17**, 4302–4311 (1997).
45. Kanwisher, N. Functional specificity in the human brain: A window into the functional architecture of the mind. *Proc. Natl. Acad. Sci.* **107**, 11163–11170 (2010).
46. Hasson, U., Levy, I., Behrmann, M., Hendler, T. & Malach, R. Eccentricity Bias as an Organizing Principle for Human High-Order Object Areas. *Neuron* **34**, 479–490 (2002).
47. Mahon, B. Z. & Caramazza, A. What drives the organization of object knowledge in the brain? *Trends Cogn. Sci.* **15**, 97–103 (2011).
48. Martin, A. Shades of Déjerine—Forging a Causal Link between the Visual Word Form Area and Reading. *Neuron* **50**, 173–175 (2006).
49. Fedorenko, E., Behr, M. K. & Kanwisher, N. Functional specificity for high-level linguistic processing in the human brain. *Proc. Natl. Acad. Sci.* **108**, 16428–16433 (2011).
50. Wang, X. *et al.* Representational similarity analysis reveals task-dependent semantic influence of the visual word form area. *Sci. Rep.* **8**, 3047 (2018).
51. Starrfelt, R. & Gerlach, C. The Visual What For Area: Words and pictures in the left fusiform gyrus. *NeuroImage* **35**, 334–342 (2007).
52. Mano, Q. R. *et al.* The Role of Left Occipitotemporal Cortex in Reading: Reconciling Stimulus, Task, and Lexicality Effects. *Cereb. Cortex* **23**, 988–1001 (2013).
53. Moore, C. J. & Price, C. J. Three Distinct Ventral Occipitotemporal Regions for Reading and Object Naming. *NeuroImage* **10**, 181–192 (1999).

54. Phillips, J. A., Humphreys, G. W., Noppeney, U. & Price, C. J. The neural substrates of action retrieval: An examination of semantic and visual routes to action. *Vis. Cogn.* **9**, 662–685 (2002).
55. Grill-Spector, K., Kushnir, T., Hendler, T. & Malach, R. The dynamics of object-selective activation correlate with recognition performance in humans. *Nat. Neurosci.* **3**, 837–843 (2000).
56. Thorpe, S., Fize, D. & Marlot, C. Speed of processing in the human visual system. *Nature* **381**, 520–522 (1996).
57. Binder, J. R., Desai, R. H., Graves, W. W. & Conant, L. L. Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. *Cereb. Cortex* **19**, 2767–2796 (2009).
58. Purcell, J. J., Shea, J. & Rapp, B. Beyond the visual word form area: The orthography–semantics interface in spelling and reading. *Cogn. Neuropsychol.* **31**, 482–510 (2014).
59. Jobard, G., Crivello, F. & Tzourio-Mazoyer, N. Evaluation of the dual route theory of reading: a metanalysis of 35 neuroimaging studies. *NeuroImage* **20**, 693–712 (2003).
60. Thompson-Schill, S. L., Aguirre, G. K., Desposito, M. & Farah, M. J. A neural basis for category and modality specificity of semantic knowledge. *Neuropsychologia* **37**, 671–676 (1999).
61. Wilson, S. M. *et al.* Transient aphasias after left hemisphere resective surgery. *J. Neurosurg.* **123**, 581–593 (2015).

62. Dehaene, S., Cohen, L., Sigman, M. & Vinckier, F. The neural code for written words: a proposal. *Trends Cogn. Sci.* **9**, 335–341 (2005).
63. Changizi, M. A., Zhang, Q., Ye, H. & Shimojo, S. The Structures of Letters and Symbols throughout Human History Are Selected to Match Those Found in Objects in Natural Scenes. *Am. Nat.* **167**, E117–E139 (2006).
64. Pitcher, D., Walsh, V. & Duchaine, B. The role of the occipital face area in the cortical face perception network. *Exp. Brain Res.* **209**, 481–493 (2011).
65. Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S. & Kanwisher, N. New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. *J. Neurophysiol.* **104**, 1177–1194 (2010).
66. Fedorenko, E., Duncan, J. & Kanwisher, N. Broad domain generality in focal regions of frontal and parietal cortex. *Proc. Natl. Acad. Sci.* **110**, 16616–16621 (2013).
67. Blank, I., Kanwisher, N. & Fedorenko, E. A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *J. Neurophysiol.* **112**, 1105–1118 (2014).
68. Julian, J. B., Fedorenko, E., Webster, J. & Kanwisher, N. An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *NeuroImage* **60**, 2357–2364 (2012).