Evidence for hierarchical representations of written and spoken words from an open-science human neuroimaging dataset

Suneel Banerjee[1,2], Kimberly Jin[1,3], Plamen Nikolov[1], Philip Cho[1,4], Vishnu R. Pendri[1,5], Lillian Chang[1], Srikanth R. Damera[1,6], Maximilian Riesenhuber[1,7]

1. Department of Neuroscience, Georgetown University Medical Center, Washington, DC, United States
2. University of California San Diego School of Medicine, La Jolla, CA, United States
3. University of Massachusetts Chan School of Medicine, Worcester, MA, United States
4. Washington University School of Medicine, St. Louis, MO, United States
5. Department of Biology, University of Rochester, Rochester, NY, United States
6. Developing Brain Institute, Children's National Medical Center, Washington, DC, United States
7. Center for Neuroengineering, Georgetown University, Washington, DC, United States

Abstract

Reading and speech recognition rely on multi-level processing that builds from basic visual or sound features to complete word representations, yet details of these processing hierarchies (in particular those for spoken words) are still poorly understood. We re-analyzed the functional magnetic resonance imaging (fMRI) data provided in the Mother Of all Unification Studies (MOUS) open-science dataset by using parametric regressions of word frequency and sublexical unit (bigram or syllable) frequency during reading and speech listening tasks in order to elucidate lexical processing hierarchies in the visual and auditory modalities. We first validated our approach in the written word domain, where the technique identified significant correlations for word frequency in the left mid-fusiform cortex (at the location of the Visual Word Form Area) with a left occipital region tracking bigram frequency, compatible with prior reports. During listening, low-frequency spoken words elicited greater responses in a left mid-superior temporal region consistent with the recently-described Auditory Word Form Area (AWFA), while a more posterior region of the superior temporal gyrus was sensitive to syllable frequency. Activation in the left inferior frontal gyrus correlated with both written and spoken word frequency. These findings demonstrate parallel hierarchical organizations in the anteroventral visual and auditory streams, with modality-specific lexica and upstream sublexical representations that converge in higher-order language areas.

1. Introduction

The neural substrates underlying reading and speech recognition are two remarkable instances of hierarchical processing in the human brain. Reading leverages the ventral visual ("what") stream. This stream is thought to be organized in a posterior-to-anterior hierarchy of neural selectivity increasingly complex shape features (Hubel et al., 1977; Kravitz et al., 2013). Applied to orthography, this 'simple-to-complex' model predicts that increasingly complex sublexical representations (e.g., those of bigrams and trigrams) emerge in progressively anterior regions in the ventral occipital-temporal cortex (vOT) (Vinckier et al., 2007), culminating with full lexical representations in the Visual Word Form Area (VWFA) in the left mid-fusiform cortex(Cohen et al., 2002; Glezer et al., 2009).

Motivated by findings of lower responses to familiar vs. unfamiliar objects in fMRI (Chao et al., 2002), early studies tested the hypothesis that, similarly, brain areas containing lexical orthographic representations should respond less to familiar (i.e., high frequency) than to unfamiliar (i.e., low frequency) words. This prediction was confirmed in the VWFA (Kronbichler et al., 2004). In parallel, word frequency has been shown to modulate reaction times in lexical decision tasks in both sensory modalities (Goh et al., 2009; Keuleers, Brysbaert, et al., 2010; NEW et al., 2007) Subsequent neuroimaging and human electrocorticography (ECoG) studies have provided further evidence for the frequency effect for written words in the VWFA (Fiez et al., 1999; Graves et al., 2010; Hauk et al., 2008; Joubert et al., 2004; Woolnough et al., 2021). The existence of a 'visual dictionary' in the VWFA shaped by experience was also demonstrated through fMRI-rapid adaptation techniques (Glezer et al., 2009, 2015) and subsequently confirmed by ECoG (Glezer et al., 2015; Lochy et al., 2018; Woolnough et al., 2021).

Other studies have applied a similar approach to probe sublexical representations using bigram frequencies (Graves et al., 2010; Woolnough et al., 2021) based on the hypothesis that lifelong exposure to written language leads to adaptive learning processes that guide the tuning of feature detectors in the ventral visual stream for frequently-encountered fragments of words (e.g., "GH" vs "GZ" in English) (Vinckier et al., 2007). These studies found significant inverse correlations of blood oxygen level-dependent (BOLD) responses with bigram frequency in a region posterior to the VWFA (later labeled the "occipital word form area," OWFA) (Strother et al., 2016), thus demonstrating how the class of orthographic stimuli instantiates the general 'simple-to-complex' hierarchical processing principle of the ventral visual stream (Riesenhuber & Poggio, 1999). Just as in the case of the VWFA, this finding of sublexical selectivity in the OWFA obtained with the frequency correlation approach was also validated and confirmed through other approaches using fMRI and ECoG (Glezer et al., 2015; Seghier et al., 2012; Strother et al., 2016; Woolnough et al., 2021).

In contrast to hierarchical (i.e., lexical and sublexical) representations for written words in the visual system, the representations of spoken words in the brain's auditory system are much less well understood. As a general principle, the auditory system, like the visual system, is also thought to be organized into two streams in the cortex: The antero-ventral stream (aka the "what" stream), projects from primary auditory cortex in Heschl's gyrus to the inferior frontal gyrus via the mid and anterior superior temporal gyrus and is specialized for identifying acoustic stimuli, including spoken words. This stream is likewise organized along a simple-to-complex hierarchy (Kell et al., 2018; Norman-Haignere et al., 2022; Rauschecker & Scott, 2009), wherein progressively anterior neuronal populations in the superior temporal cortex integrate over longer timescales, permitting the representation of increasingly complex auditory "objects." Using an fMRI-rapid adaptation approach very similar to our earlier study of lexical representations in the VWFA (Glezer et al., 2009, 2015), we recently provided evidence of neural representations for individual spoken words in the mid-anterior superior temporal gyrus (STG) (Damera et al., 2023), the site of the putative "Auditory Word Form Area" (AWFA) (Cohen & Dehaene, 2004; DeWitt & Rauschecker, 2012), sharpening the parallels between the brain's encoding schemes for written and spoken words

Despite its successes in elucidating hierarchical representations for written words, word and word-part frequency have not previously been used to probe lexical or sublexical representations in the auditory system. If the recently-elucidated AWFA stores representations of spoken words, we would hypothesize stronger responses to infrequent vs. frequent auditory words (i.e., an inverse correlation between BOLD response level and word frequency), in analogy to the response profile of the VWFA. In addition (in analogy with the neural representation of bigrams posterior to the VWFA in the ventral visual stream), we would predict the existence of an area showing a frequency effect for spoken word-parts, such as syllables, posterior to the AWFA (a "mid-anterior auditory word form area", mAWFA, as it were), yet anterior to Heschl's gyrus, which performs early acoustic processing. Confirming these hypotheses would show intriguing parallels between written and spoken word processing systems in the brain and further our understanding of hierarchical processing as a general principle across sensory modalities.

These are the predictions we tested in the present study. A notable feature of our approach was its use of the Mother-Of-all Unification Studies (MOUS) dataset, a large, open-science human neuroimaging dataset that contains fMRI data collected as subjects completed single-word reading and speech listening tasks (n = 204, 102 reading and 102 speech listening) (Schoffelen et al., 2019). We re-analyzed this dataset by employing parametric regression to identify brain regions whose activity was inversely related to the frequency of words and their sublexical components, both in the written and auditory domains. Specifically, we first validated our approach using the

written word/reading dataset (which identified lexical selectivity in the VWFA and sublexical selectivity in the OWFA regions, in line with the literature), and then applied the same approach to probe the hierarchical representation of spoken words in the brain.

Our analyses revealed strong parallels between lexical processing hierarchies in the auditory and written domains, providing evidence for separate word recognition hierarchies for speech and reading as well as downstream convergence areas in the prefrontal cortex. We found that word frequency regressors localize the VWFA during reading tasks and the AWFA during speech listening. Sublexical frequency regressors (constructed from bigram frequency and syllable frequency in the visual and auditory cases respectively) revealed peaks posterior to these lexical regions, identifying sites of intermediate orthographic and phonetic selectivity, respectively. These results provide support for lexical representations in the newly-reported AWFA and syllable representations posterior to it in the STG, situating the AWFA at the top of a word-recognition hierarchy in the auditory "what" pathway and suggesting a role analogous to that of the VWFA in the visual ventral stream.

2. Methods

2.1. Dataset

We utilized the Mother-Of-all Unification Studies (MOUS), which contains fMRI (task and resting-state), DWI, and MEG data from 204 subjects (100 males, mean age 22, range 18-33 years). Text in this section is largely reproduced from Schoffelen et al., 2019, in which the dataset was first described. Half (n = 102) of the subjects completed a reading task which consisted of Dutch sentences and scrambled word lists presented one word at a time, while the other half completed a speech listening task that consisted of spoken versions of the same linguistic stimuli. The full stimulus set contained 360 sentences, from which 548 scrambled sentences were also derived for a total of 908 linguistic stimuli. Each subject was presented with 60 scrambled word lists ('Woorden') and 60 sentences ('Zinnen') while undergoing fMRI scanning. Subjects were not exposed to scrambled word lists corresponding to sentences they had been presented with. Word lists and sentences varied between 9 and 15 words in length. Sentences or word lists were presented word-by-word with a mean duration of 351 ms for each word (minimum of 300 ms and maximum of 1400 ms). Specifically, the visual presentation rate of the stimuli was determined in relation to the duration of the audio recording of spoken versions of the sentences and the word lists ('audiodur' variable in the MOUS dataset), taking into account both the number of letters ('sumnletters') and words ('nwords') in the whole sentence and the number of letters within each word ('nletters'). The duration of a single word (in ms) was determined by the MOUS study team as:

$$(\text{nletters}/\text{sumnletters}) * (\text{audiodur} + 2000 - 150 * \text{nwords})$$

The auditory versions of the stimuli were recorded by a female native Dutch speaker. The word lists were pronounced with a neutral prosody and a clear pause in between each word. In order to check for compliance, 20% of the trials were followed by a 'Yes'/'No' question about the content of the previous sentence/word list. Half of the questions on the sentences addressed the content of the sentence (e.g. *"Did grandma give a cookie to the girl*?") whereas the other half, and all the questions about the word lists, addressed one of the main content words (e.g., "*Was the word 'grandma' mentioned?"*).

2.2. MRI Data Acquisition (Schoffelen et al., 2019)

The data were acquired with a SIEMENS Trio 3 T scanner using a 32-channel head coil. The collection order of the data analyzed here was as follows: 1) resting-state fMRI, 2) task-based fMRI, pause, 3) structural image.

2.2.1. Task fMRI

T2*-weighted functional echo planar blood oxygenation level dependent (EPI-BOLD) data were acquired during the task. A single echo 2D ascending slice acquisition sequence (with partial brain coverage) was implemented with the following specifications: Volume TR = 2.00 s, TE = 35 ms, 90 degree flip-angle, 29 oblique slices (position: L.076 A6.4 H16.6, orientation: T > C-5.8 > S-1.2, Phase encoding direction: A >> P, rotation: −0.80 degrees), slice-matrix size (base resolution) = 64 × 64, slice thickness = 3.0 mm, slice gap 0.5 mm, FOV = 224 mm, voxel size ($x = FOV_x/N_x$, $y = FOV_y/N_y$, z = slice thickness, anisotropic voxel size = 3.5 × 3.5 × 3.0 mm).

2.2.2. Structural imaging

A T1-weighted magnetization-prepared rapid gradient-echo (MP-RAGE) pulse sequence was used for the structural images, with the following parameters: volume TR = 2300 ms, TE = 3.03 ms, 8 degree flip-angle, 1 slab, slice-matrix size = 256 × 256, slice thickness = 1 mm, field of view = 256 mm, isotropic voxel-size = 1.0 × 1.0 × 1.0 mm. A vitamin-E capsule was placed as fiducial behind the right ear to allow a visual identification of left-right consistency.

2.2.3. *Anatomical data preprocessing (auto-generated)*

A total of 200 T1-weighted (T1w) images were found within the input BIDS dataset. All of them were bias-field corrected and skull-stripped with ANTS 2.2.0 (Avants et al., 2008; Tustison et al., 2010). Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using FSL 5.0.9 (Bernal et al., 2021). A T1w-reference map was computed using FreeSurfer 6.0.1 (Reuter et al., 2010) Volume-based spatial normalization to two standard spaces (MNI152NLin6Asym and MNI152NLin2009cAsym) was performed with ANTs 2.2.0) (Chung et al., 2010; Evans et al., 2012).

2.2.4. *Functional data preprocessing (auto-generated)*

Results in this manuscript come from data preprocessed with fMRIPrep 20.1.1 (Esteban et al., 2019), which is based on Nipype 1.5.0 (Gorgolewski et al., 2011). Due to missing or invalid anatomical (T1-weighted) or functional scans, 4 subjects (1 from the auditory group, A2091, and 3 from the visual group, V1050, V1109, V1071) were excluded from all further analyses. For each of the single BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed: First, a reference volume and its skull-stripped version were generated using fMRIPrep. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) were estimated before any spatiotemporal filtering using FSL 5.0.9 (Jenkinson et al., 2002). BOLD runs were slice-time corrected using AFNI 20160207 (Cox & Hyde, 1997). Susceptibility distortion correction (SDC) was omitted. The BOLD reference was then co-registered to the T1w reference using FSL 5.0.9 (Greve & Fischl, 2009; Jenkinson & Smith, 2001). The BOLD time-series (including slice-timing correction) were resampled onto their original, native space by applying the transforms to correct for head-motion. The BOLD time-series were also resampled into standard space, generating a preprocessed BOLD run in MNI152NLin6Asym space. Functional volumes were resampled to isotropic 2 × 2 × 2 mm voxels and smoothed by 6mm (FWHM) in each direction. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD) (Jenkinson et al., 2002; Power et al., 2011), DVARS (D referring to temporal derivative of timecourses, VARS referring to RMS variance over voxels) and three region-wise global signals. FD and DVARS are calculated for each functional run (Power et al., 2011). The three global signals were extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction ("CompCor") (Behzadi et al., 2007). The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file (Satterthwaite et al., 2013). Frames that exceeded a threshold of 0.4 mm FD or 1.5 standardized DVARS were annotated as motion outliers. Volumetric resamplings were performed using ANTs (Lanczos, 1964) and surface resamplings were performed using FreeSurfer. Many internal operations of fMRIPrep use Nilearn* 0.6.2, mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in fMRIPrep's documentation (https://fmriprep.org/en/latest/workflows.html).

2.3. Word, bigram, and syllable frequency calculations

We sourced the frequencies of the Dutch words used in the fMRI tasks of the MOUS dataset from the SUBTLEX-NL Dutch language database (Keuleers, Brysbaert, et al., 2010). This database employs a text corpus derived from film subtitles, thus approximating everyday word usage statistics more accurately than other written corpora. To construct our word frequency regressors, we used the provided 'Zipf' frequency, which is calculated as follows:

$$\text{Zipf} = \log_{10}\left(\frac{\text{FREQcount} + 1}{44.106}\right) + 3 \quad (1)$$

Where *FREQcount* represents the number of times a word appears in the SUBTLEX-NL text corpus. Log-transformed measures of word frequency have been shown to better explain variance in reaction time during lexical decision tasks than linear measures (Brysbaert et al., 2011, 2018; Keuleers, Diependaele, et al., 2010). Most studies of the word frequency effect have employed log-transformed word frequency values from SUBTLEX or CELEX lexical databases (Graves et al., 2010; Kronbichler et al., 2004; Woolnough et al., 2021).

We calculated individual bigram frequency by summing the FREQcounts of all words containing a certain two-letter combination, then taking the log of this value (Graves et al., 2010; Kronbichler et al., 2004).

$$\text{Bigram Frequency}(b) = log_{10}\left(1 + \sum_{n \in W_b} FREQcount_n\right) \quad (2)$$

Where b is the bigram of interest (e.g., “th”, “an”), W_b is the set of word indices whose orthographic form contains *b*, and $FREQcount_n$ is the value of FREQcount for the nth word. For each word that appeared in the fMRI tasks, we determined the mean, minimum and maximum bigram frequency found in a word.

To calculate syllable frequency, every native Dutch word found in the SUBTLEX-NL database was divided into its constituent syllables. We applied a novel, validated deep learning framework (Lathouwers et al., 2025), which slides a 5-letter window across each word, summarizes the local context with a convolutional neural network, then applies a bidirectional long short-term memory (LSTM) recurrent neural network followed by a conditional random field to decide, for every character, whether it begins a new syllable. Following the same process used to calculate bigram frequency (Equation 2), we then summed the frequencies (FREQcount) of all words containing each unique syllable and log-transformed this sum (base 10) to obtain that syllable’s frequency. For each word that appeared in the fMRI tasks, we determined the mean, minimum and maximum syllable frequency found in a word.

2.4. fMRI analyses

We used SPM12 (Ashburner et al., 2014) in MATLAB R2022b (The MathWorks Company) for all fMRI analyses. Negative Zipf word frequency was mean-centered and treated as a parametric modulator of the main effect of word onset within a general

linear model (GLM), based on previous findings that increasing word frequency is associated with decreasing brain activation in the VWFA (Kronbichler et al., 2004).

For the reading task, word onset times were provided as part of the dataset. For the speech listening task, we used the Montreal ForcedAligner to generate word onset times from the provided stimulus audio files (McAuliffe et al., 2017). AlignTool was used to force-align auditory task stimuli and determine word onset and offset times (Schillingmann et al., 2018). Matching written transcripts were available within the MOUS dataset for 900 of the auditory sound files, and the remaining 8 audio files were manually transcribed. The AlignTool function 'Extract On/Offsets' returned automated onset and offset times, which were manually investigated for accuracy. No corrections of these times were required.

Motion covariates (rotation and translation along each axis) were used as nuisance regressors. When examining correlations between brain activity and word frequency during reading, we controlled for word length (in number of letters) by including it as a parametric regressor of no interest (i.e., with a weighting of zero in the first-level contrast). When examining correlations between brain activity and word frequency during speech listening, we similarly included a control for word duration in milliseconds. Each contrast was aimed at identifying a correlation between brain activity and a specific lexical or sublexical parameter. We thus refer to them as 'correlation contrasts' (word frequency during reading, word frequency during speech listening, bigram frequency during reading, and syllable frequency during speech listening).

We used the same parametric regression technique to explore the effect of bigram and syllable frequency. We used the same onset times (corresponding to those of word presentations), but instead of using Zipf word frequency, which reflects the frequency of the word in the corpus, we tested the mean, maximum (Fig. S2, S3), or minimum (Fig. 2, 5) bigram or syllable frequency found in a word as a parametric modulator in separate contrasts of brain activity during reading or speech listening, respectively.

2.5. Juxtaposition analyses

Unthresholded statistical parametric t-maps were converted into z-maps using MATLAB's tcdf function. For each voxel, the cumulative probability of the observed t statistic was computed under the Student's t-distribution with the residual degrees of freedom estimated from the design matrix. This probability was then mapped to the equivalent standard normal deviate by applying the inverse cumulative distribution function of the standard normal distribution (norminv) using the following equation:

$$Z = \Phi^{-1}\left(F_{t,\nu}(t)\right) \tag{3}$$

where t is the voxelwise t statistic, v the residual degrees of freedom, $F_{t,v}$ the t-distribution CDF (computed with tcdf), and Φ^{-1} the inverse normal CDF. This transformation preserves the sign of the original t statistic while expressing voxel intensities on the standard normal (Z) scale, allowing different correlation contrasts to be juxtaposed on the same template.

3. Results

3.1. Lexical frequency regressor reveals significant correlation in the VWFA region and IFG for written words, validating the re-analysis approach

The inverse relationship between VWFA activation in response to written words and word frequency has been established by several prior studies (Graves et al., 2010; Kronbichler et al., 2004; Woolnough et al., 2021). We thus chose to validate our re-analysis approach of the MOUS dataset by first applying it to the scans involving visually presented words. Specifically, we predicted that, in line with its status as a visual lexicon, responses in the VWFA region as well as the downstream inferior frontal gyrus (IFG) would show inverse correlations with word frequency.

Regarding the first prediction, results (Figure 1 and Table 1) revealed negative correlations of responses to written words and lexical frequency in fusiform cortex. Validating the re-analysis approach, the centroid of the peak word frequency correlation cluster (MNI: -42, -53, -13) was localized near the literature coordinates of the VWFA (MNI: -45, -57, -12) (Vogel et al., 2012). In addition, there was a secondary peak in the IFG region, located downstream from the VWFA in the ventral reading stream. Word frequency effects have been previously reported in the IFG (Fiez et al., 1999; Grande et al., 2011; Kronbichler et al., 2004). Interestingly, a third major cluster was found in the left middle temporal gyrus (MTG). This has been less of a focus in the literature, although several studies support the role of this region in reading (Lindenberg & Scheef, 2007; Shaywitz et al., 2004). Finally, we identified a posterior occipital cluster (MNI: -30, -93, -8), in agreement with earlier studies of word selectivity in the brain (Glezer et al., 2015). Supporting theories of the left-lateralization of the reading pathway, no significant clusters were found in the right hemisphere (see Fig. S5 for unthresholded results). Probing for positive correlations with word frequency, only a single cluster was identified in the posterior cingulate cortex (Fig. S1).

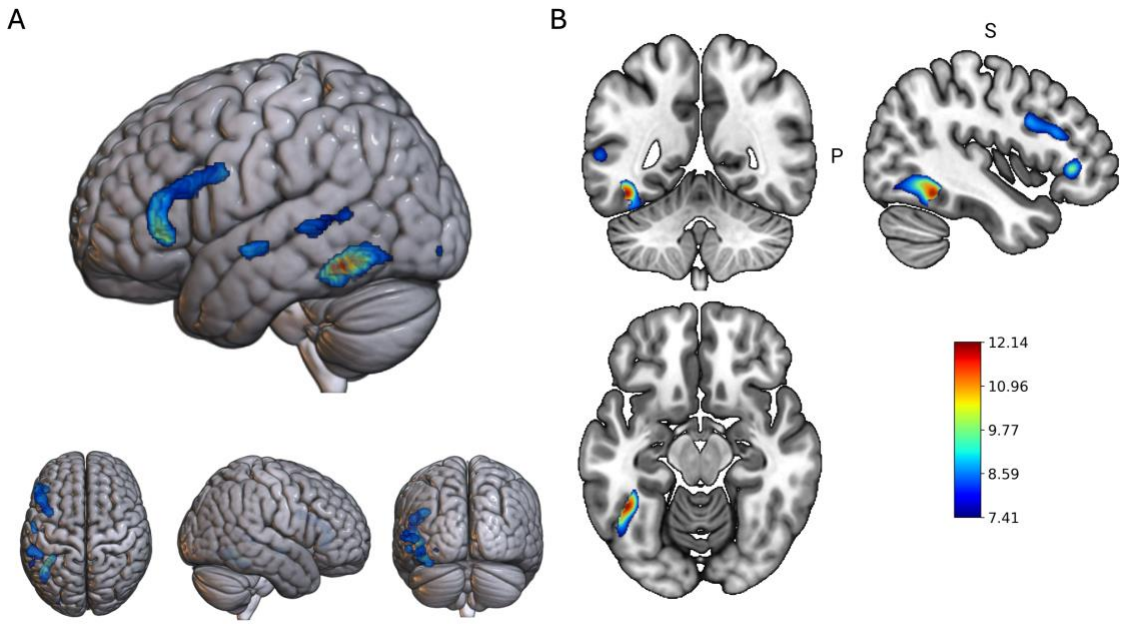

Figure 1. Significant clusters of brain activity negatively correlated with Zipf word frequency during reading. N=99 subjects**.** Colormap reflects voxel-level T-statistic values. Color bar range spans the height threshold of the t-map at pFWE < $1e^{-6}$ to the t-value of the peak voxel. Extent threshold (kE) = 20 voxels. **A)** 3D-rendering of group-level activation map in MNI space. **B)** Coronal, sagittal, and axial views of the peak cluster (MNI: -44, -50, -14). Renderings were produced using MRIcroGL (Chris & Matthew, 2000).

Cluster Centroid (MNI)	**Region (Brodmann Area)**	**Extent (voxels)**	**Cluster peak T-statistic**
-42, -54, -13	Left fusiform (37)	444	12.14
-48, 24, 15	Left inferior frontal gyrus, pars triangularis (45); Primary motor cortex (4)	761	10.6
-56, -10, -7	Left middle temporal gyrus (21)	94	9.36
-56, -40, 4	Left middle temporal gyrus (21)	278	8.99
-30, -93, -8	Left occipital pole (18)	21	8.47

Table 1. Significant clusters showing negative correlation with word frequency during reading. pFWE < $1e^{-6}$**,** extent threshold of 20 voxels. The table lists the

significant clusters of negative correlations with word frequency shown in Figure 1. Cluster centroids were calculated by taking the weighted average of the voxel-wise T-statistics within a surviving cluster in the thresholded T-map (Figure 1). In particular, the strongest correlation (first row) is found close to the literature coordinates of the VWFA.

3.2. Correlations with bigram frequency reveal evidence for a hierarchical orthographic representation in the ventral reading stream

We next tested whether our re-analysis approach could identify sublexical representations in the ventral reading stream. Specifically, we tested the hypothesis that regions posterior to the VWFA, such as the OWFA (Strother et al., 2016) a region with sublexical selectivity that lies earlier in the orthographic processing hierarchy, would show inverse correlations between activity and bigram frequency. Given that the VWFA's responses correlate inversely with lexical frequency (*i.e.,* less frequent words cause stronger responses than more frequent words), we hypothesized that the *least frequent bigram* in a word should dominate the response in sublexical areas. That is, responses in areas posterior to the VWFA should negatively correlate with the frequency of the least frequent bigram ("minimum bigram frequency") in a word.

Indeed, the minimum bigram frequency correlation contrast (Figure 2) revealed peaks posterior to the VWFA in the ventral occipital-temporal cortex (vOT), earlier in the ventral visual stream's processing hierarchy (Bruno et al., 2008; Vinckier et al., 2007; Woollams et al., 2011). Notably, in contrast to the word frequency correlation contrast, which only identified significant clusters in the left hemisphere, the bigram frequency correlation contrast identified bilateral occipital clusters.

Interestingly, as in the lexical analysis, we observed some significant correlations in the left IFG and spanning the middle temporal gyrus to superior temporal sulcus. While this could point to sublexical representations in these areas, an alternative explanation is that collinearity between minimum bigram frequency and word frequency (see Fig. S7) captures IFG activation during single-word reading. No significant clusters were found that showed a correlation between the negative maximum or mean bigram frequencies in a word (Fig. S2).

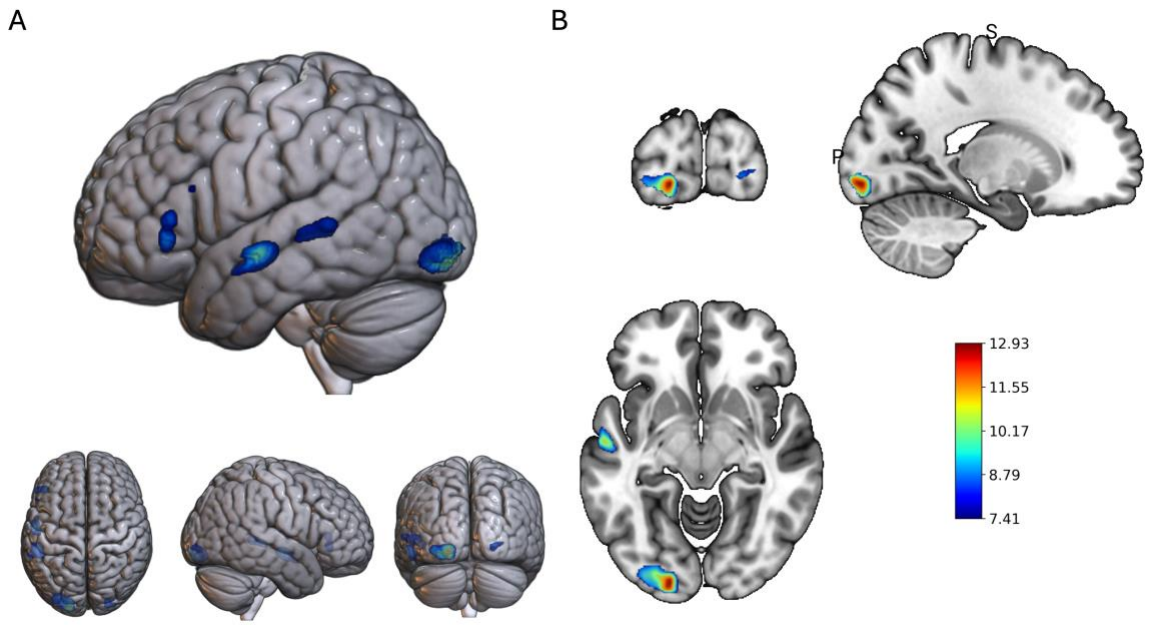

Figure 2. Brain activity negatively correlated with minimum bigram frequency during reading. N = 99 subjects. Colormap reflects voxel-level t-statistic values. Color bar range spans the height threshold of the t-map at pFWE < $1e^{-6}$ to the t-value of the peak voxel. Whole brain rendering shown at right**.** Extent threshold (kE) = 20 voxels. **A)** 3D-rendering of group-level activation map in MNI space. **B)** Coronal, sagittal, and axial views of the peak cluster, (MNI: -18, -94, -8).

Cluster Centroid (MNI)	**Region (Brodmann area)**	**Size (voxels)**	**Cluster Peak T-statistic**
-24, -92, -9	Left occipital pole (18)	424	12.93
-55, -9, -10	Left superior temporal gyrus (22)	188	10.77
-54, -36, 3	Left middle temporal gyrus (45)	195	9.32
-49, 32, 1	Left inferior frontal gyrus, pars triangularis (45)	97	9.25
24, -93, -4	Right occipital pole	66	8.83

Table 2. Significant clusters showing negative correlation with minimum bigram frequency contrast during reading. pFWE < $1e^{-6}$, extent threshold of 20 voxels. Cluster centroids were calculated by taking the weighted average of the voxel-wise T-statistics within a surviving cluster in the thresholded T-map (Figure 2).

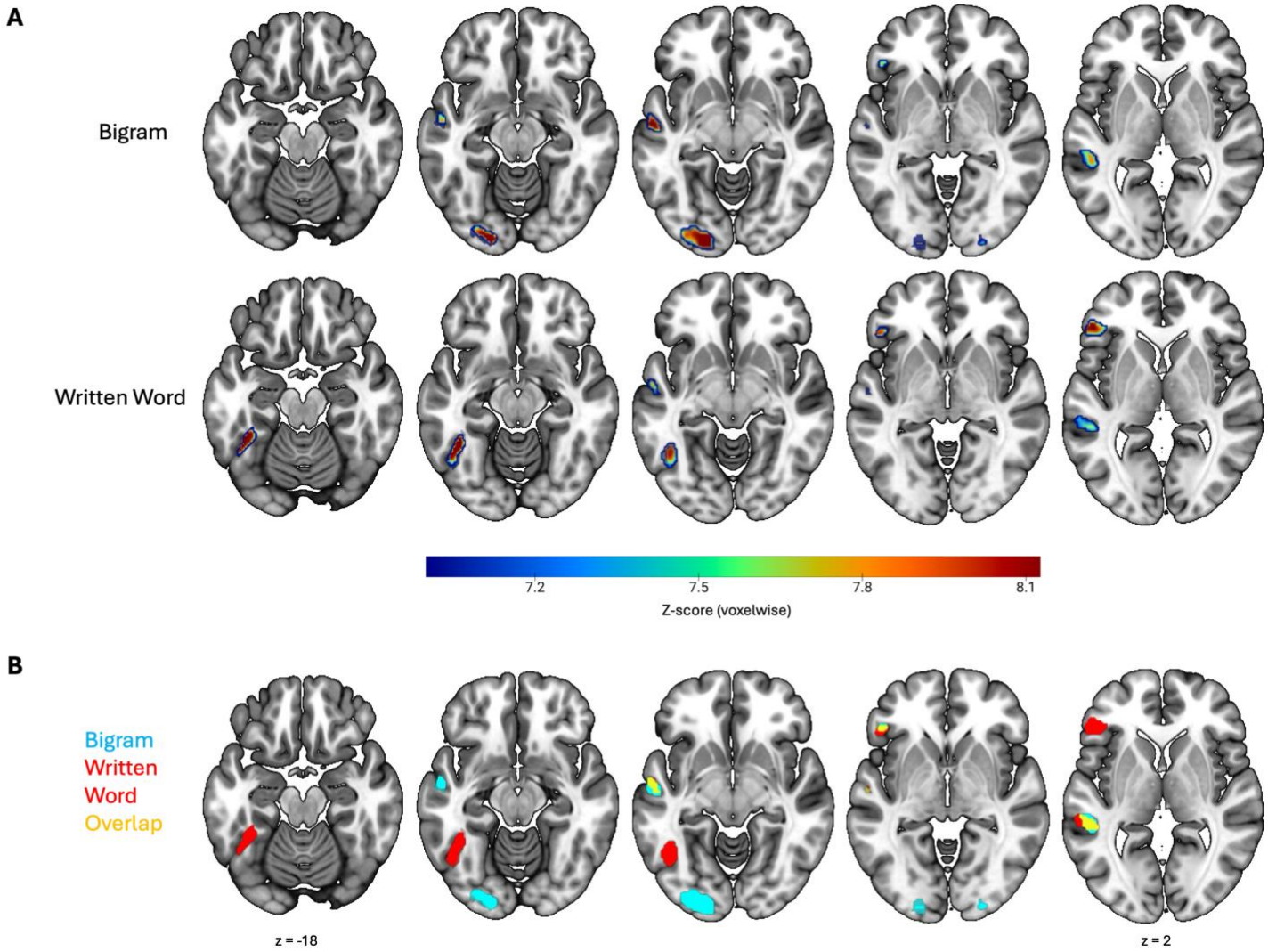

Figure 3. Posterior-to-anterior arrangement of sublexical and lexical selectivity during reading. Axial slices starting at z = -18 are spaced consecutively 5mm apart. **A)** A z-score was derived from the t-statistic of each voxel in the correlation contrast maps of minimum bigram frequency correlation (top row) and word frequency (middle row). **B)** The z-maps in the first two rows were binarized (bigrams in blue, words in red) and juxtaposed, with their spatial overlap shown in yellow.

3.3. Re-analysis of auditory scans supports the existence of auditory lexical representations in the anteroventral auditory stream

Having validated our re-analysis approach on scans recording brain activation associated with reading printed words, we next turned to the analyses of the auditory scans to probe lexical and sublexical representations for spoken words. As noted in the Introduction, the lexical and sublexical representations of auditory words are much less well understood. Recent work (Damera et al., 2023) has provided evidence for an auditory lexicon in the so-called “Auditory Word Form Area” (AWFA) in the mid-anterior superior temporal cortex, confirming earlier predictions based on a meta-analysis (DeWitt & Rauschecker, 2012)

The contrast examining parametric effect of word frequency during speech listening (Figure 4, unthresholded results in Fig. S6) revealed a group-level cluster in the left mid-anterior superior temporal gyrus (aSTG) with the centroid at (MNI: -52, 3, -16) overlapping with the average location of the AWFA (MNI: −62, −14, 2) identified in our previous study (Damera et al 2023). In addition, significant clusters were found in left parietal and frontal cortex, in line with models of speech processing in human cortex (Hickok & Poeppel, 2007; Rauschecker & Scott, 2009). In particular, there was a cluster in the left IFG that overlapped with the left IFG cluster identified in the lexical correlation analyses for *written* words (see Figure 1 and Table 1), compatible with theories that have posited the left IFG to contain the convergence point of written and spoken word processing hierarchies (Pugh et al., 2001). The analyses also identified a significant cluster in the left angular gyrus, very close to area Spt (Hickok et al., 2003a) thought to play a key role in sensorimotor transformation of speech (Hickok & Poeppel, 2007). Additionally notable, a significant cluster was also found in fusiform cortex, in line with reports of responses to auditory words in the vicinity of the VWFA (Dehaene et al., 2010; Ludersdorfer et al., 2013; Planton et al., 2019; Price & Devlin, 2003; Qin et al., 2021; Yoncheva et al., 2010).

Interestingly, there was a difference in the lateralization of these clusters: While significant clusters were found bilaterally in antero-ventral/temporal areas, significant clusters in parietal and frontal areas were found to be exclusively left-lateralized. This intriguing lateralization difference is in line with models of speech processing that have posited bilateral processing of speech in temporal cortex but left-lateralized processing in parietal and frontal cortex (Hickok & Poeppel, 2007). We observed positive correlations with word frequency during speech listening in the right and left superior temporal gyri (Fig. S1).

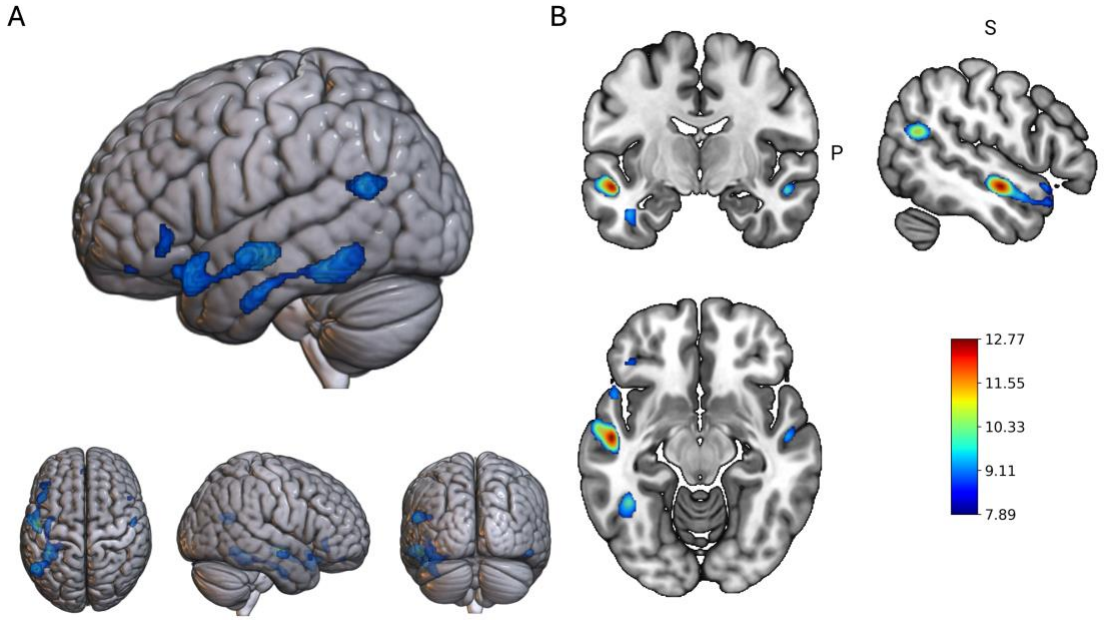

Figure 4. Group-level brain activity negatively correlated with Zipf word frequency during speech listening. N = 101 subjects. Colormap reflects voxel-level T-statistic values. Color bar range spans the height threshold of the t-map at pFWE < $1e^{-6}$ and the t-value of the peak voxel. Extent threshold (kE) = 20 voxels. **A)** 3D-rendering of group-level activation map in MNI space. **B)** Coronal, sagittal, and axial views of the peak cluster, (MNI: -52, -10, -12).

Cluster Centroid (MNI)	**Region (Brodmann Area)**	**Size (voxels)**	**Cluster Peak T-statistic**
-52, 3, -16	Left temporal pole (38)	746	12.77
-52, -59, 20	Left angular gyrus (39)	317	11.04
-39, -36, -20	Left fusiform gyrus (37)	751	10.61
53, -8, -15	Right superior temporal sulcus (21)	86	9.99
-45, 32, -6	Left anterior superior temporal gyrus (22)	128	9.23
48, 18, -18	Right temporal pole	40	8.84

-4, 47, -18	Left dorsolateral prefrontal cortex	26	8.73

Table 3. Significant clusters showing negative correlation with word frequency during speech listening. pFWE < $1e^{-6}$, extent threshold > 20 voxels. The first, most significant ROI denotes the region proximal to the previously described location of the AWFA. Cluster centroids were calculated by taking the weighted average of the voxel-wise T-statistics within a surviving cluster in the thresholded T-map (Figure 4).

3.4. Evidence for sublexical auditory speech representations in the anteroventral auditory stream supports the existence of an auditory speech processing hierarchy in the auditory "what" stream

In the auditory syllable frequency correlation contrast (Fig. 5, unthresholded results in Fig. S6), we find a cluster in the left mid-superior temporal gyrus (MNI: -61, -17, 1), superior and posterior to the main cluster observed in the word frequency contrast (MNI: -52, 3, -16). This location corresponds to bilateral sublexical phonetic processing centers identified by a prominent meta-analysis (Turkeltaub & Coslett, 2010). Notably, as in the auditory lexical case, a similar cluster was identified in the right hemisphere. Several additional regions with significant activation correlations were observed in response to negative mean (but not maximum) syllable frequency (Fig. S4).

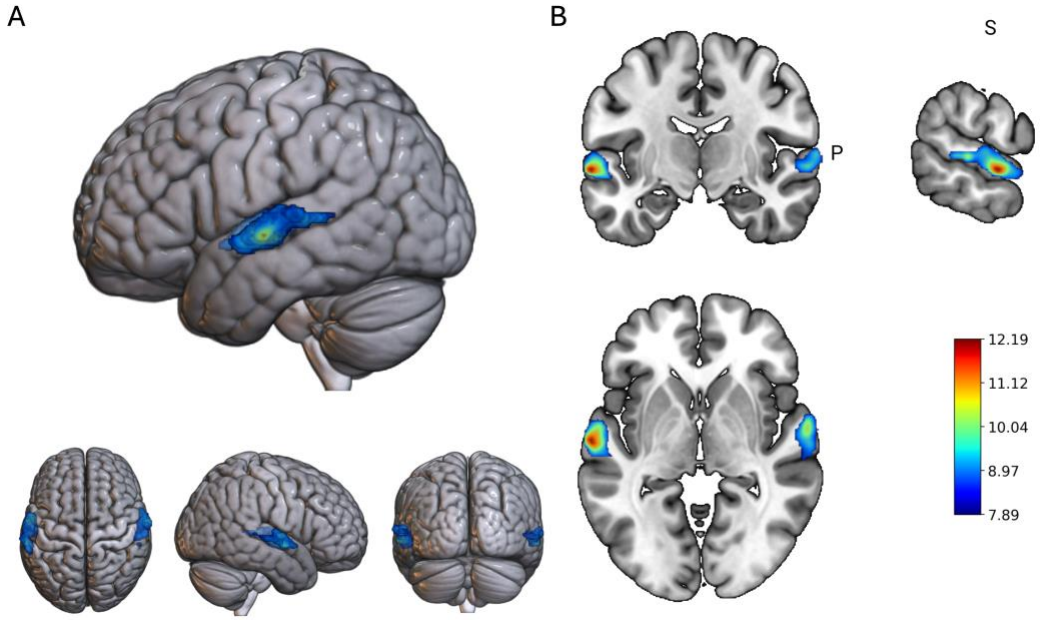

Figure 5. Brain activity negatively correlated with minimum syllable frequency during listening to speech. Color map reflects voxel-level t-statistic values. N = 101 subjects. Color bar range spans the height threshold of the t-map at pFWE< $1e^{-6}$ and the t-statistic of the peak voxel. Extent threshold (kE) = 20 voxels. **A)** 3D-rendering of group-level activation map in MNI space. **B)** Coronal, sagittal, and axial views of the peak cluster, (MNI: -64, -12, -2).

Cluster Centroid (MNI)	**Region (Brodmann Area)**	**Size (voxels)**	**Cluster Peak T-statistic**
-61, -17, 1	Left Superior Temporal Gyrus (22)	870	12.19
63, -11, 0	Right Superior Temporal Gyrus (22)	618	11.25

Table 4. Significant clusters showing negative correlation with minimum syllable frequency contrast during speech listening. pFWE < $1e^{-6}$, extent threshold > 20 voxels. This table lists clusters whose activity shows a significant negative correlation with the frequency of the least-frequent syllable in a word. Cluster centroids were calculated by taking the weighted average of the voxel-wise t-statistics within a surviving cluster in the thresholded t-map (Figure 5).

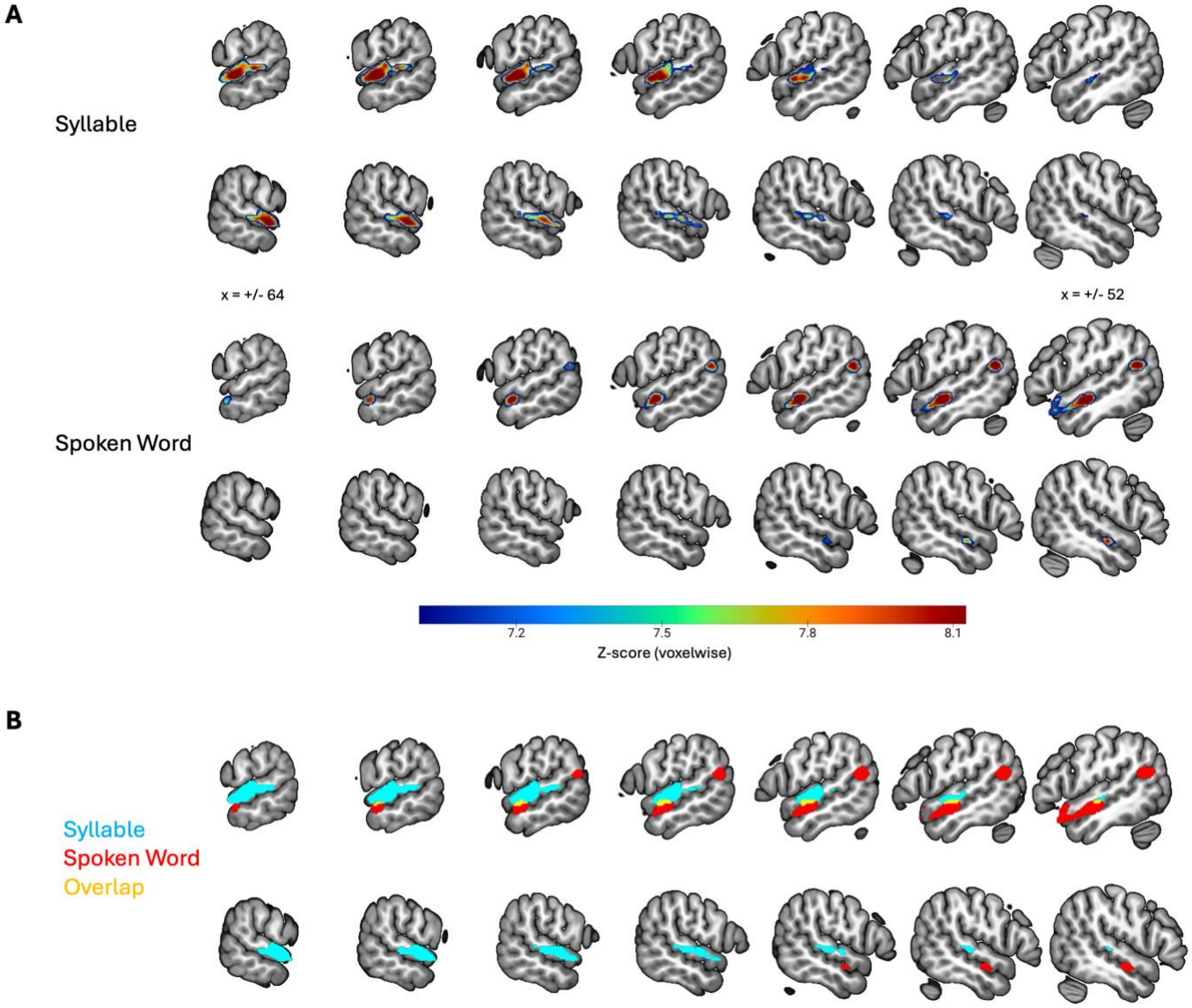

Figure 6. Posterior-to-anterior gradient of lexical selectivity during speech listening in the superior temporal gyrus. Sagittal slices are consecutively spaced 2mm apart. **A)** A z-score was derived from the t-statistic of each voxel in the correlation contrast maps of minimum syllable frequency correlation (top row) and word frequency (middle row). **B)** The z-maps in the first two rows were binarized (syllables in blue, words in red) and juxtaposed, with their spatial overlap shown in yellow.

4. Discussion

Learning has been associated with a reduction in the average neuronal response to more familiar stimuli (Anderson et al., 2009; Lim, 2019; Lim et al., 2015; Schulz et al., 2021; Ulanovsky et al., 2003). Linguistic stimuli appear to be no exception: the 'frequency effect', wherein rarely-encountered words generate stronger responses than more frequent ones, has been documented for written words across alphabetical and character-based languages (Carreiras et al., 2006; Chee et al., 2003; Graves et al., 2010; Kuo et al., 2003). Previous fMRI studies of the frequency effect during reading (Graves et al., 2010; Kronbichler et al., 2004) that employed parametric designs to explore where in the brain the magnitude of BOLD responses scales with word frequency have identified the VWFA as the site of the strongest (inverse) correlations between activity and word frequency, suggesting that it contains an "orthographic lexicon." Indeed, subsequent studies using different techniques such as fMRI rapid adaptation (Glezer et al., 2009, 2015) and ECoG (Lochy et al., 2018; Thesen et al., 2012; Woolnough et al., 2021) have provided further confirmatory evidence for the lexicon hypothesis, thus validating the approach of using negative correlations of BOLD response and lexical frequency as a probe for lexical representations in the brain. The frequency effect has previously been shown to also hold for sublexical components of written words, such as bigrams and quadrigrams, thus making it a useful tool for investigating not only lexical, but also sublexical representations of words (Grainger & Whitney, 2004; Graves et al., 2010; Schoonbaert & Grainger, 2004; Vinckier et al., 2007; Woollams et al., 2011).

In the present study, we performed a re-analysis of the MOUS dataset, using the frequency effect to probe for lexical and sublexical representations of written and, for the first time, spoken words in the brain. The MOUS dataset is a multimodal neuroimaging dataset collected from 204 subjects who read and listened to word lists and sentences (Schoffelen et al., 2019). We obtained lexical frequency metrics from the SUBTLEX-NL database, which tracks the occurrences of Dutch words across a corpus consisting of film subtitles, newspapers, books, and other forms of media (Keuleers, Brysbaert, et al., 2010). After segmenting Dutch words into their constituent bigrams and syllables, we calculated frequency metrics for these word parts as well. This enabled us to examine sublexical frequency effects in addition to lexical frequency effects for both written and spoken words

Our study of the reading group confirmed key aspects of the ventral reading pathway and validated our re-analysis approach. The site of the global peak of the correlations between word frequency and activation coincided with the VWFA. An additional cluster was found in the IFG, in agreement with models of reading in the brain (Pugh et al., 2001).

In occipital cortex, our re-analysis identified a main cluster responsive to bigram frequency in the left occipital cortex (posterior to the VWFA), supporting the notion of a simple-to-complex hierarchy for orthographic stimuli in the ventral visual stream. Unlike in our word-frequency correlation contrast, we also found a suprathreshold cluster in the right hemisphere whose activity correlated with bigram frequency. Our main bigram frequency cluster (MNI: -24, -92, -9) corresponds to the literature coordinate of the OWFA (-38,-83,-18), suggesting that bigrams found within whole words that span both hemifields are independently represented by the OWFA, and that experience-dependent tuning (acquired by learning to read) might drive the formation of sublexical, orthographic representations in this region that support the downstream development of lexical representations in the VWFA (Cohen & Dehaene, 2004; Glezer et al., 2015; Strother et al., 2016; Vinckier et al., 2007).

Employing the same parametric regression technique that we validated in our study of the reading group to auditorily presented words, we here report the first neuroimaging analysis of word frequency effects during speech listening. We identified a main cluster in the mid-anterior superior temporal gyrus (MNI: -52, 3, -16) that encompasses and extends anteriorly from the reported location of the Auditory Word Form Area, AWFA (MNI: -62, -14, 2), a possible lexicon for spoken words (Cohen et al., 2004; Damera et al., 2023). This finding supports the role of the auditory anteroventral stream in spoken word recognition and validates the use of the word-frequency correlation method for modeling task-related activations during speech listening.

We also for the first time applied the frequency correlation approach to sublexical auditory processing, focusing on the syllable level. We find a negative linear relationship between the minimum syllable frequency found in a word and BOLD activation in bilateral mid-STG. The left mid-STG cluster (MNI: -61, -17, 1) coincides with the main cluster (MNI: -62, -19, 1) of the activation likelihood map produced by DeWitt & Rauschecker's 2012 meta-analysis of phoneme repetition suppression studies. Critically, this cluster is posterior to the main cluster identified by the word frequency correlation contrast (MNI: -52, 3, -16), supporting a simple-to-complex (sublexical-to-lexical) hierarchy for processing spoken words that starts in mid-STG and proceeds anteriorly. In this framework, we speculate that the syllable-selective cluster identified by our analyses may contain intermediate acoustic-phonetic representations that are integrated by a word-selective area such as the AWFA. In line with the hypothesis that the strongest activation is associated with the most infrequent words and, by implication, their most infrequent sublexical component, we observed that at the sublexical level, correlations were strongest for the least frequent bigram and syllable, respectively.

Proceeding downstream from the word form areas, across both sensory modalities, the word frequency correlation contrast identified the left IFG. This is

consistent with its location at the putative convergence of visual and auditory language perception pathways. The left-lateralization of frontal regions during both reading and speech listening stands in contrast to the more bilateral patterns of selectivity in the temporal lobe, especially for spoken words. We also identified a cluster in the left angular gyrus during speech listening, located close to area Spt in the dorsal speech processing pathway (Hickok et al., 2003b). The auditory dorsal stream computes sensorimotor transformations (Chevillet et al., 2013; Hickok & Poeppel, 2007; Rauschecker, 2011; Rauschecker & Scott, 2009; Rauschecker & Tian, 2000) with some dual-stream models proposing that area Spt is a sensorimotor integration circuit on the basis that it is active during both perception and rehearsal of speech (Hickok et al., 2003a)

Parametric regression techniques like the one used here (that model neural response as a function of stimulus frequency) may be applied to re-analyze other datasets, provided that the frequencies or intensities of stimuli of interest can be quantified and that a sufficient number of stimuli are sampled. Given the rise of 'naturalistic' stimulus presentation paradigms (for example, those that employ movie watching) in open-science datasets (Allen et al., 2022), future research efforts may focus on parameterizing ecological variables of interest. For variables that occur with regularity or variable intensity (words, sounds, facial expressions, etc.), parametric regression may be able to localize neural populations selective for them.

Future studies may investigate frequency effects in other languages, such as those with high degrees of homophony (e.g., Japanese), as speakers must access lexical representations from a highly constrained syllabary during speech listening (Tamaoka, 2007). Character-based languages also provide the opportunity to probe word or radical frequency effects during reading. While most studies cited in this manuscript study European languages, which show an inverse relationship between length (in letters or spoken duration) and word frequency, character-based languages show no such (potentially confounding) relationship between orthographic and phonological complexity (Zhang & Xing, 2023).

5. Data and Code Availability

Code for all analyses and figures described here and in the Supplement, as well as neuroimaging workflow configuration details, can be found in a Github repository (https://github.com/suneelbanerjee/MOUS_hierarchical-representations). The original dataset is described in Schoffelen et al., 2019 (https://www.nature.com/articles/s41597-019-0020-y).

6. Ethics

Study participants whose data were analyzed in this study explicitly consented for the anonymized collected data to be used for research purposes by other researchers. In keeping with the informed consent agreement obtained from the participants and the requirements of both the Ethics Committee and the Radboud University security officer, potentially identifying data (such as imaging data) recorded at the Donders Centre for Cognitive Neuroimaging can only be shared to researchers following explicit approval of a Data Use Agreement. Author S.D. filed this request on behalf of our group.

7. Declaration of Competing Interests

The authors declare no competing interests.

8. References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126. https://doi.org/10.1038/s41593-021-00962-x

Anderson, L. A., Christianson, G. B., & Linden, J. F. (2009). Stimulus-Specific Adaptation Occurs in the Auditory Thalamus. *The Journal of Neuroscience*, *29*(22), 7359–7363. https://doi.org/10.1523/jneurosci.0793-09.2009

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*(1), 26–41. https://doi.org/10.1016/j.media.2007.06.004

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*(1), 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042

Bernal, J., Valverde, S., Kushibar, K., Cabezas, M., Oliver, A., Lladó, X., & Initiative, T. A. D. N. (2021). Generating Longitudinal Atrophy Evaluation Datasets on Brain Magnetic Resonance Images Using Convolutional Neural Networks and Segmentation Priors. *Neuroinformatics*, *19*(3), 477–492. https://doi.org/10.1007/s12021-020-09499-z

Bruno, J. L., Zumberge, A., Manis, F. R., Lu, Z.-L., & Goldman, J. G. (2008). Sensitivity to orthographic familiarity in the occipito-temporal region. *NeuroImage*, *39*(4), 1988–2001. https://doi.org/10.1016/j.neuroimage.2007.10.044

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The Word Frequency Effect: A Review of Recent Developments and Implications for the Choice of Frequency Estimates in German. *Experimental Psychology*, *58*(5), 412–424. https://doi.org/10.1027/1618-3169/a000123

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The Word Frequency Effect in Word Processing: An Updated Review. *Current Directions in Psychological Science*, *27*(1), 45–50. https://doi.org/10.1177/0963721417727521

Carreiras, M., Mechelli, A., & Price, C. J. (2006). Effect of word and syllable frequency on activation during lexical decision and reading aloud. *Human Brain Mapping*, *27*(12), 963–972. https://doi.org/10.1002/hbm.20236

Chao, L. L., Weisberg, J., & Martin, A. (2002). Experience-dependent Modulation of Category-related Cortical Activity. *Cerebral Cortex*, *12*(5), 545–551. https://doi.org/10.1093/cercor/12.5.545

Chee, M. W. L., Westphal, C., Goh, J., Graham, S., & Song, A. W. (2003). Word frequency and subsequent memory effects studied using event-related fMRI. *NeuroImage*, *20*(2), 1042–1051. https://doi.org/10.1016/s1053-8119(03)00335-5

Chevillet, M. A., Jiang, X., Rauschecker, J. P., & Riesenhuber, M. (2013). Automatic Phoneme Category Selectivity in the Dorsal Auditory Stream. *The Journal of Neuroscience*, *33*(12), 5208–5215. https://doi.org/10.1523/jneurosci.1870-12.2013

Chris, R., & Matthew, B. (2000). Stereotaxic Display of Brain Lesions. *Behavioural Neurology*, *12*(4), 191–200. https://doi.org/10.1155/2000/421719

Chung, M. K., Worsley, K. J., Nacewicz, B. M., Dalton, K. M., & Davidson, R. J. (2010). General multivariate linear modeling of surface shapes using SurfStat. *NeuroImage*, *53*(2), 491–505. https://doi.org/10.1016/j.neuroimage.2010.06.032

Cohen, L., & Dehaene, S. (2004). Specialization within the ventral stream: the case for the visual word form area. *NeuroImage*, *22*(1), 466–476. https://doi.org/10.1016/j.neuroimage.2003.12.049

Cohen, L., Jobert, A., Bihan, D. L., & Dehaene, S. (2004). Distinct unimodal and multimodal regions for word processing in the left temporal cortex. *NeuroImage*, *23*(4), 1256–1270. https://doi.org/10.1016/j.neuroimage.2004.07.052

Cohen, L., Lehéricy, S., Chochon, F., Lemer, C., Rivaud, S., & Dehaene, S. (2002). Language-specific tuning of visual cortex? Functional properties of the Visual Word Form Area. *Brain*, *125*(5), 1054–1069. https://doi.org/10.1093/brain/awf094

Cox, R. W., & Hyde, J. S. (1997). Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine*, *10*(4-5), 171–178. https://doi.org/10.1002/(sici)1099-1492(199706/08)10:4/5<171::aid-nbm453>3.0.co;2-l

Damera, S. R., Chang, L., Nikolov, P. P., Mattei, J. A., Banerjee, S., Glezer, L. S., Cox, P. H., Jiang, X., Rauschecker, J. P., & Riesenhuber, M. (2023). Evidence for a Spoken Word Lexicon in the Auditory Ventral Stream. *Neurobiology of Language*, *4*(3), 420–434. https://doi.org/10.1162/nol_a_00108

Dehaene, S., Pegado, F., Braga, L. W., Ventura, P., Filho, G. N., Jobert, A., Dehaene-Lambertz, G., Kolinsky, R., Morais, J., & Cohen, L. (2010). How Learning to Read Changes the Cortical Networks for Vision and Language. *Science*, *330*(6009), 1359–1364. https://doi.org/10.1126/science.1194140

DeWitt, I., & Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences*, *109*(8), E505–E514. https://doi.org/10.1073/pnas.1113427109

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Evans, A. C., Janke, A. L., Collins, D. L., & Baillet, S. (2012). Brain templates and atlases. *NeuroImage*, *62*(2), 911–922. https://doi.org/10.1016/j.neuroimage.2012.01.024

Fiez, J. A., Balota, D. A., Raichle, M. E., & Petersen, S. E. (1999). Effects of Lexicality, Frequency, and Spelling-to-Sound Consistency on the Functional Anatomy of Reading. *Neuron*, *24*(1), 205–218. https://doi.org/10.1016/s0896-6273(00)80833-8

Glezer, L. S., Jiang, X., & Riesenhuber, M. (2009). Evidence for Highly Selective Neuronal Tuning to Whole Words in the "Visual Word Form Area." *Neuron*, *62*(2), 199–204. https://doi.org/10.1016/j.neuron.2009.03.017

Glezer, L. S., Kim, J., Rule, J., Jiang, X., & Riesenhuber, M. (2015). Adding Words to the Brain's Visual Dictionary: Novel Word Learning Selectively Sharpens Orthographic Representations in the VWFA. *The Journal of Neuroscience*, *35*(12), 4965–4972. https://doi.org/10.1523/jneurosci.4031-14.2015

Goh, W. D., Suárez, L., Yap, M. J., & Tan, S. H. (2009). Distributional analyses in auditory lexical decision: Neighborhood density and word-frequency effects. *Psychonomic Bulletin & Review*, *16*(5), 882–887. https://doi.org/10.3758/pbr.16.5.882

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Frontiers in Neuroinformatics*, *5*, 13. https://doi.org/10.3389/fninf.2011.00013

Grainger, J., & Whitney, C. (2004). Does the huamn mnid raed wrods as a wlohe? *Trends in Cognitive Sciences*, *8*(2), 58–59. https://doi.org/10.1016/j.tics.2003.11.006

Grande, M., Meffert, E., Huber, W., Amunts, K., & Heim, S. (2011). Word frequency effects in the left IFG in dyslexic and normally reading children during picture naming and reading. *NeuroImage*, *57*(3), 1212–1220. https://doi.org/10.1016/j.neuroimage.2011.05.033

Graves, W. W., Desai, R., Humphries, C., Seidenberg, M. S., & Binder, J. R. (2010). Neural Systems for Reading Aloud: A Multiparametric Approach. *Cerebral Cortex*, *20*(8), 1799–1815. https://doi.org/10.1093/cercor/bhp245

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*(1), 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060

Hauk, O., Davis, M. H., & Pulvermüller, F. (2008). Modulation of brain activity by multiple lexical and word form variables in visual word recognition: A parametric fMRI study. *NeuroImage*, *42*(3), 1185–1195. https://doi.org/10.1016/j.neuroimage.2008.05.054

Hickok, G., Buchsbaum, B., Humphries, C., & Muftuler, T. (2003a). Auditory–Motor Interaction Revealed by fMRI: Speech, Music, and Working Memory in Area Spt. *Journal of Cognitive Neuroscience*, *15*(5), 673–682. https://doi.org/10.1162/089892903322307393

Hickok, G., Buchsbaum, B., Humphries, C., & Muftuler, T. (2003b). AuditoryMotor Interaction Revealed by fMRI: Speech, Music, and Working Memory in Area Spt. *Journal of Cognitive Neuroscience*, *15*(5), 673–682. https://doi.org/10.1162/jocn.2003.15.5.673

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402. https://doi.org/10.1038/nrn2113

Hubel, D. H., Wiesel, T. N., LeVay, S., Barlow, H. B., & Gaze, R. M. (1977). Plasticity of ocular dominance columns in monkey striate cortex. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *278*(961), 377–409. https://doi.org/10.1098/rstb.1977.0050

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, *17*(2), 825–841. https://doi.org/10.1006/nimg.2002.1132

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, *5*(2), 143–156. https://doi.org/10.1016/s1361-8415(01)00036-6

Joubert, S., Beauregard, M., Walter, N., Bourgouin, P., Beaudoin, G., Leroux, J.-M., Karama, S., & Lecours, A. R. (2004). Neural correlates of lexical and sublexical processes in reading. *Brain and Language*, *89*(1), 9–20. https://doi.org/10.1016/s0093-934x(03)00403-6

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, *98*(3), 630-644.e16. https://doi.org/10.1016/j.neuron.2018.03.044

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. https://doi.org/10.3758/brm.42.3.643

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice Effects in Large-Scale Visual Word Recognition Studies: A Lexical Decision Study on 14,000 Dutch Mono- and Disyllabic Words and Nonwords. *Frontiers in Psychology*, *1*, 174. https://doi.org/10.3389/fpsyg.2010.00174

Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & Mishkin, M. (2013). The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends in Cognitive Sciences*, *17*(1), 26–49. https://doi.org/10.1016/j.tics.2012.10.011

Kronbichler, M., Hutzler, F., Wimmer, H., Mair, A., Staffen, W., & Ladurner, G. (2004). The visual word form area and the frequency with which words are encountered: evidence from a parametric fMRI study. *NeuroImage*, *21*(3), 946–953. https://doi.org/10.1016/j.neuroimage.2003.10.021

Kuo, W.-J., Yeh, T.-C., Lee, C.-Y., Wu, Y. u-T. e, Chou, C.-C., Ho, L.-T., Hung, D. L., Tzeng, O. J. L., & Hsieh, J.-C. (2003). Frequency effects of Chinese character processing in the brain: an event-related fMRI study. *NeuroImage*, *18*(3), 720–730. https://doi.org/10.1016/s1053-8119(03)00015-6

Lanczos, C. (1964). Evaluation of Noisy Data. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, *1*(1), 76–85. https://doi.org/10.1137/0701007

Lim, S. (2019). Mechanisms underlying sharpening of visual response dynamics with familiarity. *eLife*, *8*, e44098. https://doi.org/10.7554/elife.44098

Lim, S., McKee, J. L., Woloszyn, L., Amit, Y., Freedman, D. J., Sheinberg, D. L., & Brunel, N. (2015). Inferring learning rules from distributions of firing rates in cortical neurons. *Nature Neuroscience*, *18*(12), 1804–1810. https://doi.org/10.1038/nn.4158

Lindenberg, R., & Scheef, L. (2007). Supramodal language comprehension: Role of the left temporal lobe for listening and reading. *Neuropsychologia*, *45*(10), 2407–2415. https://doi.org/10.1016/j.neuropsychologia.2007.02.008

Lochy, A., Jacques, C., Maillard, L., Colnat-Coulbois, S., Rossion, B., & Jonas, J. (2018). Selective visual representation of letters and words in the left ventral occipito-temporal cortex with intracerebral recordings. *Proceedings of the National Academy of Sciences*, *115*(32), E7595–E7604. https://doi.org/10.1073/pnas.1718987115

Ludersdorfer, P., Schurz, M., Richlan, F., Kronbichler, M., & Wimmer, H. (2013). Opposite effects of visual and auditory word-likeness on activity in the visual word form area. *Frontiers in Human Neuroscience*, *7*, 491. https://doi.org/10.3389/fnhum.2013.00491

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*(4), 661–677. https://doi.org/10.1017/s014271640707035x

Norman-Haignere, S. V., Long, L. K., Devinsky, O., Doyle, W., Irobunda, I., Merricks, E. M., Feldstein, N. A., McKhann, G. M., Schevon, C. A., Flinker, A., & Mesgarani, N. (2022). Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nature Human Behaviour*, *6*(3), 455–469. https://doi.org/10.1038/s41562-021-01261-y

Planton, S., Chanoine, V., Sein, J., Anton, J.-L., Nazarian, B., Pallier, C., & Pattamadilok, C. (2019). Top-down activation of the visuo-orthographic system during spoken sentence processing. *NeuroImage*, *202*, 116135. https://doi.org/10.1016/j.neuroimage.2019.116135

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., & Petersen, S. E. (2011). Functional Network Organization of the Human Brain. *Neuron*, *72*(4), 665–678. https://doi.org/10.1016/j.neuron.2011.09.006

Price, C. J., & Devlin, J. T. (2003). The myth of the visual word form area. *NeuroImage*, *19*(3), 473–481. https://doi.org/10.1016/s1053-8119(03)00084-3

Pugh, K. R., Mencl, W. E., Jenner, A. R., Katz, L., Frost, S. J., Lee, J. R., Shaywitz, S. E., & Shaywitz, B. A. (2001). Neurobiological studies of reading and reading disability. *Journal of Communication Disorders*, *34*(6), 479–492. https://doi.org/10.1016/s0021-9924(01)00060-0

Qin, L., Lyu, B., Shu, S., Yin, Y., Wang, X., Ge, J., Siok, W.-T., & Gao, J.-H. (2021). A Heteromodal Word-Meaning Binding Site in the Visual Word Form Area under Top-Down Frontoparietal Control. *The Journal of Neuroscience*, *41*(17), 3854–3869. https://doi.org/10.1523/jneurosci.2771-20.2021

Rauschecker, J. P. (2011). An expanded role for the dorsal auditory pathway in sensorimotor control and integration. *Hearing Research*, *271*(1–2), 16–25. https://doi.org/10.1016/j.heares.2010.09.001

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience*, *12*(6), 718–724. https://doi.org/10.1038/nn.2331

Rauschecker, J. P., & Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proceedings of the National Academy of Sciences*, *97*(22), 11800–11806. https://doi.org/10.1073/pnas.97.22.11800

Reuter, M., Rosas, H. D., & Fischl, B. (2010). Highly accurate inverse consistent registration: A robust approach. *NeuroImage*, *53*(4), 1181–1196. https://doi.org/10.1016/j.neuroimage.2010.07.020

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025. https://doi.org/10.1038/14819

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., Eickhoff, S. B., Hakonarson, H., Gur, R. C., Gur, R. E., & Wolf, D. H. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage*, *64*, 240–256. https://doi.org/10.1016/j.neuroimage.2012.08.052

Schillingmann, L., Ernst, J., Keite, V., Wrede, B., Meyer, A. S., & Belke, E. (2018). AlignTool: The automatic temporal alignment of spoken utterances in German, Dutch, and British English for psycholinguistic purposes. *Behavior Research Methods*, *50*(2), 466–489. https://doi.org/10.3758/s13428-017-1002-7

Schoffelen, J.-M., Oostenveld, R., Lam, N. H. L., Uddén, J., Hultén, A., & Hagoort, P. (2019). A 204-subject multimodal neuroimaging dataset to study language processing. *Scientific Data*, *6*(1), 17. https://doi.org/10.1038/s41597-019-0020-y

Schoonbaert, S., & Grainger, J. (2004). Letter position coding in printed word perception: Effects of repeated and transposed letters. *Language and Cognitive Processes*, *19*(3), 333–367. https://doi.org/10.1080/01690960344000198

Schulz, A., Miehl, C., Berry, M. J., & Gjorgjieva, J. (2021). The generation of cortical novelty responses through inhibitory plasticity. *eLife*, *10*, e65309. https://doi.org/10.7554/elife.65309

Seghier, M. L., Neufeld, N. H., Zeidman, P., Leff, A. P., Mechelli, A., Nagendran, A., Riddoch, J. M., Humphreys, G. W., & Price, C. J. (2012). Reading without the left ventral occipito-temporal cortex. *Neuropsychologia*, *50*(14), 3621–3635. https://doi.org/10.1016/j.neuropsychologia.2012.09.030

Shaywitz, B. A., Shaywitz, S. E., Blachman, B. A., Pugh, K. R., Fulbright, R. K., Skudlarski, P., Mencl, W. E., Constable, R. T., Holahan, J. M., Marchione, K. E., Fletcher, J. M., Lyon, G. R., & Gore, J. C. (2004). Development of left occipitotemporal systems for skilled reading in children after a phonologically- based intervention. *Biological Psychiatry*, *55*(9), 926–933. https://doi.org/10.1016/j.biopsych.2003.12.019

Strother, L., Coros, A. M., & Vilis, T. (2016). Visual Cortical Representation of Whole Words and Hemifield-split Word Parts. *Journal of Cognitive Neuroscience*, *28*(2), 252–260. https://doi.org/10.1162/jocn_a_00900

Tamaoka, K. (2007). Rebounding activation caused by lexical homophony in the processing of Japanese two-kanji compound words. *Reading and Writing*, *20*(5), 413–439. https://doi.org/10.1007/s11145-006-9036-8

Thesen, T., McDonald, C. R., Carlson, C., Doyle, W., Cash, S., Sherfey, J., Felsovalyi, O., Girard, H., Barr, W., Devinsky, O., Kuzniecky, R., & Halgren, E. (2012). Sequential then interactive processing of letters and words in the left fusiform gyrus. *Nature Communications*, *3*(1), 1284. https://doi.org/10.1038/ncomms2220

Turkeltaub, P. E., & Coslett, H. B. (2010). Localization of sublexical speech perception components. *Brain and Language*, *114*(1), 1–15. https://doi.org/10.1016/j.bandl.2010.03.008

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, *29*(6), 1310–1320. https://doi.org/10.1109/tmi.2010.2046908

Ulanovsky, N., Las, L., & Nelken, I. (2003). Processing of low-probability sounds by cortical neurons. *Nature Neuroscience*, *6*(4), 391–398. https://doi.org/10.1038/nn1032

Vinckier, F., Dehaene, S., Jobert, A., Dubus, J. P., Sigman, M., & Cohen, L. (2007). Hierarchical Coding of Letter Strings in the Ventral Stream: Dissecting the Inner Organization of the Visual Word-Form System. *Neuron*, *55*(1), 143–156. https://doi.org/10.1016/j.neuron.2007.05.031

Vogel, A. C., Miezin, F. M., Petersen, S. E., & Schlaggar, B. L. (2012). The Putative Visual Word Form Area Is Functionally Connected to the Dorsal Attention Network. *Cerebral Cortex*, *22*(3), 537–549. https://doi.org/10.1093/cercor/bhr100

Woollams, A. M., Silani, G., Okada, K., Patterson, K., & Price, C. J. (2011). Word or Word-like? Dissociating Orthographic Typicality from Lexicality in the Left Occipito-temporal Cortex. *Journal of Cognitive Neuroscience*, *23*(4), 992–1002. https://doi.org/10.1162/jocn.2010.21502

Woolnough, O., Donos, C., Rollo, P. S., Forseth, K. J., Lakretz, Y., Crone, N. E., Fischer-Baum, S., Dehaene, S., & Tandon, N. (2021). Spatiotemporal dynamics of orthographic and lexical processing in the ventral visual pathway. *Nature Human Behaviour*, *5*(3), 389–398. https://doi.org/10.1038/s41562-020-00982-w

Yoncheva, Y. N., Zevin, J. D., Maurer, U., & McCandliss, B. D. (2010). Auditory Selective Attention to Speech Modulates Activity in the Visual Word Form Area. *Cerebral Cortex*, *20*(3), 622–632. https://doi.org/10.1093/cercor/bhp129

Zhang, L., & Xing, H. (2023). The interaction of orthography, phonology and semantics in the process of second language learners' Chinese character production. *Frontiers in Psychology*, *14*, 1076810. https://doi.org/10.3389/fpsyg.2023.1076810