NeuroImage 139 (2016) 74–93

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Reliable individual-level neural markers of high-level language processing: A necessary precursor for relating neural variability to behavioral and genetic variability

Kyle Mahowald [a,*], Evelina Fedorenko [b,c]

[a] MIT, United States
[b] MGH, United States
[c] HMS, United States

ARTICLE INFO

Article history:
Accepted 27 May 2016
Available online 31 May 2016

Keywords:
Language system
Individual differences
Neural markers of language activity
Reliability

ABSTRACT

The majority of functional neuroimaging investigations aim to characterize an average human brain. However, another important goal of cognitive neuroscience is to understand the ways in which individuals differ from one another and the significance of these differences. This latter goal is given special weight by the recent reconceptualization of neurological disorders where sharp boundaries are no longer drawn either between health and neuropsychiatric and neurodevelopmental disorders, or among different disorders (e.g., Insel et al., 2010). Consequently, even the variability in the healthy population can inform our understanding of brain disorders. However, because the use of functional neural markers is still in its infancy, no consensus presently exists about which measures (e.g., effect size?, extent of activation?, degree of lateralization?) are the best ones to use. We here attempt to address this question with respect to one large-scale neural system: the set of brain regions in the frontal and temporal cortices that jointly support high-level linguistic processing (e.g., Binder et al., 1997; Fedorenko, Hsieh, Nieto-Castanon, Whitfield-Gabrieli, & Kanwisher, 2010). In particular, using data from 150 individuals all of whom had performed a language "localizer" task contrasting sentences and nonword sequences (Fedorenko et al., 2010), we: a) characterize the distributions of the values for four key neural measures of language activity (region effect sizes, region volumes, lateralization based on effect sizes, and lateralization based on volumes); b) test the reliability of these measures in a subset of 32 individuals who were scanned across two sessions; c) evaluate the relationship among the different regions of the language system; and d) evaluate the relationship among the different neural measures. Based on our results, we provide some recommendations for future studies of brain-behavior and brain-genes relationships. Although some of our conclusions are specific to the language system, others (e.g., the fact that effect-size-based measures tend to be more reliable than volume-based measures) are likely to generalize to the rest of the brain.

Introduction

The majority of studies in cognitive neuroscience seek to discover properties that are common to all individuals, to characterize an "average" human mind and brain. However, ways in which individuals differ from one another can also inform our understanding of human cognition. In psychology and cognitive science, investigations of individual differences in behavior have helped reveal the structure of – and the relationships among – many domains, including intelligence (e.g., Duncan et al., 2000; Gardner and Hatch, 1989; Kane and Engle, 2002; Spearman, 1904; Spearman, 1927), executive functions (e.g., Carlson et al., 2004; Colom, 2004; Conway, 1996; Mischel et al., 1989; Miyake et al., 2000), visual processing (e.g., Childers et al., 1985; Colombo et al., 1991; Vogel and Machizawa, 2004), social cognition (e.g., Herrmann et al., 2007; Miller and Saygin, 2013), speech perception (e.g., Surprenant and Watson, 2001), language comprehension (e.g., Daneman and Carpenter, 1980; Gernsbacher, 1991; Just and Carpenter, 1992; Pakulak and Neville, 2010; Traxler et al., 2012), music processing (e.g., Grahn and Schuit, 2012; Perrachione et al., 2013), and so on.

In addition to their importance for addressing questions in basic research, investigations of individual differences can shed light on neurological disorders. In particular, recent years have witnessed a shift in how mental illness is conceptualized, from the traditional, categorical, approach where sharp boundaries were drawn between health and neuropsychiatric and neurodevelopmental disorders, as well as among different neurological conditions (American Psychiatric Association,

* Corresponding author at: 77 Massachusetts Ave, Bldg 46-3037, Cambridge, MA 02139, United States.
E-mail address: kylemaho@mit.edu (K. Mahowald).

http://dx.doi.org/10.1016/j.neuroimage.2016.05.073
1053-8119/

2013) to a more probabilistic approach (e.g., Insel et al., 2010; Krug et al., 2010). Such a shift was inspired by a long-standing observation of variability present both within the healthy population and among individuals diagnosed with neurological disorders, combined with substantial overlap in the symptoms and genetic risk factors among disorders. This new way of thinking about mental illness calls for a shift in research practices: from group comparisons between, for example, individuals diagnosed with autism and neurotypical controls, to explorations of variability across large populations, to discover true endophenotypes.

Although a number of studies have attempted to link behavioral variability to genetic variability directly, including in the domain of language (e.g., Ocklenburg et al., 2013; Scerri et al., 2011; Whitehouse et al., 2011), *neural* markers are plausibly an important intermediate link given that genes shape the anatomy and functional organization of the brain, and these structural and functional characteristics of the brain in turn give rise to the observable behaviors. Indeed, neural markers – both anatomical and functional – are being used increasingly often in individual differences investigations of human cognition, including language (Cope et al., 2012; Krug et al., 2010; Landi et al., 2013; Pinel et al., 2012; Whalley et al., 2011). See Dubois and Adolphs (2016) for a thorough discussion of this approach.

At present, the use of *anatomical* markers is more common, plausibly due to the availability of large datasets, with hundreds, and sometimes thousands, of participants. Such datasets accumulate because most cognitive neuroscience labs routinely collect high-resolution structural scans from every participant. However, anatomical markers based purely on macroanatomy (e.g., the cortical thickness and/or volume of a macroanatomically defined brain area) have their limitations. In particular, the relationship between structure and function is a complex one, especially in the higher-order association cortices, where functional activations do not align well with the macroanatomical landmarks (e.g., Fischl et al., 2008; Frost and Goebel, 2012; Tahmasebi et al., 2012). For example, a well-characterized face-selective brain region – the fusiform face area (FFA; Kanwisher et al., 1997) – cannot be defined anatomically (e.g., Frost and Goebel, 2012). Consequently, markers of brain activity may provide a stronger link between genes and behavior, especially for higher-level cognitive processes. Furthermore, they can increase the power of anatomical investigations (e.g., studies examining cortical thickness) by enabling researchers to delineate the relevant brain regions more accurately than sulcal/gyral landmarks alone allow.

To successfully relate functional neural markers to genetic and behavioral variability, however, it is important – for each relevant cognitive function – to determine a) which markers are reliable (i.e., stable within individuals over time), and b) how different markers relate to one another. At present, in the domain of language research, different groups use different language tasks (e.g., semantic verbal fluency, verb generation, sentence completion, rhyme judgment tasks), focus on different brain regions (e.g., inferior frontal regions, regions in the middle temporal gyrus, or even regions outside of the core fronto-temporal language network), and examine different markers of neural activity (e.g., effect size for the relevant contrast in some region of interest, volume of an activated region, degree of lateralization of a region). Any one of prior studies individually can potentially reveal something important about language or cognition more broadly. However, the real power would come from the ability *to compare and replicate findings* across studies and research groups, to discover truly robust relationships. This could only be achieved if we, as a field, agreed on a set of tasks and measures that are reasonable, and adopted a set of guidelines for how to use those. For example, in increasingly more domains of study researchers use "functional localizer" tasks, which quickly and reliably identify a subset of the brain engaged in a particular mental activity (e.g., face-selective regions, Kanwisher et al., 1997; voice-selective regions, Belin et al., 2000; or regions engaged in theory of mind, Saxe and Kanwisher, 2003). Because labs that use functional localizers include a localizer scan in every participant, large datasets are eventually accumulated, as needed for brain-genes investigations. Further, because the same or comparable localizers are used across research groups, findings can be replicated across groups in a straightforward way.

Indeed, it has become increasingly clear that, in order to begin to link functional and behavioral data to underlying genetic variation, we will need large datasets involving hundreds or even thousands of participants. For instance, Stein et al. (2012), in a large-scale meta-analysis ($n = 7795$) of how genomic variation affects total brain volume, intracranial volume, and hippocampal volume, found that the largest observed effect (which was for hippocampal volume) explained only a tiny fraction of the variance. Hoogman et al. (2014) found similarly small effect sizes for the FOXP2 gene. Of course, it is possible that part of the difficulty in detecting these relationships between genetic and neural variation stems from the reliance on macroanatomical landmarks, which may fail to identify the "natural kinds" of the mind and brain, as discussed above. For example, the hippocampus is structurally and functionally diverse (e.g., Poppenk et al., 2013; Schoene-Bake et al., 2014; Travis et al., 2014) and perhaps detecting relationships between genetic variability and the volumes of its different subregions would be easier. However, even setting this issue aside, the effect sizes of the relationships between genetic and neural (anatomical or functional) variation are likely to be small because any given trait is a product of a vast number of genetic factors. Between small effect sizes and the huge space of possible variation in the genome, a well-powered study needs a large number of participants, such as the data now available from widely used functional localizer tasks.

We have recently developed methods for identifying the fronto-temporal system engaged in high-level linguistic processing using a contrast between sentences and sequences of nonwords (Fedorenko et al., 2010). This and similar contrasts have been used in many prior studies (e.g., words vs. fixation or tones: Binder et al., 1997; Diaz and McCarthy, 2009; words vs. pseudowords: Petersen et al., 1990; sentences vs. fixation: Kuperberg et al., 2003; sentences vs. false font or consonant strings: Bavelier et al., 1998; Noppeney and Price, 2004; Robertson et al., 2000; sentences vs. lists of words: Fedorenko and Kanwisher, 2011; Fedorenko et al., 2010; Snijders et al., 2009; speech vs. backwards or degraded speech: Bedny et al., 2011; Scott et al., in press), and we established that this contrast works robustly at the individual-subject level. We also demonstrated that this fronto-temporal language system exhibits a high degree of functional specificity: its regions respond robustly during language processing, but not during other complex cognitive tasks, like arithmetic processing, general working memory tasks or music perception (Fedorenko et al., 2011, 2012b). This system is thus functionally distinct from another large-scale brain network, which has a strong presence in the left prefrontal cortex: the bilateral fronto-parietal executive, or cognitive control, system (Duncan, 2013; Fedorenko et al., 2012a), and this dissociation holds even during naturalistic language comprehension (Blank et al., 2014).

The goal of the current study, which targets the fronto-temporal language system, is three-fold. *First*, using a large dataset of healthy adult participants ($n = 79$), we characterize activity in the language system in a number of ways: focusing on eight key language regions (Fig. 1) and their right-hemisphere homologs, we report the distributions of values for effect sizes, volumes, and lateralization (computed based on either effect sizes or volumes). These distributions clearly show that there is substantial variability to be explained even in the healthy population with respect to language activations. In addition, any new population can now be evaluated with respect to these normative distributions, be it older or younger individuals, left handers, learners of English as a second language, bi/multi-linguals, or individuals with neurodevelopmental or acquired disorders. The data for this set of participants are available at https://web.archive.org/web/20160608155930/https://evlab.mit.edu/papers/Mahowald_NI. *Second*, we evaluate the reliability of these functional measures in a subset of

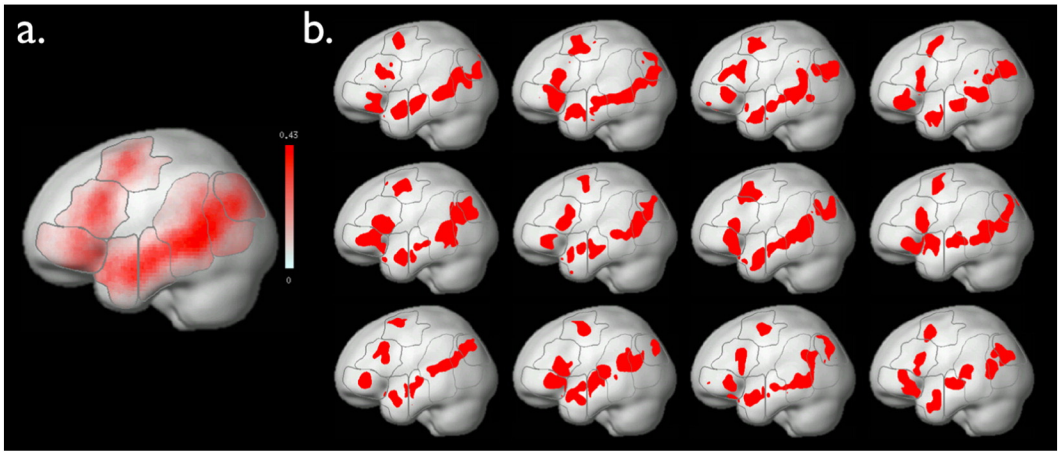


Fig. 1. a) Probabilistic activation overlap map and the parcels used to constrain the selection of individual functional ROIs. The overlap map was created by i) taking the 150 individual activation maps for the localizer (*Sentences* > *Nonwords*) contrast, ii) thresholding them at $p < 0.001$, uncorrected level, and iii) overlaying them on top of one another. The value in each voxel represents the proportion of individuals that show a significant effect (at the $p < 0.001$, uncorrected at the whole-brain level) in that voxel. The parcels were derived from a similar overlap map in a set of 25 individuals via a watershed image parcellation algorithm, as described in Fedorenko et al. (2010). b) Individual functional ROIs in a set of 12 sample participants. The fROIs were created by taking the top 10% of voxels for the localizer contrast within each parcel, as described in Methods.

32 individuals who were scanned across two sessions (as well as in the full set of 150 participants by looking at across-runs reliability). As Rombouts et al. (1998) showed for visual cortex, we find high within-subject reliability of language activations (see also Gorgolewski et al., 2013), and, like Cohen and Dubois (1999), we find that effect sizes are more stable than volume measures. And *third*, we examine the relationship in the full set of 150 participants i) among the different *brain regions* of the language network with respect to the key functional measures, and ii) among the different *functional measures*, in an effort to reduce the number of measures. Based on this examination, we provide some guidelines for future studies of brain-behavior and brain-genes relationships.

Methods

Participants

One hundred and fifty adult participants (105 females), aged 18 to 52 (average age: 23.5 years; standard deviation: 4.9 years) – students at MIT and members of the larger Boston community – participated for payment, between September 2007 and September 2012. A subset (n = 32) was scanned twice on the language localizer task, on different days, between 0.5 and 58 months apart (average time between sessions: 8.5 months; standard deviation: 11.2 months). Participants were right-handed (by self report) native speakers of English, naïve to the purposes of the study. All participants gave informed consent in accordance with the requirements of the MIT's Committee on the Use of Humans as Experimental Subjects (COUHES).

Language localizer task

Participants read sentences (e.g., A RUSTY LOCK WAS FOUND IN THE DRAWER) and lists of pronounceable nonwords (e.g., DAP DRELLO SMOP UL PLID KAV CRE REPLODE) in a blocked design. Six slightly different versions of the localizer task were used across the 150 participants, as summarized in Tables 1 and 2 (see also SI-2). In previous work, we established that the localizer contrast is robust to changes in materials, task and modality of presentation (Fedorenko et al., 2010; Fedorenko, 2014). Each participant saw between 12 and 32 blocks per condition. The task was a memory probe task for all but six participants: a probe word/nonword appeared at the end of each trial and participants had to decide whether it was present in the trial. The remaining six participants, who did the *SWJN_v1_ips252* version (see Table 2), performed a simple button-pressing task ("press a button at the end of each sentence/sequence"). We include these participants because there is no reason to expect meaningful differences based on the task (Fedorenko, 2014).

fMRI data acquisition

Structural and functional data were collected on the whole-body 3 Tesla Siemens Trio scanner with a 12-channel head coil (n = 26/150; n = 4/32 of the 2nd session of the participants scanned twice and

Table 1
Information on which subsets of participants in the set of n = 150 performed which version of the language localizer (see Table SI-1 for the details of the three scanning sequences used; see Table SI-2 for information on the localizer versions for the 1st vs. 2nd session of the subset of 32 participants scanned twice and examined in the across-session reliability analyses).

Number of participants	Language localizer version	Runs/blocks per condition	Coil	Sequence
n = 8	SWJN_v1_ips252	n = 7: 8 runs/32 blocks; n = 1: 7 runs/28 blocks	12	n = 6: #1; n = 2: #2

n = 18	SWJN_v2_ips232	n = 5: 8 runs/32 blocks;
		n = 6: 6 runs/24 blocks;
		n = 7: 4 runs/16 blocks
			12	#2

n = 69	SWNloc_ips168	n = 18: 5 runs/20 blocks;
		n = 48: 4 runs/16 blocks;
		n = 3: 3 runs/12 blocks
			32	#2

n = 5	SNloc_ips232	2 runs/16 blocks	32	#2
n = 33	SNloc_ips189	2 runs/16 blocks	32	#3
n = 17	SWNloc_ips198	n = 15: 3 runs/18 blocks; n = 2: 2 runs/12 blocks	32	n = 2: #2; n = 15: #3

Table 2
Procedure and timing details for the six different versions of the language localizer.

SWJN_v1_ips252	
• Conditions	Sentences, word lists, Jabberwocky sentences, and nonword lists
• Materials	12-Word/nonword-long sequences (see Expt 1 in Fedorenko et al., 2010)
• Expt block duration	24 s
• Trials per block	5
• Trial duration	4.8 s
• Trial structure	* 600 ms of trial-initial fixation; * 12 words/nonwords presented for 350 ms each
• Expt blocks per run	16 (4 per condition)
• Fix block duration	24 s
• Fix blocks per run	5
• Run duration	504 s
SWJN_v2_ips232	
• Conditions	Sentences, word lists, Jabberwocky sentences, and nonword lists
• Materials	8-Word/nonword-long sequences (see Expt 2 in Fedorenko et al., 2010)
• Expt block duration	24 s
• Trials per block	5
• Trial duration	4.8 s
• Trial structure	* 300 ms of trial-initial fixation; * 8 words/nonwords presented for 350 ms each; * 350 ms probe * 1000 response window * 350 ms trial-final fixation
• Expt blocks per run	16 (4 per condition)
• Fix block duration	16 s
• Fix blocks per run	5
• Run duration	464 s
SWNloc_ips168	
• Conditions	Sentences, word lists, and nonword lists
• Materials	8-Word/nonword-long sequences (same as in SWJN_v2_ips232)
• Expt block duration	24 s
• Trials per block	5
• Trial duration	4.8 s
• Trial structure	* 300 ms trial-initial fixation * 8 words/nonwords presented for 350 ms each; * 1350 ms probe * 350 ms trial-final fixation
• Expt blocks per run	12 (4 per condition)
• Fix block duration	16 s
• Fix blocks per run	3
• Run duration	336 s
SNloc_ips232	
• Conditions	Sentences and nonword lists
• Materials	8-Word/nonword-long sequences (same as in SWJN_v2_ips232)
• Expt block duration	24 s
• Trials per block	5
• Trial duration	4.8 s
• Trial structure	* 300 ms trial-initial fixation * 8 words/nonwords presented for 350 ms each; * 1350 ms probe * 350 ms trial-final fixation
• Expt blocks per run	16 (8 per condition)
• Fix block duration	16 s
• Fix blocks per run	5
• Run duration	464 s
SNloc_ips189	
• Conditions	Sentences and nonword lists
• Materials	12-Word/nonword-long sequences (same as in SWJN_v1_ips252) [NB: 7 participants did a version with the sentences taken from the Brown corpus]
• Expt block duration	18 s
• Trials per block	3
• Trial duration	6 s
• Trial structure	* 300 ms trial-initial fixation * 12 words/nonwords presented for 350 ms each; * 1000 ms probe * 500 ms trial-final fixation
• Expt blocks per run	16 (8 per condition)
• Fix block duration	18 s
• Fix blocks per run	5
• Run duration	378 s
SWNloc_ips198	
• Conditions	Sentences, word lists, and nonword lists
• Materials	12-Word/nonword-long sequences (same as in SWJN_v1_ips252)
• Expt block duration	18 s
• Trials per block	3
• Trial duration	6 s
• Trial structure	* 300 ms trial-initial fixation * 12 words/nonwords presented for 350 ms each; * 1000 ms probe * 500 ms trial-final fixation
• Expt blocks per run	18 (6 per condition)
• Fix block duration	18 s
• Fix blocks per run	4
• Run duration	396 s

examined in the across-session reliability analyses) or a 32-channel head coil (n = 124/150; 28/32) at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT. Functional, blood oxygenation level dependent (BOLD), data were acquired using one of three similar EPI sequences (see Table SI-1 for details). All three sequences had a 90 degree flip angle and had the following acquisition parameters: 33 (sequence #1) or 31 (sequences #2 and 3) 4-mm-thick near-axial slices acquired in the interleaved order (with 10% distance factor), 3 mm × 3 mm (sequence #1) or 2.1 mm × 2.1 mm (sequences #2 and 3) in-plane resolution, FoV in the phase encoding (A ≫ P) direction 192 mm (sequence #1) or 200 mm (sequences #2 and 3) and matrix size 64 mm × 64 mm (sequence #1) or 96 mm × 96 mm (sequences #2 and 3), TR = 2000 ms and TE = 30 ms. The first several seconds of each run were excluded to allow for steady state magnetization.

Preprocessing and first-level analyses

MRI data were preprocessed and analyzed using SPM5 and custom Matlab scripts (available – in the form of an SPM toolbox – from http://www.nitrc.org/projects/spm_ss). Each subject's data were motion corrected and then normalized into a common brain space (the Montreal Neurological Institute, MNI template) and resampled into 2 mm isotropic voxels. The data were then smoothed with a 4 mm Gaussian filter and high-pass filtered (at 200 s). The effects were estimated using a General Linear Model (GLM) in which each experimental condition was modeled with a boxcar function (scaled between 0 and 1) convolved with the canonical hemodynamic response function (HRF, scaled to have unit integral). The boxcar function modeled entire blocks (the fixation condition was modeled implicitly). More specifically, for each condition, we estimated from the experimental design the expected BOLD-response changes (up to a scaling factor), and then used a GLM to get the values of the scaling factors (the beta volumes associated with each condition). The BOLD signal entered in the GLM estimation step has been scaled by the average global signal (SPM's session-specific grand mean scaling procedure) so that the associated scaling factors can be interpreted in percent signal change (PSC) units.

Critical analyses

Regions of interest (ROIs) were defined functionally in each individual participant using the *Sentences* > *Nonwords* contrast. To do so, we used the Group-constrained Subject-Specific (GSS) analysis method developed in Fedorenko et al. (2010). In this analysis, a group-level representation of the activations is used to divide up the activation landscape into regions (what we refer to as "parcels"). These parcels are subsequently used to constrain the selection of individual-level

functional ROIs. Here, we used the parcels (Fig. 1a) generated for the *Sentences* > *Nonwords* contrast in a set of 25 participants as described in Fedorenko et al. (2010). (Nineteen of these 25 participants are included in our set of 150 participants.). These parcels were intersected with each individual subject's activation map for the *Sentences* > *Nonwords* contrast, as described in more detail below. We here focus on the "core" set of eight regions on the lateral surfaces of the left frontal, temporal and parietal cortices (Fig. 1a) – which most robustly and consistently emerge in the investigations of the language system – and their right-hemisphere (RH) homologs, for a total of 16 regions.

Using the individual activation maps for the *Sentences* > *Nonwords* contrast, we extracted two measures from each of the 16 ROIs:

- effect size for the *Sentences* > *Nonwords* contrast;
- volume based on the *Sentences* > *Nonwords* contrast.

To do so, we defined individual functional ROIs in two ways. In particular, to compute the *effect size measure*, for each subject we sorted the voxels within each parcel based on their *t*-values for the localizer contrast, and the top 10% of voxels were chosen as that subject's functional ROI (see Fig. 1b for sample fROIs). To ensure the independence of the data used for fROI definition and for response estimation (Kriegeskorte et al., 2009), we used an across-runs cross-validation procedure. In particular, each subject's activation map was computed for the *Sentences* > *Nonwords* contrast using all but one run of data, and the 10% of voxels with the highest *t*-value within a given parcel were selected as that subject's fROI. The response to the two conditions (sentences and nonwords) was then estimated using the left-out run. This procedure was iterated across all possible partitions of the data, and the responses were then averaged across the left-out runs to derive a single response magnitude for each condition for a given region and subject. This *n*-fold cross-validation procedure (where *n* is the number of functional runs) allows one to use all of the data for defining the ROIs and for estimating the responses (Nieto-Castañón and Fedorenko, 2012). Statistical tests described below were performed on the percent BOLD signal change (PSC) values extracted from these fROIs.

To compute the *volume measure*, for each subject we counted the number of voxels that fell within each parcel and that were significant for the localizer, *Sentences* > *Nonwords*, contrast at a fixed threshold ($p < 0.001$, uncorrected at the whole-brain level). Statistical tests described below were performed on these voxel count values.

We further computed two *measures of lateralization* for each of the eight regions: an effect-size-based measure and a volume-based measure. For the latter, more traditional, measure (Binder et al., 1997; Hinke et al., 1993), we used the following formula (e.g., Seghier et al., 2008): (number of voxels in the LH − number of voxels in the RH) / (number of voxels in the LH + number of voxels in the RH). For the former measure, we simply subtracted the *Sentences* > *Nonwords* effect size in the right hemisphere from the *Sentences* > *Nonwords* effect size in the left hemisphere. We did not divide by effect size since, in cases where effect size was very small, lateralization would be far too large. The measure used is roughly normally distributed.

Results

Inter-individual variability in the language activations

Starting with the full set of 150 participants, we analyzed whether there were meaningful differences among participants' activations based on a) the head coil (12-channel or 32-channel) and b) the sequence (Table SI-1). We found potentially meaningful effects for both, as discussed in the Appendix. Therefore, we restricted our analysis in this section to 76 participants scanned using the 32-channel coil and the sequence most frequently used with that coil (sequence #3; see Table SI-1). In evaluating any new individual or set of individuals relative to these distributions, we therefore recommend using data collected with a 32-channel coil and a sequence comparable to our sequence #3. (We use the full n = 150 dataset in our analysis of the reliability of functional measures because potential coil- and sequence-based differences could only lower the reliability making our estimates conservative, and in the analyses or both inter-region correlations and inter-measure correlations because the differences are not relevant to the questions asked). Having said that, because of the possibility of different scanners producing different effect sizes (Friedman et al., 2006, 2008), we recommend that new scanners be "calibrated" using the existing data to make sure that the data obtained is comparable to the reference distributions provided here.

We characterized the functional language activations of 76 participants in 16 ROIs (8 in each hemisphere), for the following measures (see Fig. 2):

a) effect size for the *Sentences* > *Nonwords* contrast;
b) volume based on the *Sentences* > *Nonwords* contrast;
c) lateralization based on the *Sentences* > *Nonwords* effect size; and
d) lateralization based on the *Sentences* > *Nonwords* volume.

The values for the effect size measure are roughly normally distributed in each region and show large variation across individuals. For example, in the LIFG fROI, a representative ROI, we observed a roughly normal distribution, with a mean effect of .69 and a standard deviation of .46. At the group level, the *Sentences* > *Nonwords* effect was significantly different from 0 in every ROI in both the LH ($ts > 16$; $ps < 0.0001$; $df = 75$) and the RH ($ts > 6$; $ps < 0.0001$; $df = 75$), although the effect size was substantially lower in the RH ROIs (2.7 times lower on average across regions).

The values for the volume measure (which cannot be below 0) are roughly exponentially distributed in the LH regions. Furthermore, we observed an asymmetry between the hemispheres, with larger regions of activation in the LH. However, individuals vary widely with respect to their RH activations: whereas the majority of individuals show a near-0 activation across regions in the right hemisphere, some individuals show large activations, as can be seen in the long tails in the right column of the Volume plot (Fig. 2b). This variability in the amount of RH activation leads to substantial variability in the volume-based lateralization scores (Fig. 2d).

For the effect-sized measure, there are also some participants who show more extreme right-hemisphere activation, but the values are roughly normally distributed. Thus, the effect-size-based lateralization values are not as sharply skewed as those for volume-based lateralization, as can be seen in Fig. 2c.

The reliability of different neural measures of language activity

In addition to characterizing the distributional properties of the key functional measures of language activity in the population, it is critical to measure their reliability, because it is only sensible to try to relate a neural measure to some aspect of behavior or genetics if it is stable in an individual over time.

We began by examining *broad similarity in the activation patterns* within each of the sixteen parcels. Such similarity is apparent when visually examining whole-brain activation maps (e.g., Fig. 3). To quantify this similarity, we examined the correlations between the contrast values for the *Sentences* > *Nonwords* contrast across i) odd- and even-numbered runs, for the full set of 150 participants, and ii) the two sessions, for the subset of 32 participants who were scanned twice on the language localizer task, on different days, 8.5 months apart on average. As can be seen in Fig. 4, activation patterns were highly similar both across runs and across sessions within participants. For the *across-runs comparison*, the Fisher-transformed correlations were above 0.66 and as high as 1 for the LH regions ($ts > 19$; $ps < 0.0001$; $df = 149$) and

above 0.44 for the RH regions ($ts > 11$; $ps < 0.0001$; df = 149). The values were extremely similar for the *across-sessions comparison* within participants: the Fisher-transformed correlations were above 0.64 for the LH regions ($ts > 9$; $ps < 0.0001$; df = 31) and above 0.45 for the RH regions ($ts > 6$; $ps < 0.0001$; df = 31). These results suggest that the activation patterns are highly stable within individuals over time both within and across scanning sessions. (See Fig. 5.)

For the control analysis where the values from the two sessions were paired randomly across participants (e.g., session 1 in participant 1 vs. session 2 in participant 2; cf. session 1 vs. 2 in the same participant),

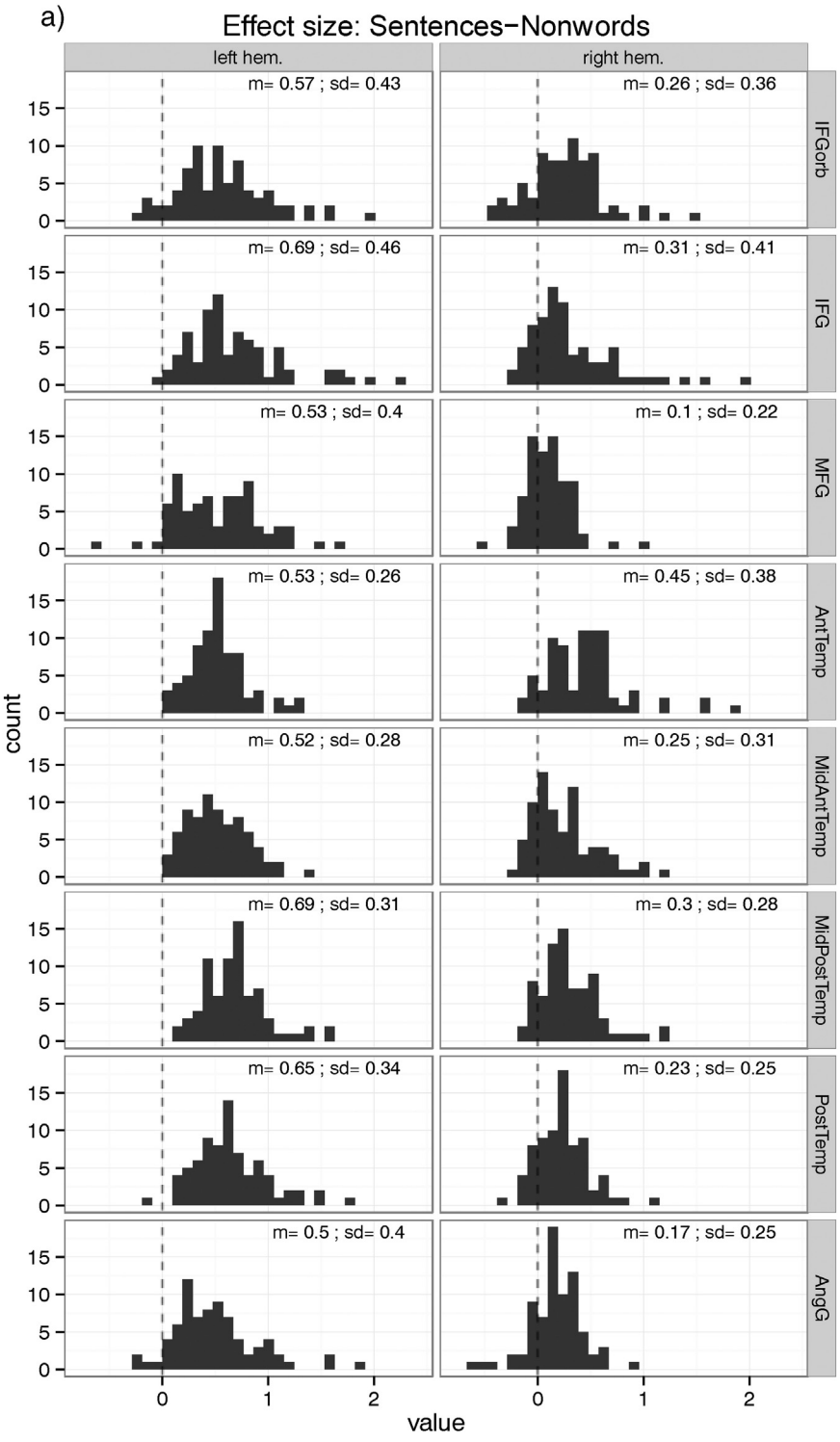


Fig. 2. a: The distribution of effect size measures for the *Sentences* > *Nonwords* contrast. The x-axis shows effect size (in percent BOLD signal change), and the y-axis shows the number of subjects showing each effect size. b: The distribution of volume measures for the *Sentences* > *Nonwords* contrast. The x-axis shows volume (in number of 2 mm^3 voxels), and the y-axis shows the number of subjects with each volume. c: The distribution of lateralization measures based on the *Sentences* > *Nonwords* effect size. The x-axis shows the strength of lateralization (positive values = left lateralization, negative values = right lateralization), and the y-axis shows the number of subjects with each lateralization value. d: The distribution of lateralization measures based on the *Sentences* > *Nonwords* volume. The x-axis shows the strength of lateralization (positive values = left lateralization, negative values = right lateralization), and the y-axis shows the number of subjects with each lateralization value.

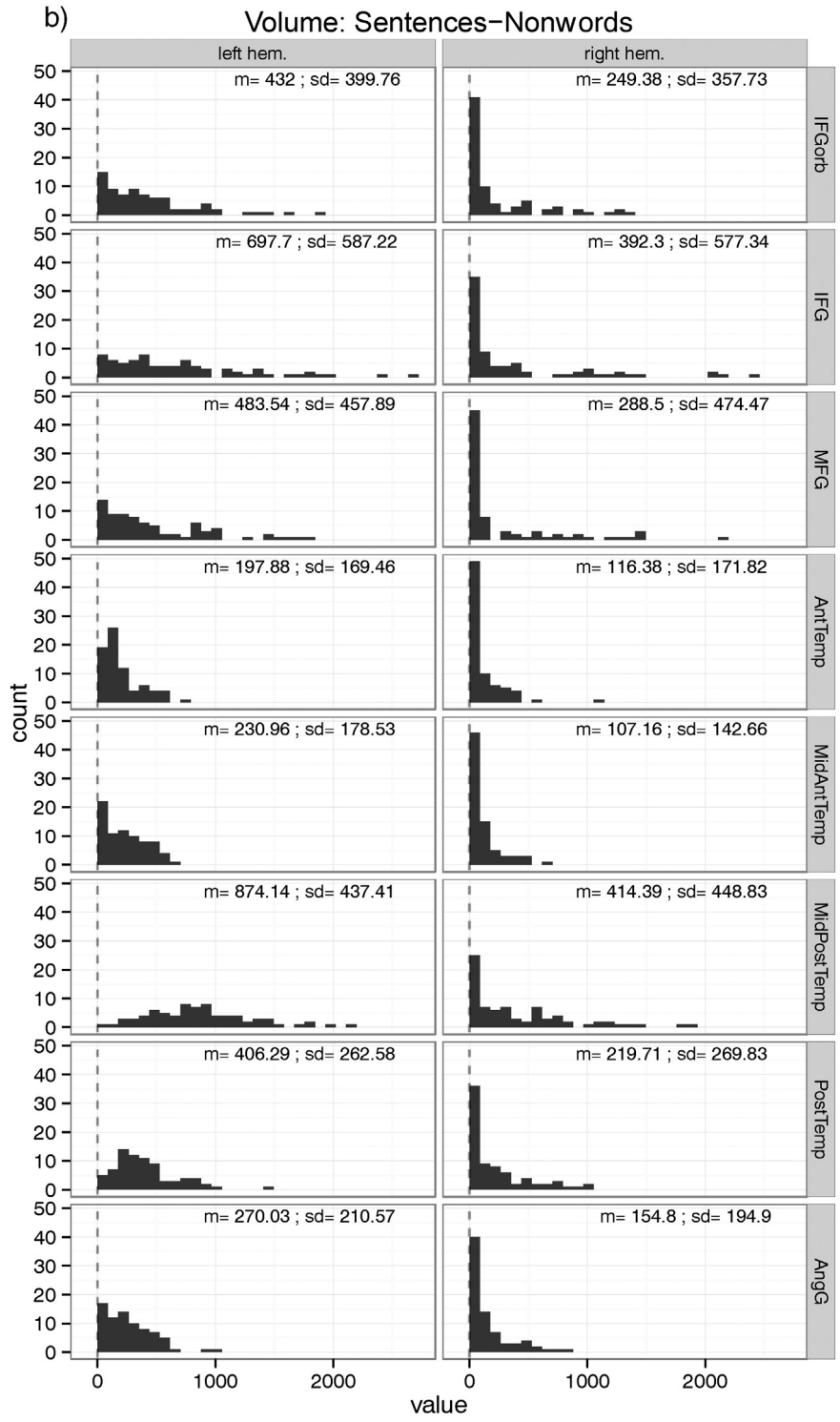


Fig. 2 (*continued*).

the correlations were drastically lower (<0.2 for the LH regions, and <0.16 for the RH regions). Indeed, a linear regression revealed that the correlations between participants were significantly lower than the correlations within participants between sessions (beta = .51, t = 10.4, p < .0001) and within participants between runs of the same session (beta = .54, 11.0, p < .0001). In a post-hoc direct comparison of the correlations within participants within session (across runs) versus the correlations within participants across sessions, there was no significant difference (beta = .03, t = .47, p = .64). Even in the across-subjects analyses, the correlations were still reliable in most regions, reflecting some degree of across-subjects similarity in the activation patterns (LH regions: ts > 2.4; ps < 0.05; df = 31; RH regions: ts > 1.2; ps < 0.01, except for the RMFG region, where the correlation did not reach significance; df = 31). The much lower correlations for the randomly paired sessions reflect the inter-individual variability apparent in Figs. 1 and 3 and documented previously (e.g., Fedorenko et al., 2010).

Next, we examined the reliability of our four functional measures (effect size, volume, and the two lateralization measures) in more detail by correlating each participant's first-session values and second-session

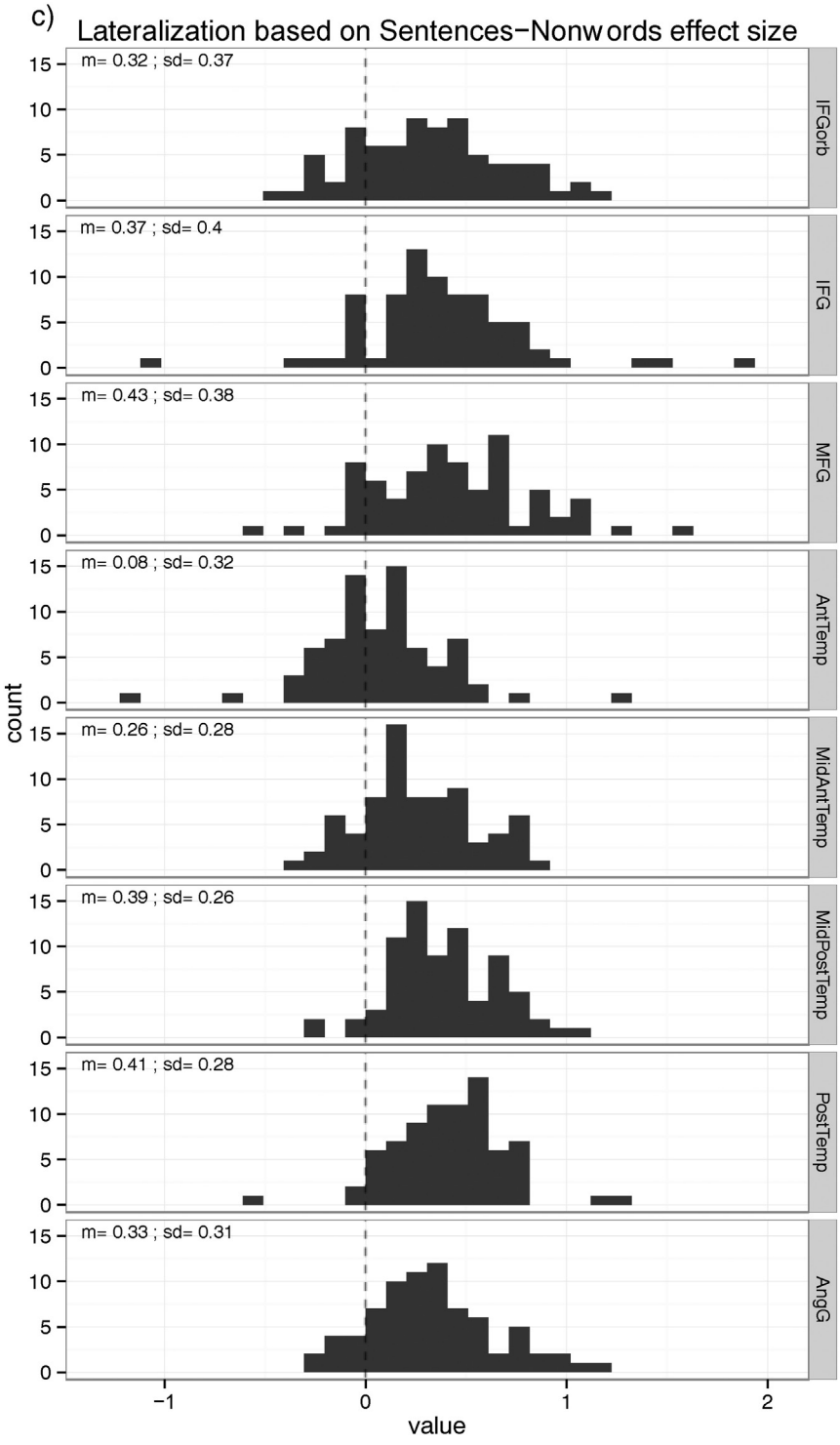


Fig. 2 (*continued*).

values for each region separately. Reliable measures are ones that have similar values in the two sessions. For the *Sentences > Nonwords effect size*, seven of the eight LH regions showed a correlation greater than 0.4, and these correlations were significantly different from 0 ($ps < 0.05$; 3 surviving multiple comparison correction for the 16 regions tested). The only exception was the LAngG region with a correlation of 0.29, which did not reach significance. A similar pattern was observed for the *lateralization based on the effect size* measure: seven of the eight regions (all but LAngG) were significantly positive – with *r*s above 0.5 and as high as 0.86 – at $p < 0.05$ and were still significant after Bonferroni correction for the 8 regions tested. In contrast, the *volume* measure only showed a significant correlation in two of the eight regions: LIFG and LPostTemp ($ps < 0.05$). And the *lateralization based on the volume* measure showed a significant correlation in five of the eight regions ($ps < 0.05$). It therefore appears that effect-size-based measures are more reliable within subjects than volume-based measures (see Cohen and Dubois (1999), for a similar conclusion). Time between the two sessions (which varied from 0.5 to 58 months) did not

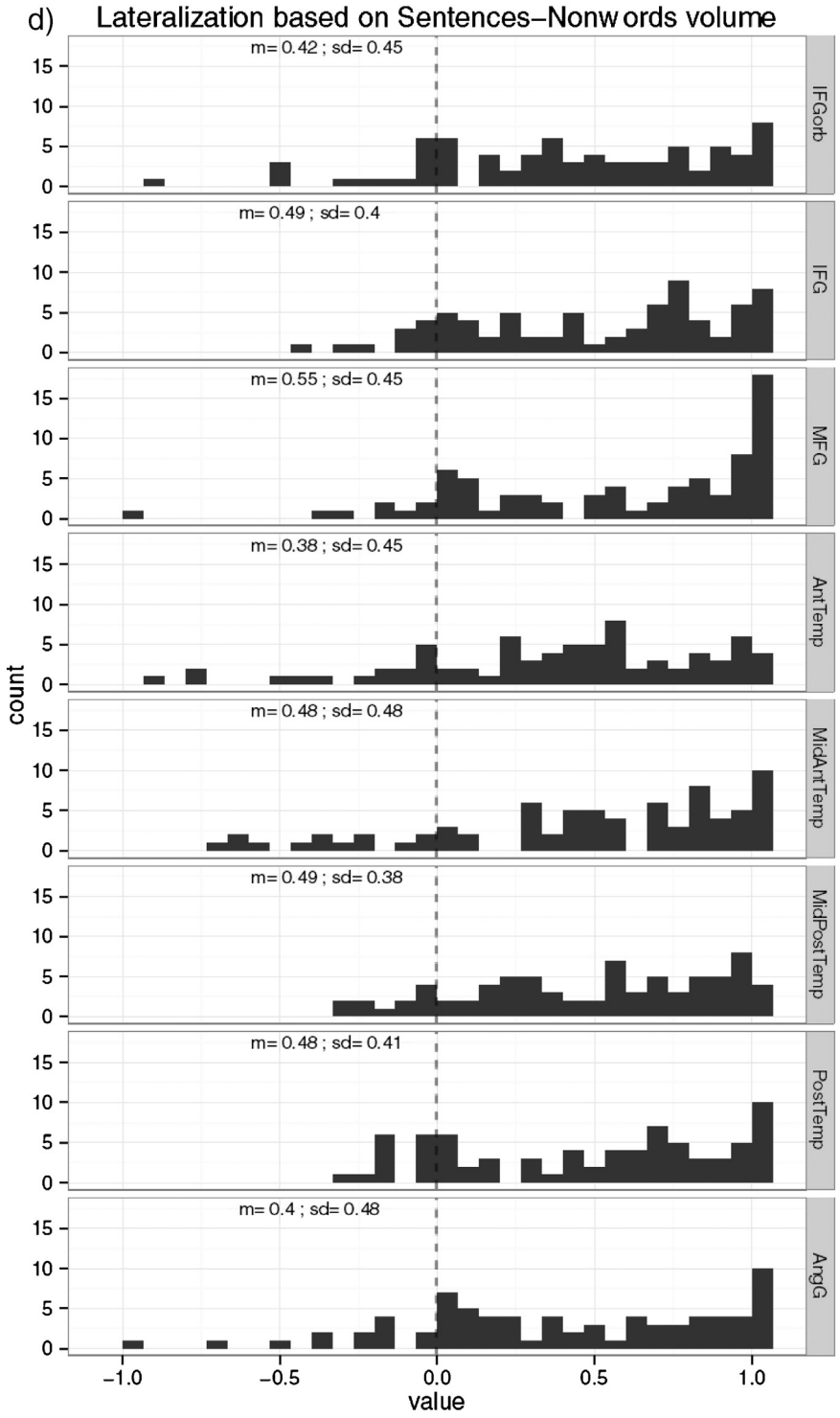


Fig. 2 (*continued*).

appear to affect the reliability of the measures (all *ps* > .05, with numerically small estimated betas), suggesting that it is not the case that activation patterns become more dissimilar with longer intervals between scans. This result is consistent with the fact that activation patterns were highly similar both between the runs within a session as well as between sessions (see Fig. 4). It is also worth noting that Gorgolewski et al., 2013 have previously established that several other factors do not much affect within-subject reliability measures, including scanner noise and co-registration errors. By far the most variance in reliability was explained by the paradigm, with some paradigms being more robust than others. It is also bound to be the case that the amount of data collected from each participant plays a large role, which is why we always have at least 8–10 blocks per condition in our experiments.

Some proposals exist for more sophisticated measures of lateralization (e.g., Wilke and Lidzba, 2007). We leave it to future work to evaluate whether some other measures may prove to be more stable across time.

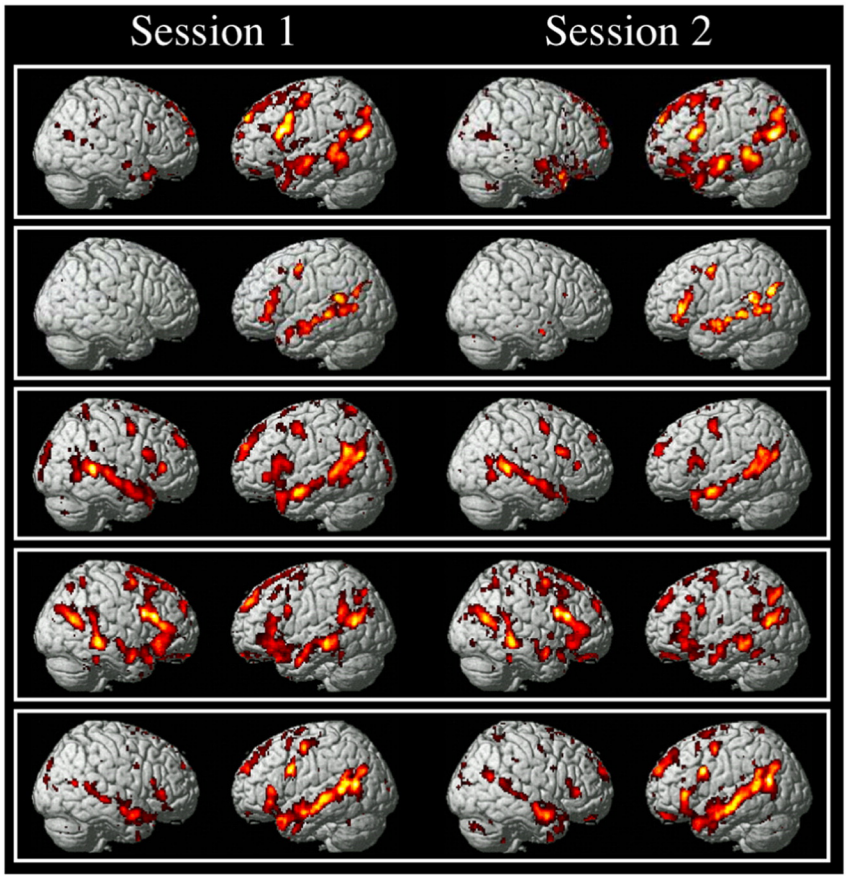


Fig. 3. Sample activation maps from the two sessions of 5 participants scanned twice on the language localizer task.

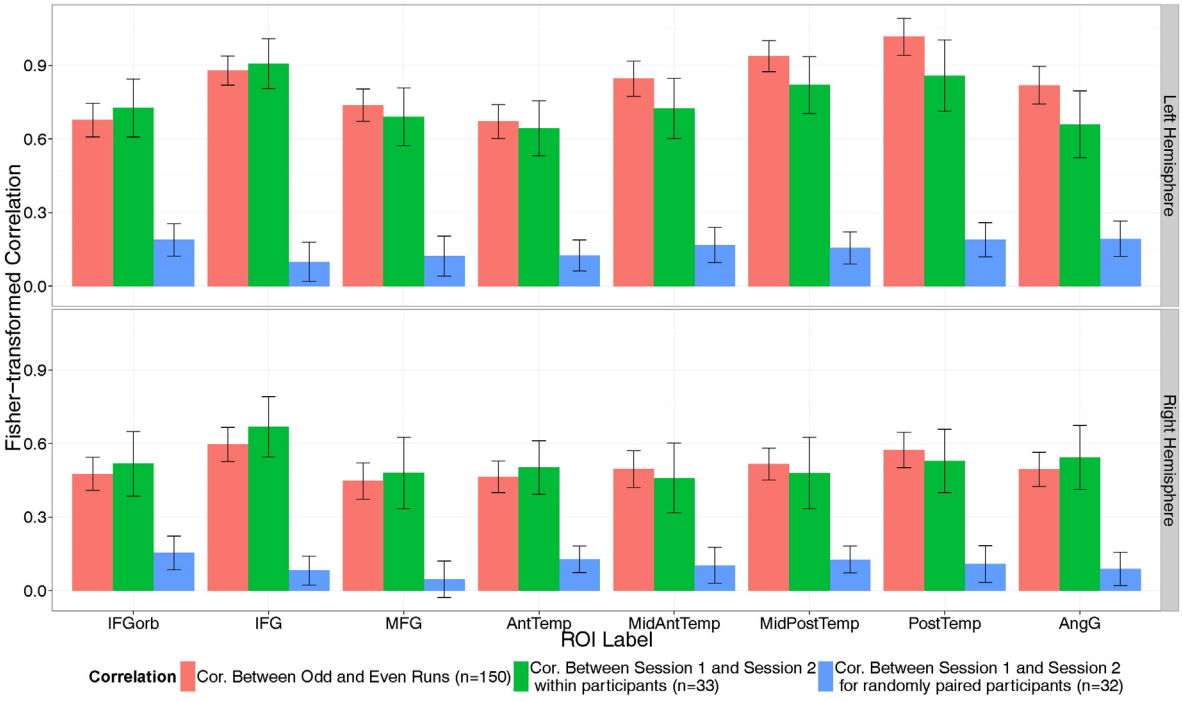


Fig. 4. Correlations between the activation patterns (i.e., the vectors of contrast values with one value per voxel) in each of the LH and RH parcels for the *Sentences* > *Nonwords* contrast between odd- and even-numbered runs within a session in the full set of 150 participants (red bars), between the two sessions in the 32 participants who were scanned twice on the language localizer task (green bars), and between the two sessions taken from different participants (e.g., session 1 in participant 1 vs. session 2 in participant 2; blue bars).

The relationship among the different language regions

In the analyses above, following prior work (Fedorenko et al., 2010), we used a set of eight parcels to constrain the selection of subject-specific functional ROIs. As discussed above, these parcels were derived from a group-level representation of language activity. In particular, a watershed algorithm was applied to the probabilistic overlap map for the *Sentences* > *Nonwords* contrast in a set of 25 participants (see Fedorenko et al., 2010 for details; see Julian et al., 2012, for an application of this approach to high-level visual cortical regions) to derive a set of parcels corresponding to regions of activation that are spatially consistent across individuals. We here focus on eight "core" cortical language regions in the frontal and temporal/temporo-parietal cortices (and their RH homologs; cf. Fedorenko et al., 2010, for discussion of several additional regions that emerge as consistently active across people). However, the question of how to carve up the language network

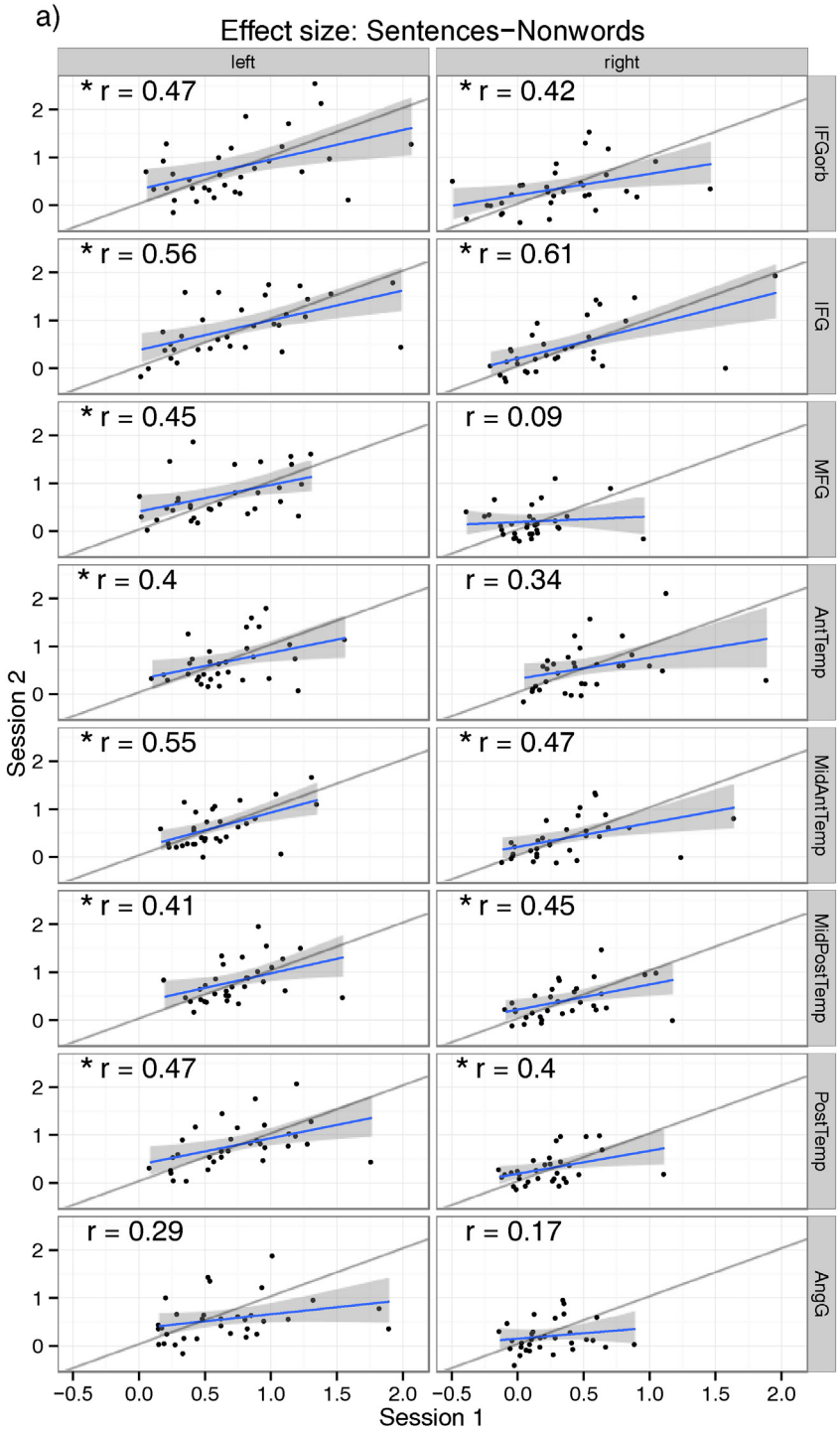


Fig. 5. a: The reliability of the *Sentences* > *Nonwords* effect size measure in a subset of 32 participants scanned across two sessions. An asterisk before the *r*-value indicates (uncorrected) statistical significance ($p < 0.05$). b: The reliability of the *Sentences* > *Nonwords* volume measure in a subset of 32 participants scanned across two sessions. An asterisk before the *r*-value indicates (uncorrected) statistical significance ($p < 0.05$). c: The reliability of the effect-size-based lateralization measure in a subset of 32 participants scanned across two sessions. An asterisk before the *r*-value indicates (uncorrected) statistical significance ($p < 0.05$). d: The reliability of the volume-based lateralization measure in a subset of 32 participants scanned across two sessions. An asterisk before the *r*-value indicates (uncorrected) statistical significance ($p < 0.05$).

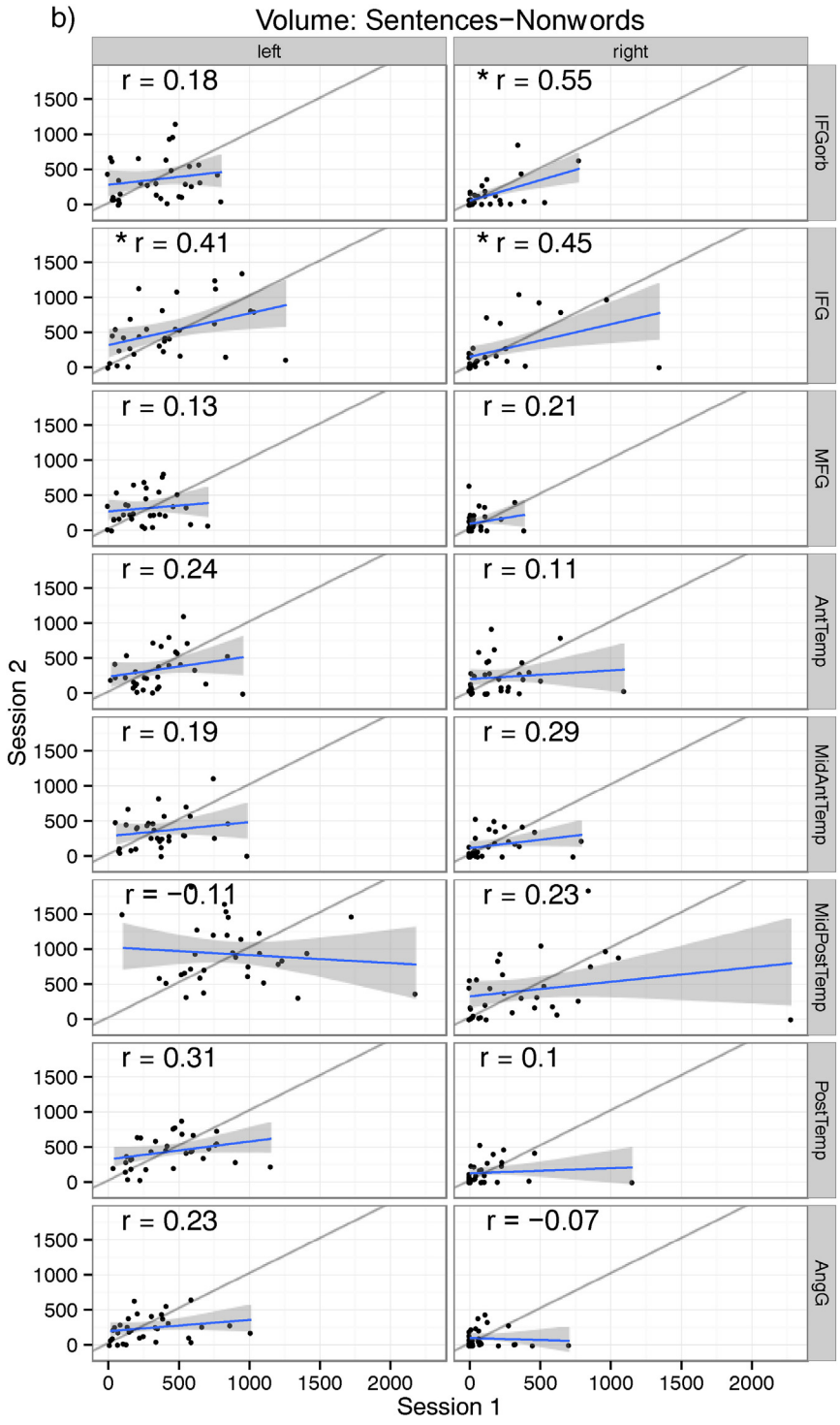


Fig. 5 (*continued*).

into regions – or whether such division is even warranted – remains open (Fedorenko and Thompson-Schill, 2014). Moreover, a number of studies have suggested that regions within the language network are correlated in their activity both during rest (e.g., Blank et al., 2014; Tie et al., 2014; Turken and Dronkers, 2011) and during naturalistic language comprehension (Blank et al., 2014; Yue et al., 2013). It is therefore important to know how correlated the different fROIs are with respect to our four measures. This issue is important for the problem of multiple comparisons in brain-behavior and brain-genetics investigations: in particular, if a set of language regions show little or no correlation in their functional profiles, they should be treated as independent, and hypotheses tested should be corrected for the number of regions examined; if, on the other hand, a set of language regions show highly correlated functional profiles, they should not be treated as independent, and a milder form or the multiple-comparisons correction may be appropriate.

To evaluate the relationship among our sixteen regions, we calculated the correlation across 150 subjects between each pair of regions for each functional measure, except for region volume, which is omitted due to its low across-session reliability as shown in our section on

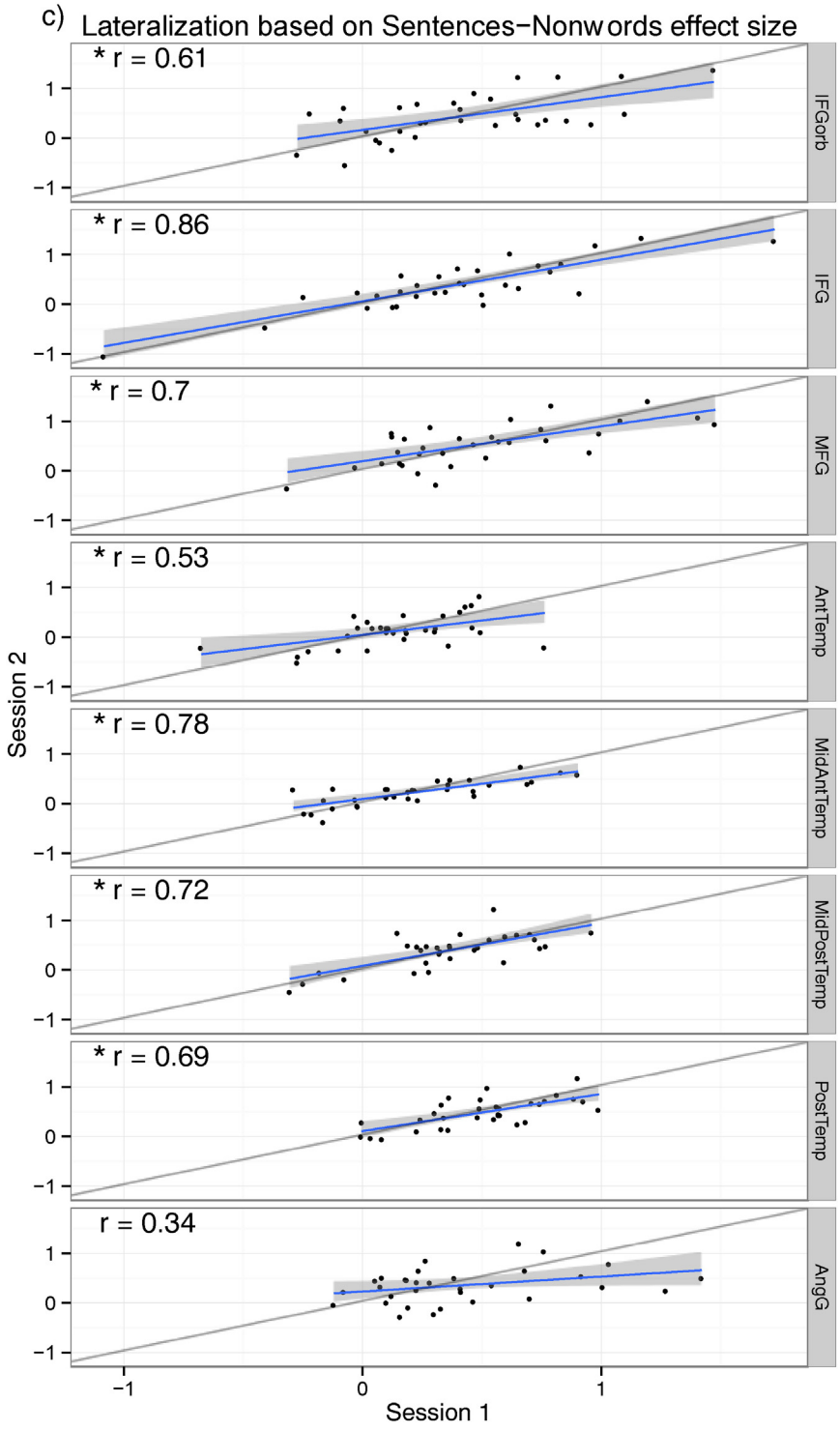


Fig. 5 (*continued*).

"The reliability of different neural measures of language activity." In line with prior findings of correlated time-courses among the different regions (e.g., Blank et al., 2014), we found correlations across subjects between each fROI and every other fROI, as shown in Fig. 6. (Note that we do not correct for multiple comparisons here given that we are not trying to make claims about the significance for any particular pair of regions; in a dataset consisting of pure noise, we would expect 5% of our correlations to be significantly different from 0.). The three measures show comparable mean levels of correlation across regions: the mean correlation among regions for the *Sentences* > *Nonwords* effect size is 0.53; for the effect-size-based lateralization is .34; and for the volume-based lateralization is 0.44.

Notably, however, it is clear that some pairs of regions are more correlated than others. To further assess the relationship among regions, we ran a clustering algorithm on the three functional measures that showed high across-sessions reliability (effect size, effect-size-based lateralization, and volume-based lateralization) to determine which pairs of ROIs show the most similar responses

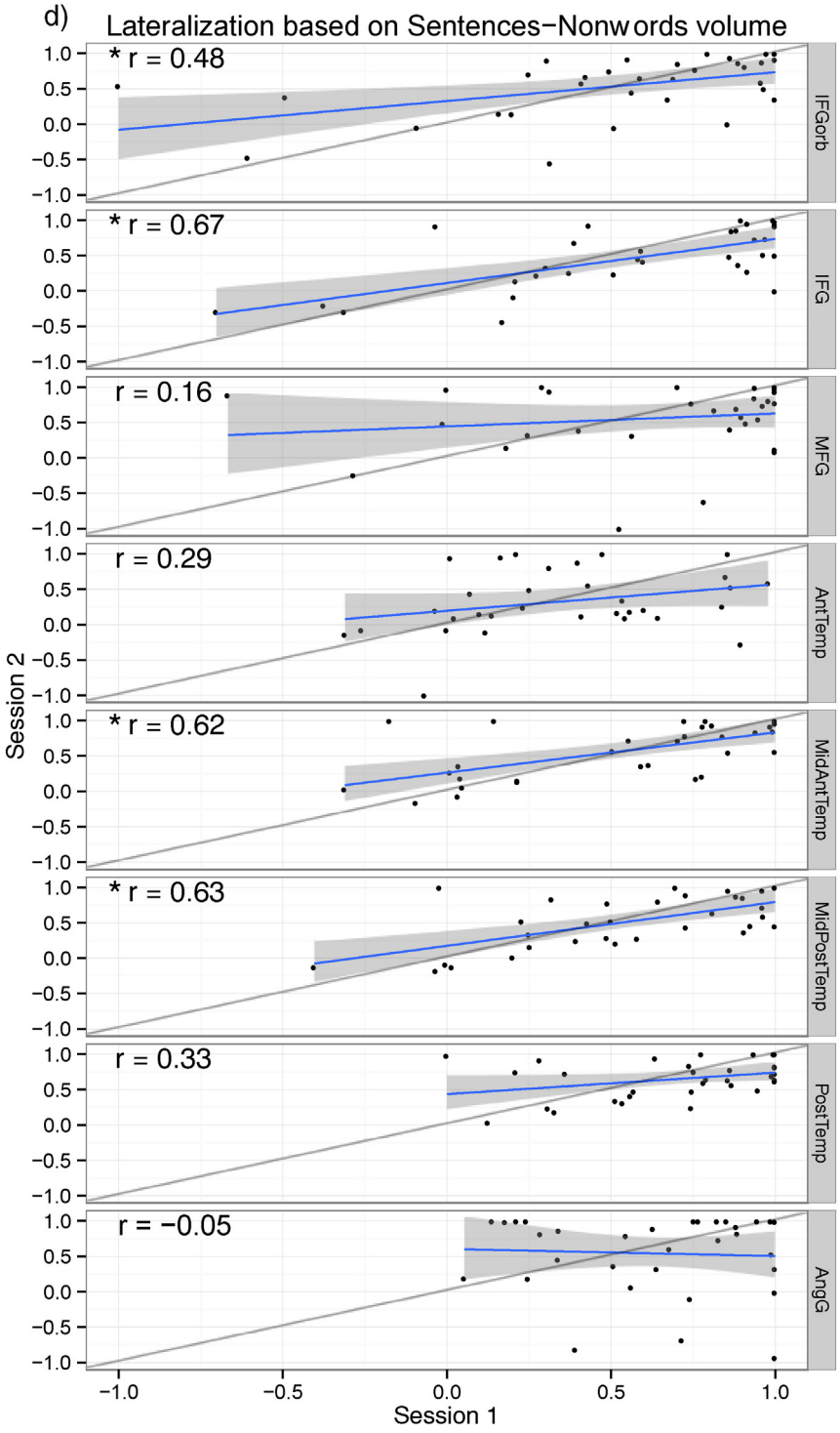
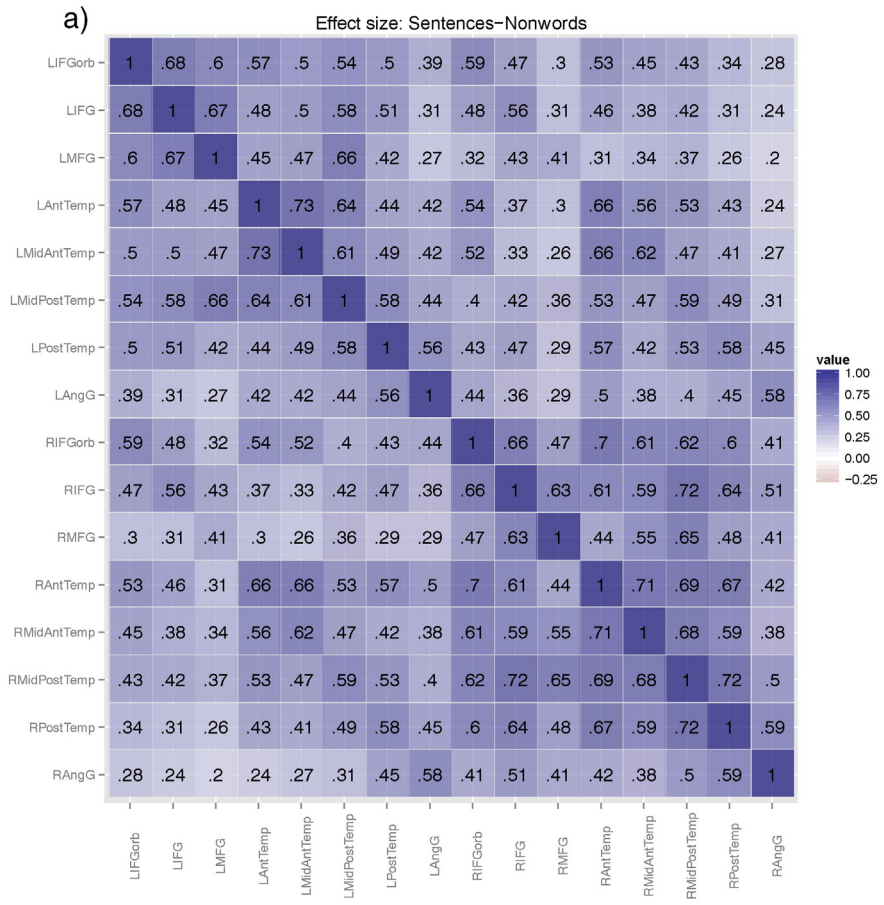
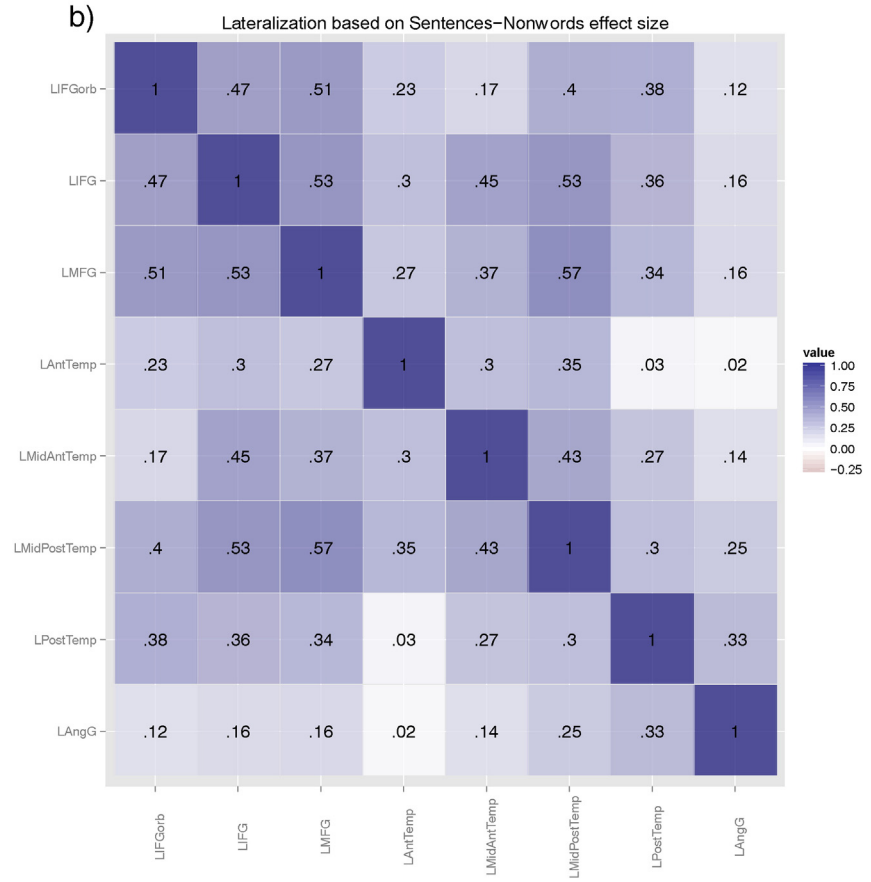


Fig. 5 (*continued*).

across the 150 participants. To do so, we first normalized the functional measures for each ROI across participants such that each functional measure for each ROI had a mean value of 0 across participants. Without this normalization procedure, the clustering algorithm would cluster regions largely based on the overall strength of activation (such that regions with the highest effect sizes and strongest lateralizations would be clustered). By scaling the values, the algorithm instead focuses on the *relative* functional profile across participants. We then used these values to create a vector for each ROI. We used a package `pvclust` (Suzuki and Shimodaira, 2006) in the R statistical programming language (R Core Team and others, 2012) with average clustering. Average clustering is a hierarchical clustering procedure that seeks to minimize the average distance between elements in a cluster. A correlation-based distance measure was used to quantify the distance between the ROI vectors. To assess the significance of the resulting clusters, we computed an AU (Approximately Unbiased) p-value (the numbers shown in black in Fig. 7). In effect, these p-values measure how consistent the clusters are with the data. An AU p-value of 1.0 would mean that the cluster is perfectly supported by the data.


a)
Effect size: Sentences−Nonwords
LIFGorb 1 .68 .6 .57 .5 .54 .5 .39 .59 .47 .3 .53 .45 .43 .34 .28
LIFG .68 1 .67 .48 .5 .58 .51 .31 .48 .56 .31 .46 .38 .42 .31 .24
LMFG .6 .67 1 .45 .47 .66 .42 .27 .32 .43 .41 .31 .34 .37 .26 .2
LAntTemp .57 .48 .45 1 .73 .64 .44 .42 .54 .37 .3 .66 .56 .53 .43 .24
LMidAntTemp .5 .5 .47 .73 1 .61 .49 .42 .52 .33 .26 .66 .62 .47 .41 .27
LMidPostTemp .54 .58 .66 .64 .61 1 .58 .44 .4 .42 .36 .53 .47 .59 .49 .31
LPostTemp .5 .51 .42 .44 .49 .58 1 .56 .43 .47 .29 .57 .42 .53 .58 .45
LAngG .39 .31 .27 .42 .42 .44 .56 1 .44 .36 .29 .5 .38 .4 .45 .58
RIFGorb .59 .48 .32 .54 .52 .4 .43 .44 1 .66 .47 .7 .61 .62 .6 .41
RIFG .47 .56 .43 .37 .33 .42 .47 .36 .66 1 .63 .61 .59 .72 .64 .51
RMFG .3 .31 .41 .3 .26 .36 .29 .29 .47 .63 1 .44 .55 .65 .48 .41
RAntTemp .53 .46 .31 .66 .66 .53 .57 .5 .7 .61 .44 1 .71 .69 .67 .42
RMidAntTemp .45 .38 .34 .56 .62 .47 .42 .38 .61 .59 .55 .71 1 .68 .59 .38
RMidPostTemp .43 .42 .37 .53 .47 .59 .53 .4 .62 .72 .65 .69 .68 1 .72 .5
RPostTemp .34 .31 .26 .43 .41 .49 .58 .45 .6 .64 .48 .67 .59 .72 1 .59
RAngG .28 .24 .2 .24 .27 .31 .45 .58 .41 .51 .41 .42 .38 .5 .59 1
LIFGorb LIFG LMFG LAntTemp LMidAntTemp LMidPostTemp LPostTemp LAngG RIFGorb RIFG RMFG RAntTemp RMidAntTemp RMidPostTemp RPostTemp RAngG
value
1.00
0.75
0.50
0.25
0.00
−0.25


b)
Lateralization based on Sentences−Nonwords effect size
LIFGorb 1 .47 .51 .23 .17 .4 .38 .12
LIFG .47 1 .53 .3 .45 .53 .36 .16
LMFG .51 .53 1 .27 .37 .57 .34 .16
LAntTemp .23 .3 .27 1 .3 .35 .03 .02
LMidAntTemp .17 .45 .37 .3 1 .43 .27 .14
LMidPostTemp .4 .53 .57 .35 .43 1 .3 .25
LPostTemp .38 .36 .34 .03 .27 .3 1 .33
LAngG .12 .16 .16 .02 .14 .25 .33 1
LIFGorb LIFG LMFG LAntTemp LMidAntTemp LMidPostTemp LPostTemp LAngG
value
1.00
0.75
0.50
0.25
0.00
−0.25

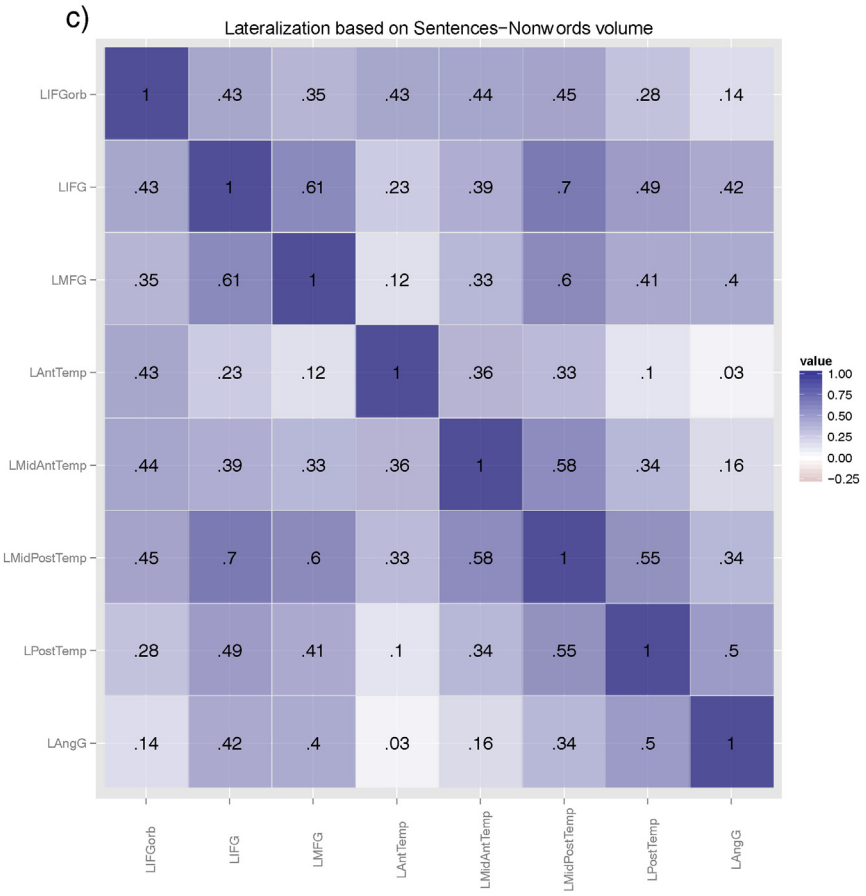


Fig. 6. a: Correlations between pairs or regions for the *Sentences* > *Nonwords* effect size measure. The color of the box reflects the strength and direction of the correlation (darker blue boxes are ones with higher positive correlations, darker red boxes are ones with higher negative correlations, and lighter boxes have values close to 0).b: Correlations between pairs or regions for the effect-size-based lateralization measure. The color of the box reflects the strength and direction of the correlation (darker blue boxes are ones with higher positive correlations, darker red boxes are ones with higher negative correlations, and lighter boxes have values close to 0).c: Correlations between pairs or regions for the volume-based lateralization measure. The color of the box reflects the strength and direction of the correlation (darker blue boxes are ones with higher positive correlations, darker red boxes are ones with higher negative correlations, and lighter boxes have values close to 0).

We observe a significant (AU = 0.95) cluster, which included the three frontal regions (LIFGorb, LIFG and LMFG fROIs) and the LMidPostTemp region (highlighted with a red box in Fig. 7). Intriguingly, the same cluster was recently reported in the analysis of the time-courses of the language regions during naturalistic story comprehension (Blank et al., 2014; Fig. 6). Jointly, these results strongly suggest that the LMidPostTemp fROI forms an integrated system with the frontal language regions, in spite of the fact that it is more spatially proximal to the language regions in the temporal and parietal cortices.

Another quite reliable cluster (AU = 0.91) emerged in the anterior temporal cortex, comprising LAntTemp and LMidAntTemp fROIs.

The relationship among the different neural measures of language activity

Finally, it is important to know how different functional measures relate to one another. In other words, do these different measures capture (at least somewhat) distinct aspects of the phenotype? For example, can we predict lateralization scores from the *Sentences* > *Nonwords* effect size? Or does each measure capture something unique about individual subjects relative to other measures? To evaluate the relationship among our functional measures, we calculated the correlation across 150 subjects between each pair of functional measures – except for volume due to its low across-session reliability as shown in the previous section – for each ROI.

As can be seen in Fig. 8, the two lateralization measures are quite strongly inter-correlated (*r*s between 0.31 and 0.53; all significantly different from 0 at $p < .001$ and still significant after Bonferroni correction). Furthermore, the effect-size-based lateralization measure is (trivially) correlated with the *Sentences* > *Nonwords* effect size measure in the LH (*r*s between 0.34 and 0.84; all significantly different from 0 at $p < .001$ and still significant after Bonferroni correction), such that individuals with a larger LH *Sentences* > *Nonwords* effect show a bigger difference in effect sizes between the LH and RH (i.e., our effect-size-based lateralization measure). However, the volume-based lateralization measure is almost entirely uncorrelated with the size of the *Sentences* > *Nonwords* effect in the LH regions (*r*s between −0.09 and 0.17; 0.02 on average across regions; no regions showing a correlation significantly different from 0 at $p < .05$), and it is negatively correlated with the size of the *Sentences* > *Nonwords* effect in the RH regions (*r*s between −0.34 and −0.58; all significantly different from 0 at $p < .001$ and still significant after Bonferroni correction). The latter is likely because individuals with larger *Sentences* > *Nonwords* effects in the RH also have bigger RH regions, and are thus less strongly left-lateralized in the volume-based measure.

Summary and conclusions

Here, we have characterized neural activity in the fronto-temporal language system – which has been shown to support high-level linguistic processing in a relatively selective manner (Fedorenko et al., 2010, 2011; Scott et al., in press) – in a dataset comprised of 150 right-handed native-English-speaking individuals. We examined four functional measures: the *Sentences* > *Nonwords* effect size, the

Sentences > *Nonwords* volume, and two lateralization measures (one based on the effect size, and one — on volume). To summarize the key results:

1. We have observed that all four measures exhibit substantial inter-individual variability.
2. Activations for the language "localizer" task (based on the contrast between sentences and nonword sequences; Fedorenko et al., 2010) are *highly stable* within individuals both within and across sessions. We suspect that similar contrasts (based on comparisons between linguistic stimuli and degraded versions of those stimuli) would show similar reliability as long as sufficient amount of data is collected from each individual. We found, however, that effect sizes tend to be more reliable than volume measures, consistent with some prior findings (e.g., Cohen and Dubois, 1999), although both effect-size-based and volume-based lateralization measures showed quite high reliability.
3. We observed strong positive correlations across regions with respect to all the functional measures, suggesting that these regions should not be treated independently. Furthermore, a clustering analysis using a combination of functional measures suggested that some sets of ROIs are especially strongly related, including a) LMidPostTemp and the three frontal ROIs, b) LAntTemp and LMidAntTemp, and possibly c) LPostTemp and LAngG (see Blank et al., 2014, for converging evidence).
4. Finally, we found that the two lateralization measures are correlated with each other, but the volume-based lateralization measure is not correlated with the size of the *Sentences* > *Nonwords* effect in the LH ROIs.

Based on these results, we offer several *guidelines* for researchers who wish to examine relationships between language activations – as assessed with this or similar language localizers – and a) behavior, or b) genetic variability.

- We recommend focusing on the measures of *effect size* rather than volumes, because the former prove to be more reliable within individuals.
- The inter-region correlations strongly suggest that different language regions form *an integrated functional system*, consistent with prior claims from the analyses of low-frequency fluctuations at rest and during language comprehension (Blank et al., 2014; Tie et al., 2014). Thus future studies may want to consider the fronto-temporal high-level language system as a whole, instead of focusing on one or two regions within it. Doing so may yield greater experimental power for detecting relationships between neural markers on the one hand, and behavioral or genetic markers on the other hand. (NB: Of course, these high-level language-processing regions are distinct from e.g., the lower-level perceptual speech regions, the visual word-form area, the lower-level articulatory motor regions, and the domain-general cognitive control regions, implicated in some aspects of language comprehension and production (e.g., Fedorenko and Thompson-Schill, 2014). Functional properties of those regions/systems can thus be examined independently from those of the fronto-temporal language system and may relate to distinct aspects of behavioral and genetic variability.).
- The ROI clustering results suggest that some divisions of the language system into regions adopted here and based on the most common topographic patterns across subjects (Fedorenko et al., 2010) may not be warranted. In particular, instead of the division into eight regions, our clustering suggests a tripartite division, with a) the LMidPostTemp region being grouped with the frontal ROIs, b) the LAntTemp and LMidAntTemp ROIs forming a single region, and possibly c) the LPostTemp and LAngG ROIs forming a single region.
- Finally, it appears that two of the functional measures examined here are not correlated with one another and thus may reflect distinct phenotypic characteristics. These are i) the size of the *Sentences* > *Nonwords* effect, and ii) volume-based lateralization. These measures can thus be examined independently in future studies, with a proper correction for the number of measures.

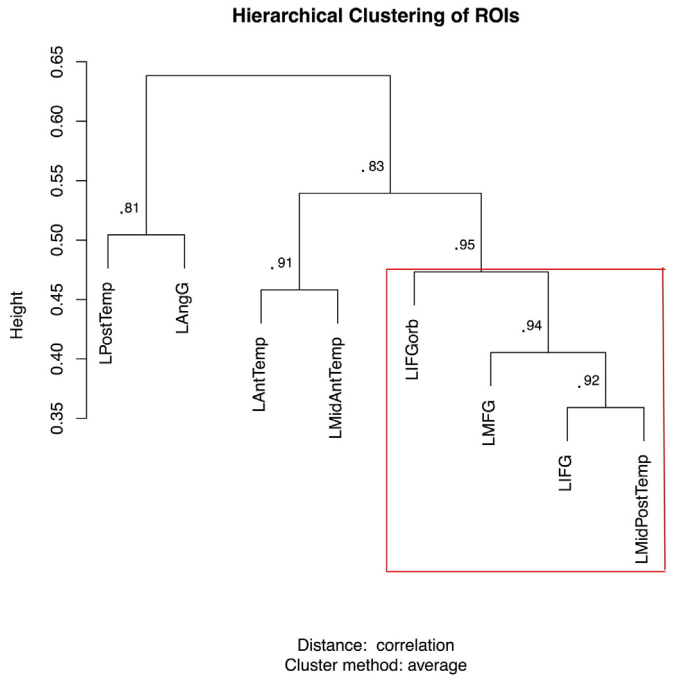


Fig. 7. Hierarchical ROI clustering results based on three functional measures: the *Sentences* > *Nonwords* effect size, the effect-size-based lateralization measure, and the volume-based lateralization measure. The numbers denote AU (Approximately Unbiased) p-values. Red boxes highlight significant clusters at p ≥ 0.95.

In addition to the measures examined here, a number of other measures – both functional and structural – may be worth considering in future studies. For example, instead of considering the functional properties of different regions/subsets of the language system separately, one may examine the *relationships* among them. For example, we have observed in this dataset that individuals vary in terms of which regions show the largest *Sentences* > *Nonwords* effect or have the most *Sentences* > *Nonwords* voxels. For example, some individuals show the strongest effect in the inferior frontal regions, others in the anterior temporal regions, and yet others in the posterior temporal regions. To the extent that these different components of the language system are somewhat functionally distinct, these inter-individual differences in which region is "dominant" (i.e., most robustly active) may be important. Using our subset of 32 participants who were scanned twice, we assessed the reliability of this measure, asking whether a participant who shows more activation in Region A than Region B in session 1 also shows more activation in Region A than Region B in session 2. For the *Sentences* > *Nonwords* effect size measure, 30/32 participants showed a positive correlation across ROIs (mean $r = 0.63$ for the LH ROIs), and for the volume measure, all 32 participants showed a positive correlation (mean $r = 0.79$). Thus, these relative patterns of greater or lesser activation in different parts of the language system appear to be highly stable within individuals across time and may serve as an additional useful characteristic of individuals.

Furthermore, as we briefly mentioned in the Introduction, having reliable functional activations in individual subjects can allow researchers to obtain better structural measures than those based purely on macroanatomy. For example, we can examine the cortical thickness of *language-responsive* parts of the left inferior frontal

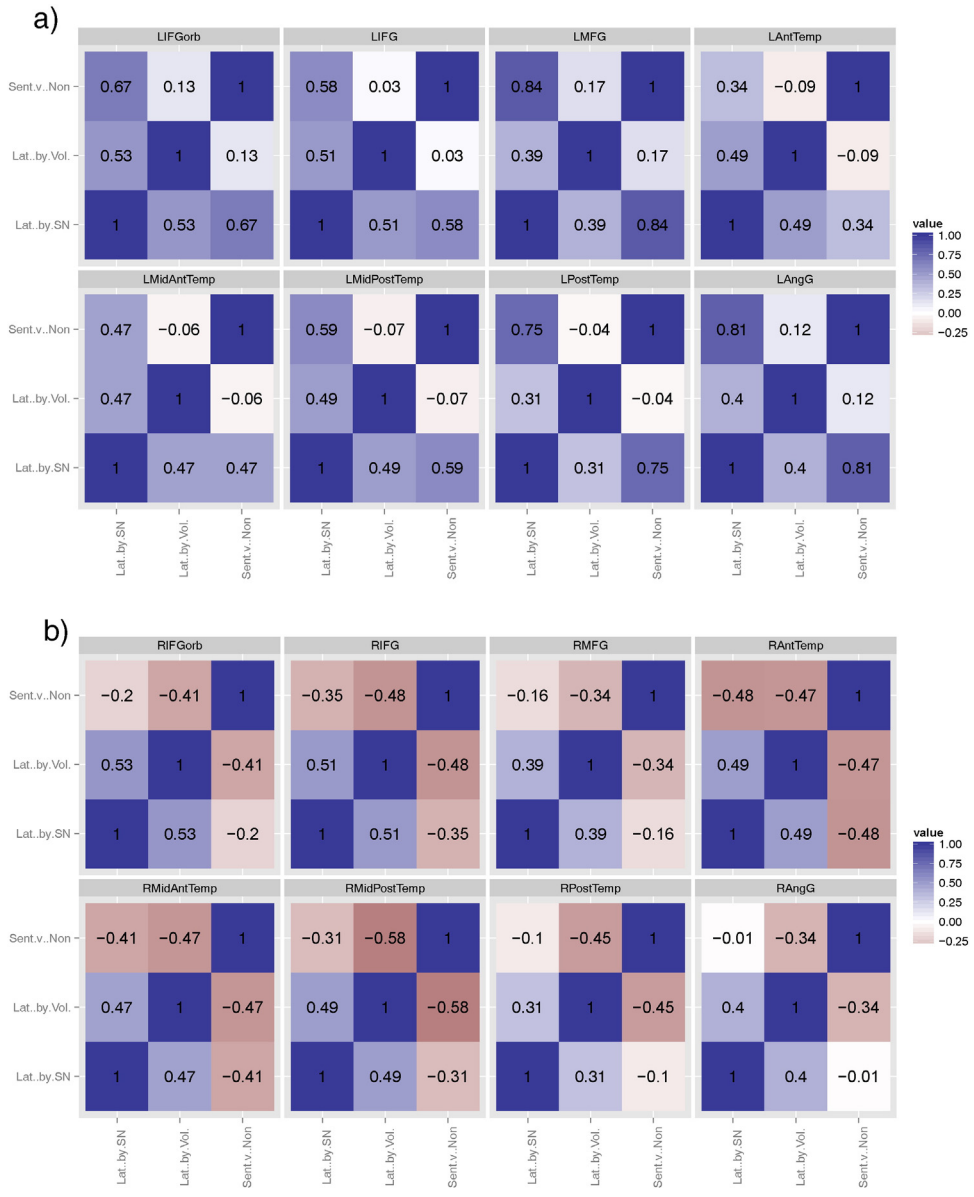


Fig. 8. a: Correlations among the three functional measures – the *Sentences* > *Nonwords* effect size in the LH, the effect-size-based lateralization measure, and the volume-based lateralization measure – for each LH ROI. The color of the box reflects the strength and direction of the correlation (darker blue boxes are ones with higher positive correlations, darker red boxes are ones with higher negative correlations, and lighter boxes have values close to 0). b: Correlations among the three functional measures – the *Sentences* > *Nonwords* effect size in the RH, the effect-size-based lateralization measure, and the volume-based lateralization measure – for each RH ROI. (The correlation values for the relationship between the two lateralization values are the same as in Fig. 8.). The color of the box reflects the strength and direction of the correlation (darker blue boxes are ones with higher positive correlations, darker red boxes are ones with higher negative correlations, and lighter boxes have values close to 0).

gyrus instead of using anatomical sulcal/gyral boundaries, and this functional-activation-based measure is likely to work better than pure anatomy-based measures because we know that these cortical regions contain structurally (e.g., Amunts et al., 2010) and functionally (e.g., Fedorenko et al., 2012a, 2012b) distinct sub-regions.

To conclude, we report a number of reliable functional markers of language activity based on fMRI activation patterns for a robust contrast that activates the fronto-temporal language system long implicated in high-level language processing (e.g., Binder et al., 1997; Fedorenko et al., 2010). Although determining which markers prove to work best in explaining behavioral and genetic variability is likely to be a long and iterative process, we offer some recommendations based on the patterns observed in a large dataset of 150 individuals, which should constrain the space of possibilities, at least somewhat, in future investigations of brain-behavior and brain-genes relationships.

Acknowledgments

We thank all our participants over the years. We thank Brown Hsieh for his help in collecting the data, and (in alphabetical order) Zuzanna Balewski, Michael Behr, Idan Blank, Emile Bruneau, Eyal Dechter, Danny Dilks, Mike Frank, Tanya Goldhaber, Josh Julian, Alex Kell, Kami Koldewyn, Sam Norman-Haignere, Alex Paunov, David Pitcher, Liz Ryan, Terri Scott, Steve Shannon, Irina Shklyar, Veronica Smith, Todd Thompson, Jason Webster, and Anna Wexler for their help with seconding during scanning. We thank Zach Mineroff for helping with the data maintenance and website. We thank Christina Triantafyllou, Atsushi Takahashi, Steve Shannon and Sheeba Arnold for technical support, and Walid Bendris and Jenelle Feather for their help with extracting and organizing all the relevant information. We thank the audience at the Neurobiology of Language 2014 conference (in Amsterdam, Netherlands) for helpful discussions of this work. For

comments on the earlier drafts of the manuscript we are grateful to David Carey, Simon Fisher and Alex Kell. This work was supported by a K99/R00 award HD 057522 to E.F. from NICHD. K.M. was supported by a National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2016.05.073.

References

Amunts, K., Lenzen, M., Friederici, A.D., Schleicher, A., Morosan, P., Palomero-Gallagher, N., Zilles, K., 2010. Broca's Region: Novel Organizational Principles and Multiple Receptor Mapping. PLoS Biol. 8 (9), e1000489. http://dx.doi.org/10.1371/journal.pbio.1000489.

Association, A. P., & others, 2013. The Diagnostic and Statistical Manual of Mental Disorders: DSM 5. bookpointUS.

Bavelier, D., Corina, D.P., Neville, H.J., 1998. Brain and language: a perspective from sign language. Neuron 21 (2), 275–278.

Bedny, M., Pascual-Leone, A., Dodell-Feder, D., Fedorenko, E., Saxe, R., 2011. Language processing in the occipital cortex of congenitally blind adults. Proc. Natl. Acad. Sci. 108, 4429–4434. http://dx.doi.org/10.1073/pnas.1014818108.

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403 (6767), 309–312.

Binder, J.R., Frost, J.A., Hammeke, T.A., Cox, R.W., Rao, S.M., Prieto, T., 1997. Human brain language areas identified by functional magnetic resonance imaging. J. Neurosci. 17 (1), 353–362.

Blank, I.A., Kanwisher, N., Fedorenko, E., 2014. A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. J. Neurophysiol. (jn–00884).

Carlson, S.M., Moses, L.J., Claxton, L.J., 2004. Individual differences in executive functioning and theory of mind: an investigation of inhibitory control and planning ability. J. Exp. Child Psychol. 87 (4), 299–319.

Childers, T.L., Houston, M.J., Heckler, S.E., 1985. Measurement of individual differences in visual versus verbal information processing. J. Consum. Res. 125–134.

Cohen, M.S., DuBois, R.M., 1999. Stability, repeatability, and the expression of signal magnitude in functional magnetic resonance imaging. J. Magn. Reson. Imaging: JMRI 10 (1), 33–40.

Colom, R., 2004. Working memory is (almost) perfectly predicted by g. Intelligence 32 (3), 277–296. http://dx.doi.org/10.1016/j.intell.2003.12.002.

Colombo, J., Mitchell, D.W., Coldren, J.T., Freeseman, L.J., 1991. Individual differences in infant visual attention: are short lookers faster processors or feature processors? Child Dev. 62 (6), 1247–1257.

Conway, A.R.A., 1996. Individual differences in working memory capacity: more evidence for a general capacity theory. Memory 4 (6), 577–590. http://dx.doi.org/10.1080/741940997.

Cope, N., Eicher, J.D., Meng, H., Gibson, C.J., Hager, K., Lacadie, C., ... Gruen, J.R., 2012. Variants in the DYX2 locus are associated with altered brain activation in reading-related brain regions in subjects with reading disability. NeuroImage 63 (1), 148–156. http://dx.doi.org/10.1016/j.neuroimage.2012.06.037.

Daneman, M., Carpenter, P.A., 1980. Individual differences in working memory and reading. J. Verbal Learn. Verbal Behav. 19 (4), 450–466. http://dx.doi.org/10.1016/S0022-5371(80)90312-6.

Diaz, M.T., McCarthy, G., 2009. A comparison of brain activity evoked by single content and function words: an fMRI investigation of implicit word processing. Brain Res. 1282, 38–49. http://dx.doi.org/10.1016/j.brainres.2009.05.043.

Dubois, J., Adolphs, R., 2016. Building a science of individual differences from fMRI. Trends Cogn. Sci. 20 (6), 425–443. http://dx.doi.org/10.1016/j.tics.2016.03.014 (June).

Duncan, J., 2013. The structure of cognition: attentional episodes in mind and brain. Neuron 80 (1), 35–50.

Duncan, J., Seitz, R.J., Kolodny, J., Bor, D., Herzog, H., Ahmed, A., ... Emslie, H., 2000. A neural basis for general intelligence. Science 289 (5478), 457–460.

Fedorenko, E., 2014. The role of domain-general cognitive control in language comprehension. Front. Psychol. 5. http://dx.doi.org/10.3389/fpsyg.2014.00335.

Fedorenko, E., Kanwisher, N., 2011. Some regions within Broca's area do respond more strongly to sentences than to linguistically degraded stimuli: a comment on Rogalsky and Hickok (2011). J. Cogn. Neurosci. 23 (10), 2632–2635.

Fedorenko, E., Thompson-Schill, S.L., 2014. Reworking the language network. Trends Cogn. Sci. 18 (3), 120–126.

Fedorenko, E., Hsieh, P.-J., Nieto-Castanon, A., Whitfield-Gabrieli, S., Kanwisher, N., 2010. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. J. Neurophysiol. 104 (2), 1177–1194. http://dx.doi.org/10.1152/jn.00032.2010.

Fedorenko, E., Behr, M.K., Kanwisher, N., 2011. Functional specificity for high-level linguistic processing in the human brain. Proc. Natl. Acad. Sci. 108 (39), 16428–16433. http://dx.doi.org/10.1073/pnas.1112937108.

Fedorenko, E., Duncan, J., Kanwisher, N., 2012a. Language-selective and domain-general regions lie side by side within Broca's area. Curr. Biol. 22 (21), 2059–2062.

Fedorenko, E., McDermott, J.H., Norman-Haignere, S., Kanwisher, N., 2012b. Sensitivity to musical structure in the human brain. J. Neurophysiol. 108 (12), 3289–3300.

Fischl, B., Rajendran, N., Busa, E., Augustinack, J., Hinds, O., Yeo, B.T.T., ... Zilles, K., 2008. Cortical folding patterns and predicting cytoarchitecture. Cereb. Cortex 18 (8), 1973–1980. http://dx.doi.org/10.1093/cercor/bhm225.

Friedman, L., Glover, G.H., The FBIRN Consortium, 2006. Reducing interscanner variability of activation in a multicenter fMRI study: controlling for signal-to-fluctuation-noise-ratio (SFNR) differences. NeuroImage 33 (2), 471–481. http://dx.doi.org/10.1016/j.neuroimage.2006.07.012 (November).

Friedman, L., Stern, H., Brown, G.G., Mathalon, D.H., Turner, J., Glover, G.H., Gollub, R.L., et al., 2008. Test–retest and between-site reliability in a multicenter fMRI study. Hum. Brain Mapp. 29 (8), 958–972. http://dx.doi.org/10.1002/hbm.20440 (August).

Frost, M.A., Goebel, R., 2012. Measuring structural–functional correspondence: spatial variability of specialised brain regions after macro-anatomical alignment. NeuroImage 59 (2), 1369–1381. http://dx.doi.org/10.1016/j.neuroimage.2011.08.035.

Gardner, H., Hatch, T., 1989. Educational implications of the theory of multiple intelligences. Educ. Res. 18 (8), 4–10. http://dx.doi.org/10.3102/0013189X018008004.

Gernsbacher, M.A., 1991. Cognitive processes and mechanisms in language comprehension: the structure building framework. Psychology of Learning and Motivation Vol. 27. Elsevier, pp. 217–263 (Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0079742108601255).

Gorgolewski, K.J., Storkey, A., Bastin, M.E., Whittle, I.R., Wardlaw, J.M., Pernet, C.R., 2013. A test-retest fMRI dataset for motor, language and spatial attention functions. GigaScience 2 (1), 6. http://dx.doi.org/10.1186/2047-217X-2-6.

Grahn, J.A., Schuit, D., 2012. Individual differences in rhythmic ability: behavioral and neuroimaging investigations. Psychomusicology Music Mind Brain 22 (2), 105–121. http://dx.doi.org/10.1037/a0031188.

Herrmann, E., Call, J., Hernandez-Lloreda, M.V., Hare, B., Tomasello, M., 2007. Humans have evolved specialized skills of social cognition: the cultural intelligence hypothesis. Science 317 (5843), 1360–1366. http://dx.doi.org/10.1126/science.1146282.

Hinke, R.M., Hu, X., Stillman, A.E., Kim, S.-G., Merkle, H., Salmi, R., Ugurbil, K., 1993. Functional magnetic resonance imaging of Broca's area during internal speech. NeuroReport 4 (6), 675–678. http://dx.doi.org/10.1097/00001756-199306000-00018.

Hoogman, M., Guadalupe, T., Zwiers, M.P., Klarenbeek, P., Francks, C., Fisher, S.E., 2014. Assessing the effects of common variation in the FOXP2 gene on human brain structure. Front. Hum. Neurosci. 8. http://dx.doi.org/10.3389/fnhum.2014.00473.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., Quinn, K., ... Wang, P., 2010. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am. J. Psychiatr. 167 (7), 748–751. http://dx.doi.org/10.1176/appi.ajp.2010.09091379.

Julian, J.B., Fedorenko, E., Webster, J., Kanwisher, N., 2012. An algorithmic method for functionally defining regions of interest in the ventral visual pathway. NeuroImage 60 (4), 2357–2364. http://dx.doi.org/10.1016/j.neuroimage.2012.02.055.

Just, M.A., Carpenter, P.A., 1992. A capacity theory of comprehension: individual differences in working memory. Psychol. Rev. 99 (1), 122–149. http://dx.doi.org/10.1037/0033-295X.99.1.122.

Kane, M.J., Engle, R.W., 2002. The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: an individual-differences perspective. Psychon. Bull. Rev. 9 (4), 637–671.

Kanwisher, N., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. 17 (11), 4302.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. 12 (5), 535–540. http://dx.doi.org/10.1038/nn.2303.

Krug, A., Nieratschker, V., Markov, V., Krach, S., Jansen, A., Zerres, K., ... Kircher, T., 2010. Effect of CACNA1C rs1006737 on neural correlates of verbal fluency in healthy individuals. NeuroImage 49 (2), 1831–1836. http://dx.doi.org/10.1016/j.neuroimage.2009.09.028.

Kuperberg, G.R., Sitnikova, T., Caplan, D., Holcomb, P.J., 2003. Electrophysiological distinctions in processing conceptual relationships within simple sentences. Cogn. Brain Res. 17 (1), 117–129. http://dx.doi.org/10.1016/S0926-6410(03)00086-7.

Landi, N., Frost, S.J., Mencl, W.E., Sandak, R., Pugh, K.R., 2013. Neurobiological bases of reading comprehension: insights from neuroimaging studies of word-level and text-level processing in skilled and impaired readers. Read. Writ. Q. 29 (2), 145–167. http://dx.doi.org/10.1080/10573569.2013.758566.

Miller, L.E., Saygin, A.P., 2013. Individual differences in the perception of biological motion: links to social cognition and motor imagery. Cognition 128 (2), 140–148. http://dx.doi.org/10.1016/j.cognition.2013.03.013.

Mischel, W., Shoda, Y., Rodriguez, M.I., 1989. Delay of gratification in children. Science 244 (4907), 933–938.

Miyake, A., Friedman, N.P., Emerson, M.J., Witzki, A.H., Howerter, A., Wager, T.D., 2000. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. Cogn. Psychol. 41 (1), 49–100.

Nieto-Castañón, A., Fedorenko, E., 2012. Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. NeuroImage 63 (3), 1646–1669. http://dx.doi.org/10.1016/j.neuroimage.2012.06.065.

Noppeney, U., Price, C.J., 2004. An fMRI study of syntactic adaptation. J. Cogn. Neurosci. 16 (4), 702–713. http://dx.doi.org/10.1162/089892904323057399.

Ocklenburg, S., Arning, L., Gerding, W.M., Epplen, J.T., Güntürkün, O., Beste, C., 2013. Cholecystokinin A receptor (CCKAR) gene variation is associated with language lateralization. PLoS One 8 (1), e53643. http://dx.doi.org/10.1371/journal.pone.0053643.

Pakulak, E., Neville, H.J., 2010. Proficiency differences in syntactic processing of monolingual native speakers indexed by event-related potentials. J. Cogn. Neurosci. 22 (12), 2728–2744. http://dx.doi.org/10.1162/jocn.2009.21393.

Perrachione, T.K., Fedorenko, E.G., Vinke, L., Gibson, E., Dilley, L.C., 2013. Evidence for shared cognitive processing of pitch in music and language. PLoS One 8 (8), e73372. http://dx.doi.org/10.1371/journal.pone.0073372.

Petersen, S., Fox, P., Snyder, A., Raichle, M., 1990. Activation of extrastriate and frontal cortical areas by visual words and word-like stimuli. Science 249 (4972), 1041–1044. http://dx.doi.org/10.1126/science.2396097.

Pinel, P., Fauchereau, F., Moreno, A., Barbot, A., Lathrop, M., Zelenika, D., ... Dehaene, S., 2012. Genetic variants of FOXP2 and KIAA0319/TTRAP/THEM2 locus are associated with altered brain activation in distinct language-related regions. J. Neurosci. 32 (3), 817–825. http://dx.doi.org/10.1523/JNEUROSCI.5996-10.2012.

Poppenk, J., Evensmoen, H.R., Moscovitch, M., Nadel, L., 2013. Long-axis specialization of the human hippocampus. Trends Cogn. Sci. 17 (5), 230–240. http://dx.doi.org/10.1016/j.tics.2013.03.005.

Robertson, D.A., Gernsbacher, M.A., Guidotti, S.J., Robertson, R.R.W., Irwin, W., Mock, B.J., Campana, M.E., 2000. Functional neuroanatomy of the cognitive process of mapping during discourse comprehension. Psychol. Sci. 11 (3), 255–260. http://dx.doi.org/10.1111/1467-9280.00251.

Rombouts, S.A., Barkhof, F., Hoogenraad, F.G., Sprenger, M., Scheltens, P., 1998. Within-subject reproducibility of visual activation patterns with functional magnetic resonance imaging using multislice echo planar imaging. Magn. Reson. Imaging 16 (2), 105–113.

Saxe, R., Kanwisher, N., 2003. People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind. NeuroImage 19 (4), 1835–1842. http://dx.doi.org/10.1016/S1053-8119(03)00230-1.

Scerri, T.S., Morris, A.P., Buckingham, L.-L., Newbury, D.F., Miller, L.L., Monaco, A.P., ... Paracchini, S., 2011. DCDC2, KIAA0319 and CMIP are associated with reading-related traits. Biol. Psychiatry 70 (3), 237–245. http://dx.doi.org/10.1016/j.biopsych.2011.02.005.

Scott, T. Gallée, J., and Fedorenko, E. A new fun and robust version of an fMRI localizer for the fronto-temporal language system, (in press).

Scott, T. Gallée, J., and Fedorenko, E. A new fun and robust version of an fMRI localizer for the fronto-temporal language system, (in press). Schoene-Bake, J.-C., Keller, S.S., Niehusmann, P., Volmering, E., Elger, C., Deppe, M., Weber, B., 2014. In vivo mapping of hippocampal subfields in mesial temporal lobe epilepsy: relation to histopathology: hippocampal subfield mapping in mTLE. Hum. Brain Mapp. 35 (9), 4718–4728. http://dx.doi.org/10.1002/hbm.22506.

Seghier, M.L., Lazeyras, F., Pegna, A.J., Annoni, J.-M., Khateb, A., 2008. Group analysis and the subject factor in functional magnetic resonance imaging: analysis of fifty right-handed healthy subjects in a semantic language task. Hum. Brain Mapp. 29 (4), 461–477. http://dx.doi.org/10.1002/hbm.20410.

Snijders, T.M., Vosse, T., Kempen, G., Van Berkum, J.J.A., Petersson, K.M., Hagoort, P., 2009. Retrieval and unification of syntactic structure in sentence comprehension: an fMRI study using word-category ambiguity. Cereb. Cortex 19 (7), 1493–1503. http://dx.doi.org/10.1093/cercor/bhn187.

Spearman, C., 1904. "General intelligence," objectively determined and measured. Am. J. Psychol. 15 (2), 201. http://dx.doi.org/10.2307/1412107.

Spearman, C., 1927. The Abilities of Man.

Stein, J.L., Medland, S.E., Vasquez, A.A., Hibar, D.P., Senstad, R.E., Winkler, A.M., ... Thompson, P.M., 2012. Identification of common variants associated with human hippocampal and intracranial volumes. Nat. Genet. 44 (5), 552–561. http://dx.doi.org/10.1038/ng.2250.

Surprenant, A.M., Watson, C.S., 2001. Individual differences in the processing of speech and nonspeech sounds by normal-hearing listeners. J. Acoust. Soc. Am. 110 (4), 2085. http://dx.doi.org/10.1121/1.1404973.

Suzuki, R., Shimodaira, H., 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22 (12), 1540–1542.

Tahmasebi, A.M., Davis, M.H., Wild, C.J., Rodd, J.M., Hakyemez, H., Abolmaesumi, P., Johnsrude, I.S., 2012. Is the link between anatomical structure and function equally strong at all cognitive levels of processing? Cereb. Cortex 22 (7), 1593–1603. http://dx.doi.org/10.1093/cercor/bhr205.

Team, R. C., & others, 2012. R: A Language and Environment for Statistical Computing.

Tie, Y., Rigolo, L., Norton, I.H., Huang, R.Y., Wu, W., Orringer, D., ... Golby, A.J., 2014. Defining language networks from resting-state fMRI for surgical planning-a feasibility study: resting-state fMRI for language mapping. Hum. Brain Mapp. 35 (3), 1018–1030. http://dx.doi.org/10.1002/hbm.22231.

Travis, S.G., Huang, Y., Fujiwara, E., Radomski, A., Olsen, F., Carter, R., ... Malykhin, N.V., 2014. High field structural MRI reveals specific episodic memory correlates in the subfields of the hippocampus. Neuropsychologia 53, 233–245. http://dx.doi.org/10.1016/j.neuropsychologia.2013.11.016.

Traxler, M.J., Johns, C.L., Long, D.L., Zirnstein, M., Tooley, K.M., Jonathan, E., 2012. Individual differences in eye-movements during reading: working memory and speed-of-processing effects. J. Eye Mov. Res. 5 (1), 5.

Turken, A.U., Dronkers, N.F., 2011. The neural architecture of the language comprehension network: converging evidence from lesion and connectivity analyses. Front. Syst. Neurosci. 5. http://dx.doi.org/10.3389/fnsys.2011.00001.

Vogel, E.K., Machizawa, M.G., 2004. Neural activity predicts individual differences in visual working memory capacity. Nature 428 (6984), 748–751.

Whalley, H.C., O'Connell, G., Sussmann, J.E., Peel, A., Stanfield, A.C., Hayiou-Thomas, M.E., ... Hall, J., 2011. Genetic variation in CNTNAP2 alters brain function during linguistic processing in healthy individuals. Am. J. Med. Genet. B Neuropsychiatr. Genet. 156 (8), 941–948. http://dx.doi.org/10.1002/ajmg.b.31241.

Whitehouse, A.J.O., Bishop, D.V.M., Ang, Q.W., Pennell, C.E., Fisher, S.E., 2011. *CNTNAP2* variants affect early language development in the general population. Genes Brain Behav. 10 (4), 451–456. http://dx.doi.org/10.1111/j.1601-183X.2011.00684.x.

Wilke, M., Lidzba, K., 2007. LI-tool: a new toolbox to assess lateralization in functional MR-data. J. Neurosci. Methods 163 (1), 128–136. http://dx.doi.org/10.1016/j.jneumeth.2007.01.026.

Yue, X., Nasr, S., Devaney, K.J., Holt, D.J., Tootell, R.B.H., 2013. FMRI analysis of contrast polarity in face-selective cortex in humans and monkeys. NeuroImage 76, 57–69. http://dx.doi.org/10.1016/j.neuroimage.2013.02.068.