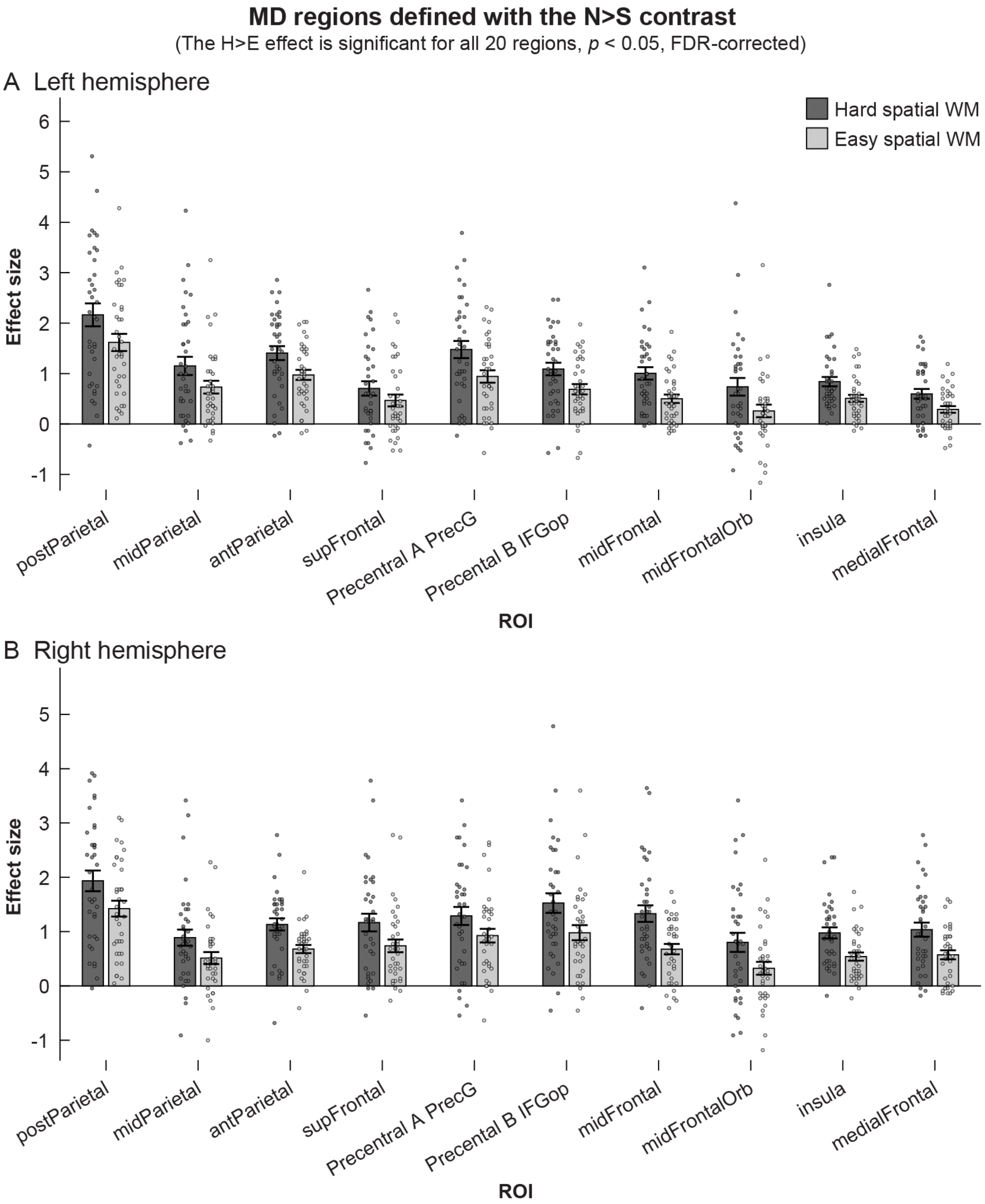
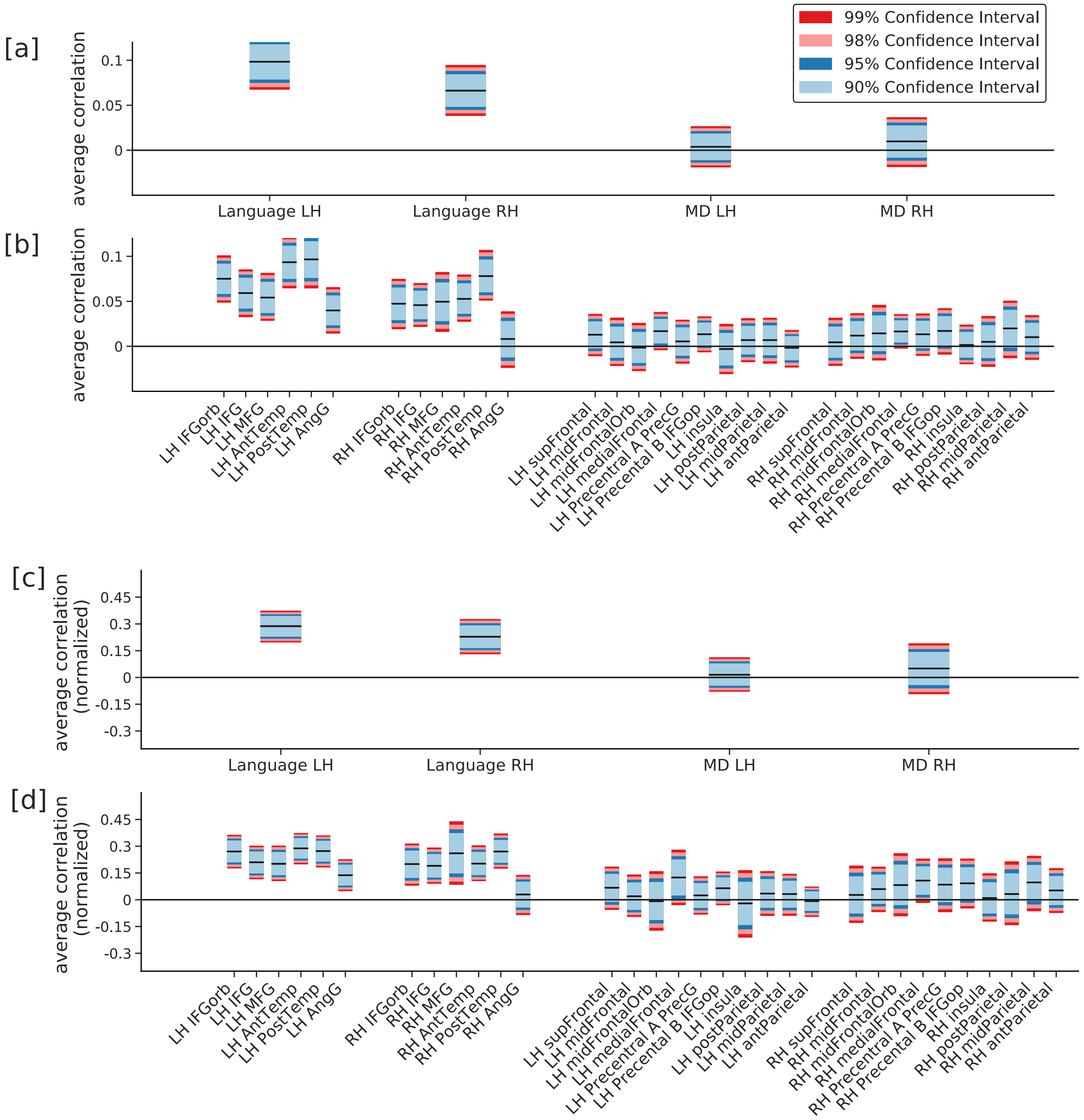
Supplementary Table 1: Studies that used naturalistic linguistic materials with the goal of relating brain responses to properties of the materials.

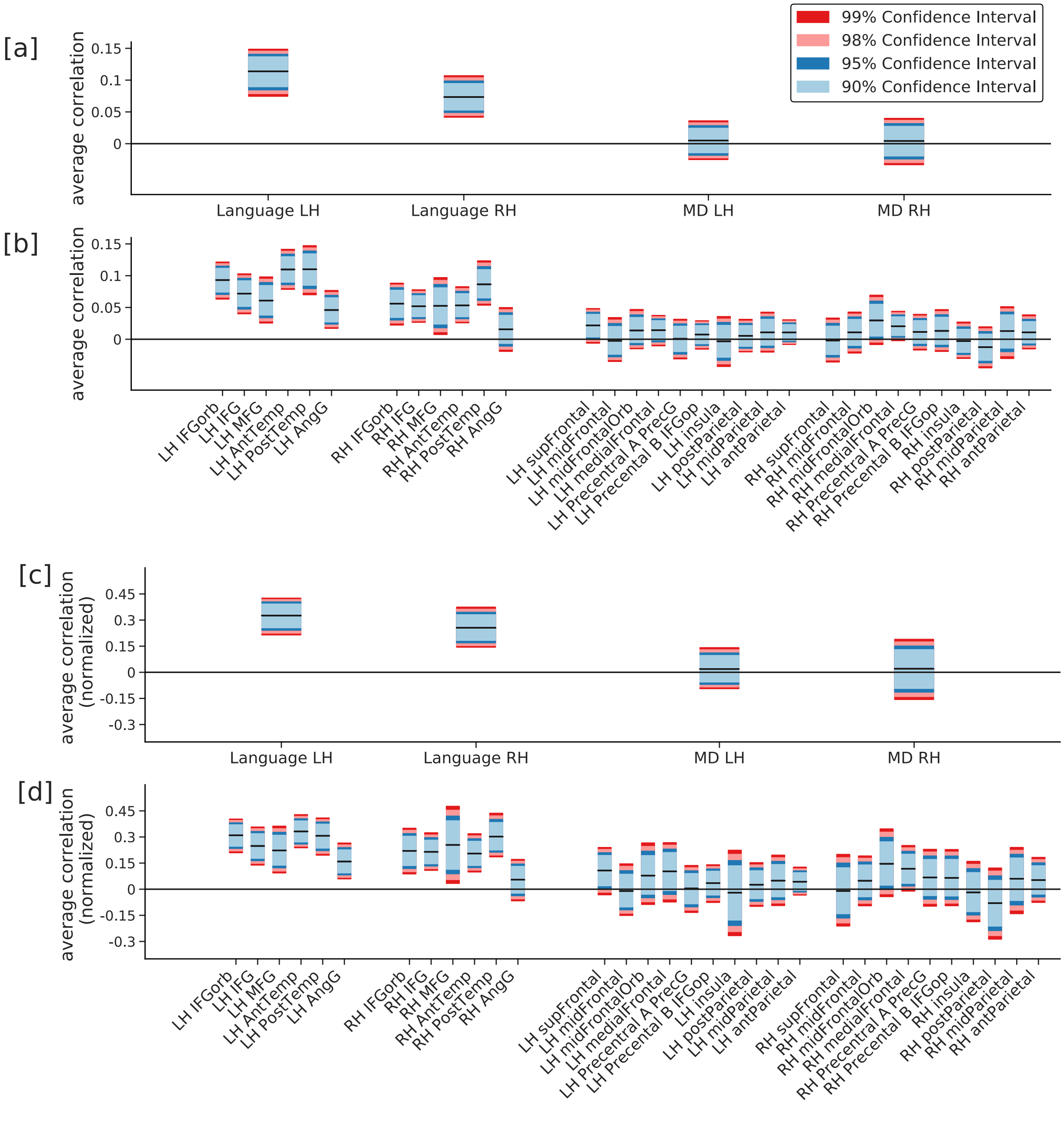
Author	Description	N	Statistical procedure	Controls (not of interest)	Predictors of interest	Held-Out Evaluation
Bhattasali et al., (2018)	Evidence of brain areas engaged in memory retrieval vs. parsing.	42	2-step GLM	word-rate, unigram, sound power, pitch	parser operations number, Is Last Word of Multiword Expression	NO
Brennan et al., (2012)	Evidence of structure building in Anterior Temporal Lobe.	9	2-step GLM	word-rate, unigram, sound power, pitch	syntactic node count	NO
Brennan et al., (2016)	Evidence of different types of structure building throughout the language network.	26	LME/LRT	prosodic-breaks, head movement, unigram, sound-power	syntactic node count, POS surprisal	NO
de Heer et al., (2017)	Evidence of increasing layers of abstraction for linguistic processing.	7	Ridge regression + held-out eval.		spectral features space, phonetic feature space, semantic feature space	YES
Dehghani et al., (2017)	Evidence that story embeddings can support story classification during naturalistic reading, even across languages.	90	Ridge regression + decoder held-out eval.		narrative features	YES
(Deniz et al., 2019)	Evidence that semantic selectivity is similar during listening and reading		Ridge regression + held-out eval.	word-rate, visual, syntactic and phonetic feature spaces	semantic feature space	YES
Desai et al., (2016)	Evidence that semantic representations are grounded in sensorimotor representations.	31	Linear regression + generalized linear test	head movement, mean CSF and white matter signal	fixation-duration, fixation to other words, word length, is noun, noun-concreteness, noun manipulability, unigram	NO
Hale et al., (2015)	Evidence of different types of structure building throughout the language network.	13	Mixed effect model, likelihood ratio test	prosodic-breaks, unigram, head movement, heart rate, lung action	syntactic node count, POS surprisal, PCFG surprisal	NO
Henderson et al., (2015)	Evidence of association between fixation duration and activity in the language network during reading and not pseudo-reading.	29	2-step GLM	head movement and CSF signal	Fixation onset, fixation duration, fixation number	NO
Henderson et al., (2016)	Evidence of sensitivity to syntactic surprisal in IFG and AntTemp.	40	Linear regression + generalized linear test	CSF and white matter signal, head movement	word-length, unigram, PCFG surprisal	NO
Huth et al., (2016)	Evidence of semantic selectivity in patterns of cortical regions.	7	Ridge regression + held-out eval.	word-rate, phonetic feature space,	semantic feature space	YES
Lopopolo et al.,(2017)	Evidence for distinct brain regions predicted by statistical structure of lexical, syntactic, and phonological information.	22	2-step GLM	word-rate, unigram, POS frequency, Phoneme Frequency	POS surprisal, lexical surprisal, phonetic surprisal	NO
Murphy et al., (2016)	Evidence for grammatical relation processing in the superior and middle temporal gyrus, using fMRI	22	Logistic regression classification		narrative features	YES
Speer et al., (2009)	Evidence of different brain regions tracking different narrative features such as character identity, goal changes, location and time change etc.	28	Hierarchical regression		narrative features	NO
Speer et al., (2007)	Evidence of sensitivity of a number of brain regions to narrative event boundaries.	28	GLM+ANOVA		narrative features	NO
Wehbe et al.,(2014)	Evidence that different areas in the language system are involved in representing semantic, syntax, and discourse level features.	8	Ridge regression + decoder held-out eval.		word-length, syntactic feature space, semantic feature space, narrative feature space	YES
Whitney et al., (2009)	Evidence that the right precuneus and cingulate cortex are sensitive for narrative shifts.	16	GLM+ANOVA		narrative features	NO
Willems et al., (2016)	Evidence of sensitivity of brain areas to entropy of next word probability distribution and surprisal.	24	2-step GLM	word-rate, unigram	lexical surprisal, next word entropy	NO
Present study	Evidence that the language network is predicted by measures of comprehension difficulty	42	Ridge regression + held-out eval.		self-paced reading times, eye-tracking measures	YES



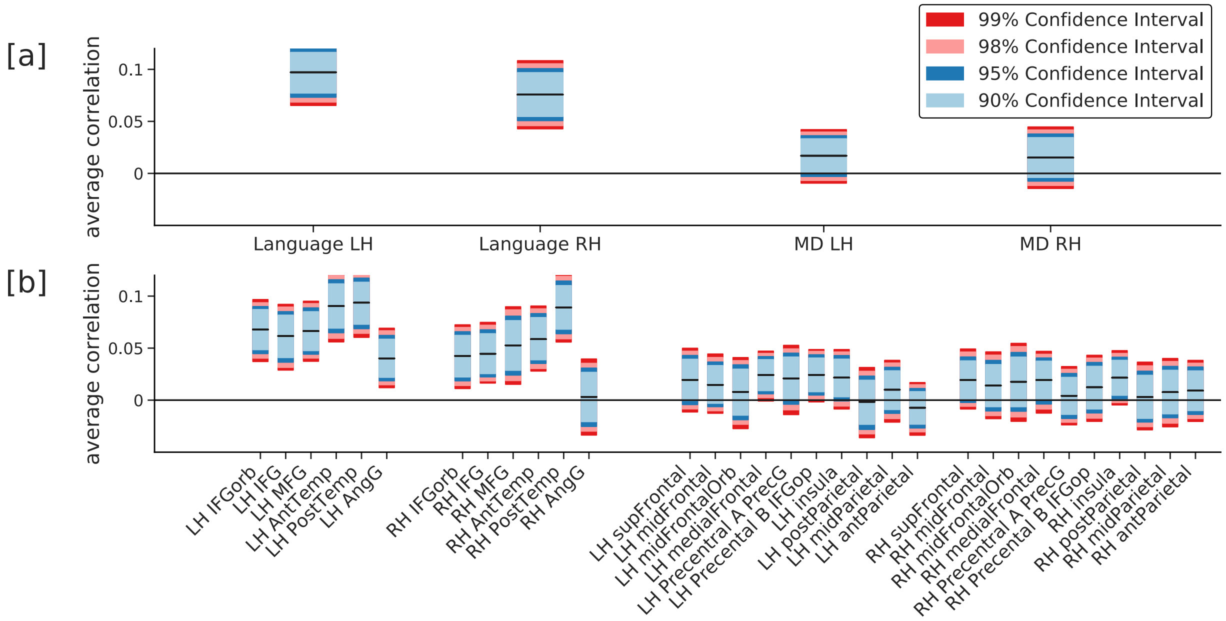
Supplementary Figure 1. Response of MD regions defined with the Nonwords > Sentences contrast to the Hard and Easy conditions of the visuo-spatial working memory MD localizer.



Supplementary Figure 2. Average (unnormalized and normalized) correlation between activity predicted as a function of comprehension difficulty (estimated using self-paced reading times and eye-tracking measures) and real held-out activity, normalized by the estimated reliability of the signal for each fROI group ([a] unnormalized and [c] normalized) and each fROI ([b] unnormalized and [d] normalized). **The MD fROIs were localized using the Nonwords>Sentences localizer which was available for all participant, allowing us to include all 42 participants in the analysis.** Performance was averaged across the 42 participants and bootstrap confidence intervals were constructed. Reading times predict the activity in left and right language fROIs, but not in MD fROIs.



Extended data fig. 3. Average (unnormalized and normalized) correlation between activity predicted as a function of comprehension difficulty (estimated using self-paced reading times and eye-tracking measures) and real held-out activity, normalized by the estimated reliability of the signal for each fROI group ([a] unnormalized and [c] normalized) and each fROI ([b] unnormalized and [d] normalized). **The analysis is restricted here to the 24 participants with the best performance.** The MD fROIs were localized using the Nonwords>Sentences localizer which was available for all participant with the best performance. Performance was averaged across these 24 participants and bootstrap confidence intervals were constructed. Reading times predict the activity in left and right language fROIs, but not in MD fROIs.



Extended data fig. 4. Average **unnormalized** correlation between activity predicted as a function of comprehension difficulty (estimated using a combination of self-paced reading times and eye-tracking measures) and real held-out activity, for [a] each fROI group and [b] each fROI. The MD fROIs were localized using the visuo-spatial memory task (available for 35 subjects). Performance was averaged across the 35 participants and bootstrap confidence intervals were constructed. Reading times predict the activity in left and right language fROIs, but not in MD fROIs.