nature research

Corresponding author(s): Leila Wehbe

Last updated by author(s): 10/10/2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The stimulus was displayed in the fMRI and MEG scanner using the Psychtoolbox-3 extension for MATLAB.
Data analysis	All data was analyzed using custom functions programmed in python 3.7. For preprocessing, we used SPM-8 for realigning the data and Freesurfer for reconstructing the subject surfaces. All our analysis custom scripts are available without restrictions at https://github.com/brainML/supraword. DOI=10.5281/ZENODO.7178795 We used pycortex for visualizations of the fMRI data and MNE for visualizations of the MEG data.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Two of the three datasets analyzed during this study can be found at http://www.cs.cmu.edu/~fmri/plosone/ for the fMRI dataset and at https://kilthub.cmu.edu/articles/dataset/RSVP_reading_of_book_chapter_in_MEG/20465898 for the MEG dataset. The remaining dataset is available from the Courtois Neuromod group at https://docs.cneuromod.ca/en/latest/ACCESS.html. Source data for figures 2, 3 and 4 is available with this manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We use fMRI and MEG data from previous studies. We have 9 subjects in fMRI and 9 subjects in MEG each reading a 5000 words long chapter. This data was collected between 2012 and 2015. At the time, the number of subjects was in keep with other experiments that use encoding models (Huth et al 2016 Nature, Kay et al 2008 Nature, and Sudre et al 2012 Neuroimage, etc.). We chose to collect a larger amount of data per participant than to collect a fewer amount for more participants. The fMRI dataset was one of the first naturalistic fMRI datasets shared and has been widely used and cited in the literature.
Data exclusions	No fMRI data was excluded from the analyses. One MEG subject was excluded due to too much noise in the recording session which resulted in problems with preprocessing.
Replication	Two independent researchers performed the analyses with independent code and reproduced the findings. Both the first author and the corresponding author repeated the experiments at various points in time and found the same pattern.
Randomization	A randomization is typically done in an interventional study to test the effect of a treatment. In a passive neuroimaging setup such as ours, the intervention is the choice of the stimuli. The stimuli will present different language features at different points to the participants and we build models to capture the effect of those features on the brain response. Thus our randomization is an implicit step that happens through the complex properties of the text stimuli.
Blinding	Blinding is not relevant to our study because there was no group allocation, as our study does not test an intervention.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	fMRI data was collected from 9 subjects (5 females and 4 males) recruited through Carnegie Mellon University, aged 18 to 40 years. MEG data was collected from 9 subjects (5 females and 4 males) recruited through Carnegie Mellon University and the University of Pittsburgh, aged 18 to 40 years.
Recruitment	Participants were recruited through Carnegie Mellon University and University of Pittsburgh. Written informed consent was obtained from all subjects.
Ethics oversight	Carnegie Mellon University and University of Pittsburgh IRB.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Magnetic resonance imaging

Experimental design

Design type	Task, naturalistic design (RSVP of text from an entire chapter)
Design specifications	Each subject read the entire chapter, divided into 4 runs. The total reading time was 45 min.
Behavioral performance measures	NA

Acquisition

Imaging type(s)	functional, structural
Field strength	3T
Sequence & imaging parameters	EPI with repetition time (TR)=2s, echo time=29 ms, flip angle=79◦, 36 slices and 3 × 3 × 3mm voxels
Area of acquisition	Whole brain

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software	The data for each subject was slice-time and motion corrected using SPM8, then detrended and smoothed with a 3mm full-width-half-max kernel. The brain surface of each subject was reconstructed using Freesurfer (Fischl, 2012), and a grey matter mask was obtained.
Normalization	Data was analyzed in the subject native space
Normalization template	NA
Noise and artifact removal	Data was detrended and smoothed with a 3mm full-width-half-max kernel
Volume censoring	NA

Statistical modeling & inference

Model type and settings	Predictive model (i.e. encoding model that predicts brain recordings as a function of a representation of the stimulus)
Effect(s) tested	We test whether a specific representation of the stimulus is predictive of fMRI and MEG recordings.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☒ Both

Anatomical location(s)	The ROI masks were obtained from previous work by different researchers who found that these ROI are involved in language processing (Fedorenko et al. 2010, Fedorenko and Thompson-Schill 2014) and word semantics (Binder et al. 2009). These masks are entirely independent of our analyses and data.
Statistic type for inference (See Eklund et al. 2016)	voxel-level and ROI-level. The ROI were predetermined using different data that is not part of our analyses (see above).
Correction	FWE control using Holm-Bonferroni procedure and FDR control using Benjamini-Hochberg procedure

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Multivariate modeling and predictive analysis	Feature extraction: We passed sequences of 25 consecutive stimulus words through a pre-trained ELMo model (Peters et al. 2018). For each word, we obtained the token-level representations and the representations from the first hidden layer. Full details are available in the paragraphs titled “Obtaining full stimulus representations” and “Obtaining residual stimulus representations” under Materials and Methods. Dimension reduction: We reduced the dimensionality of the obtained stimulus representations to the first 10 principal components, obtained using PCA. Model: We estimate an encoding model that takes a stimulus representation e_t as input and predicts the brain recording associated with reading the same words that were used to derive the stimulus

representation. We estimate a function f, such that f(e_t)=b, where b is the brain activity recorded with either MEG or fMRI. We follow previous work (Sudre et al. 2012, Wehbe et al. 2014a, Wehbe et al. 2014b, Nishimoto et al. 2011, Huth et al. 2016) and model f as a linear function, regularized by the ridge penalty. Full details about the models that are specific to the fMRI and MEG data are available in the paragraphs titled "fMRI encoding models" and "MEG encoding models" under Materials and Methods.

Training: All encoding models are trained via four-fold cross-validation and the regularization parameters are chosen via nested cross-validation.

Evaluation: We evaluate the predictions of each encoding model by computing the Pearson correlation between the held-out brain recordings and the corresponding predictions in the four-fold cross-validation setting. We compute one correlation value for each of the 4 cross-validation folds and report the average value as the final encoding model performance.