nature portfolio

Corresponding author(s): Alexander G. Huth

Last updated by author(s): Mar 13, 2023

Reporting Summary

Nature Portfolio 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 Portfolio 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	All functional data were motion corrected using the FMRIB Linear Image Registration Tool (FLIRT) from FSL (5.0). Cortical surface meshes were generated from T1-weighted anatomical scans using FreeSurfer software (6.0). Flat maps were created using pycortex software (1.3.0). The Penn Phonetics Lab Forced Aligner (P2FA) (1.003) was used to automatically align stimulus audio with manual transcripts. Praat (6.2.01) was used to manually check and correct the alignments
Data analysis	All model fitting and analysis was performed using custom software written in Python (3.9.5), making heavy use of NumPy (1.20.2), SciPy (1.6.2), PyTorch (1.9.0), Transformers (4.6.1), and pycortex (1.3.0). Custom decoding code is available at https://github.com/HuthLab/semantic-decoding.

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data collected during the decoder resistance experiment are available upon reasonable request, but were not publicly released due to concern that the data could be used to discover ways to bypass subject resistance. All other data are available at https://openneuro.org/datasets/ds003020 and https://openneuro.org/datasets/ds004510.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	fMRI data were collected from 3 female subjects and 4 male subjects. Behavioral data were collected from 50 female subjects, 49 male subjects, and 1 non-binary subject. Sex and gender were not considered in the study design because we do not expect language decoding performance or language comprehension performance to depend on sex or gender. Sex and gender were determined based on self-reporting.
Population characteristics	fMRI data were collected from 7 healthy human subjects (3 female, 4 male) between 23 and 36 years of age, with normal hearing, normal or corrected-to-normal vision, and native English language proficiency. Behavioral data were collected from 100 human subjects (50 female, 49 male, 1 non-binary) between 19 and 70 years of age, with native or fluent English language proficiency.
Recruitment	fMRI subjects were recruited from the lab. Lab members were asked if they would like to participate in the experiment. The subjects used in this experiment were the lab members that volunteered to get scanned for at least 6 sessions. Because there are no experimental manipulations in this study and subjects are only performing naturalistic tasks, we do not expect this to have an effect on our results in any way. Behavioral subjects were recruited through the Prolific online platform. The only inclusion criteria were age (at least 18 years) and native or fluent English language proficiency. While the subjects chose whether to participate in the experiment, we do not expect this self-selection to have an effect on our results in any way.
Ethics oversight	The experimental protocol was approved by the Institutional Review Board at the University of Texas at Austin. All subjects gave written informed consent.

Note that full information on the approval of the study protocol must also be provided in the 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	For the fMRI experiment, no sample-size calculation was performed. Due to the large amount of data collected for each fMRI subject, the experimental approach does not rely on a large sample-size. The sample size of the fMRI experiment is comparable to those of previous decoding publications For the behavioral experiment, control group participants were expected to perform close to ceiling accuracy, so a sample size of 100 provides sufficient power to detect significance differences with test accuracies as high as 70% (G*Power, exact test of proportions with independent groups).
Data exclusions	No data were excluded from the analysis.
Replication	For the fMRI experiment, language decoding models were independently fit and evaluated for each subject, and the results of the study were consistent across subjects. This is effectively 3 separate and independent replications of the fMRI experiment. These sample sizes are similar to those reported in previous decoding publications. Exploratory decoding analyses were restricted to data collected while subject S3 listened to one test story. All exploratory analyses were qualitative. The analysis pipeline was frozen before we viewed results for the remaining

	subjects, stimuli, and experiments. For the behavioral experiment, multiple-choice answers were written by a separate researcher who had never seen the decoder predictions to remove researcher bias. Bootstrap re-sampling was used to test for reproducibility.
Randomization	For the fMRI experiment, subjects were not allocated into experimental groups. For the behavioral experiment, subjects were randomly allocated into experimental and control groups and blinded to group assignment.
Blinding	For the fMRI experiment, subjects were not allocated into experimental groups. For the behavioral experiment, investigators were blinded to group allocation during data collection and analysis.

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
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Magnetic resonance imaging

Experimental design

Design type	Naturalistic task design
Design specifications	Training data collection was broken up into 16 different scanning sessions, the first session involving the anatomical scan and localizers, and each successive session consisting of 5 or 6 spoken narrative stories from The Moth Radio Hour or Modern Love. Testing data was collected in 2 different scanning sessions for subjects S2 and S3, and 1 scanning session for subject S1.
Behavioral performance measures	The study does not involve behavioral performance.

Acquisition

Imaging type(s)	Functional
Field strength	3T
Sequence & imaging parameters	Gradient echo EPI sequence, field of view = 220mm, matrix size = 84x84, slice thickness = 2.6mm, flip angle = 71°
Area of acquisition	Whole brain

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software	Functional data were motion corrected using the FMRIB Linear Image Registration Tool (FLIRT) from FSL 5.0. FLIRT was used to align all data to a template that was made from the average across the first functional run in the first story session for each subject. These automatic alignments were manually checked for accuracy.
Normalization	Data were not normalized as we were looking for effects within individual subjects.
Normalization template	Data were not normalized as we were looking for effects within individual subjects.
Noise and artifact removal	Low frequency voxel response drift was identified using a 2nd order Savitzky-Golay filter with a 120 second window and then subtracted from the signal. To avoid onset artifacts and poor detrending performance near each end of the scan, responses were trimmed by removing 20 seconds (10 volumes) at the beginning and end of each scan.
Volume censoring	There were no volumes with sufficient motion to warrant censoring. All subjects wore customized headcases to prevent excessive movement.

Statistical modeling & inference

Model type and settings	Our study follows a Bayesian decoding framework. In the model estimation stage, predictive voxel-wise encoding models were separately estimated for each subject using brain recordings collected while that subject listened to narrative stories. In the language reconstruction stage, brain recordings were collected while the subject performed a range of naturalistic tasks. A language model generates candidate word sequences, and the subject's encoding model evaluates the candidates against the brain recordings.
Effect(s) tested	Decoding performance for each task was calculated by comparing decoder predictions to reference stimulus words. Paired t-tests were used to compare decoding performance across tasks.

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

Anatomical location(s)	Anatomical ROIs were defined for each subject using Freesurfer. Functional ROIs were defined for each subject using an auditory cortex localizer and a motor localizer.
Statistic type for inference (See Eklund et al. 2016)	Decoding performance was calculated by comparing decoder predictions to reference stimulus words.
Correction	All tests were corrected for multiple comparisons when necessary using the false discovery rate (FDR).

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	Voxel-wise encoding models were fit on a training dataset using L2-regularized linear regression to predict BOLD responses from semantic stimulus features. Semantic features of each stimulus word were extracted from a pre-trained GPT language model. Encoding models were used to decode novel BOLD responses in a test dataset. Prediction performance was computed using standard language similarity metrics on the decoder predictions and reference stimulus words.