scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Introducing MEG-MASC a high-quality magneto-encephalography dataset for evaluating natural speech processing

Laura Gwilliams[1,2,3] ✉, Graham Flick[2,3,4,5], Alec Marantz[2,3,4], Liina Pylkkänen[2,3,4], David Poeppel[2,6] & Jean-Rémi King[2,7]

The "MEG-MASC" dataset provides a curated set of raw magnetoencephalography (MEG) recordings of 27 English speakers who listened to two hours of naturalistic stories. Each participant performed two identical sessions, involving listening to four fictional stories from the Manually Annotated Sub-Corpus (MASC) intermixed with random word lists and comprehension questions. We time-stamp the onset and offset of each word and phoneme in the metadata of the recording, and organize the dataset according to the 'Brain Imaging Data Structure' (BIDS). This data collection provides a suitable benchmark to large-scale encoding and decoding analyses of temporally-resolved brain responses to speech. We provide the Python code to replicate several validations analyses of the MEG evoked responses such as the temporal decoding of phonetic features and word frequency. All code and MEG, audio and text data are publicly available to keep with best practices in transparent and reproducible research.

Background & Summary

Humans have the unique ability to produce and comprehend an infinite number of novel utterances. This capacity of the human brain has been the subject of vigorous studies for decades. Yet, the core computational mechanisms upholding this feat remain largely unknown[1–3].

To tackle this issue, a common experimental approach has been to decompose language processing into elementary computations using highly controlled factorial designs. This approach allows experimenters to compare average brain responses to carefully chosen stimuli and make inferences based on the select ways that those stimuli were designed to differ. The field has learnt a lot about the neurobiology of language by taking this approach; however, factorial designs also face several key challenges[4]. First, this method has led the community to study language processing in atypical scenarios (*e.g.* using unusual text fonts[5], meaningless syntactic constructs[6,7], or words and phrases isolated from context[8,9]). Presenting language in this unconventional manner runs the risk of studying phenomena that are not representative of how language is naturally processed. Second, high-level cognitive functions can be difficult to fully orthogonalize in a factorial design. For instance, comparing brain responses to words and sentences matched in length, syntactic structure, plausibility and pronunciation is often close to impossible. In the best case, experimenters will be forced to make concessions on how well the critical contrasts are controlled. In the worst case, unidentified confounds may drive differences associated with experimental contrasts, leading to incorrect conclusions.

During the past decade, several studies have complemented the factorial paradigm with more natural experiments. In these studies, participants listen to continuous speech[10–12], read continuous prose[13,14] or watch videos that include verbal communication[15]. This approach is more likely to recruit neural computations that are representative of day-to-day language processing. Complications arising from correlated language features can be overcome by explicitly modeling properties of interest, in tandem with potential confounds. This allows variance belonging to either source to be appropriately distinguished.

[1]Department of Psychology, Stanford University, Stanford, USA. [2]Department of Psychology, New York University, New York, USA. [3]NYU Abu Dhabi Institute, Abu Dhabi, United Arab Emirates. [4]Department of Linguistics, New York University, New York, USA. [5]Rotman Research Institute, Baycrest Hospital, Toronto, Canada. [6]Ernst Struengmann Institute for Neuroscience, Frankfurt, Germany. [7]LSP, École normale supérieure, PSL University, CNRS, 75005, Paris, France. ✉e-mail: laura.gwilliams@stanford.edu

To analyze the brain responses to the complex stimulation that natural language provides, a variety of encoding and decoding methods have proved remarkably effective[10,16–21]. Consequently, language studies based on naturalistic designs have since flourished[11,22]. The popularity of this approach has some of its roots in the rise of natural language processing (NLP) algorithms, which map remarkably onto brain responses to written and spoken sentences[23–30]. Such tools also allow experimenters to annotate the language stimuli for features of interest, without relying on time-consuming annotations done by hand. These data have allowed researchers to identify the main semantic components[10], recover the hierarchy of integration constants in the language network[31], distinguish syntax and semantics hubs[32] and to track the hierarchy of predictions elicited during speech processing[28,33,34]. More generally, brain responses to natural stories have proved useful in keeping participants engaged, while studying the neural representations of phonemes, word surprisal and entropy[22,35,36].

While large and high-quality functional Magnetic Resonance Imaging (fMRI) datasets related to language processing have recently been released[37,38], there is currently little publicly available high-quality temporally-resolved brain recordings acquired during story listening. The most extensive databases of such data include:

- van Essen *et al.*[39]: 72 subjects recorded with fMRI and MEG as part of the Human Connectome Project, listening to 10 minutes of short stories, no repeated session[39]
- Brennan and Hale[40]: 33 subjects recorded with EEG, listening to 12 min of a book chapter, no repeated session[40]
- Broderick *et al.*[11]: 9–33 subjects recorded with EEG, conducting different speech tasks, no repeated sessions[11]
- Schoffelen *et al.*[38]: 100 subjects recorded with fMRI and MEG, listening to de-contextualised sentences and word lists, no repeated session[38]
- Armeni *et al.*[41]: 3 subjects recorded with MEG, listening to 10 h of Sherlock Holmes, no repeated session[41]

Until the present release of our dataset, there existed no public magneto-encephalography (MEG) with (1) several hours of story listening (2) multiple sessions (3) a systematic audio, phonetic and word annotations (4) a standardized data structure. Thus, our dataset offers a powerful resource to the scientific community.

In the present study, 27 English-speaking subjects performed ~two hours of story listening, punctuated by random word lists and comprehension questions in the MEG scanner. Except if stated otherwise, each subject listened to four distinct fictional stories twice.

Methods

Participants. Twenty-seven English-speaking adults were recruited from the subject pool of NYU Abu Dhabi (15 females; age: M = 24.8, SD = 6.4). All participants provided a written informed consent and were compensated for their time. Participants reported having normal hearing and no history of neurological disorders. All participants were right-handed, as evaluated using the Edinburgh Handedness Inventory questionnaire[42]. All but one participant (S21) were native English speakers - this person was a native speaker of Hindi, and learned English at 10 years old. All but five participants (S3, S12, S16, S20, S21) performed two identical one-hour-long sessions. These two recording sessions were separated by at least one day and at most two months depending on the availability of the experimenters and of the participants. The study was approved by the Institutional Review Board (IRB) ethics committee of New York University Abu Dhabi.

Procedure. Within each ~1 h recording session, participants were recorded with a 208 axial-gradiometer MEG scanner built by the Kanazawa Institute of Technology (KIT), and sampled at 1,000 Hz, and online band-pass filtered between 0.01 and 200 Hz while they listened to four distinct stories through binaural tube earphones (Aero Technologies), at a mean level of 70 dB sound pressure level.

Before the experiment, participants were exposed to 20 sec of each of the distinct speaker voices used in the study to (i) clarify the structure of the session and (ii) familiarize the participants with these voices. The sound files and scripts are available in ('/stimuli/exp_intro/').

The order in which the four stories were presented was assigned pseudo-randomly, thanks to a "Latin-square design" across participants. The story order for each participant can be found in 'participants.tsv'. This participant-specific order was used for both recording sessions. Our motivation for running two identical sessions was to (i) give researchers the ability to average the data across the two recordings to boost signal-to-noise; (ii) provide a like-for-like data reliability measure; (iii) give the opportunity for matched train and test datasets if attempting to run cross validated analyses.

To ensure that the participants were attentive to the stories, they responded, every ~3 min and with a button press, to a two-alternative forced-choice question relative to the story content (e.g. 'What precious material had Chuck found? Diamonds or Gold'). Participants performed this task with an average accuracy of 98%, confirming their engagement with and comprehension of the stories. The questions and answers are provided in ('stimuli/task/question_dict.py').

Participants who did not already have a T1-weighted anatomical scan usable for the present study were scanned in a 3 T Magnetic-Resonance-Imaging (MRI) scanner after the MEG recording to avoid magnetic artefacts. Twelve participants returned for their T1 scan.

Before each MEG session, the head shape of each participant was digitized with a hand-held FastSCAN laser scanner (Polhemus), and co-registered with five head-position coils. The positions of these coils with regard to the MEG sensors were collected before and after each recording and stored in the 'marker' file, following the KIT's system. The experimenter continuously monitored head position during the acquisition to ensure that the participants did not move.

Stimuli. Four English fictional stories were selected from the Manually Annotated Sub-Corpus (MASC) which is part of the larger Open American National Corpus[43]. MASC is distributed without license or other restrictions (https://anc.org/data/masc/corpus/577-2/):

- 'LW1': a 861-word story narrating an alien spaceship trying to find its way home (5 min, 20 sec)
- 'Cable Spool Boy': a 1,948-word story narrating two young brothers playing in the woods (11 min)
- 'Easy Money': a 3,541-word fiction narrating two friends using a magical trick to make money (12 min, 10 sec)
- 'The Black Willow': a 4,652-word story narrating the difficulties an author encounters during writing (25 min, 50 sec)

An audio track corresponding to each of these stories was synthesized using Mac OS Mojave © version 10.14 text-to-speech. To help decorrelate language features from acoustic representations, we varied both voices and speech rate every 5–20 sentences. Specifically, we used three distinct synthetic voices:'Ava', 'Samantha' and 'Allison' speaking between 145 and 205 words per minute. Additionally, we varied the silence between sentences between 0 and 1,000 ms. Both speech rate and silence duration were sampled from a uniform distribution between the min and max values.

Each story was divided into ~3 min sound files. In between these sounds – approximately every 30 s – we played a random word list generated from the unique content words (nouns, proper nouns, verbs, adverbs and adjectives) selected from the preceding 5 min segment presented in random order. We decided to include word lists to allow data users to compare brain responses to content words within and outside of context, following experimental paradigms of previous studies[38,44]. In addition, a very small fraction (<1%) of non-words were inserted into the natural sentences, on average every 30 words. We decided to include non-words to allow comparisons between phonetic sequences that do and do not have an associated meaning.

Hereafter, and following the BIDS labeling[45], each "task" corresponds to the concatenation of these sentences and word lists. Each subject listened to the exact same set of four tasks, in a different block order.

Preprocessing. *MEG.* The MEG dataset and its annotations are shared raw (i.e. not preprocessed) organized according to the Brain Imaging Data Structure[45] MNE-BIDS[46].

MRI. Structural MRIs were collected with separate averages of the T1w images using 3D MPRAGE sequence with 0.8 mm isotropic resolution (FOV = 256 mm, matrix = 320,208 sagittal slices in a single slab), TR = 2400 ms, TE = 2.22 ms, TI = 1000 ms, FA = 8 degrees, Bandwidth (BW) = 220 Hz per pixel, Echo Spacing (ES) = 7.5 ms, phase encoding undersampling factor GRAPPA = 2, no phase encoding oversampling.

To avoid subject identification, the T1-weighted MRI anatomical scan was defaced using PyDeface[47] (https://github.com/poldracklab/pydeface) and manually checked.

For four subjects (02, 06, 07, 19) we were unable to record structural MRIs, and so instead we provide the scaled FreeSurfer average MRIs in their place.

The alignment between the spaces of (1) the head-position coils, (2) the MEG sensors and (3) the T1 MRI was co-registered manually with MNE-Python[48].

Stimuli. We include in the dataset: the original stories ('stimuli/text'), the stories intertwined with the word lists ('stimuli/text_with_wordlists') and their corresponding audio tracks ('stimuli/audio').he alignment between the MEG data and the words and phonemes is provided for each participant separately (e.g., /sub-01/ses-0/meg/sub-01_ses-0_task-1_events.tsv').

Both sentences and word lists were annotated for phoneme boundaries and labels (107 phoneme labels, detailing phoneme category and its location in the word (Beginning; Internal; End) using the 'Gentle aligner' from the Python module lowerquality https://lowerquality.com/gentle/. However, the inclusion of the original audio leaves the possibility for future research to develop more advanced alignment technique and recover additional features.

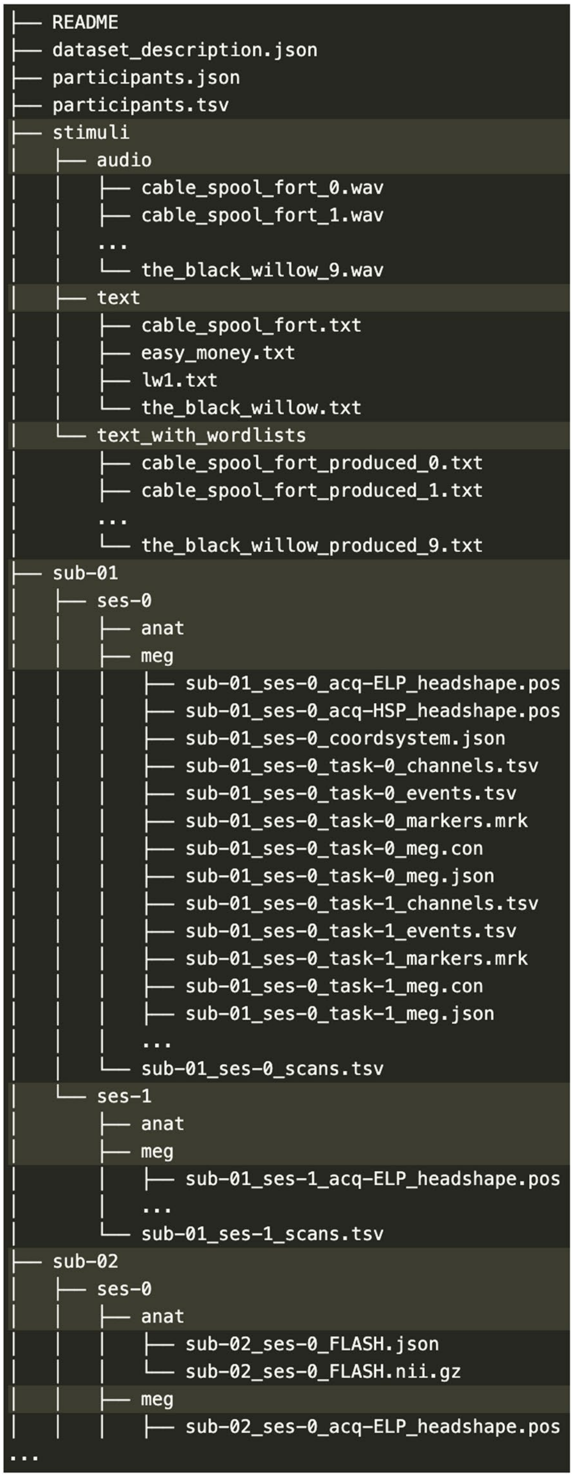
For each phoneme and word, we indicate the corresponding voice, speech rate, wav file, story, word position within the sequence, and sequence position within the story, and whether the sequence is a word list or a sentence.

Computing environment. In addition to the packages mentioned in this manuscript, the processing of the present data is based on the free and open-source ecosystem of the neuroimaging community. In particular, we used:

- MNE BIDS[46] (https://mne.tools/mne-bids)
- Bids-Validator (https://github.com/bids-standard/bids-validator)
- Nibabel[49] (https://nipy.org/nibabel/)
- Scikit-Learn[50] (https://scikit-learn.org/)
- Pandas[51] (https://pandas.pydata.org/)

Data Records

The dataset is organized according to Brain Imaging Data Structure (BIDS) 1.2.1[45] and publicly available on the Open Science Framework data repository[52] https://doi.org/10.17605/OSF.IO/AG3KJ under a Creative Common Licence 0. An image of the folder structure is provided in Fig. 1. The detailed description of the BIDS file system is available at http://bids.neuroimaging.io/. In summary,



```
├── README
├── dataset_description.json
├── participants.json
├── participants.tsv
├── stimuli
│   ├── audio
│   │   ├── cable_spool_fort_0.wav
│   │   ├── cable_spool_fort_1.wav
│   │   ...
│   │   └── the_black_willow_9.wav
│   ├── text
│   │   ├── cable_spool_fort.txt
│   │   ├── easy_money.txt
│   │   ├── lw1.txt
│   │   └── the_black_willow.txt
│   └── text_with_wordlists
│       ├── cable_spool_fort_produced_0.txt
│       ├── cable_spool_fort_produced_1.txt
│       ...
│       └── the_black_willow_produced_9.txt
├── sub-01
│   ├── ses-0
│   │   ├── anat
│   │   ├── meg
│   │   │   ├── sub-01_ses-0_acq-ELP_headshape.pos
│   │   │   ├── sub-01_ses-0_acq-HSP_headshape.pos
│   │   │   ├── sub-01_ses-0_coordsystem.json
│   │   │   ├── sub-01_ses-0_task-0_channels.tsv
│   │   │   ├── sub-01_ses-0_task-0_events.tsv
│   │   │   ├── sub-01_ses-0_task-0_markers.mrk
│   │   │   ├── sub-01_ses-0_task-0_meg.con
│   │   │   ├── sub-01_ses-0_task-0_meg.json
│   │   │   ├── sub-01_ses-0_task-1_channels.tsv
│   │   │   ├── sub-01_ses-0_task-1_events.tsv
│   │   │   ├── sub-01_ses-0_task-1_markers.mrk
│   │   │   ├── sub-01_ses-0_task-1_meg.con
│   │   │   ├── sub-01_ses-0_task-1_meg.json
│   │   │   ...
│   │   └── sub-01_ses-0_scans.tsv
│   └── ses-1
│       ├── anat
│       ├── meg
│       │   ├── sub-01_ses-1_acq-ELP_headshape.pos
│       │   ...
│       └── sub-01_ses-1_scans.tsv
├── sub-02
│   ├── ses-0
│   │   ├── anat
│   │   │   ├── sub-02_ses-0_FLASH.json
│   │   │   └── sub-02_ses-0_FLASH.nii.gz
│   │   ├── meg
│   │   │   ├── sub-02_ses-0_acq-ELP_headshape.pos
...
```

Fig. 1 Dataset file structure.

- ‘./dataset_description.json’ describes the dataset
- ‘./participants.tsv’ indicates the age and gender of each participant, the order in which they heard the stories, whether they have an anatomical MRI scan, and how many recording sessions they completed
- ‘./stimuli/’ contains the original texts, the modified texts (*i.e*. with word lists), the synthesized audio tracks.

- Each’./sub-SXXX’ contains the brain recordings of a unique participant divided by session (*e.g*.’ses-0’ and’ses-1’)
- In each session folder lies the anatomical and the meg data, and the timestamp annotations (see Fig. 4).
- Sessions are numbered by temporal order (s0 is first; s1 is second).

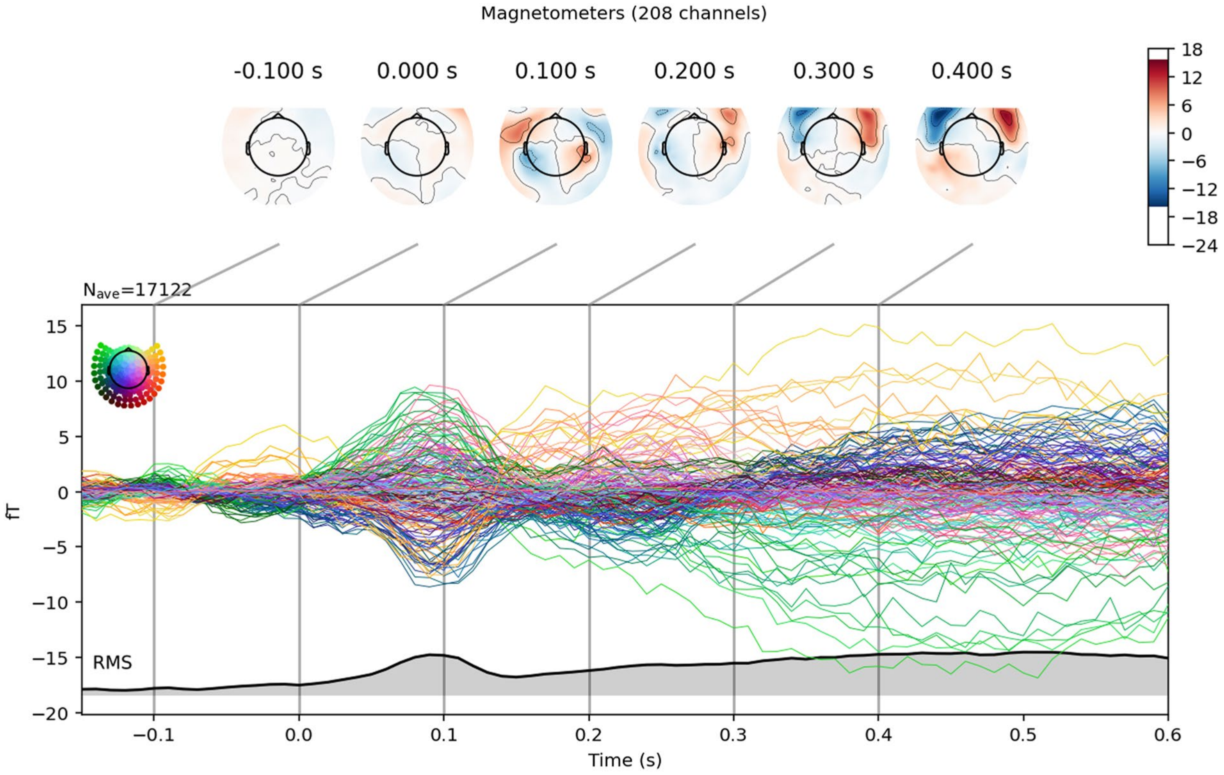


Fig. 2 Median (across subjects) evoked response to all words. The gray area indicates the global field power (GFP).

- Tasks are numbered by a unique integer common across participants.
- The dataset can be read directly with MNE-BIDS[46].

Technical Validation

We checked that the present dataset complies with the standardized brain imaging data structure by using the Bids-Validator (https://github.com/bids-standard/bids-validator).

MEG recordings are notoriously noisy and thus challenging to validate empirically. In particular, MEG can be corrupted by environmental noise (nearby electronic systems) and physiological noise (eye movement, heart activity, facial movements)[53]. To address this issue, several labs have proposed a myriad of preprocessing techniques based on temporal and spatial filtering[54] and trial and channel rejection[55]. However, there is currently no accepted standard for the selection and ordering these preprocessing steps. Consequently, we here opted for (1) a minimalist preprocessing pipeline derived from MNE-Python's default pipeline[48] followed by (2) median evoked responses and (3) standard single-trial linear decoding analyses.

Minimal preprocessing. For each subject separately, and using the default parameters of MNE-Python, we:

- bandpass filtered the MEG data between 0.5 and 30.0 Hz with `raw.load_data().filter(0.5, 30.0, n_jobs=1)`,
- temporally-decimate the data 10x, segment these continuous signals between −200 ms and 600 ms after word and phoneme onset, and apply a baseline correction between −200 ms to 0 ms with mne.Epochs(tmin = −0.2, tmax = 0.6, decim = 10, baseline = (−0.2, 0.0)),
- and clip the MEG data between fifth and ninety-fifth percentile of the data across channels.

Evoked. Figure 2 displays the median evoked responses across participants and words onset and after phoneme onsets, respectively. Both of these topographies are typical of auditory activity in MEG[36].

Decoding. For each recording independently, our objective was to verify the alignment between the word annotations and the MEG recordings. To this end, we trained a linear classifier $W \in Rd$ across all $d = 208$ magnetometers ($X \in Rn \times d$), for each time sample relative to word (or phoneme) onset independently, and for each subject separately. The classifier consisted of a standard scaler, followed by a linear discriminant regression implemented by scikit-learn[50] using `model = make_pipeline(StandardScaler(), LinearDiscriminantAnalysis())`

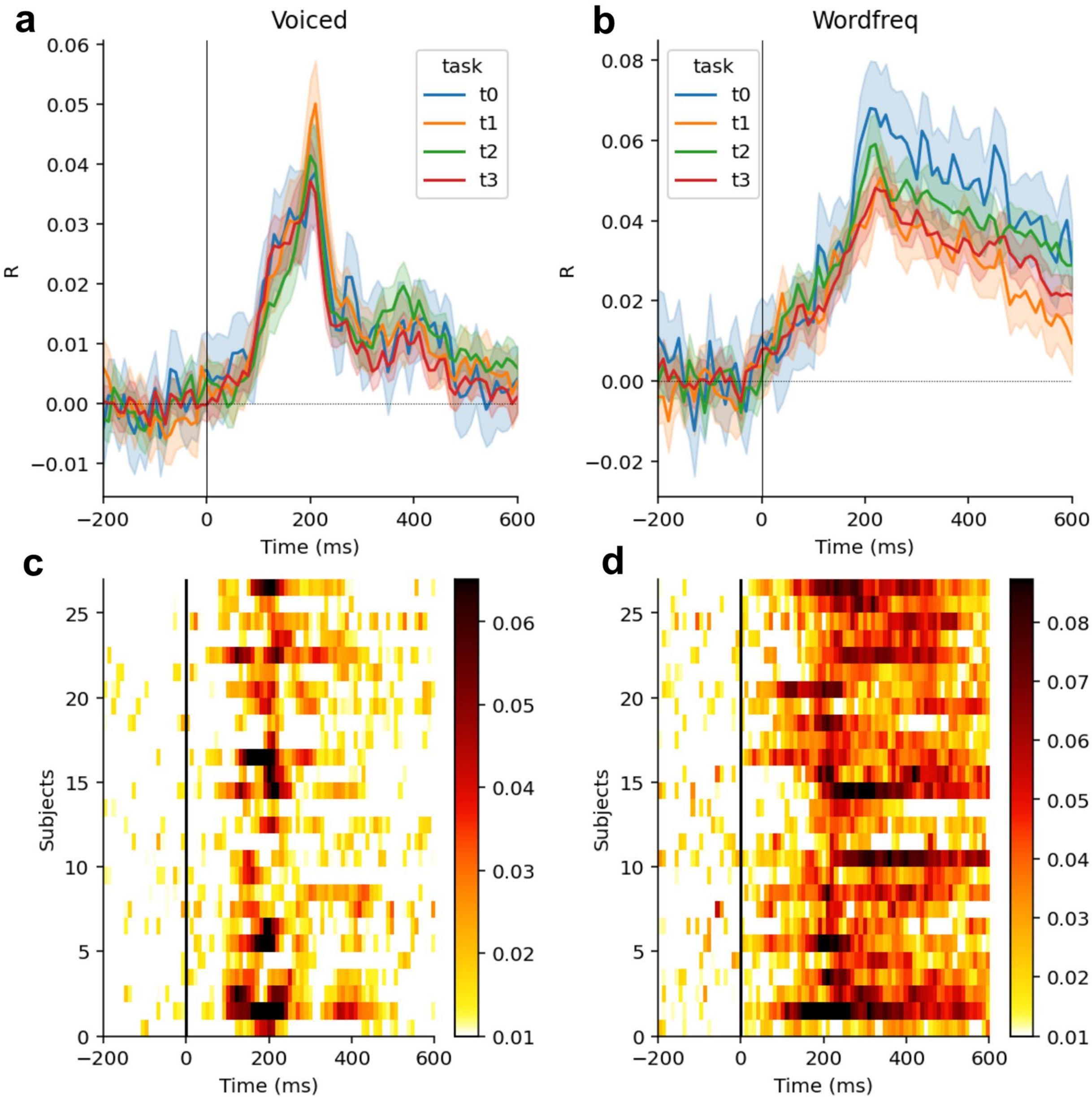


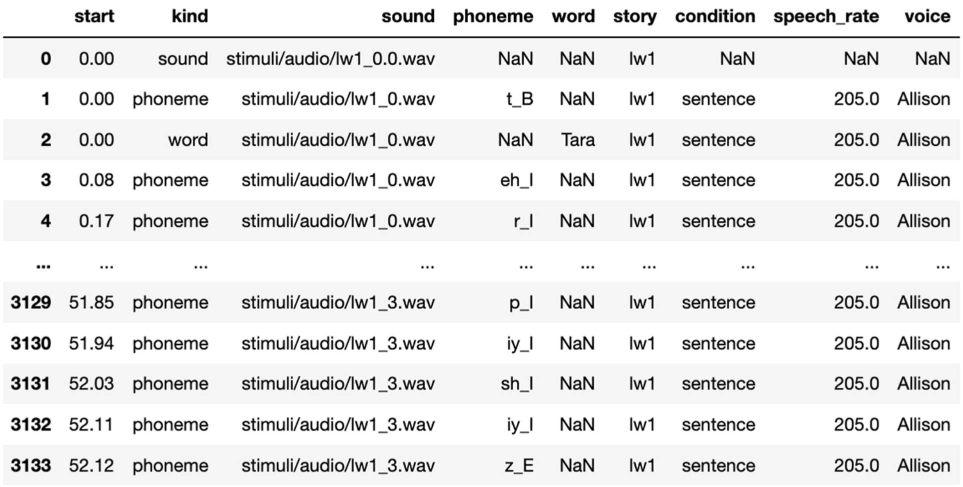
Fig. 3 (**a**) Average (mean) decoding of whether the phoneme is voiced or not as a function of time following phoneme onset. The four colors refer to the four tasks (stories + word lists). Error bar are SEM across subjects. (**b**) Same as A for the decoding of words' zipf frequency as a function of word onset. (**c**) Decoding of voicing (average across all tasks) for each participant, as a function of time following phoneme onset. (**d**) Same as C for decoding of word frequency (average across all tasks) for each participant, as a function of time following word onset.

- decode high versus low median zipf-frequency of each word, as defined by the WordFreq package[56].
- decode whether the phoneme is voiced or not.

The decoding pipeline was trained and evaluated using a five-split cross-validation scheme (with shuffling) using `cv = KFold(5, shuffle = True, random_state = 0)` The scoring metric reported is Pearson R correlation between the continuous probabilistic output of the classifier on each trial, and the ground truth label (high vs. low for word frequency; voiced vs. voiceless for voicing). The full decoding pipeline can be found in the script check_decoding.py.

The results displayed in Fig. 3 show a reliable decoding at the phoneme and at the word level, across both subjects and tasks (*i.e.* stories).

The success of our decoding analysis demonstrates: (i) the data have been correctly time-stamped relative to phoneme and word onset, in order to elicit a zero-aligned decoding timecourse; (ii) the data contain reliable signals that contain speech-related properties, suitable for further investigation; (iii) information at multiple levels



	start	kind	sound	phoneme	word	story	condition	speech_rate	voice
0	0.00	sound	stimuli/audio/lw1_0.0.wav	NaN	NaN	lw1	NaN	NaN	NaN
1	0.00	phoneme	stimuli/audio/lw1_0.wav	t_B	NaN	lw1	sentence	205.0	Allison
2	0.00	word	stimuli/audio/lw1_0.wav	NaN	Tara	lw1	sentence	205.0	Allison
3	0.08	phoneme	stimuli/audio/lw1_0.wav	eh_I	NaN	lw1	sentence	205.0	Allison
4	0.17	phoneme	stimuli/audio/lw1_0.wav	r_I	NaN	lw1	sentence	205.0	Allison
...	...	...	...	...	...	...	...	...	...
3129	51.85	phoneme	stimuli/audio/lw1_3.wav	p_I	NaN	lw1	sentence	205.0	Allison
3130	51.94	phoneme	stimuli/audio/lw1_3.wav	iy_I	NaN	lw1	sentence	205.0	Allison
3131	52.03	phoneme	stimuli/audio/lw1_3.wav	sh_I	NaN	lw1	sentence	205.0	Allison
3132	52.11	phoneme	stimuli/audio/lw1_3.wav	iy_I	NaN	lw1	sentence	205.0	Allison
3133	52.12	phoneme	stimuli/audio/lw1_3.wav	z_E	NaN	lw1	sentence	205.0	Allison

Fig. 4 MEG data annotations: Pandas DataFrame of sound, phoneme and word time-stamps.

(phonetic and lexical) are present in the data, allowing users to test hypotheses at different linguistic levels of description. We anticipate that encoding models would provide equally compelling results. Note that a decoding performance of Pearson R = 0.08 is typical for single-trial MEG data of continuous listening, and is of the same magnitude that has been reported in previous studies[38]. We have a large number of events (tens of thousands of phonemes; thousands of words), and this dataset has been demonstrated to provide sufficient statistical power to yield significant results, despite small effect sizes[36,57].

Usage Notes

```
import pandas as pd
import mne bids

bids_path=mne_bids.BIDSPATH(
        subject=’01’,
        session=’0’,
        task=’0’,
        datatype=”meg”,
        root=’my/data/path’)

raw=mne_bids.read_raw_bids(bids_path)
raw.load_data().data # channels X times

df=raw.annotations.to_data_frame()
```

Accessing all sound, word and phoneme annotation is directly readable in a Pandas[58] DataFrame format:

```
df=pd.DataFrame(df.description.apply(eval).to_list())
```

Code availability

The code is available on https://github.com/kingjr/meg-masc/.

Received: 29 July 2022; Accepted: 16 November 2023;
Published online: 04 December 2023

References

1. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nat. reviews neuroscience* **8**, 393–402 (2007).
2. Berwick, R. C., Friederici, A. D., Chomsky, N. & Bolhuis, J. J. Evolution, brain, and the nature of language. *Trends cognitive sciences* **17**, 89–98 (2013).
3. Dehaene, S., Meyniel, F., Wacongne, C., Wang, L. & Pallier, C. The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. *Neuron* **88**, 2–19 (2015).
4. Hamilton, L. S. & Huth, A. G. The revolution will not be controlled: natural stimuli in speech neuroscience. *Lang. cognition neuroscience* **35**, 573–582 (2020).
5. Gwilliams, L. & King, J.-R. Recurrent processes support a cascade of hierarchical decisions. *ELife* **9**, e56603 (2020).
6. Pallier, C., Devauchelle, A.-D. & Dehaene, S. Cortical representation of the constituent structure of sentences. *Proc. Natl. Acad. Sci.* **108**, 2522–2527 (2011).
7. Petersson, K.-M., Folia, V. & Hagoort, P. What artificial grammar learning reveals about the neurobiology of syntax. *Brain language* **120**, 83–95 (2012).
8. Gwilliams, L., Linzen, T., Poeppel, D. & Marantz, A. In spoken word recognition, the future predicts the past. *J. Neurosci.* **38**, 7585–7599 (2018).

9. Bemis, D. K. & Pylkkänen, L. Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *J. Neurosci.* **31**, 2801–2814 (2011).
10. Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).
11. Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J. & Lalor, E. C. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Curr. Biol.* **28**, 803–809 (2018).
12. Brodbeck, C. & Simon, J. Z. Continuous speech processing. *Curr. Opin. Physiol.* **18**, 25–31 (2020).
13. Schuster, S., Hawelka, S., Hutzler, F., Kronbichler, M. & Richlan, F. Words in context: The effects of length, frequency, and predictability on brain responses during natural reading. *Cereb. Cortex* **26**, 3889–3904 (2016).
14. Wehbe, L. *et al.* Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one* **9**, e112575 (2014).
15. Haxby, J. V., Connolly, A. C. & Guntupalli, J. S. Decoding neural representational spaces using multivariate pattern analysis. *Annu. review neuroscience* **37**, 435–456 (2014).
16. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fmri. *Neuroimage* **56**, 400–410 (2011).
17. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Front. systems neuroscience* **2**, 4 (2008).
18. King, J.-R. & Dehaene, S. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends cognitive sciences* **18**, 203–210 (2014).
19. King, J.-R. *et al.* Encoding and decoding neuronal dynamics: Methodological framework to uncover the algorithms of cognition (2018).
20. King, J.-R., Charton, F., Lopez-Paz, D. & Oquab, M. Back-to-back regression: Disentangling the influence of correlated factors from multivariate observations. *NeuroImage* **220**, 117028 (2020).
21. Chehab, O., Defossez, A., Loiseau, J.-C., Gramfort, A. & King, J.-R. Deep recurrent encoder: A scalable end-to-end network to model brain signals. *arXiv preprint arXiv:2103.02339* (2021).
22. Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006–1010 (2014).
23. Qian, P., Qiu, X. & Huang, X. Bridging lstm architecture and the neural dynamics during reading. *arXiv preprint arXiv:1604.06635* (2016).
24. Jain, S. & Huth, A. Incorporating context into language encoding models for fmri. *Adv. neural information processing systems* **31** (2018).
25. Toneva, M. & Wehbe, L. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Adv. Neural Inf. Process. Syst.* **32** (2019).
26. Millet, J. & King, J.-R. Inductive biases, pretraining and fine-tuning jointly account for brain responses to speech. *arXiv preprint arXiv:2103.01032* (2021).
27. Caucheteux, C. & King, J.-R. Brains and algorithms partially converge in natural language processing. *Commun. Biology* **5**, 1–10 (2022).
28. Goldstein, A. *et al.* Shared computational principles for language processing in humans and deep language models. *Nat. neuroscience* **25**, 369–380 (2022).
29. Schrimpf, M. *et al.* The neural architecture of language: Integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci.* **118** (2021).
30. Caucheteux, C., Gramfort, A. & King, J.-R. Gpt-2's activations predict the degree of semantic comprehension in the human brain (2021).
31. Caucheteux, C., Gramfort, A. & King, J.-R. Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects. *arXiv preprint arXiv:2110.06078*, (2021).
32. Caucheteux, C., Gramfort, A. & King, J.-R. Disentangling syntax and semantics in the brain with deep networks. In *International Conference on Machine Learning*, 1336–1348 (PMLR, 2021).
33. Caucheteux, C., Gramfort, A. & King, J.-R. Long-range and hierarchical language predictions in brains and algorithms. *arXiv preprint arXiv:2111.14232* (2021).
34. Heilbron, M., Ehinger, B., Hagoort, P. & De Lange, F. P. Tracking naturalistic linguistic predictions with deep neural language models. *arXiv preprint arXiv:1909.04400*, (2019).
35. Gillis, M., Vanthornhout, J., Simon, J. Z., Francart, T. & Brodbeck, C. Neural markers of speech comprehension: measuring eeg tracking of linguistic speech representations, controlling the speech acoustics. *J. Neurosci.* **41**, 10316–10329 (2021).
36. Gwilliams, L., King, J. R., Marantz, A. & Poeppel, D. Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nature Communications* **13**(1), 1–14 (2022).
37. Nastase, S. A. *et al.* The "narratives" fmri dataset for evaluating models of naturalistic language comprehension. *Sci. data* **8**, 1–22 (2021).
38. Schoffelen, J. *et al.* Mother of unification studies, a 204-subject multimodal neuroimaging dataset to study language processing (2019).
39. Van Essen, D. C. *et al.* The WU-Minn human connectome project: an overview. *Neuroimage* **80**, 62–79 (2013).
40. Brennan, J. R. & Hale, J. T. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PloS one* **14**(1), e0207741 (2019).
41. Armeni, K. *et al.* A 10-hour within-participant magnetoencephalography narrative dataset to test models of language comprehension. *Sci Data* **9**, 278, https://doi.org/10.1038/s41597-022-01382-7 (2022).
42. Veale, J. F. Edinburgh handedness inventory–short form: a revised version based on confirmatory factor analysis. *Laterality: Asymmetries of Body, Brain and Cognition* **19**(2), 164–177 (2014).
43. Ide, N. & Macleod, C. The american national corpus: A standardized resource of american english. In Proceedings of corpus linguistics (Vol. 3, pp. 1–7). Lancaster, UK: Lancaster University Centre for Computer Corpus Research on Language (2001).
44. Fedorenko, E. *et al* Neural correlate of the construction of sentence meaning. Proceedings of the National Academy of Sciences, **113**(41), E6256-E6262. Chicago (2016).
45. Gorgolewski, K. J. *et al.* The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. data* **3**, 1–9 (2016).
46. Appelhoff, S. *et al.* Mne-bids: Organizing electrophysiological data into the bids format and facilitating their analysis. *The J. Open Source Software*. **4** (2019).
47. Gulban, OF. *et al.* poldracklab/pydeface: v2. 0.0, *Zenodo*, https://doi.org/10.5281/zenodo.3524401 (2019).
48. Gramfort, A. *et al.* Meg and eeg data analysis with mne-python. *Front. neuroscience* **267** (2013).
49. Brett, M. *et al.* nipy/nibabel: 3.2.1, *Zenodo*, https://doi.org/10.5281/zenodo.4295521 (2020).
50. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. machine Learn. research* **12**, 2825–2830 (2011).
51. W McKinney. Data Structures for Statistical Computing in Python. In van der Walt, S. & Millman, J. (eds.) *Proceedings of the 9th Python in Science Conference*, 56–61, https://doi.org/10.25080/Majora-92bf1922-00a (2010).
52. King, J.-R. & Gwilliams, L. MASC-MEG. *OSF* https://doi.org/10.17605/OSF.IO/AG3KJ (2022).

53. Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J. & Lounasmaa, O. V. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Rev. modern Phys.* **65**, 413 (1993).
54. de Cheveigné, A. & Nelken, I. Filters: when, why, and how (not) to use them. *Neuron* **102**, 280–293 (2019).
55. Jas, M., Engemann, D. A., Bekhti, Y., Raimondo, F. & Gramfort, A. Autoreject: Automated artifact rejection for meg and eeg data. *NeuroImage* **159**, 417–429 (2017).
56. Speer, R., Chin, J., Lin, A., Jewett, S. & Nathan, L. Luminosoinsight/wordfreq: v2.2, *Zenodo*, https://doi.org/10.5281/zenodo.1443582 (2018).
57. Gwilliams, L., Marantz, A., Poeppel, D. & King, J. R. Top-down information shapes lexical processing when listening to continuous speech. Language, Cognition and Neuroscience, 1–14 (2023).
58. pandas development team, T. pandas-dev/pandas: Pandas. *Zenodo* https://doi.org/10.5281/zenodo.3509134 (2020).

Acknowledgements

This work was funded in part by FrontCog grant ANR-17-EURE-0017 (JRK); Abu Dhabi Research Institute (G1001) (AM; LP); Dingwall Foundation (LG).

Author contributions

L.G. and J.-R.K. conceived the experiment. L.G. conducted the experiment. G.F. assisted with MEG and MRI data collection. L.G. and J-R.K. analyzed the results. A.M., L.P., D.P. and J.-R.K. contributed to the funding of the study. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.