HAL open science

Le Petit Prince multilingual naturalistic fMRI corpus

Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R. Nathan Spreng, Jonathan R Brennan, Yiming Yang, Christophe Pallier, John Hale

To cite this version:

Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, et al.. Le Petit Prince multilingual naturalistic fMRI corpus. Scientific Data , 2022, 9 (1), pp.530. 10.1038/s41597-022-01625-7 . hal-03849524

HAL Id: hal-03849524

https://hal.science/hal-03849524v1

Submitted on 11 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

scientific data

OPEN

DATA DESCRIPTOR

Le Petit Prince multilingual naturalistic fMRI corpus

Jixing Li[1,2,10]✉, Shohini Bhattasali[3,10], Shulin Zhang[4], Berta Franzluebbers[4], Wen-Ming Luh[5], R. Nathan Spreng[6], Jonathan R. Brennan[7], Yiming Yang[8], Christophe Pallier[9] & John Hale[4]✉

Neuroimaging using more ecologically valid stimuli such as audiobooks has advanced our understanding of natural language comprehension in the brain. However, prior naturalistic stimuli have typically been restricted to a single language, which limited generalizability beyond small typological domains. Here we present the *Le Petit Prince* fMRI Corpus (LPPC–fMRI), a multilingual resource for research in the cognitive neuroscience of speech and language during naturalistic listening (OpenNeuro: ds003643). 49 English speakers, 35 Chinese speakers and 28 French speakers listened to the same audiobook *The Little Prince* in their native language while multi-echo functional magnetic resonance imaging was acquired. We also provide time-aligned speech annotation and word-by-word predictors obtained using natural language processing tools. The resulting timeseries data are shown to be of high quality with good temporal signal-to-noise ratio and high inter-subject correlation. Data-driven functional analyses provide further evidence of data quality. This annotated, multilingual fMRI dataset facilitates future re-analysis that addresses cross-linguistic commonalities and differences in the neural substrate of language processing on multiple perceptual and linguistic levels.

Background & Summary

In the cognitive neuroscience of language, there is a growing consensus that using more ecologically valid stimuli such as audiobooks might extend our understanding of language processing in the brain[1–3]. Compared to traditional factorial designs with a large number of repetitive trials, naturalistic paradigms use stories and dialogues with a rich context and produce results that are generalizable to everyday language use[3,4]. However, prior naturalistic studies have typically been restricted to a single language, which limited neurobiological frameworks for language processing to small typological domains. Here we present *Le Petit Prince* fMRI Corpus (LPPC-fMRI)[5], a multilingual fMRI dataset where English, Chinese and French speakers listened to the same audiobook *Le Petit Prince (The Little Prince)* in their native language (see Fig. 1 for a Schematic overview of the LPPC-fMRI data collection, preprocessing, technical validation and annotation procedures). Our parallel corpus facilitates future research on cross-linguistic commonalities and differences in the neural processes for language comprehension.

In naturalistic designs such as story listening, linguistic processes on multiple levels (e.g., word, phrase, sentence, discourse) unfold naturally at different timescales. Such a rich contextual setting extends the range of linguistic phenomena that can be examined in parallel, and allows for testing assumptions on the neural mechanisms of language processing. For example, whether different linguistic levels coincide with different frequencies of oscillatory activity in the brain[6,7], and whether these levels correspond to a hierarchically organized predictive coding architecture[8]. In addition, naturalistic approaches to neurolinguistics are in synergy with natural language processing (NLP), where using ecologically valid language corpora for training models has been common practice for the past quarter-century. Accordingly, NLP models can be leveraged to understand linguistic

[1]New York University Abu Dhabi, Neuroscience of Language Lab, Abu Dhabi, UAE. [2]Department of Linguistics and Translation, City University of Hong Kong, Hong Kong, Hong Kong. [3]University of Maryland, Department of Linguistics & Institute of Advanced Computer Studies, College Park, MD, 20742, USA. [4]University of Georgia, Department of Linguistics, Athens, GA, 30602, USA. [5]National Institute on Aging, Baltimore, MD, 21225, USA. [6]Laboratory of Brain and Cognition, Montreal Neurological Institute, Department of Neurology and Neurosurgery, Faculty of Medicine, McGill University, Montreal, QC, H3A 2B4, Canada. [7]Department of Linguistics, University of Michigan, Ann Arbor, MI48109, USA. [8]Jiangsu Key Laboratory of Language and Cognitive Neuroscience, Jiangsu Normal University, Xuzhou, 221116, China. [9]Cognitive Neuroimaging Unit, INSERM, CEA, CNRS, Universit Paris-Saclay, NeuroSpin center, Gif-sur-Yvette, 91191, France. [10]These authors contributed equally: Jixing Li, Shohini Bhattasali. ✉e-mail: jixingli@cityu.edu.hk; jthale@uga.edu

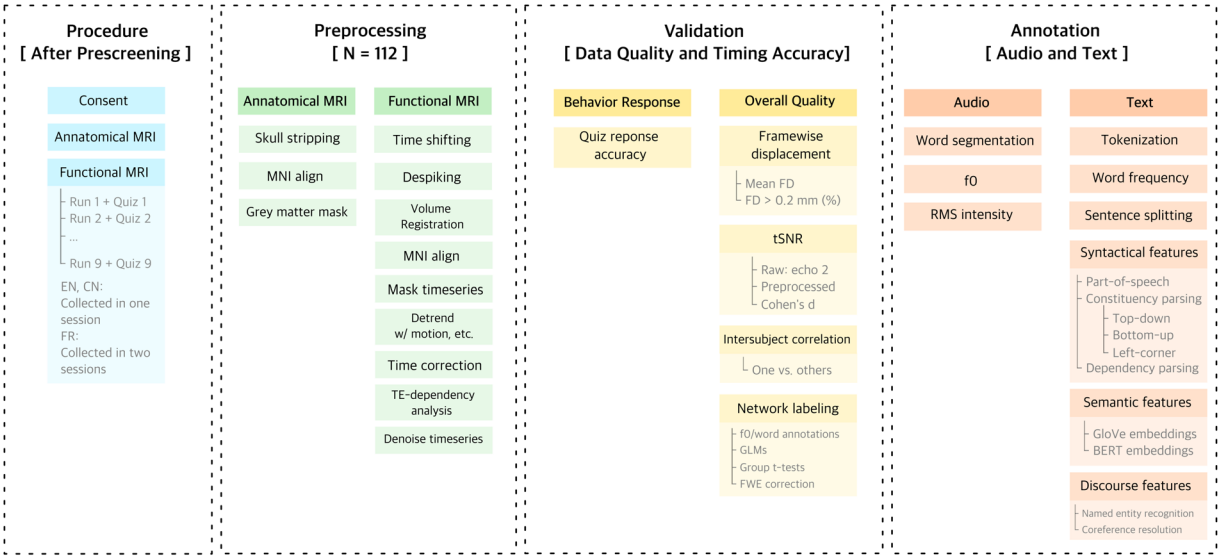


Fig. 1 Schematic overview of the LPPC-fMRI data collection procedures, preprocessing, technical validation and annotation. During data collection (blue), anatomical MRI was first acquired, followed by functional MRI while participants listened to 9 sections of the audiobook. After preprocessing the data (green), behavioral and overall data quality were examined (yellow). Audio and text annotations were extracted using NLP tools.

processes at an algorithmic level by comparing model predictions against brain data during naturalistic comprehension. For example, syntactic structure-building as predicted by the bottom-up or left-corner parsing strategies[9–11] and recurrent neural network grammars (RNNG)[12] has been shown to fit well with left temporal activity. Recent neural network architectures such as bidirectional LSTMs[13] and Transformers[14,15] have also been shown to correlate with neural responses during naturalistic comprehension, suggesting construction-specific variations in the understanding of linguistic expressions.

While naturalistic designs opened up a host of new research questions that are not possible to study under tightly controlled experimental designs, the majority of prior naturalistic studies have been restricted to a single language. This limited our understanding of the neural processes of language comprehension to small typological domains. To complement monolingual datasets such as the Narrative Brain Dataset (NBD)[16], the Alice Dataset[17], the Narratives dataset[18] and the Mother of Unification Studies[19], we collected a multilingual fMRI dataset consisted of Antoine de Saint-Exupéry's *The Little Prince* in English, Chinese and French. A total of 112 subjects (49 English speakers, 35 Chinese speakers and 28 French speakers) listened to the whole audiobook for about 100 minutes in the scanner (see Tables 2 and 4 for the demographics of the participants, data collection procedures, and stimuli information for the English, Chinese, and French datasets).

This stimulus is considerably longer than other datasets (i.e., 6 minutes on average for the NBD dataset and 12 minutes for the Alice dataset), allowing for testing linguistic phenomena that may not be sufficiently attested in smaller samples. This dataset includes time-aligned speech segmentation, prosodic information and word-by-word predictors obtained using natural language processing tools, ranging from lexical semantics to syntax to discourse information (see Fig. 2 for the annotations available for an example sentence from the English audiobook). The neuroimaging data, as well as the annotations and information about the experimental procedure are shared in a standardized BIDS format on OpenNeuro[5].

The LPPC-fMRI facilitates cross-linguistic generalization and helps overcome current statistical and typological limitations in the neurobiology of language. We stress the importance of considering multiple languages when building and testing neurobiological models of language processing, assuming that the neural substrates and processes of language are shared among speakers of all languages. As shown in previous work examining coreference resolution using the English and Chinese subset of this corpus, the computational model that best explains the neural signature for pronoun processing is generalizable for both English and Chinese[20]. These data can be reused to address different research questions with a variety of analytical methods. Future work envisions an expanded LPPC, one that incorporates data from additional neuroimaging modalities, such as electrocorticography (EEG) and magnetoencephalography (MEG). For instance, LPPC-EEG dataset aspires to 26 languages[4]. Our vision is for the LPPC to become an open infrastructure to which researchers from various communities can contribute by adding further modalities, languages and annotations.

Methods

Participants. A total of 112 subjects listened to the whole audiobook for about 100 minutes in the scanner. Tables 2 and 4 show the summary of the data collection procedure, the stimuli and participants information for the three datasets.

English participants were 49 young adults (30 females, mean age = 21.3, SD = 3.6) with no history of psychiatric, neurological or other medical illness that might compromise cognitive functions. (A subset of prior work

Original file	Renamed file
sub-EN084_task-lppEN_run-09_echo-1_bold.nii.gz	sub-EN084_task-lppEN_run-01_echo-1_bold.nii.gz
sub-EN084_task-lppEN_run-09_echo-2_bold.nii.gz	sub-EN084_task-lppEN_run-01_echo-2_bold.nii.gz
sub-EN084_task-lppEN_run-09_echo-3_bold.nii.gz	sub-EN084_task-lppEN_run-01_echo-3_bold.nii.gz
sub-EN084_task-lppEN_run-10_echo-1_bold.nii.gz	sub-EN084_task-lppEN_run-02_echo-1_bold.nii.gz
sub-EN084_task-lppEN_run-10_echo-2_bold.nii.gz	sub-EN084_task-lppEN_run-02_echo-2_bold.nii.gz
sub-EN084_task-lppEN_run-10_echo-3_bold.nii.gz	sub-EN084_task-lppEN_run-02_echo-3_bold.nii.gz
sub-EN084_task-lppEN_run-13_echo-1_bold.nii.gz	sub-EN084_task-lppEN_run-03_echo-1_bold.nii.gz
sub-EN084_task-lppEN_run-13_echo-2_bold.nii.gz	sub-EN084_task-lppEN_run-03_echo-2_bold.nii.gz
sub-EN084_task-lppEN_run-13_echo-3_bold.nii.gz	sub-EN084_task-lppEN_run-03_echo-3_bold.nii.gz
sub-EN084_task-lppEN_run-14_echo-1_bold.nii.gz	sub-EN084_task-lppEN_run-04_echo-1_bold.nii.gz
sub-EN084_task-lppEN_run-14_echo-2_bold.nii.gz	sub-EN084_task-lppEN_run-04_echo-2_bold.nii.gz
sub-EN084_task-lppEN_run-14_echo-3_bold.nii.gz	sub-EN084_task-lppEN_run-04_echo-3_bold.nii.gz
sub-EN084_task-lppEN_run-15_echo-1_bold.nii.gz	sub-EN084_task-lppEN_run-05_echo-1_bold.nii.gz
sub-EN084_task-lppEN_run-15_echo-2_bold.nii.gz	sub-EN084_task-lppEN_run-05_echo-2_bold.nii.gz
sub-EN084_task-lppEN_run-15_echo-3_bold.nii.gz	sub-EN084_task-lppEN_run-05_echo-3_bold.nii.gz
sub-EN084_task-lppEN_run-16_echo-1_bold.nii.gz	sub-EN084_task-lppEN_run-06_echo-1_bold.nii.gz
sub-EN084_task-lppEN_run-16_echo-2_bold.nii.gz	sub-EN084_task-lppEN_run-06_echo-2_bold.nii.gz
sub-EN084_task-lppEN_run-16_echo-3_bold.nii.gz	sub-EN084_task-lppEN_run-06_echo-3_bold.nii.gz
sub-EN084_task-lppEN_run-17_echo-1_bold.nii.gz	sub-EN084_task-lppEN_run-07_echo-1_bold.nii.gz
sub-EN084_task-lppEN_run-17_echo-2_bold.nii.gz	sub-EN084_task-lppEN_run-07_echo-2_bold.nii.gz
sub-EN084_task-lppEN_run-17_echo-3_bold.nii.gz	sub-EN084_task-lppEN_run-07_echo-3_bold.nii.gz
sub-EN084_task-lppEN_run-18_echo-1_bold.nii.gz	sub-EN084_task-lppEN_run-08_echo-1_bold.nii.gz
sub-EN084_task-lppEN_run-18_echo-2_bold.nii.gz	sub-EN084_task-lppEN_run-08_echo-2_bold.nii.gz
sub-EN084_task-lppEN_run-18_echo-3_bold.nii.gz	sub-EN084_task-lppEN_run-08_echo-3_bold.nii.gz
sub-EN084_task-lppEN_run-19_echo-1_bold.nii.gz	sub-EN084_task-lppEN_run-09_echo-1_bold.nii.gz
sub-EN084_task-lppEN_run-19_echo-2_bold.nii.gz	sub-EN084_task-lppEN_run-09_echo-2_bold.nii.gz
sub-EN084_task-lppEN_run-19_echo-3_bold.nii.gz	sub-EN084_task-lppEN_run-09_echo-3_bold.nii.gz

Table 1. Example of renaming convention using symbolic links to keep run numbers consistent across participants.

	Participants			Data Collection	Stimuli			
Language	Number	Mean Age	Female	Location	Material	Length (s)	N Words	N Sentences
English	49	21.3	30	Cornell University, United States	The little prince EN audiobook	5632	15376	1499
Chinese	35	19.9	15	Jiangsu Normal University, China	The little prince CN audiobook	5954	16009	1577
French	28	24.4	15	NeuroSpin, France	The little prince FR audiobook	5828	15391	1480

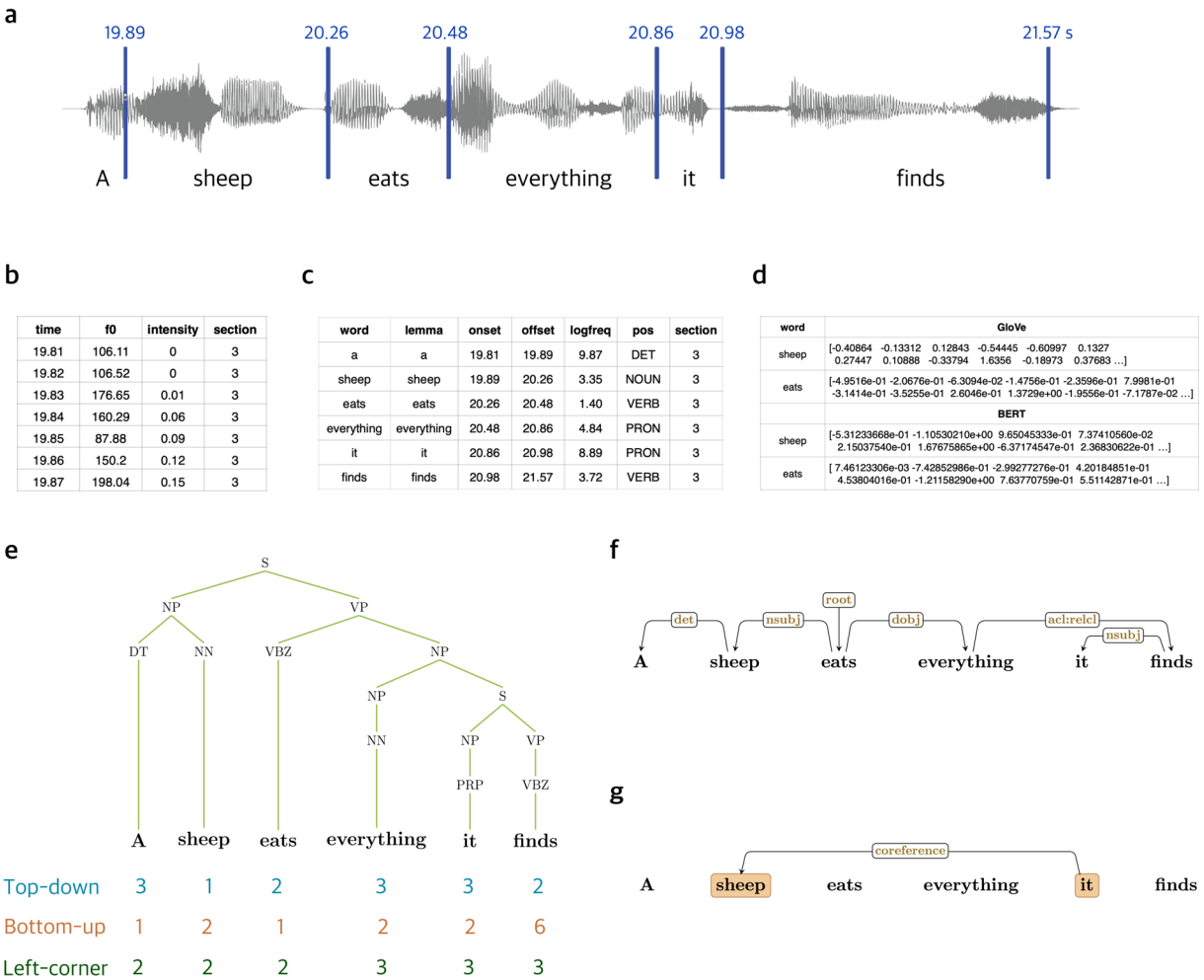
Table 2. Demographics of the participants, data collection procedures, and stimuli information for the English, Chinese, and French datasets.

using the LPP English fMRI dataset used 51 participants' data[21–23]. Due to concerns about head movement, only 49 participants' data is released in this corpus.) They self-identified as native English speakers, and strictly qualified as right-handed on the Edinburgh handedness inventory[24]. All participants were paid, and gave written informed consent prior to participation, in accordance with the IRB guidelines of Cornell University.

Chinese participants were 35 healthy, right-handed young adults (15 females, mean age = 19.3, SD = 1.6). They self-identified as native Chinese speakers, and had no history of psychiatric, neurological, or other medical illness that could compromise cognitive functions. All participants were paid, and gave written informed consent prior to participation, in accordance with the IRB guidelines of Jiangsu Normal University.

French participants were 28 healthy, right-handed adults (15 females, mean age = 24.4, SD = 4.6). They self-identified as native French speakers and had no history of psychiatric, neurological, or other medical illness that could compromise cognitive functions. All participants gave written informed consent prior to participation, in accordance with the Regional Committee for the Protection of Persons involved in Biomedical Research.

Procedures. After giving their informed consent, participants were familiarized with the MRI facility and assumed a supine position on the scanner. They were instructed to not move as best as they could throughout scanning as movement would make the scans unusable. Next, participants were put in the head-coil with pillows under and on the sides of their head and under the knees for comfort and to reduce movement over the scanning



time	f0	intensity	section
19.81	106.11	0	3
19.82	106.52	0	3
19.83	176.65	0.01	3
19.84	160.29	0.06	3
19.85	87.88	0.09	3
19.86	150.2	0.12	3
19.87	198.04	0.15	3

word	lemma	onset	offset	logfreq	pos	section
a	a	19.81	19.89	9.87	DET	3
sheep	sheep	19.89	20.26	3.35	NOUN	3
eats	eats	20.26	20.48	1.40	VERB	3
everything	everything	20.48	20.86	4.84	PRON	3
it	it	20.86	20.98	8.89	PRON	3
finds	finds	20.98	21.57	3.72	VERB	3

word	GloVe
sheep	[-0.40864 -0.13312 0.12843 -0.54445 -0.60997 0.1327 0.27447 0.10888 -0.33794 1.6356 -0.18973 0.37683 ...]
eats	[-4.9516e-01 -2.0676e-01 -6.3094e-02 -1.4756e-01 -2.3596e-01 7.9981e-01 -3.1414e-01 -3.5255e-01 2.6046e-01 1.3729e+00 -1.9556e-01 -7.1787e-02 ...]
	BERT
sheep	[-5.31233668e-01 -1.10530210e+00 9.65045333e-01 7.37410560e-02 2.15037540e-01 1.67675865e+00 -6.37174547e-01 2.36830622e-01 ...]
eats	[7.46123306e-03 -7.42852986e-01 -2.99277276e-01 4.20184851e-01 4.53804016e-01 -1.21158290e+00 7.63770759e-01 5.51142871e-01 ...]

Fig. 2 Annotation information for the stimuli. (**a**) Word boundaries in the audio files, included in files: `lpp<EN/CN/FR>_section[1-9].TextGrid`. (**b**) f0 and RMS intensity for every 10 ms of the audios, included in files: `lpp<EN/CN/FR>_prosody.csv` (**c**) Tokenization, lemmatization, log-tranformed word frequency and POS tagging, included in files: `lpp<EN/CN/FR>_word_information.csv`. (**d**) GloVe and BERT embeddings for every word in the audiobooks, included in files: `lpp<EN/CN/FR>_word_embeddings_GloVe.csv` and `lpp<EN/CN/FR>_word_embeddings_BERT.csv` (**e**) Parsed syntactic trees based on constituency grammar with node counts using top-down, bottom-up, and left-corner parsing strategies[31], included in files: `lpp<EN/CN/FR>_trees.csv`. (**f**) Dependency relations for each words in each sentence, included in files: lpp < EN/CN/FR > _dependency.csv. (**g**) Named entity recognition and coreference relations for the English and Chinese texts, included in files: `lpp<EN/CN>_coreference.csv`.

session. Participants were given a bulb in their right hand and told to squeeze if something was wrong or they needed a break during scanning. Once in place, participants chose an optimal stimulus volume by determining a level that was loud but comfortable. Auditory stimuli were delivered through MRI-safe, high-fidelity headphones inside the head coil (English: Confon HP-VS01, MR Confon, Magdeburg, Germany; Chinese: Ear Bud Headset, Resonance Technology, Inc, California, USA; French: Magnacoil TIM headset, Siemens, Germany). The headphones were secured against the plastic frame of the coil using foam blocks.

The English and Chinese participants went through one scanning session, which was divided into 9 runs, and each lasted for about 10 minutes. Participants listened passively to 1 section of the audiobook in each run and completed 4 quiz questions after each run (36 questions in total). These questions were used to confirm their comprehension and were viewed by the participants via a mirror attached to the head coil and they answered through a button box. During scanning, participants were monitored by a camera over their left eye. If they appeared drowsy or seemed to move too much during the movie, the operator of the scanner gave them a warning over the intercom by producing a beep or speaking to them. During breaks between the runs, participants were told that they could relax but not move. Finally, participants were paid and sent home. The entire session lasted for around 2.5 hours. In French, due to a legal limitation, participants could not stay for longer than 1.5 hours inside the scanner; therefore, the acquisition was split into two sessions separated by a period of 1 to 2 hours out of the scanner.

			Anatomical/Structural Scans			Functional Scans									
Language	Scanner	Head coil	Pulse sequence	in-plane resolution	slice thickness	Pulse sequence	TRs	TEs	Flip angle	Matrix size	FoV	Image acceleration	N axial slices	in-plane resolution	slice thickness
English	3 T MRI GE Discovery MR750	32 channel	T1W MPRAGE	1.0 mm × 1.0 mm	1.0 mm	ME-EPI	2000 ms	2.8, 27.5, 43 ms	77	72 × 72	240.0 mm × 240.0 mm	2x	33	3.75 mm × 3.75 mm	3.8 mm
Chinese	3 T MRI GE Discovery MR750	32 channel	T1W MPRAGE	1.0 mm × 1.0 mm	1.0 mm	ME-EPI	2000 ms	2.8, 27.5, 43 ms	77	72 × 72	240.0 mm × 240.0 mm	2x	33	3.75 mm × 3.75 mm	3.8 mm
French	3 T Siemens Magnetom Prisma Fit 230	64 channel	T1W MPRAGE	1.0 mm × 1.0 mm	1.0 mm	ME-EPI	2000 ms	10, 25, 38 ms	77	72 × 72	240.0 mm × 240.0 mm	2x	34	3.75 mm × 3.75 mm	3.8 mm

Table 3. Scanner parameters for structural and functional scans across English, Chinese, and French datasets.

Stimuli. The English *The Little Prince* audiobook is 94 minutes long, translated by David Wilkinson and read by Karen Savage. The Chinese audiobook http://www.xiaowangzi.org/ is 99 minutes long, read by a professional female Chinese broadcaster hired by the experimenter. The French audiobook is 97 minutes long, read by Nadine Eckert-Boulet and published by the now-defunct Omilia Languages Ltd. The original French text is copyrighted by Gallimard 1946.

One of the central themes in the story is the difference between adults and children, especially the lack of imagination in the former. The narrator uses the visual cues of different drawings to emphasize this message and these drawings are present in the original text. In the English and Chinese study, to help the participants understand this point, these visual cues were incorporated during the audio presentation for the first chapter and are included in the OpenNeuro repository. In order to control for the visual stimuli and its associated neural activation, "picture events" conditions and "picture blocks" conditions are also included in the analysis to account for the visual stimuli presented to participants and its associated neural activation. The "picture events" occur at the 10 s, 35 s, and 60 s timepoints in the first section of the story while the "picture blocks" also occur at the 10 s, 35 s, and 60 s timepoints in the first section and last for 15 s, 20 s, and 15 s respectively. These conditions match the presentation and duration of the visual stimuli and are aligned with particular plot points in the story.

Acquisition. Data acquisition parameters are listed in Table 3 for ease of comparison across English, Chinese, and French. The scanner parameters were the same for English and Chinese with some differences for French. There was a trigger at the beginning of each section and a delay of 8 s (4 TRs) between the trigger and onset of stimulus presentation for all three languages.

Preprocessing. MRI data files were converted from DICOM to NIfTI format and preprocessed using AFNI version 16[25].

Anatomical. The anatomical/structural MRI scans were deskulled using *3dSkullStrip*. The resulting anatomical images were nonlinearly aligned to the Montreal Neurological Institute (MNI) N27 template brain. Resulting anatomical images were used to create grey matter masks.

Functional. The first 4 volumes in each run were excluded from analyses to allow for T1-equilibration effects. The fMRI timeseries were then corrected for slice-timing differences (*3dTshift*) and despiked (*3dDespike*). Next, volume registration was done by aligning each timepoint to the mean functional image of the centre timeseries (*3dvolreg*). Then the volume-registered and anatomically-aligned functional data were nonlinearly aligned to the MNI template brain. Multi-echo independent components analysis (ME-ICA)[26] were used to denoise data for motion, physiology and scanner artifacts. Images were then resampled at 2 mm cubic voxels (*3dresample*).

Annotations. Apart from the fMRI timeseries data, we also provide audio and text annotations ranging from time-aligned speech segmentation and prosodic information to word-by-word predictors obtained using natural language processing tools, including lexical semantics, syntax and discourse-level information. See Fig. 2 for a summary of our annotations. These annotations are available on OpenNeuro too (see the Data records section).

Speech segmentation. Word boundaries in the audio were identified and aligned to the transcripts using Forced Alignment and Vowel Extraction (FAVE) (https://www.research.ed.ac.uk/portal/en/publications/fave-forced-alignment-and-vowel-extraction-suite-version-113(bbc2046d-6768-47c5-b574-2987895b0307).html) and were manually checked by two native speakers each of the three languages.

Prosodic information. Root mean square intensity and the fundamental frequency (f0) for every 10 ms of each audio section of the three languages were extracted using the Voicebox toolbox (http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html).

English			Chinese			French		
Participant ID	Age	Sex	Participant ID	Age	Sex	Participant ID	Age	Sex
sub-EN057	20	F	sub-CN001	18	F	sub-FR001	40	M
sub-EN058	22	M	sub-CN002	18	F	sub-FR002	23	M
sub-EN059	21	F	sub-CN003	22	F	sub-FR003	26	F
sub-EN061	25	F	sub-CN004	18	M	sub-FR004	20	M
sub-EN062	23	M	sub-CN005	18	F	sub-FR005	23	F
sub-EN063	22	M	sub-CN006	19	F	sub-FR006	30	M
sub-EN064	19	M	sub-CN007	20	F	sub-FR007	20	M
sub-EN065	21	F	sub-CN008	21	F	sub-FR008	23	M
sub-EN067	21	F	sub-CN009	20	M	sub-FR009	18	F
sub-EN068	19	M	sub-CN010	22	M	sub-FR010	28	F
sub-EN069	21	F	sub-CN011	20	M	sub-FR011	26	F
sub-EN070	20	F	sub-CN013	20	F	sub-FR012	28	F
sub-EN072	18	F	sub-CN014	19	M	sub-FR013	23	F
sub-EN073	19	F	sub-CN015	19	F	sub-FR014	20	F
sub-EN074	18	F	sub-CN016	18	F	sub-FR015	23	F
sub-EN075	18	M	sub-CN017	22	M	sub-FR016	22	M
sub-EN076	20	M	sub-CN018	21	M	sub-FR017	24	M
sub-EN077	22	M	sub-CN019	20	M	sub-FR018	23	F
sub-EN078	19	F	sub-CN020	21	M	sub-FR019	25	F
sub-EN079	21	F	sub-CN021	19	F	sub-FR020	25	F
sub-EN081	22	F	sub-CN022	20	F	sub-FR022	20	F
sub-EN082	28	F	sub-CN023	20	F	sub-FR023	19	M
sub-EN083	20	F	sub-CN024	19	F	sub-FR024	20	M
sub-EN084	28	F	sub-CN025	18	M	sub-FR025	22	M
sub-EN086	19	M	sub-CN026	20	M	sub-FR026	32	F
sub-EN087	22	M	sub-CN027	18	M	sub-FR028	22	M
sub-EN088	21	M	sub-CN028	24	M	sub-FR029	30	F
sub-EN089	33	M	sub-CN029	19	M	sub-FR030	27	M
sub-EN091	20	M	sub-CN030	19	M			
sub-EN092	21	M	sub-CN031	21	M			
sub-EN093	20	F	sub-CN032	21	M			
sub-EN094	21	F	sub-CN033	22	M			
sub-EN095	20	F	sub-CN034	18	F			
sub-EN096	18	F	sub-CN036	22	M			
sub-EN097	21	F	sub-CN037	22	M			
sub-EN098	24	F						
sub-EN099	37	F						
sub-EN100	19	F						
sub-EN101	23	M						
sub-EN103	18	F						
sub-EN104	19	F						
sub-EN105	19	F						
sub-EN106	20	M						
sub-EN108	18	M						
sub-EN109	19	M						
sub-EN110	21	F						
sub-EN113	21	F						
sub-EN114	20	M						
sub-EN115	23	F						

Table 4. List of subjects in the data collection with basic demographic information.

Word frequency. Log-transformed unigram frequency of each word in *The Little Prince* in English, Chinese and French was estimated using Google Books Ngram Viewer, Version 20120701 (http://storage.googleapis.com/books/ngrams/books/datasetsv2.html).

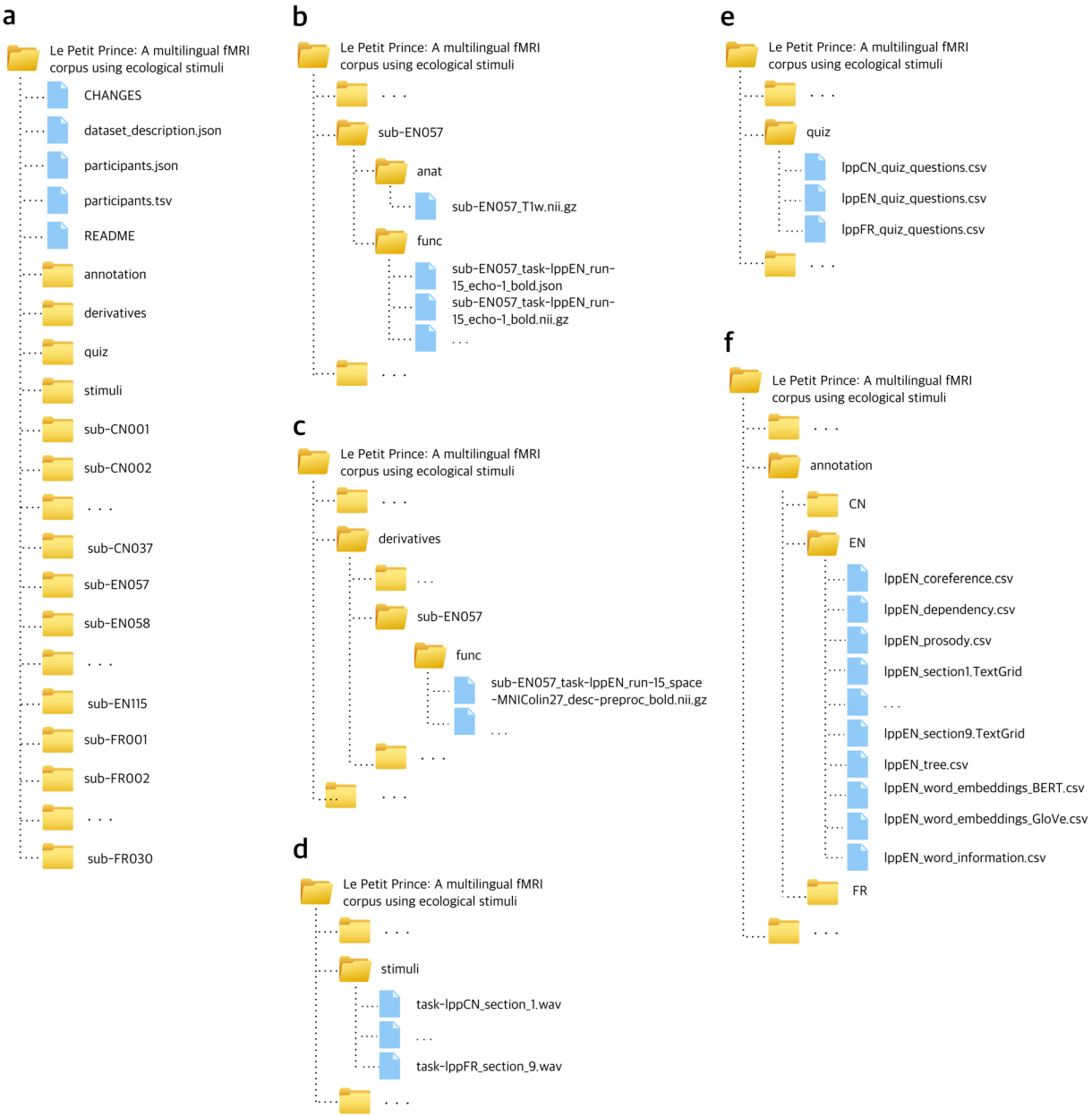


Fig. 3 Organization of the data collection. (**a**) General overview of directory structure. (**b**) Content of subject-specific anatomical and raw data directories. (**c**) Content of subject-specific preprocessed data directories. (**d**) Content of the stimuli directory. (**e**) Content of the quiz directory. (**f**) Content of the language-specific annotation directory.

Word embeddings. Static GloVe embeddings[27] and contextualized BERT embeddings for each word (given its sentential context) in the *The Little Prince* in the three languages were extracted using the SpaCy package (https://spacy.io/). Words that are divided into subwords by BERT used the average embedding of the subwords.

Part-of-speech tagging. Part-of-speech (POS) tagging for each word in the book in the three languages was extracted using the Stanford parser for English[28], Chinese[29] and French[30].

Constituency parsing. Syntactic tree structures of each sentence in the audiobooks was parsed using the Stanford parser for English[28], Chinese[29] and French[30].

Parser actions. Syntactic node counts for each word in the audiobooks based on bottom-up, top-down and left-corner parsing strategies[31] as applied to the Stanford-derived constituency trees described above. These word-by-word counts are the number of parser actions that would be taken (on a given strategy) before moving on to the next word in the sentence. They were calculated using custom tree-walking software.

	FD (mm)		FD > 0.2 mm (%)	
	Mean	SD	Mean	SD
English	0.11	0.05	9.3	10.6
Chinese	0.08	0.05	5.0	8.2
French	0.10	0.02	4.6	5.0

Table 5. Summary of framewise displacement information for the English, Chinese and French data.

Dependency parsing. Dependency relations of words in each sentence of the audiobooks were parsed using the Stanford dependency parser for English[32], Chinese[33] and French[30].

Coreference resolution. Antecedents for each third person pronoun in the English and Chinese audiobooks were manually annotated using the annotation tool brat[34].

Data Records

Information and anatomical data that could be used to identify participants has been removed from all records. Resulting files are available from the OpenNeuro repository at https://doi.org/10.18112/openneuro.ds003643.v2.0.0[5]. See Fig. 3 for the organization of the data collection. A README file there provides a description of the available content. The scripts used for this manuscript are available on the repository and GitHub (https://github.com/jixing-li/lpp_data).

Participant responses. **Location** `participants.json, participants.tsv.`

File format tab-separated value.

Participants' sex, age and responses to quiz questions in tab-separated value (tsv) files. Data is structured as one line per participant.

Audio files. **Location** `stimuli/task-lpp<EN/CN/FR>_section_0[1-9].wav`

File format wav.

The English, Chinese and French audiobooks divided into nine sections.

Anatomical MRI. **Location** `sub-<EN/CN/FR><ID>/anat/sub-<EN/CN/FR><ID>_T1w.nii.gz`

File format NIfTI, gzip-compressed.

The defaced raw high-resolution anatomical image.

Functional MRI. **Location** `sub-<EN/CN/FR><ID>/func/sub-<EN/CN/FR><ID>_task-lpp<EN/CN/FR>_run-0[1-9]_echo-[1-3]_bold.nii.gz.`

File format NIfTI, gzip-compressed.

Sequence protocol `sub-<EN/CN/FR><ID>/func/sub-<EN/CN/FR><ID>_task-lpp<EN/CN/FR>_run-0[1-9]_echo-[1-9]_bold.json.`

The mutli-echo fMRI data are available as individual timeseries files, stored as:

`sub-<EN/CN/FR><ID>/func/sub-<EN/CN/FR><ID>_task-lpp<EN/CN/FR>_run-0[1-9]_echo-[1-3]_bold.nii.gz.`

The MEI-CA preprocessed timeseries are also available as:

`derivatives/sub<EN/CN/FR><ID>/func/sub-<EN/CN/FR><ID>_task-lpp<EN/CN/FR>_run-0[1-9]_space-MNIColin27_desc-preproc_bold.nii.gz.`

Annotations. **Location** `annotation/<EN/CN/FR>/lpp<EN/CN/FR>_section[1-9].TextGrid,`

File format TextGrid (requires Praat software; http://www.praat.org/).

Location `annotation/<EN/CN/FR>/lpp<EN/CN/FR>_prosody.csv,`

`annotation/<EN/CN/FR>/lpp<EN/CN/FR>_word_information.csv, annotation/<EN/CN/FR>/lpp<EN/CN/FR>_word_embeddings_GloVe.csv, annotation/<EN/CN/FR>/lpp<EN/CN/FR>_word_embeddings_BERT.csv, annotation/<EN/CN/FR>/lpp<EN/CN/FR>_tree.csv, annotation/<EN/CN/FR>/lpp<EN/CN/FR>_dependency.csv,annotation/<CN/EN>/lpp<CN/EN>_coreference.csv.`

File format comma-separated value.

Speech and linguistic annotations for the audio and text of the three languages.

Quiz questions. **Location** `quiz/lpp<EN/CN/FR>_quiz_questions.csv.`

File format comma-separated value.

The 36 comprehension quiz questions used in the English, Chinese and French experiments.

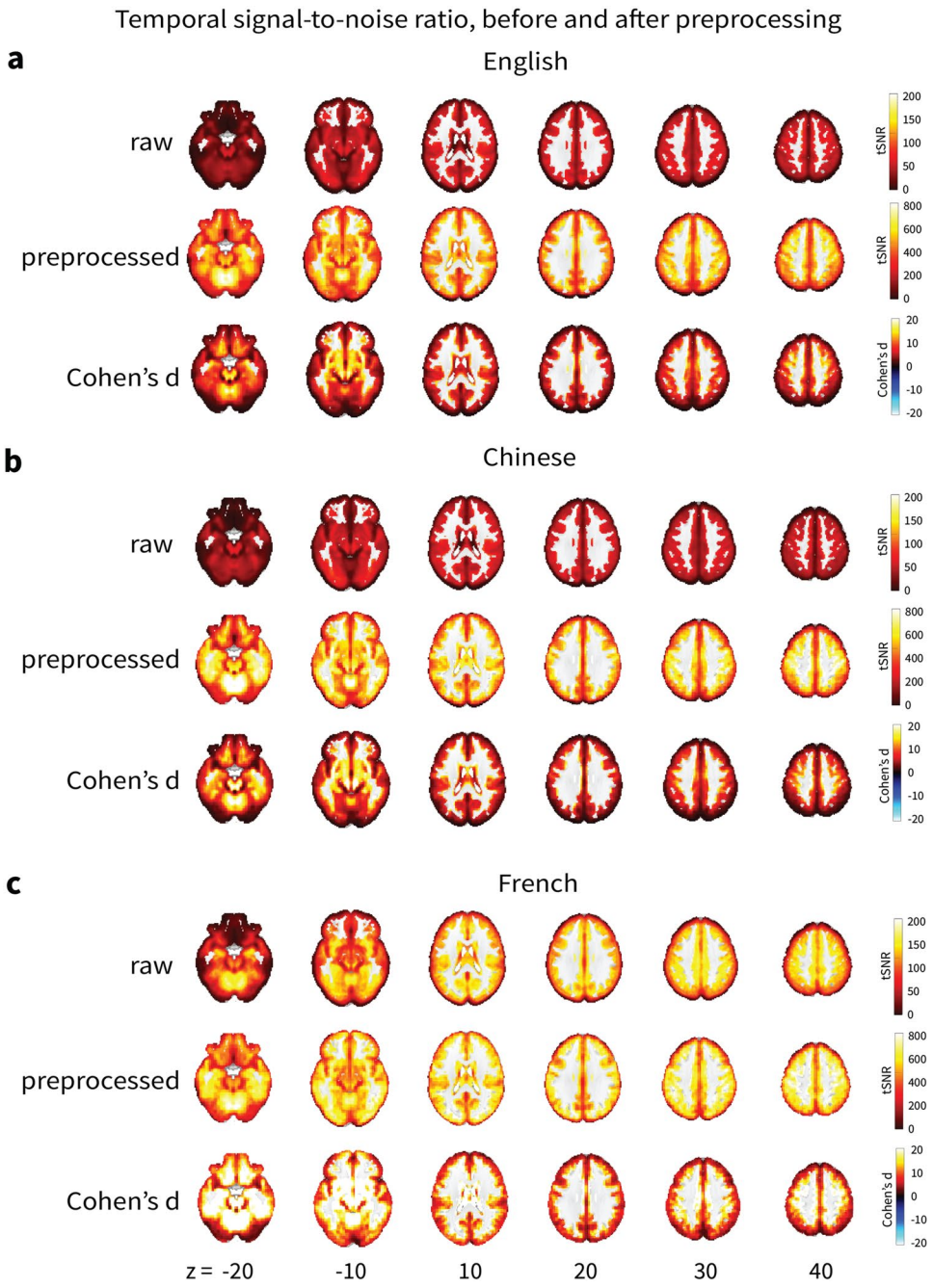


Fig. 4 Voxel-wise temporal signal-to-noise ratio analysis before and after preprocessing. Cohen's d effect sizes showed increase in tSNR after preprocessing.

Technical Validation

Accuracy of participants' responses to the quizzes after each section was calculated to ensure adequate comprehension. To assess fMRI scan quality, we calculated framewise displacement (FD), temporal signal-to-noise ratio (tSNR) and inter-subject correlation (ISC). We also did two whole-brain functional analyses using pitch (f0) and word annotations. These serve to show data quality similar to past work and provide evidence for timing accuracy between fMRI timeseries for participants.

Behavioral results. Participants answered four four-choice comprehension questions after each section (36 questions in total). An example question is shown below. Participants performed well with a mean accuracy of 89.5% (SD = 3.8) and 86.4% (SD = 2.7) for English and Chinese participants, respectively. French participants' responses were noted on paper by the experimenters during recording and were unfortunately unable to locate now. But the experimenters did not notice any French participant with an abnormally low accuracy (<75%) for the quiz questions.

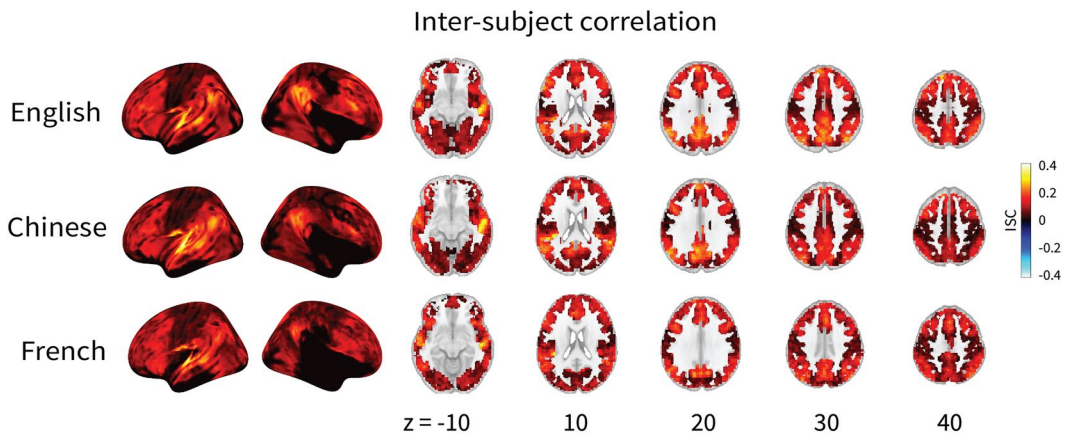


Fig. 5 Results of inter-subject correlation (ISC) demonstrating data quality and timing synchrony between participants. As expected, the temporal regions showed the largest correlation in brain responses across subjects.

Why was the little prince difficult to talk to?
(a) He spoke a foreign language.
(b) He was mute.
(c) He didn't ask enough questions.
(d) He didn't answer questions directly.
Key: (d)

Framewise displacement. Framewise displacement is a measure of the frame-to-frame movement, assessed in millimetres. The six motion parameters (3 translation parameters and 3 rotation parameters) generated by MEI-CA.py were used to calculate FD, defined as the sum of the absolute temporal derivatives of the six motion parameters, following conversion of rotational parameters to distances by computing the arc length displacement on the surface of a sphere with radius 50 mm[35,36]:

$$FD(t) = \sum |d(t-1) - d(t)| + 50 \cdot (\pi/180) \cdot \sum |r(t-1) - r(t)|$$

where d denotes translation distances x, y, z, and r denotes rotation angles α, β, γ. For each participant, a single (scalar) estimate of overall motion, the mean FD, can be calculated by averaging the FD time series.

For the English data, the average FD was 0.11 mm (SD = 0.05); for the Chinese data, the average FD was 0.08 mm (SD = 0.05), and for the French data, the average FD was 0.10 mm (SD = 0.02). FD values greater than 0.20 mm are conventionally considered high motion[36], we therefore also calculated the percentage of frames for each subject where FD exceeded 0.20 mm. The average percentage of frames where FD was greater than 0.20 mm was 9.3% (SD = 10.6%), 5.0% (SD = 8.2%) and 4.6% (SD = 5.0%) for the English, Chinese and French data, respectively (see Table 5).

Temporal signal-to-noise ratio. tSNR is a measure of signal strength at the voxel level, defined as the mean signal intensity of a voxel across the timeseries divided by its standard deviation. We calculated tSNR both before preprocessing using the middle echo image which most closely approximates standard single echo collection, and after the optimal combination of the echo images with MEI-CA denoising. We compared the tSNR values before and after extensive preprocessing using Cohen's d:

$$Cohen's\ d = \frac{M_1 - M_2}{\sqrt{\frac{(SD_1^2 + SD_2^2)}{2}}}$$

where M and SD are the mean and standard deviation of the tSNR in a voxel for the more (subscript one) minus the less preprocessed timeseries (subscript two). We applied a grey matter mask with most white matter and ventricle voxels removed. The tSNR values showed a clear increase after MEI-CA denoising across the three language groups, suggesting clearer signal compared to standard single echo acquisition (see Fig. 4).

Inter-subject correlation. To estimate what proportion of the brain signal in response to the audiobook was consistent across subjects, we computed the inter-subject correlation (ISC) for each voxel's timeseries across subjects in each language group. Each subject's data in a voxel was correlated to the average timeseries of the other subjects in the same voxel. This generated a map that quantifies the similarity of an individual subject's response with the group response. The procedure was repeated for all subjects, and a median ISC map was computed at the group level. The ISC results showed largest correlation in brain responses across subjects in the temporal regions, the brain regions implicated for speech and language processing (see Fig. 5).

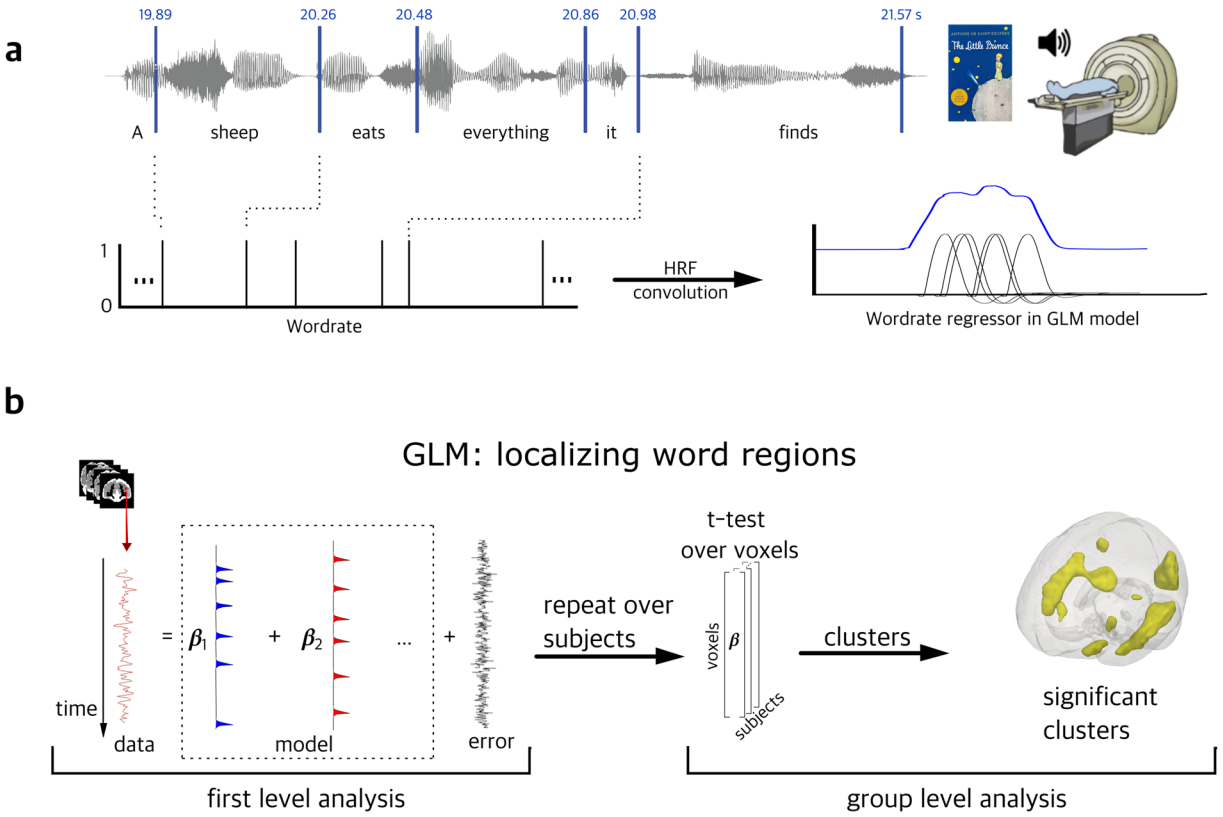


Fig. 6 GLM analyses to localize the *wordrate* regressor. (**a**) Offest of each word in the audiobook was marked 1 and was convolved with the canonical hemodynamic response function. (**b**) The timecourse of each voxel's BOLD signals was modeled using our designmatrix at the first level At the group level, a one-sample t-test was performed on the distribution of the beta values for the *wordrate* regressor across subjects at each voxel for the fMRI data. Statistical significance was held at $p < 0.05$ FWE with a cluster size greater than 50.

Network labeling. Besides demonstrating data and timing quality, here we also illustrate the general linear model (GLM) methods to derive the prosody and word regions using our pitch and word annotations. In particular, we calculated the *f0* for every 10 ms of the audio in each language and marked 1 at the offset of each word in the audio (*wordrate*). We then convolved the *f0* and *wordrate* annotations with a canonical hemodynamic response function and regressed them against the preprocessed fMRI timecourses using GLMs. At the group level, the contrast images for the *f0* and *wordrate* regressors were examined by a one-sample *t*-test. An 8 mm full-width at half-maximum (FWHM) Gaussian smoothing kernel was applied on the contrast images from the first-level analysis to counteract inter-subject anatomical variation. Statistical significance was held at $p < 0.05$ FWE with a cluster size greater than 50. Figure 6 illustrates the GLM methods to localize the pitch and word regions.

To illustrate the precise anatomical correspondence of our results with prior data, we overlaid fMRI term-based meta-analysis from Neurosynth[37] (Retrieved September 2021) for the "pitch" area (https://neurosynth.org/analyses/terms/pitch/; from 102 studies) and the "words" area (https://neurosynth.org/analyses/terms/words/; from 944 studies). Our results are highly consistent with prior literature (see Fig. 7). MNI coordinates of the significant clusters and their statistics are shown in Table 6.

Usage Notes

The LPPC-fMRI can advance our understanding of speech and language processing in the human brain during naturalistic listening. However, there are several limitations and usage bottlenecks, including annotations and analyses that we now discuss to help others use the LPPC-fMRI to make new discoveries.

Annotation bottleneck. Most of the linguistic annotations were done automatically using existing NLP tools, which may contain errors and affect downstream annotations. For example, syntactic node counts for each word in the audiobooks based on bottom-up, top-down and left-corner parsing strategies were applied to the Stanford-derived constituency trees, and the accuracy of the tree structures will affect the number of node counts.

Analysis bottleneck. Although GLM or encoding models have been commonly applied to fMRI data using long naturalistic stimuli like audiobooks[9,10,12,23,38–40], there are no standardised approaches for analysing complex and high dimensional naturalistic fMRI data. Machine learning approaches are becoming an increasingly

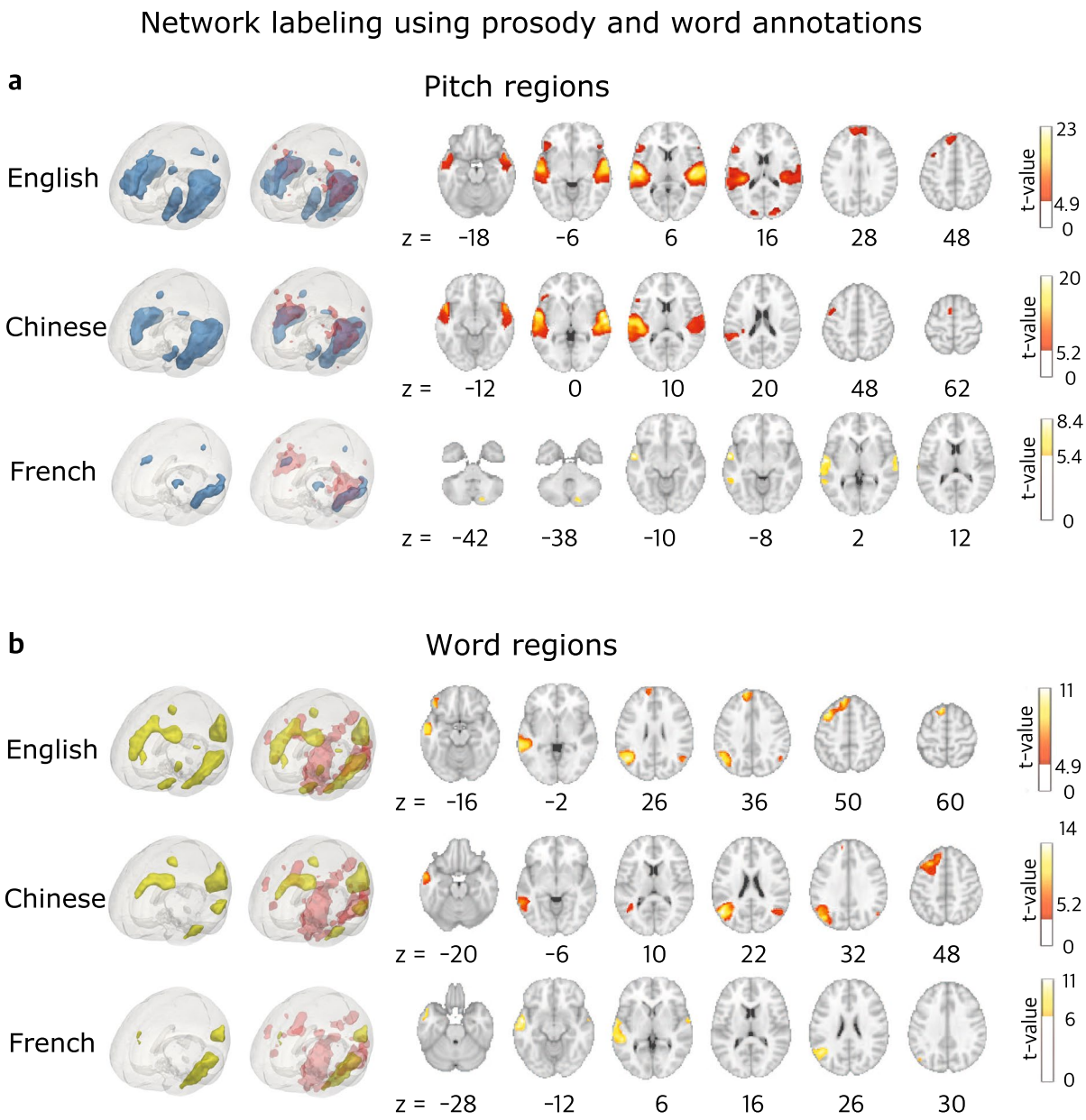


Fig. 7 GLM results showing the significant clusters for (**a**) the pitch and (**b**) word regions in the English, Chinese and French data using *f0* and *wordrate* annotations. Red areas in the second column of the 3D brains shows meta-analyses of pitch and word regions from Neurosynth[37]. Statistical significance was thresholded at $p < 0.05$ FWE and $k > 50$.

common way to analyze fMRI data, and we encourage the development of innovative analysis approaches by running machine learning competitions on the LPPC-fMRI corpus.

Cross-linguistic analyses. This multilingual fMRI data is a novel cognitive neuroscience resource since it enables cross-linguistic research. However, there are two points we would like to highlight. Firstly, for each language the dataset was acquired at different sites and we look at interaction effects between sites, not main effects (as seen in Fig. 7). Therefore, any specific baseline effects of acquisition would be controlled for (except for potential differences in SNR). Secondly, a group-level analysis, pooling together the data across the three languages would be infeasible. Although English, Chinese, and French follow the same underlying word order (SVO), given the structural, lexical, and prosodic differences between them, it would not be possible to align the same words along a temporal pattern cross-linguistically. However, within each language it is possible to investigate the same research question and compare the neural correlates cross-linguistically, as it has been done for semantic number[22] and antecedent tracking[41].

Miscellaneous. The file name patterns reported in the Data Records are meant to be a template. In the actual dataset, some of the runs for a single participant have non-consecutive numbering due to scanning issues or

Condition	Language	Cluster	MNI Coordinates	k-size	*t*-value	*p*-value
Prosody	Chinese	RSTG	62, −14, 0	3566	20.22	<0.001
		L Heschl's Gyrus	−56, −6, 4	5036	19.97	<0.001
		L Frontal Lobe	−4, 0, 62	73	7.40	<0.001
		LMFG	−52, −2, 48	64	6.15	0.0005
	English	L Heschl's Gyrus	−50, −18, 6	5330	22.98	<0.001
		RSTG	58, −20, 4	5053	22.64	<0.001
		LIFG	−52, 26, 10	864	10.09	<0.001
		LSFG	−8, 58, 26	1272	9.55	<0.001
		LMFG	−34, 12, 42	145	6.88	<0.001
		RIFG	52, 26, −8	153	6.88	<0.001
	French	LSTG	−62, −12, 4	1349	8.92	<0.001
		RSTG	68, −22, 2	218	7.09	<0.001
		L Precuneus	−4, −70, 32	53	6.63	<0.001
		LMFG	−42, 18, 28	150	6.14	<0.001
Word	Chinese	LAG	−50, −64, 22	2040	14.45	<0.001
		LMFG	−28, 22, 48	1194	10.39	<0.001
		LMTG	−56, 0, −20	358	9.65	<0.001
		LMTG	−60, −46, −6	511	8.61	<0.001
		RAG	54, −64, 26	289	8.03	<0.001
	English	LMTG	−52, −4, −28	1683	11.10	<0.001
		LAG	−48, −60, 26	1561	10.54	<0.001
		LMFG	−36, 16, 50	1770	10.09	<0.001
		LIFG	−46, 32, −12	288	9.10	<0.001
		RMTG	60, −4, −30	171	6.92	<0.001
		LIFG	−52, 26, 8	86	6.57	<0.001
		RAG	52, −64, 28	191	6.55	<0.001
	French	LSTG	−54, −4, −12	1674	9.45	<0.001
		LAG	−50, −60, 24	516	8.88	<0.001
		RSTG	62, −2, 2	72	7.01	<0.001

Table 6. GLM results for the *f0* and *wordrate* regressors for the Chinese, English and French fMRI data: MNI coordinates, cluster size and their peak level statistics, thresholded at $p < 0.05$ FWE and $k > 50$.

participants needing a break. As a workaround, we created symbolic links for each of the participants' runs by using the Unix `in` command. As an example, Table 1 illustrates how the runs were renamed for subject 84 in the LPP English dataset to be consistent with the runs[1–9] pattern specified and execute our scripts across all participants.

Code availability

The code for LPP-fMRI corpus is publicly available at the OpenNeuro repository under `code/` subdirectory, and also at the following GitHub repositories: https://github.com/jixing-li/lpp_data, https://github.com/chrplr/lpp-paradigm.

The code includes the presentation scripts for all three languages, the scripts used in technical validation and for preparing this data paper (e.g., `compute_tsnr.py`), in addition to code for obtaining annotations (e.g. `count_parser_actions.py`). Code for certain annotations like word embeddings and POS tagging is not included since there are several publicly available toolkits available to researchers.

Received: 7 October 2021; Accepted: 10 August 2022;
Published online: 29 August 2022

References

1. Alday, P. M. M/EEG analysis of naturalistic stories: A review from speech to language processing. *Language. Cognition and Neuroscience* **34**, 457–473 (2019).
2. Brennan, J. Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass* **10**, 299–313 (2016).
3. Kandylaki, K. D. & Bornkessel-Schlesewsky, I. From story comprehension to the neurobiology of language. *Language. Cognition and Neuroscience* **34**, 405–410 (2019).
4. Stehwien, S., Henke, L., Hale, J., Brennan, J. & Meyer, L. The Little Prince in 26 languages: Towards a multilingual neuro-cognitive corpus. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, 43–49 (European Language Resources Association, 2020).
5. Li, J., Hale, J. & Pallier, C. Le Petit Prince: A multilingual fMRI corpus using ecological stimuli. *OpenNeuro* https://doi.org/10.18112/openneuro.ds003643.v2.0.0 (2022).
6. Ding, N., Melloni, L., Zhang, H., Tian, X. & Poeppel, D. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience* **19**, 158–164 (2016).

7. Giraud, A. L. & Poeppel, D. Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience* **15**, 511–517 (2012).
8. Bornkessel-Schlesewsky, I. & Schlesewsky, M. Towards a neurobiologically plausible model of language-related, negative event-related potentials. *Frontiers in Psychology* **10** (2019).
9. Brennan, J. *et al*. Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language* **120**, 163–173 (2012).
10. Brennan, J., Stabler, E., Van Wagenen, S., Luh, W. & Hale, J. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language* **157-158**, 81–94 (2016).
11. Nelson, M. J. *et al*. Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E3669–E3678 (2017).
12. Hale, J., Dyer, C., Kuncoro, A. & Brennan, J. R. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2727–2736 (Association for Computational Linguistics, Melbourne, Australia, 2018).
13. Reddy, A. J. & Wehbe, L. Can fMRI reveal the representation of syntactic structure in the brain? *Advances in Neural Information Processing Systems* **34**, 9843–9856 (2021).
14. Schrimpf, M. *et al*. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences* **118**, e2105646118 (2021).
15. 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).
16. Lopopolo, A., Frank, S. L., van den Bosch, A., Nijhof, A. & Willems, R. M. The Narrative Brain Dataset (NBD), an fMRI dataset for the study of natural language processing in the brain. *Linguistic and Neuro-Cognitive Resources (LiNCR)* 8 (2018).
17. Bhattasali, S., Jonathan, B., Luh, W.-M., Franzluebbers, B. & Hale, J. The Alice Datasets: fMRI & EEG observations of natural language comprehension. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 120–125 (European Language Resources Association, 2020).
18. Nastase, S. A. *et al*. The Narratives fMRI dataset for evaluating models of naturalistic language comprehension. *Scientific data* **8**, 1–22 (2021).
19. Schoffelen, J.-M. *et al*. A 204-subject multimodal neuroimaging dataset to study language processing. *Scientific Data* **6**, 17 (2019).
20. Li, J., Fabre, M., Luh, W.-M. & Hale, J. Modeling brain activity associated with pronoun resolution in English and Chinese. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, 87–96 (2018).
21. Stanojević, M. *et al*. Modeling incremental language comprehension in the brain with Combinatory Categorial Grammar. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 23–38 (2021).
22. Dunagan, D. *et al*. Neural correlates of semantic number: A cross-linguistic investigation. *Brain and Language* **229**, 105110 (2022).
23. Bhattasali, S. *et al*. Localising memory retrieval and syntactic composition: an fMRI study of naturalistic language comprehension. *Language, Cognition and Neuroscience* 2327–3801 (2019).
24. Oldfield, R. C. The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia* **9**, 97–113 (1971).
25. Cox, R. W. AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, an International Journal* **29**, 162–173 (1996).
26. Kundu, P., Inati, S. J., Evans, J. W., Luh, W.-M. & Bandettini, P. A. Differentiating BOLD and non-BOLD signals in fMRI time series using multi-echo EPI. *NeuroImage* **60**, 1759–1770 (2012).
27. Pennington, J., Socher, R. & Manning, C. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543 (Association for Computational Linguistics, Doha, Qatar, 2014).
28. Klein, D. & Manning, C. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the association for computational linguistics.*, 423–430 (2003).
29. Levy, R. & Manning, C. D. Is it harder to parse Chinese, or the Chinese treebank? In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, 439–446 (Association for Computational Linguistics, Sapporo, Japan, 2003).
30. Green, S., de Marneffe, M.-C., Bauer, J. & Manning, C. D. Multiword expression identification with tree substitution grammars: A parsing tour de force with French. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 725–735 (Association for Computational Linguistics, Edinburgh, Scotland, UK., 2011).
31. Hale, J. *Automaton theories of human sentence comprehension* (CSLI Publications, 2014).
32. de Marneffe, M., MacCartney, B. & Manning, C. Generating typed dependency parses from phrase structure parses. In *LREC 2006* (2006).
33. Chang, P.-C., Tseng, H., Jurafsky, D. & Manning, C. D. Discriminative reordering with Chinese grammatical relations features. In *Proceedings of the third workshop on syntax and structure in statistical translation* (2009).
34. Stenetorp, P. *et al*. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 102–107 (Association for Computational Linguistics, 2012).
35. Patel, A. X. *et al*. A wavelet method for modeling and despiking motion artifacts from resting-state fMRI time series. *NeuroImage* **95**, 287–304 (2014).
36. Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L. & Petersen, S. E. Spurious but systematic correlations in functional connectivity mri networks arise from subject motion. *NeuroImage* **59**, 2142–2154 (2012).
37. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods* **8**, 665–670 (2011).
38. 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).
39. Li, J. *et al*. *Cortical processing of reference in language revealed by computational models* BioRxiv 2020.11.24.396598: https://doi.org/10.1101/2020.11.24.396598 (2021).
40. Wehbe, L. *et al*. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one* **9**, e112575 (2014).
41. Li, S. *et al*. Analogical reasoning on Chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (*Volume 2*: *Short Papers)*, 138–143 (2018).

Acknowledgements

This material is based upon work supported by the National Science Foundation under grant numbers 1903783 and 1607251, the French Agence Nationale pour la Recherche under grant NCM-NL ANR 16-NEUC-0005-02, and the Jeffrey Sean Lehman Fund for Scholarly Exchange with China at Cornell University. J.L. is supported by the NYU Abu Dhabi Institute under Grant G1001. S.B. is supported by the ONR MURI Award N000141812670. W.-M.L. is supported in part by the Intramural Research Program of the NIA/NIH.

Author contributions

J.H. and S.B. designed the study. J.H., S.B. and J.L. collected and preprocessed the English data with the help of N.S. and W.-M.L. J.L. collected and preprocessed the Chinese data with the help of Y.Y. C.P. collected and preprocessed the French data. W.-M.L. provided data acquisition and preprocessing methods. J.L. and S.Z. did the technical validation of the data. B.F. and S.B. prepared the OpenNeuro archive. J.L. and S.B. wrote the manuscript with the help of J.H.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.L. or J.H.

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.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022