Imaging Neuroscience
an open access journal

Research Article

Semantic composition in experimental and naturalistic paradigms

Jixing Li[a], Marco Lai[b], Liina Pylkkänen[b]

[a]Department of Linguistics and Translation, City University of Hong Kong, Kowloon Tong, Hong Kong
[b]Department of Linguistics and Department of Psychology, New York University, New York, USA

Corresponding Author: Jixing Li (jixingli@cityu.edu.hk)

ABSTRACT

Naturalistic paradigms using movies or audiobooks have become increasingly popular in cognitive neuroscience, but connecting them to findings from controlled experiments remains rare. Here, we aim to bridge this gap in the context of semantic composition in language processing, which is typically examined using a "minimal" two-word paradigm. Using magnetoencephalography (MEG), we investigated whether the neural signatures of semantic composition observed in an auditory two-word paradigm can extend to naturalistic story listening, and vice versa. Our results demonstrate consistent differentiation between phrases and single nouns in the left anterior and middle temporal lobe, regardless of the context. Notably, this distinction emerged later during naturalistic listening. Yet this latency difference disappeared when accounting for various factors in the naturalistic data, such as prosody, word rate, word frequency, surprisal, and emotional content. These findings suggest the presence of a unified compositional process underlying both isolated and connected speech comprehension.

Keywords: semantic composition, two-word paradigm, naturalistic paradigm, classification, MEG

1. INTRODUCTION

Naturalistic paradigms utilizing movies or audiobooks have gained considerable popularity in the field of cognitive neuroscience. Within the realm of language studies, these approaches provide valuable insights into language processing in real-world contexts, allowing for the examination of a broader range of linguistic phenomena (Alday, 2019; Brennan, 2016; Kandylaki & Bornkessel-Schlesewsky, 2019). During the comprehension of narratives, linguistic processes unfold naturally across multiple levels, including words, phrases, sentences, and discourse, each operating on distinct timescales. Computational models have often been employed to isolate these sub-processes and target specific linguistic levels (Brennan & Pylkkanen, 2012; Brennan et al., 2016; Goldstein et al., 2022; Huth et al., 2016; Schrimpf et al., 2021; Wehbe et al., 2021). For instance, relevant neural signals for semantic features of words were identified using word embedding models (Huth et al., 2016). Nevertheless, it is important to recognize that narrative comprehension involves a multitude of processes beyond the domain of language, including attention, emotion, social-cognitive functions, and memory encoding and retrieval (Hasson & Egidi, 2015). Consequently, it is possible to misattribute regions involved in these non-linguistic processes as core language regions.

Controlled experiments, on the other hand, are designed to isolate relevant cognitive processes by comparing conditions that differ solely in the component of interest. Early neurolinguistic experiments typically compared sentences with simple and complex syntactic structures, such as center-embedded and object-relative clauses (Stromswold et al., 1996), garden-path sentences (Bever, 1970), or implausible sentence completions (Kutas & Hillyard, 1980). This work was later complemented by research on basic meaning composition using

Received: 25 July 2023 Revision: 18 November 2023 Accepted: 22 December 2023 Available Online: 4 January 2024

a "minimal" two-word paradigm, where compositional phrases such as "red boat" were contrasted with single nouns such as "xkq boat" (Bemis & Pylkkanen, 2011, 2013a, b, c; reviewed in Pylkkänen, 2019). The underlying rationale behind these experimental manipulations is based on the concept of subtraction, although this approach has faced criticism as the brain is unlikely to behave like a linear system (Friston et al., 1996). Moreover, the experimental stimuli often diverge from everyday language use (Brennan, 2016). Thus, while controlled experiments have been widely embraced in neurolinguistics, their applicability to language processing in real-world contexts remains uncertain.

To compare the findings from experimental and naturalistic paradigms, we conducted a study incorporating both designs, with a specific focus on meaning composition, a fundamental function underlying human language's expressive capacity. The left anterior temporal lobe (LATL) has been consistently implicated in the effects of semantic composition, as demonstrated in studies using a two-word design (e.g., Blanco-Elorrieta et al., 2018; Li & Pylkkänen, 2021; Westerlund & Pylkkänen, 2014, reviewed in Pylkkänen, 2019). However, the generalizability of these findings to naturalistic settings has received limited exploration. Here, we trained feed-forward neural network (FFNN) classifiers to differentiate between MEG source estimates for adjective-noun phrases and single nouns, in both the two-word (e.g., "green glass" vs. "glass") and naturalistic settings (e.g., "…soft music…" vs. "…a bath…"). To examine the generalizability of the classifiers, we tested the classifiers trained in the experimental setting on the naturalistic data using the temporal generalization method (TGM; King & Dehaene, 2014), and vice versa. Note that we included both school-age children and adults in our sample to test whether language proficiency and the development of cognitive functions such as social and emotional processing may affect language comprehension in experimental and naturalistic contexts. We chose children in the school-age range of 7-15 years because they can more readily follow the experimenter's instructions and complete the tasks. Furthermore, research has shown that children within this age bracket exhibit cognitive profiles that are distinct from adults on a range of neuropsychological tests (Ardila & Rosselli, 1994). However, given the minimal differences observed between the behavioral and neural data of children and adults, and considering the relatively small number of child participants due to the pandemic, we merged the data from both demographics into a single group for analysis. Our results revealed that the left anterior and middle temporal lobe consistently differentiated between phrases and single nouns in both the experimental and naturalistic contexts, aligning with previous findings concerning semantic composition (see Pylkkänen, 2019 for a review).

The combinatory effect occurred much later in the naturalistic setting, which may be attributed to additional processing demands imposed by other information present in the naturalistic data, such as prosody, word rate, word frequency, surprisal of incoming words, and emotional content. To examine this possibility, we conducted further analyses by regressing out these effects and re-evaluating the classification results. The revised analyses revealed an earlier composition effect in the naturalistic setting, closely resembling the pattern observed in the two-word setting. These findings provide compelling evidence for a unified compositional process underlying both the experimental and naturalistic contexts, once the confounding effects are accounted for.

2. MATERIALS AND METHODS

2.1. Experimental design

The MEG experiment consists of a two-word session and a naturalistic listening session and was presented within a larger protocol that also included production tasks. Fitting multiple tasks into a single recording session manageable for children was a major design constraint. While most of the prior comprehension literature has used reading, the current study was auditory, as we wanted the paradigm to be suitable even for children who cannot read yet. Reading and listening were contrasted in Bemis and Pylkkänen (2013b) who did observe an LATL sensitivity to a composition effect for both reading and listening.

In the two-word session, participants listened to both adjective-noun phrases (e.g., "green glass") and single nouns that were preceded by a non-lexical "mmm" sound, chosen for naturalness in a speech context ("mmm glass"). After the auditory stimulus, participants selected a matching picture from a set of eight pictures. This task differed from the prior minimal composition studies which have only used one matching or mismatching task picture (Bemis & Pylkkanen, 2011). The reason for our larger set of pictures was that this decreased the chance of an accurate response by chance, making the behavioral data more informative if the task were to be used in, say, a clinical setting. There were six unique color words ("red, pink, blue, green, black, white") and six unique nouns ("glass, comb, door, sword, heart, house"), and they were

randomly combined to form adjective-noun phrases. Each participant received a unique randomisation. A total of 50 phrases and 50 single nouns were presented. Some adjective-noun combinations were presented more than once, and each noun was repeated eight to nine times.

2.2. Participants

Participants were 20 healthy adults (15 females, *M* = 27.8 years, SD = 13.2) and 11 school-age children (6 females, *M* = 9.4 years, SD = 2.3) with normal hearing and normal or corrected-to-normal vision. We included children in our sample to test whether language proficiency and the development of cognitive functions such as social and emotional functions may affect language processing in natural and unnatural contexts. The sample size of children is relatively small due to the pandemic. The aggregate sample size of 31 for both groups aligns with the norm for MEG studies of similar scope (e.g., Bemis & Pylkkanen, 2011; Blanco-Elorrieta et al., 2018; Flick et al., 2018; Law & Pylkkänen, 2021; Li & Pylkkänen, 2021; Zhang & Pylkkänen, 2015). We also performed a power analysis to determine whether our dataset of 31 participants was adequate to detect a medium-sized effect (Cohen's d = 0.6, as referenced by Cohen, 1988) when contrasting adjective-noun phrases with single noun MEG data. Our results suggest a power of 0.9, which exceeds the conventionally acceptable minimum power of 0.8. We excluded data from two children who did not complete the entire naturalistic listening task from the naturalistic dataset; their data were retained in the two-word dataset analysis. Consequently, the two-word dataset comprises 31 participants (21 females, *M* = 21.3 years, SD = 13.9), whereas the naturalistic dataset includes data from 29 participants (21 females, *M* = 22.1 years, SD = 13.9). All of the participants were strictly qualified as right-handed on the Edinburgh handedness inventory (Oldfield, 1971). They self-identified as native English speakers and gave their written informed consent prior to participation, in accordance with New York University.

2.3. Experiment procedures

Before recording, each subject's head shape was digitized using a Polhemus dual source handheld FastSCAN laser scanner. Participants then completed the experiment while lying supine in a dimly lit, magnetically shielded room (MSR). MEG data were collected using a whole-head 156-channel axial gradiometer system (Kanazawa Institute of Technology, Kanazawa, Japan). The two words were presented for 875 ms each, and an image with eight objects appeared on screen 600 ms after the second word. Subjects then selected the correct object that matched the auditory stimuli. No feedback was provided. The inter-stimulus interval was normally distributed with a mean of 300 ms (SD = 100 ms). Order of stimulus presentation was randomized, and each participant received a unique randomisation. After the two-word session, participants completed a naturalistic listening session where they passively listened to an audio excerpt consisted of four stories from the YouTube channel "SciShow Kids." The two-word session lasted around 20 minutes, and the naturalistic listening session lasted about 12 minutes. After the MEG recording, participants completed four picture-matching questions on the contents of the stories (See Fig. 1A for the experiment procedure).

2.4. MEG data acquisition and pre-processing

MEG data were recorded continuously at a sampling rate of 1000 Hz with an online 0.1 to 200 Hz band-pass filter. The raw data were first noise reduced via the Continuously Adjusted Least-Squares Method (Adachi et al., 2001) and low-pass filtered at 40 Hz. Independent component analysis (ICA) was then applied to remove artifacts such as eye blinks, heartbeats, movements, and well-characterized external noise sources. MEG data from the two-word task were segmented into epochs spanning 100 ms pre-stimulus onset to 1750 ms post-stimulus onset. MEG data from the naturalistic task were segmented into epochs from the onset to 875 ms after the target word. The target words include words at the boundary of single nouns and adjective-noun phrases in the naturalistic stimuli. Single nouns and adjective-noun phrases were annotated based on the Stanford part-of-speech tagger (Toutanova et al., 2003).

Epochs containing amplitudes greater than an absolute threshold of 2000 fT were automatically removed. Additional artifact rejection was performed through manual inspection of the data, removing trials that were contaminated with movement artifacts or extraneous noise. The whole epoch rejection procedure results in an average rejection rate of 7.6% (SD = 5.1%) for the adult participants and an average rejection rate of 11.1% (SD = 5%) for the child participants.

We then computed the cortically constrained minimum-norm estimates (Hämäläinen & Ilmoniemi, 1994) for each epoch for each participant. To perform source localization, the location of the participant's head was first coregistered

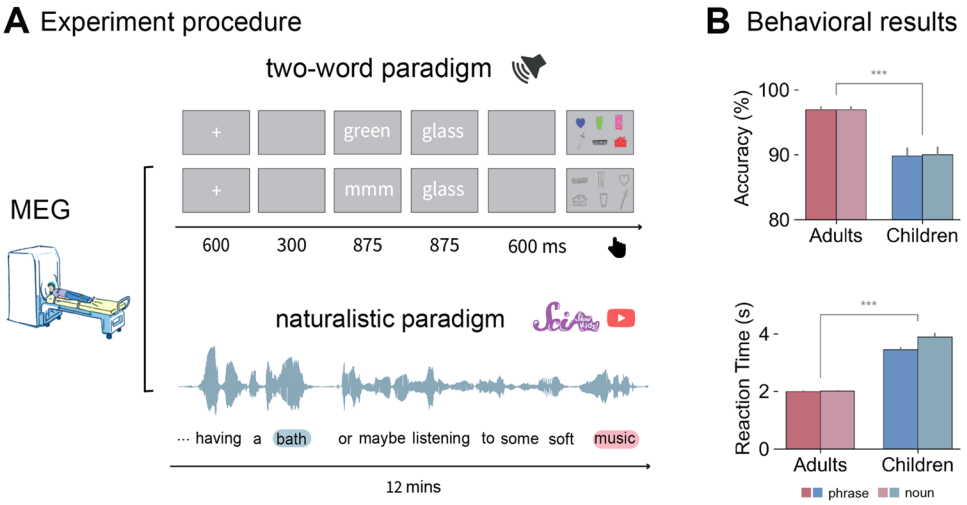


Fig. 1. Experimental design and behavioral results. (A) Experimental design and trial structure. In the two-word session, participants selected a picture from eight pictures that matched the preceding words in each trial. Half of the target pictures matched and half did not. Activities recorded from the onset of the second word to 875 ms after the second word were analyzed. In the naturalistic listening session, participants passively listened to a 12-minute audio excerpt from the YouTube channel "SciShow Kids." Participants completed a picture-matching task after the listening session to ensure comprehension. (B) Behavioral results on the two-word task. Mean predicted accuracy and reaction time for the phrase and noun conditions across the adults and children group. A two-way mixed ANOVA revealed significant differences between the groups in both accuracy ($p = 0.002$) and reaction time ($p < 0.001$). Composition was not significant for either accuracy or reaction time for either group. Error bars indicate 1 standard error. *** indicates $p < 0.001$.

with respect to the sensor array in the MEG helmet. We used the FreeSurfer (http://surfer.nmr.mgh.harvard.edu/) "fsaverage" brain with rotation and translation and then scaling the average brain to match the size of the head scan. A source space of 2562 source points per hemisphere was generated on the cortical surface for each participant. The Boundary Element Method (BEM) was employed to compute a forward solution, explaining the contribution of activity at each source to the magnetic flux at the sensors. We applied the BEM protocols as per MNE-Python's default configurations, following previous publications (e.g., Bemis & Pylkkanen, 2011; Flick & Pylkkänen, 2020; Law & Pylkkänen, 2021; Li & Pylkkänen, 2021). Specifically, we used the head surface triangulation computed by the watershed algorithm (Ségonne et al., 2004), which yielded the inner, outer skull triangulations and the head surface triangulation. We then set up the boundary-element model with a conductivity value of 0.3 S/m for the scalp and the brain compartments, and 0.006 S/m for the skull. These values were the default set in MNE and were corroborated by prior literature (Goncalves et al., 2003; Lew et al., 2009; Oostendorp et al., 2000). We next aligned the head and the sensors in a common coordinate system by identifying the fiducial landmark locations. Following coregistration, we computed the forward solution using MNE's mne.make_forward_solution() function, which calculates the magnetic fields and electric potentials that sensors and electrodes detect from cortical dipole sources in each subject. For the two-word data, channel-noise covariance was estimated based on the 100 ms intervals prior to each artifact-free trial, aligning with previous literature on phrasal composition in a two-word setting (Bemis & Pylkkanen, 2011; Li & Pylkkänen, 2021). The naturalistic data were baseline-corrected using the mean of the whole epoch. We acknowledge that this may lead to different SNRs for the noise covariance estimates for the two-word and the naturalistic data. However, since the main goal of our study is to examine whether phrase processing in controlled laboratory settings can be generalized to naturalistic settings, we would like to maintain consistency with prior analytical methods. We have also reprocessed the two-word data, based-lined corrected using the whole epoch. The results are very similar to our initial findings (see Supplementary Fig. 1). The inverse solution was computed from the forward solution and the grand average activity across conditions with "free" orientation, meaning that the inverse operator places three orthogonal dipoles at each location defined by the source space. However, when computing the source estimate, only activity from the dipoles perpendicular to the cortex was included. This

approach is equivalent to estimating the inverse solution with fixed orientation dipoles, however, it affords greater flexibility by allowing us to evaluate results under both fixed and loose orientations. For each trial, the same inverse operator was applied to yield dynamic statistical parameter maps (dSPM) units (Dale et al., 1999). This was done using a regularization parameter computed as $\frac{1}{SNR^2}$, with the SNR value set at 3. The final source estimates were downsampled to 200 Hz to save computing time. All data preprocessing steps were performed using MNE-python (v.0.24.0; Gramfort et al., 2014).

2.5. Behavioral data analyses

Accuracies were analyzed using a generalized linear mixed-effects model (GLMM) with binomial error distribution, and the log-transformed RTs were analyzed using a linear mixed-effects model. Our fixed effects include the binary variables Composition (single nouns vs. phrases) and Age (adults vs. children). Subject-level variability was included as random intercepts. The GLMM analyses were conducted via the "lme4" package (Bates et al., 2015) in R (v4.2.1) and RStudio (v022.12.0+353). The statistical significance of fixed effects was estimated using the "lmerTest" package (Kuznetsova et al., 2017), in which Satterthwaite's approximation was applied to estimate degrees of freedom (see Fig. 1B for the results).

2.6. Phrasal and noun representations in LLMs

To gain insights into the neural representations of phrases and single nouns in the two-word and naturalistic contexts, we first examined phrasal and noun representations in isolated two words and narratives in a large language model (LLM). Recent LLMs have achieved extraordinary performance in language comprehension tasks and have been suggested to share some computational principles with the human brain (e.g., Caucheteux & King, 2022; Goldstein et al., 2022; Schrimpf et al., 2021). Here, we first extracted each layer's embeddings from the pre-trained GPT2-large model (Radford et al., 2019) for the nouns in single nouns and adjective-noun phrases in the two-word (e.g., "green glass" vs. "glass") and narrative contexts (e.g., "…soft music…" vs. "…a bath…"). We then applied multidimensional scaling (MDS), a dimensionality reduction technique to visualize the last layer's embedding of each adjective-noun phrase and single noun in the two-word and naturalistic contexts to two dimensions (see Fig. 2A). We also computed the cosine distance between each layer's embeddings for single nouns and adjective-noun embeddings (see Fig. 2B). The pretrained GPT2-large model was obtained from the transformers (v4.10.2) package in python.

2.7. Classification on LLM embeddings for phrases and nouns

We trained a feed-forward neural network (FFNN) classifier to distinguish the nouns in single nouns and adjective-noun phrases using the two-word stimuli and tested the classifier on the nouns in single nouns and adjective-noun phrases in the naturalistic text. Adjective-noun phrases were annotated using the Stanford part-of-speech tagger (Toutanova et al., 2003) and were manually checked. Conversely, we also trained an FFNN classifier on the naturalistic data and tested it on the two-word data. The FFNN contains one hidden layer with two units (see Fig. 2C). To control for the confounding factor that the nouns in single nouns were the initial token whereas the nouns in adjective-noun phrases were not, we performed a linear regression model using the binary variable "word position" to predict each layer's embeddings. We took the residuals of the model for the classification analyses. The classification analyses were performed using the python package scikit-learn (v0.22.1).

2.8. Searchlight multivariate pattern classification on MEG data

We conducted searchlight multivariate pattern classification analyses on the source-localized MEG data within a left-lateralized language mask for each subject. The language mask (see the pink region in Fig. 3A) covered regions including the whole left temporal lobe, the left inferior frontal gyrus (LIFG; defined as the combination of BAs 44 and 45), the left ventromedial prefrontal cortex (LvmPFC; defined as BA11), the left angular gyrus (LAG; defined as BA39), and the left supramarginal gyrus (LSMA; defined as BA 40). The left AG and vmPFC have also been implicated in previous literature on conceptual combination (Bemis & Pylkkanen, 2011; Price et al., 2015) and the LIFG and the LMTG have been suggested to underlie syntactic combination (Flick & Pylkkänen, 2020; Hagoort, 2005; Lyu et al., 2019; Matchin & Hickok, 2020; Matchin et al., 2019).

We trained feedforward neural network (FFNN) classifiers to pairwise combinations of the MEG data for single nouns and phrases in the two-word and naturalistic experiments (see Fig. 3A). The FFNN contains one hidden layer with two units. The binary classifiers were separately applied to all spatiotemporal timepoints, with a radius of 20 sources. The same analysis pipeline was

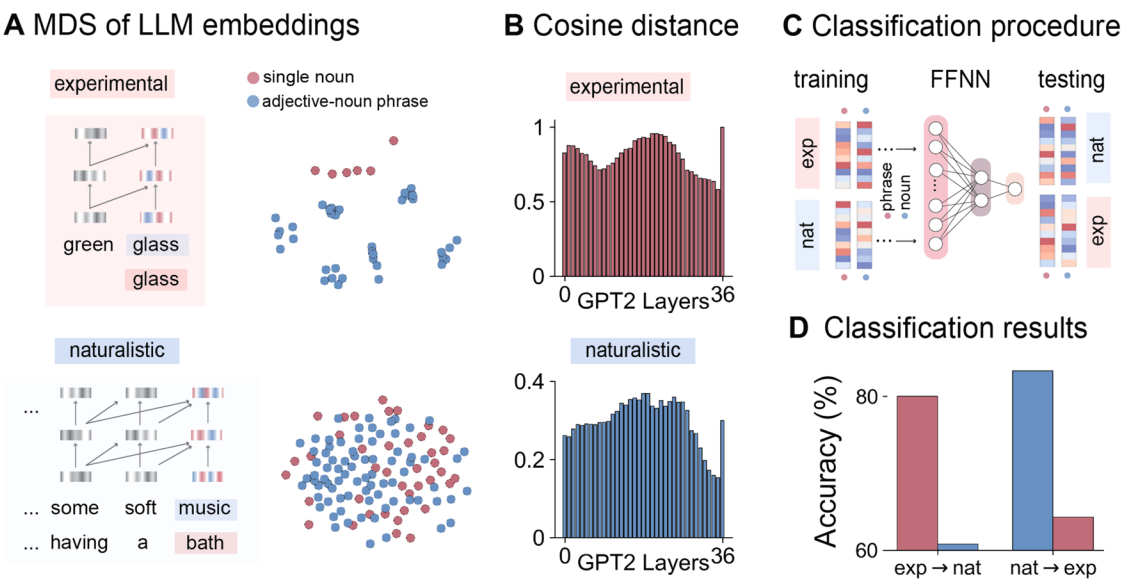
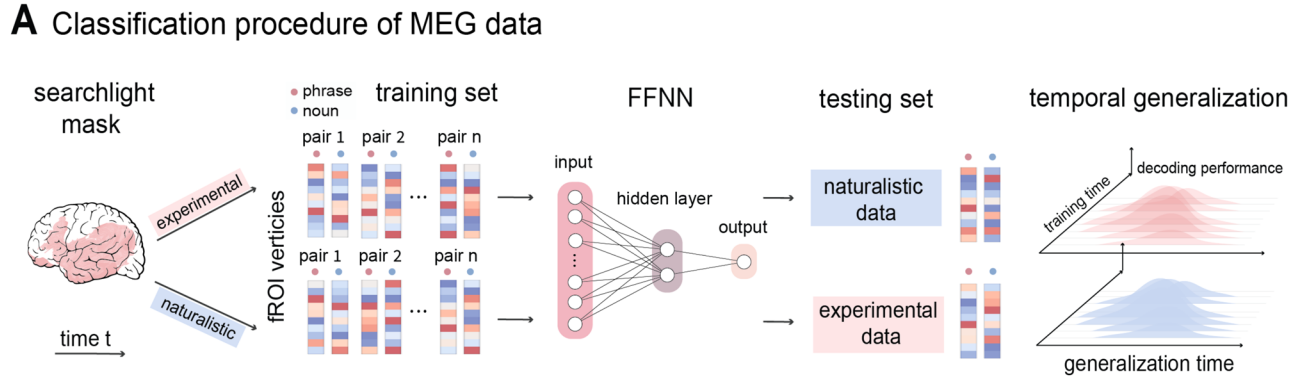
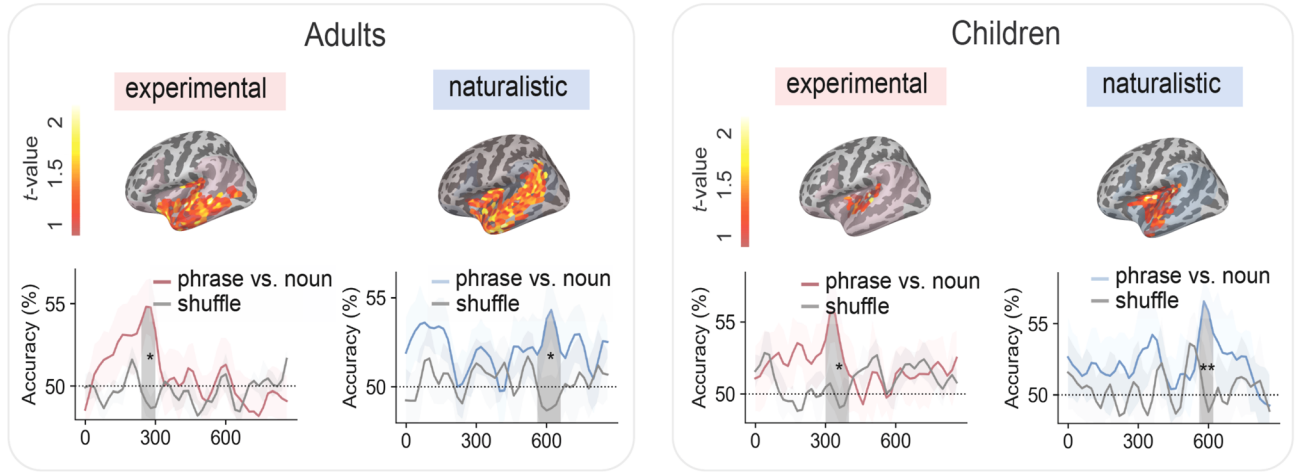
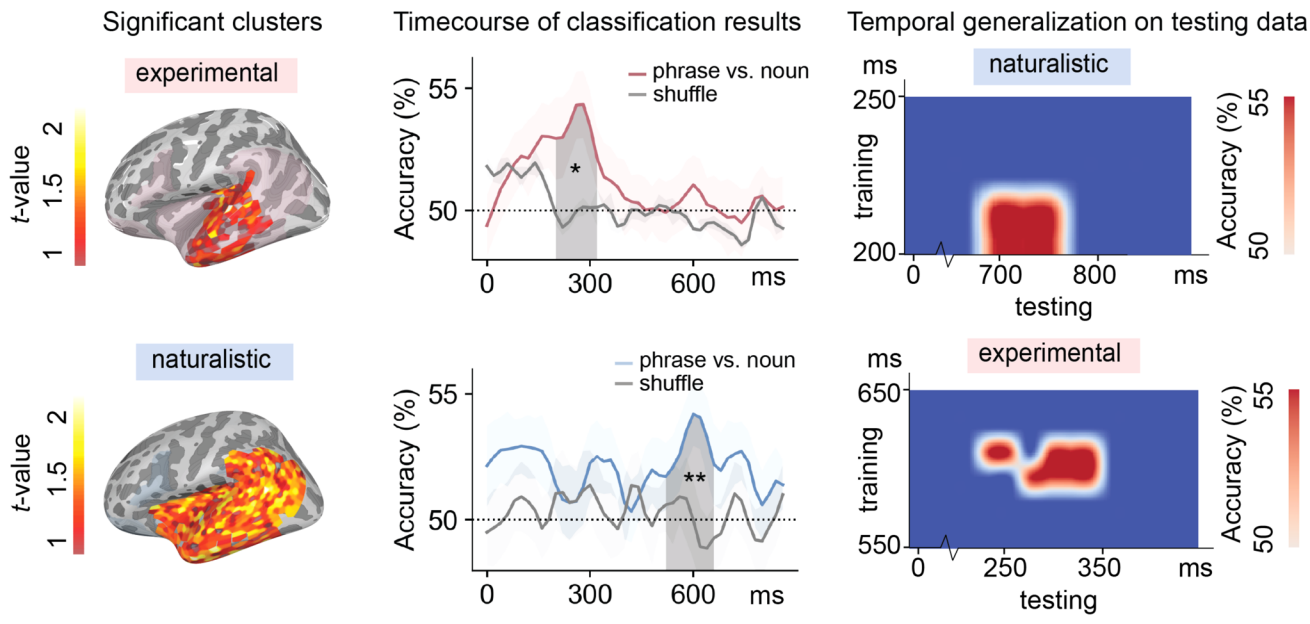


Fig. 2. Phrasal and noun representations in two-word and naturalistic contexts in an LLM. (A) Multidimensional scaling (MDS) of the last layer's embeddings of adjective-noun phrases and single nouns in two-word and naturalistic contexts in GPT-2. (B) Cosine distance of each layer's embeddings of adjective-noun phrases and single nouns in two-word and naturalistic contexts in GPT-2. (C) A feed-forward neural network classifier was trained to distinguish the last layer's embeddings of nouns in single nouns and adjective-noun phrases in the two-word context, and tested on the last layers of nouns in single nouns and adjective-noun phrases in the naturalistic context. Conversely, a classifier was trained on the naturalistic context and tested on the two-word context. (D) Classification results on the LLM's embeddings. The classifier trained in the two-word context achieved an accuracy of 80% in distinguishing phrases from nouns and an accuracy of 60.8% when applied to the naturalistic context. The classifier trained in the naturalistic context achieved an accuracy of 83.3% and an accuracy of 64.3% when tested in the experimental context.

Fig. 3. Classification analyses procedure and results on the MEG data. (A) Following the classification analysis of phrases versus nouns in LLMs, we trained feed-forward neural network (FFNN) classifiers to distinguish phrases from nouns in one context and tested in another context. The same classification was applied independently with a searchlight radius of 20 sources within a language mask and at every timepoint. Classification accuracies for the training data were averaged over subjects at source and time point minus the chance level of 50% was submitted to a one-sample *t*-test and the statistical significance was determined by a TFCE correction with 10,000 permutations. At the testing time, we applied the temporal generalization method and tested the classifiers' performance at each time point on every timepoint in the testing data. (B) Classification results of adults' and children's MEG data. For adults, the classifiers trained on the experimental data can distinguish phrases from single nouns in the left anterior and middle temporal lobe from 240-320 ms (N sources = 214, $t = 5.58$, Cohen's d = 1.28, $p = 0.025$) after the onset of the target word. The classifiers trained on the naturalistic data can distinguish phrases from single nouns in the similar left anterior and middle temporal regions from 560-680 ms (N sources = 222, $t = 3.59$, Cohen's d = 0.82, $p = 0.03$) after the onset of the target word. For children, the classifiers trained on the experimental data can distinguish phrases from single nouns in the left middle temporal lobe from 300-420 ms (N sources = 43, $t = 2.24$, Cohen's d = 0.71, $p = 0.014$) after the onset of the target word. The classifiers trained on the naturalistic data can distinguish phrases from single nouns in the left anterior and middle temporal regions from 560-640 ms (N sources = 112, $t = 3.55$, Cohen's d = 1.12, $p = 0.001$) after the onset of the target word. (C) Classification results of all MEG data. When trained on experimental data, the classifiers can distinguish phrases from single nouns in the left anterior and middle temporal lobe from 200-340 ms (N sources = 136, $t = 5.08$, Cohen's d = 0.93, $p = 0.005$) after the onset of the word. When trained on the naturalistic data, the classifiers can distinguish phrases from single nouns in the whole left temporal lobe from 520-680 ms (N sources = 532, $t = 4.54$, Cohen's d = 0.83, $p = 0.001$) after the onset of the word. The grey lines represented shuffled classification results. * indicates $p < 0.05$; ** indicates $p < 0.01$.


A Classification procedure of MEG data
searchlight mask
experimental
naturalistic
time t
fROI verticies
phrase
noun
training set
pair 1
pair 2
pair n
FFNN
input
hidden layer
output
testing set
naturalistic data
experimental data
temporal generalization
training time
decoding performance
generalization time


B Classification results of adults' and children's MEG data
Adults
Children
experimental
naturalistic
t-value
phrase vs. noun
shuffle
Accuracy (%)


C Classification results of all MEG data
Significant clusters
Timecourse of classification results
Temporal generalization on testing data
experimental
naturalistic
t-value
phrase vs. noun
shuffle
Accuracy (%)
ms
training
testing

applied to each subject. At the group level, the classification accuracy averaged over subjects at each timepoint minus the chance level of 50% was submitted to a one-sample one-tailed t-test with threshold-free cluster enhancement (TFCE) correction (Smith & Nichols, 2009) for 10,000 permutations (see the first two columns of Fig. 3B for the results). The analysis time window was between 0-875 ms after the onset of the second word.

2.9. Testing the classifiers using the temporal generalization method (TGM)

The classifiers trained to distinguish the MEG data for single nouns and phrases in the two-word task were tested on the MEG data for single nouns and phrases in the naturalistic task using TGM. TGM allows us to probe compositional processing in the brain over time by training the classifier using data from one time period and testing the classifier on data from all time periods. This method is particularly useful for neuroimaging data with high temporal resolution (e.g., EEG, MEG), and it has been successfully applied in other domains of cognitive neuroscience such as memory (Meyers, 2018), vision (Dobs et al., 2019), audition (King et al., 2014), etc. The results of TGM is a 2D matrix, where the color at point *i, j* indicated prediction accuracy when the model is trained using data at time *i* and tested with data at time *j* (see Fig. 3A for the classification procedure).

Similarly, the classifiers trained on the naturalistic data were tested on the experimental data using TGM. During testing, each classifier trained from training data at a timepoint was applied to testing data at all timepoints. This procedure led to two TGM matrices of classification performance, one for training on experimental data and testing on naturalistic data, and one for training on naturalistic data and testing on experimental data. Statistical significance is decided based on a cluster-based one-sample one-tailed *t*-test with 10,000 permutations (Maris & Oostenveld, 2007), comparing the 2D matrix to a chance level of 0.5 (see the last column of Fig. 3B for the results). The classification analyses were performed using the python package scikit-learn (v0.22.1), and the statistical analyses were performed using the python package eelbrain (v0.38).

2.10. MDS of MEG data of phrases and nouns

We extracted the MEG data from the significant clusters derived from the classification analyses (see the first column in Fig. 3B). We then applied MDS to the MEG source estimates of each target word in the two-word and naturalistic contexts. We also plotted the temporal dynamics of the 2D representations of the single-nouns and adjective-noun phrases in the "experimental" and "naturalistic" state space (see Fig. 4).

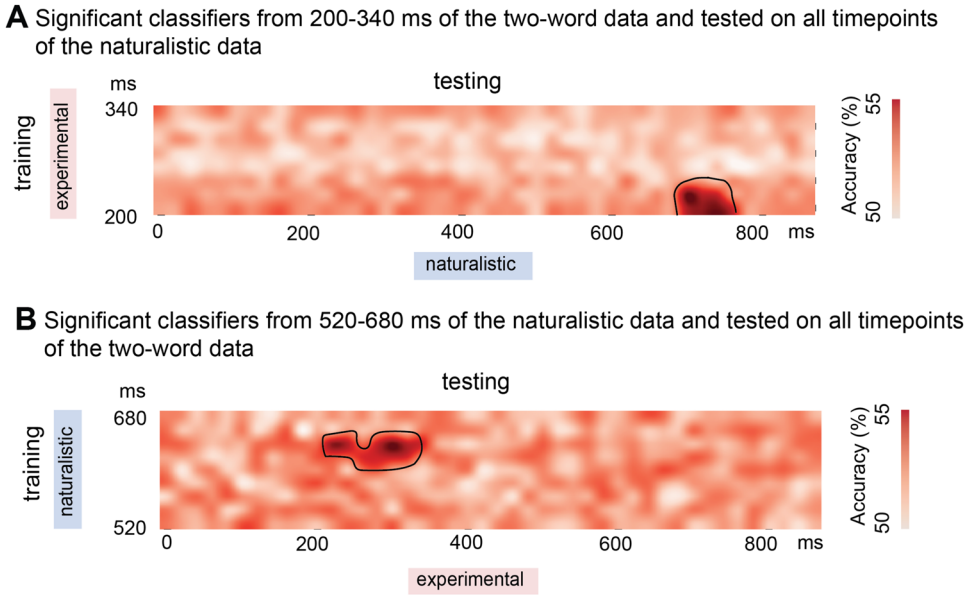


Fig. 4. The temporal generalization results. (A) The classifiers trained on the two-word data from 200-220 ms can significantly distinguish phrases from nouns from 700-760 ms in the testing data. (B) The classifiers trained on the naturalistic data from 620-640 ms can significantly distinguish phrases from nouns from 220-340 ms in the testing data.

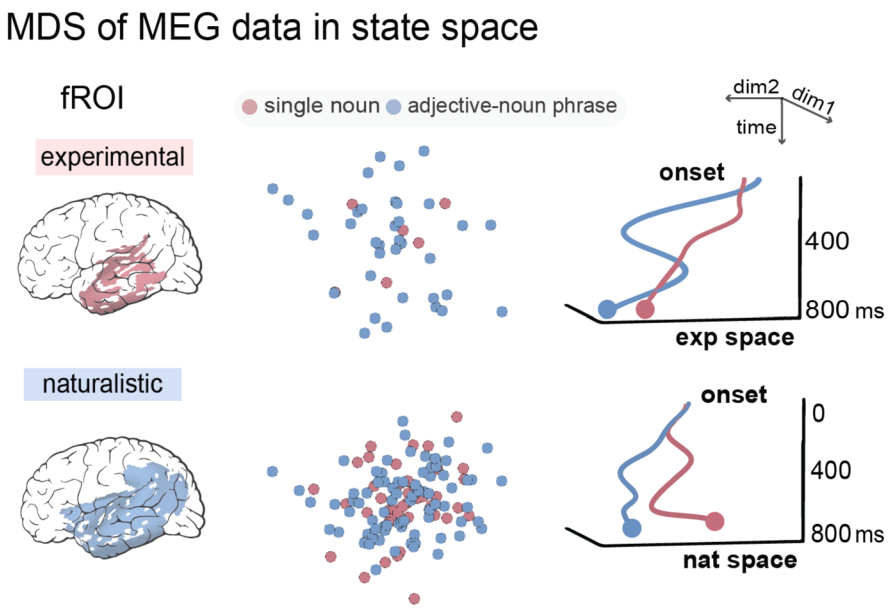


Fig. 5. MDS of the neural representations for phrases and nouns in the experimental and naturalistic contexts. The MEG source estimates from the significant spatiotemporal clusters in the classification analyses were extracted and reduced to two dimensions using MDS. The timecourses of the MDS representations of phrases and nouns in the experimental state space suggested larger distance in an earlier time window. For the naturalistic data, the timecourses of the MDS representations diverted from the middle to the end of the whole epoch.

2.11. Multiple regression on the naturalistic MEG data

Naturalistic stimuli differ from two-word stimuli in many dimensions. For example, the stimuli in the two-word task had lower surprisal as they were repeatedly presented during the experiment. Surprisal evoked by an incoming word indicates the amount of information that was not predictable from the context (Hale, 2001; Levy, 2008), and is calculated as the negative logarithm of the probability assigned to the actual next word. A slower presentation rate of words (875 ms) in the two-word task may also facilitate faster composition compared to words that are much faster during naturalistic speaking. Other linguistic factors such as richer prosodic information and different word frequency may also induce additional processes that delayed the composition effect. In addition, processes beyond the language domain may be involved during narrative understanding. Emotional arousal and valence, for example, have been shown to also evoke activity in the language network (Wallentin et al., 2011).

To understand whether these factors that are underlying the "naturalness" of the narrative stimuli contributed to the late composition effect, we conducted a multiple regression model to regress out these factors (see Fig. 6A). Our dependent variable is the source estimates of each subject's naturalistic data. Our regressors included the peak intensity and f0 of the target words, word rate, word frequency, word surprisal based on the GPT-2 language model (Radford et al., 2019), emotional valence and arousal indicated by human judgment on Amazon Mechanical Turk (see details of the regressors below). Both the dependent and independent variables were z-scored. Pearson's *r* correlations among the regressors were examined to ensure no collinearity among the regressors (see Fig. 6C).

2.12. Intensity and pitch

Root mean square (RMS) intensity and the fundamental frequency (f0) for every 10 ms of the audio were extracted using the Voicebox toolbox (http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html). Peak RMS intensity and peak f0 within the during of each word in the naturalistic stimuli were used to represent the intensity and pitch information for each word.

2.13. Word rate

Since word duration is largely determined by the length of the word, we computed the presentation rate of each word as the duration of each word in milliseconds divided by the number of letters in the word. A slow presentation

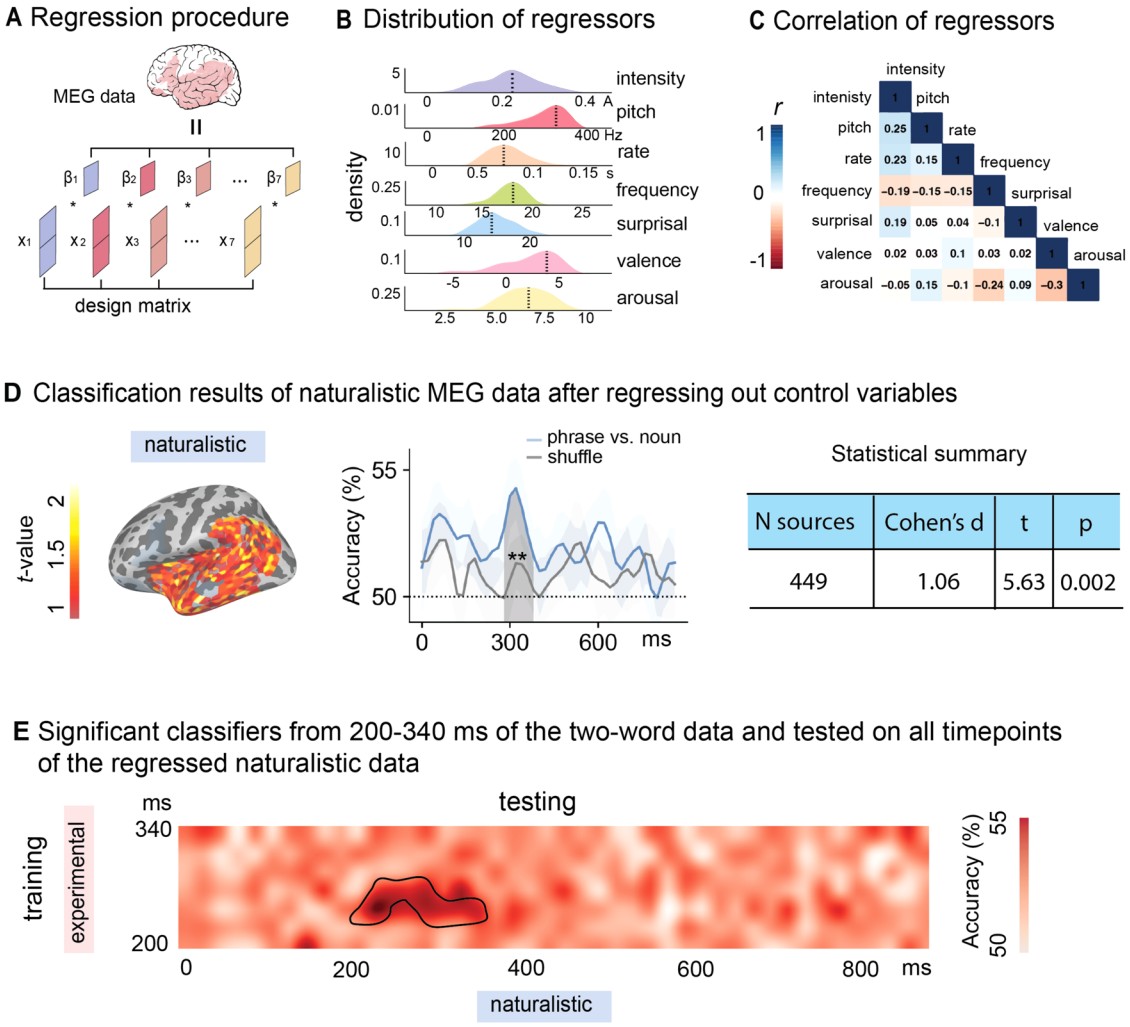


Fig. 6. Regression analyses procedure and the classification results of the naturalistic data after regressing out control variables. (A) We applied a linear regression model to predict the source estimates of the target words in the naturalistic data for each subject. (B) Distribution of the regressors. Our regressors include intensity and pitch for the audio, word frequency, presentation rate, surprisal of the word given previous context, and emotional arousal and emotional valence of the text. (C) Correlations among the regressors. The correlation matrix suggested low correlation among the predictors. (D) Classification results after controlling for the regressors. We performed the same classification analyses on the residuals of the source estimates after regression. When trained on the naturalistic data, the classifiers can distinguish phrases from single nouns in a large cluster in the left temporal lobe from 280-400 ms (N sources = 449, t = 5.63, Cohen's d = 1.06, p = 0.002) after the onset of the word. (E) When tested on the experimental data using TGM, the classifiers from 320-340 ms in the training data can significantly distinguish phrases from nouns from 220-360 ms in the testing data. The grey lines represented shuffled classification results. ** indicates $p < 0.01$.

rate indicates words with longer duration and fewer letters, while a fast presentation rate suggests a shorter presentation of long words.

2.14. Word frequency

Log-transformed unigram frequency of each word in the naturalistic stimuli was estimated using Google Books Ngram Viewer, Version 2012070129 (http://storage.googleapis.com/books/ngrams/books/datasetsv2.html).

2.15. Surprisal

The predictability of each word in the naturalistic stimuli given the previous context was indexed by the surprisal of all the words in the naturalistic stimuli. Surprisal evoked

by an incoming word indicates the amount of information that was not predictable from the context (Hale, 2001; Levy, 2008), and is calculated as the negative logarithm of the probability assigned to the actual next word. The probability of each word in the naturalistic stimuli given the previous words within the same sentence was derived from the pretrained GPT2-large model. This model uses a transformer architecture and has been shown to successfully capture human performance on next-word prediction (e.g., Goldstein et al., 2022; Schrimpf et al., 2021). The analyses was performed using the python package transformers (v4.10.2).

2.16. Emotion arousal and emotional valence

Emotional arousal and emotional valence of each sentence in the naturalistic stimuli were rated by participants on Amazon Mechanical Turk (MTurk). Following a prior study (Wallentin et al., 2011), arousal was rated on a scale from 0 to 10 indicating extreme boredom to extreme arousal. Emotional valence was rated on a scale from -5 to 5, where -5 indicates strong negative emotions and 5 indicates strong positive emotions. A total of 30 participants completed the survey. The mean valence and arousal ratings for each sentence were computed, and words in the same sentence have the same emotional arousal and emotional valence. Inter-subject correlations (ISC) among each subject's ratings for arousal and valence were computed as the mean of the Pearson's *r* coefficients between each subject's ratings and the overall mean ratings. The statistical significance of subjects' ISC coefficients was determined by comparing the observed values with randomly generated ratings using paired two-sample *t*-tests.

3. RESULTS

3.1. Behavioral results for the two-word task

Overall, participants achieved an accuracy of 94.4% (SD = 23%) with a mean reaction time (RT) of 2.6 s (SD = 2.07 s). The mean accuracy for adults was 96.9% (SD = 17.3%), and the mean accuracy for children was 89.9% (SD = 30.1%). The mean RT for adults was 2 s (SD = 1.1 s), and the mean RT for children was 3.67 s (SD = 2.83 s; see Fig. 1B). Compared to prior studies (e.g., Bemis & Pylkkanen, 2011), these RTs seem longer. This is because the task was more difficult as the participants needed to use two buttons to select from eight pictures. The reason for the more complex task was to reduce the possibility of correct responses by chance, which makes the task more applicable for possible clinical uses.

The binary variable Accuracy was analyzed using a generalized linear mixed-effects model (GLMM) with binomial error distribution, and RTs were log-transformed and analyzed using a linear mixed-effects model (LMM). Composition (single nouns vs. phrases) and Age (adults vs. children) were included as fixed effects and subjects as random intercepts. The results revealed a significant effect of Age on both accuracy ($p < 0.001$) and RT ($p < 0.001$). Composition was significant for RT ($p = 0.0003$) but not accuracy ($p = 0.94$).

3.2. Phrasal and noun representation in LLMs

To gain insights into the neural representations of phrases and single nouns in the two-word and naturalistic contexts, we first examined the pretrained GPT2-large model's embeddings of adjective-noun phrases and single nouns in a two-word setting (e.g., "green glass" vs. "glass") and a naturalistic setting (e.g., "…soft music…" vs. "…a bath…"). The MDS results showed that in the two-word context, there is a clear separation of noun and phrasal representations in the LLM. In the naturalistic setting, however, the last layer's representations of nouns and phrases were both highly distributed (see Fig. 2A). The cosine distances between each layer's embeddings of adjective-noun phrases and single nouns in the two contexts were shown in Figure 2B. We can see a larger distance in the middle and final layers of the LLM.

3.3. Classification results on LLM embeddings

To understand whether the LLM has learned the contrast between single nouns and adjective-noun phrases, we trained two feed-forward neural network classifiers to distinguish phrases from nouns in the two-word context and tested the trained classifiers in the naturalistic context, and vice versa (see Fig. 2C). The classifier trained in the two-word context achieved an accuracy of 80% in distinguishing phrases from nouns and an accuracy of 60.8% when applied to the naturalistic context. The classifier trained in the naturalistic context achieved an accuracy of 83.3% and an accuracy of 64.3% when tested in the experimental context (see Fig. 2D). Although the testing accuracies were much lower than the training accuracy, the results in the two-word and naturalistic settings were comparable and were well above the chance level of 50%, suggesting that the LLM has learned different representations for single nouns and adjective-noun phrases and can be generalized across contexts.

3.4. Classification results on MEG data

We applied the same classification methods to the MEG data to examine the generalizability of the neural reflections of semantic composition. Figure 3B shows the classification results of adults' and children's MEG data. For adults, the classifiers trained on the experimental data can distinguish phrases from single nouns in the left anterior and middle temporal lobe from 240-320 ms (N sources = 214, t = 5.58, Cohen's d = 1.28, p = 0.025) after the onset of the target word. The classifiers trained on the naturalistic data can distinguish phrases from single nouns in the similar left anterior and middle temporal regions from 520-680 ms (N sources = 222, t = 3.59, Cohen's d = 0.82, p = 0.03) after the onset of the target word. For children, the classifiers trained on the experimental data can distinguish phrases from single nouns in the left middle temporal lobe from 300-420 ms (N sources = 43, t = 2.24, Cohen's d = 0.71, p = 0.014) after the onset of the target word. The classifiers trained on the naturalistic data can distinguish phrases from single nouns in the left anterior and middle temporal regions from 560-640 ms (N sources = 112, t = 3.55, Cohen's d = 1.12, p = 0.001) after the onset of the target word. Since the adults' children's results exhibited similar spatiotemporal patterns, we collapsed their data together for future analyses.

For all subjects' data, we found that when trained on the two-word data, the classifiers can distinguish phrases from single nouns in the left anterior and middle temporal lobe from 200-340 ms (N sources = 136, t = 5.08, Cohen's d = 0.93, p = 0.005) after the onset of the second word. When tested on the naturalistic data, the TGM results suggest that the classifiers from 200-220 ms in the training data can significantly distinguish phrases from nouns from 700-760 ms in the testing data. When trained on the naturalistic data, the classifiers can distinguish phrases from single nouns in the whole left temporal lobe from 520-680 ms (N sources = 532, t = 4.54, Cohen's d = 0.83, p = 0.001) after the onset of the word. When tested on the experimental data using TGM, the classifiers from 620-640 ms in the training data can significantly distinguish phrases from nouns from 220-340 ms in the testing data (see Figs. 3C and 4). The generalization effects observed in the training data are indeed brief, with only a 20 ms segment generalizing to the testing data. This is likely due to our methodology of only selecting the classifiers within significant spatiotemporal clusters from our classification analyses, resulting in relatively short analysis windows for the training data—140 ms for the two-word data and 160 ms for the naturalistic data. This 20 ms window constitutes approximately 14% of the training data timeframe. Moreover, the classification analysis identified distinct spatial clusters for the two-word and naturalistic data. Such differences might influence the TGM outcomes, considering we applied classifiers from significant clusters in the training data to the testing data.

3.5. Neural dynamics of phrasal and noun representations

We used MDS to visualize the neural codes associated with each adjective-noun phrase and single nouns in the two-word and naturalistic contexts. Within the significant spatiotemporal clusters derived from the classification analyses, we plotted the averaged MEG data of each phrase and noun in a two-dimensional space. We also plotted the temporal dynamics of the mean 2D neural codes for all phrases and nouns in the two contexts. The results suggested reliable segregation of multivariate neural signals associated with adjective-noun phrases and single nouns in both experimental and naturalistic contexts. However, the temporal dynamics of the MDS representations also showed different patterns in the two contexts: In the two-word setting, the neural distance between phrases and nouns was larger in an earlier time window at around 100-400 ms and converged near the end of 800 ms. In the naturalistic setting, the neural codes for phrases and nouns remained distant from around 400 ms to the end of the epoch (see Fig. 5). This is consistent with the classification results where the composition effect occurred later in the naturalistic context.

3.6. Regression model of the naturalistic MEG data

Figure 6B shows the distributions of these regressors for the naturalistic stimuli. The mean root-mean-squared (RMS) intensity and mean f0 for all target words in the naturalistic stimuli were 0.21 A (SD = 0.08 A) and 317.9 Hz (SD = 40.02 Hz), respectively. The mean presentation rate of the target words in the naturalistic stimuli, calculated as the duration of the word divided by the number of letters in the word, was 72.8 ms (SD = 23.5 ms). The mean log frequency and surprisal of the target words based on GPT2 in the naturalistic stimuli were 18.04 (SD = 1.83) and 14.04 (SD = 2.66). The mean emotional valence and arousal indicated whether the sentences containing the target words induced positive or negative emotion (-5 is very negative and 5 is very positive), and how strong the emotion was (on a scale of 0-10). Their mean values were

1.83 (SD = 2.65) and 6.48 (SD = 2.66). The mean inter-subject correlations (ISC) among the participants' ratings on valence and arousal were 0.74 (SD = 0.13) and 0.57 (SD = 0.2) and were both significantly greater than randomly generated ratings (t = 20.24, p < 0.0001 and t = 13.66, p < 0.0001, respectively), suggesting high agreement among the subjects on the two emotional dimensions associated with sentences in the naturalistic stimuli. We also examined the correlation coefficients among the regressors, and the results suggested no collinearity among the regressors. The correlation coefficient between emotional valence and emotional arousal is -0.3, which is the highest absolute r value among all the regressor pairs (see Fig. 6C).

3.7. Classification results of the naturalistic data after regressing out other factors

We took the residuals of the source estimates for the target words in the naturalistic stimuli for each subject, and re-conducted the same classification analyses on the residuals. Our results confirmed that the late composition effect observed in the naturalistic data was due to additional processing efforts of these factors: The classifiers trained on the naturalistic data distinguished phrases from single nouns in a large cluster in the left temporal lobe from 280-400 ms (N sources = 449, t = 5.63, Cohen's d = 1.06, p = 0.002) after the onset of the word. When tested on the experimental data using TGM, the classifiers from 320-340 ms in the training data significantly distinguished phrases from nouns from 220-360 ms in the testing data (see Fig. 6D).

4. DISCUSSION

Traditional experimental paradigms in cognitive neuroscience of language aim to isolate specific cognitive processes by comparing conditions that differ in the component of interest. In contrast, recent naturalistic paradigms use audiobooks or movies to mimic everyday language experiences. However, both paradigms have limitations. Controlled experimental stimuli may deviate from natural language use, and subtraction methods assume linearity in a brain that is likely non-linear (Friston et al., 1996). Naturalistic stimuli contain diverse linguistic and non-linguistic information, making it challenging to isolate specific subprocesses (Hasson & Egidi, 2015). Direct comparisons of neural signals for linguistic processes between the two paradigms are rare, leaving it unclear if results from traditional experiments generalize to naturalistic settings and vice versa. According to existing neurolinguistic models (e.g., Hickok & Poeppel, 2000), brain areas associated with specific functions should not vary with the research context. For example, the left anterior temporal regions' involvement in semantic composition should be consistent during phrasal processing, regardless of the paradigm or modality of stimuli presentation.

This study investigates the generalizability of meaning composition across traditional experimental and naturalistic paradigms, focusing on the core function of human language. We examined whether semantic composition observed in experimental paradigms extends to a naturalistic setting, and vice versa. Our classification results revealed similar neural activity for meaning composition in the left anterior and middle temporal regions in both experimental and naturalistic contexts. Notably, the spatial distribution of the combinatory activity reported here is wider than the LATL, which has been the focus of most prior literature on basic composition using the red-boat paradigm (Bemis & Pylkkanen, 2011). To understand the wider distribution, it is relevant to keep in mind that most prior basic composition studies have been conducted in the visual modality, with the exception of Bemis and Pylkkänen (2013a), which used both auditory and visual modalities. That study identified both an LATL effect and a posterior temporo-parietal effect, with the latter being more robust in the auditory modality. This finding broadly conforms with the current, auditory results. Further, a recent replication by Flick and Pylkkänen (2020) of the original visual red-boat study (Bemis & Pylkkanen, 2011) also revealed wider left temporo-parietal sensitivity to basic composition. Thus it is likely that the LATL is the most consistent locus of such effects, with the highest rate of replication, but there are now several indications of the participatory role of surrounding temporal cortex as well. Studies probing the functional details of the LATL have revealed a conceptual, non-syntactic role for it (Bemis & Pylkkänen, 2013c; Li & Pylkkänen, 2021; Parrish & Pylkkänen, 2022; Zhang & Pylkkänen, 2015). For example, LATL can combine concepts like "boat red" even when the two words do not syntactically combine (Bemis & Pylkkänen, 2013c; Parrish & Pylkkänen, 2022). Conversely, the posterior temporal cortex is more syntactically sensitive (Flick & Pylkkänen, 2020; Hagoort, 2005; Li & Pylkkänen, 2021; Lyu et al., 2019; Matchin & Hickok, 2020; Matchin et al., 2019). As discussed in Pylkkänen, (2019), composition may involve syntactic, logico-semantic, and conceptual subroutines. In the present study, we most likely are observing the contributions of both conceptual

and syntactic composition. Overall, the various aspects of combinatory processing are thought to engage multiple areas of temporal, parietal and frontal cortex beyond the LATL (see Pylkkänen, 2019 for a review). For the naturalistic data, the classification performance extended beyond the significant clusters observed in the two-word data (as depicted in Fig. 3C), indicating the involvement of a larger network during the naturalistic task.

One line of research suggests that there is a hierarchy of increasing temporal receptive windows from lower sensory to higher perceptual and cognitive brain areas, and different levels of linguistic units are encoded at different cortical regions (e.g., Blank & Fedorenko, 2020; Hasson et al., 2008; Lerner et al., 2011; Schmitt et al., 2021). It is possible that phrasal processing in the naturalistic context encompasses longer temporal receptive windows, considering the richer contextual information, thus engaging more anterior or posterior temporal regions compared to isolated phrases.

Consistent with the hypothesis of longer temporal receptive windows, our findings revealed a delayed distinction between single nouns and adjective-noun phrases in the naturalistic MEG data, occurring from 520-680 ms after the onset of the target word, compared to the effect observed in the two-word task from 200-340 ms. Both our Temporal Generalization Mapping (TGM) and Multidimensional Scaling (MDS) analyses on the MEG data supported this latency contrast for composition in both paradigms. As naturalistic stimuli encompass richer information, including diverse prosodic cues, word rate, word frequency, and surprisal evoked by incoming words (Hale, 2001; Levy, 2008), as well as non-linguistic factors like emotional arousal and valence (Wallentin et al., 2011), prior neurolinguistic studies employing a naturalistic design have commonly controlled for these factors using regression models (e.g., Brennan et al., 2016; Caucheteux & King, 2022; Huth et al., 2016; Nelson et al., 2017). In our study, we accounted for these factors by regressing them out and then conducted the classification analyses using the residuals. Interestingly, after controlling for these factors, we observed an earlier composition effect that closely resembled the effect observed in the two-word data. This suggests that the composition effect observed in both experimental and naturalistic approaches reflects the same underlying processes, rather than being distinct processes.

Similarly, the classification results on the embeddings of single nouns and adjective-noun phrases in both the two-word and narrative contexts of the large language models (LLMs) indicate the presence of generalized patterns for these word types. While the question of whether these patterns reflect composed meaning in LLMs remains open, the results demonstrate the existence of specific features that differentiate single nouns from adjective-noun phrases and can be generalized across different contexts. It is important to note that the two-word stimuli introduce a confounding factor, as the nouns in the single-nouns condition are the initial tokens, while the nouns in the adjective-noun phrases condition are the second tokens. To mitigate this factor, we deliberately removed the word position effect from each layer's embeddings, ensuring that the classifier cannot rely solely on word position to distinguish between the two conditions.

To sum up, we observed the composition effect in both the experimental and naturalistic designs in similar brain regions and similar temporal windows when controlled for additional factors in the naturalistic stimuli, suggesting a single compositional process during both isolated and connected speech comprehension. One limitation of our study is that we only focused on a specific linguistic subprocess, and further research is needed to examine whether other subprocesses, such as morphological or syntactic processing, can be replicated across different research paradigms. Conducting meta-analyses using existing experimental and naturalistic fMRI datasets from open data platforms could serve as a valuable starting point for future investigations in this direction.

DATA AND CODE AVAILABILITY

The data and the codes for analyses are available at https://osf.io/7c58j/.

AUTHOR CONTRIBUTIONS

J.L. designed research, analyzed data, and wrote the paper. M.L. analyzed data. L.P. designed research and wrote the paper.

FUNDING

This work was supported by the NSF Award BCS-1923144 (L.P.) and award G1001 from NYUAD Institute, New York University Abu Dhabi (L.P.).

DECLARATION OF COMPETING INTEREST

The authors declare no competing interests.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available with the online version here: https://doi.org/10.1162/imag_a_00072.

REFERENCES

Adachi, Y., Shimogawara, M., Higuchi, M., Haruta, Y., & Ochiai, M. (2001). Reduction of non-periodic environmental magnetic noise in MEG measurement by continuously adjusted least squares method. *IEEE Transactions on Applied Superconductivity*, *11*(1), 669–672. https://doi.org/10.1109/77.919433

Alday, P. M. (2019). M/EEG analysis of naturalistic stories: A review from speech to language processing. *Language, Cognition and Neuroscience*, *34*(4), 457–473. https://doi.org/10.1080/23273798.2018.1546882

Ardila, A., & Rosselli, M. (1994). Development of language, memory, and visuospatial abilities in 5- to 12-year-old children using a neuropsychological battery. *Developmental Neuropsychology*, *10*(2), 97–120. https://doi.org/10.1080/87565649409540571

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bemis, D. K., & Pylkkanen, L. (2011). Simple composition: An MEG investigation into the comprehension of minimal linguistic phrases. *Journal of Neuroscience*, *31*(8), 2801–2814. https://doi.org/10.1523/jneurosci.5003-10.2011

Bemis, D. K., & Pylkkänen, L. (2013a). Combination across domains: An MEG investigation into the relationship between mathematical, pictorial, and linguistic processing. *Frontiers in Psychology*, *3*, 583. https://doi.org/10.3389/fpsyg.2012.00583

Bemis, D. K., & Pylkkänen, L. (2013b). Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cerebral Cortex (New York, N.Y.: 1991)*, *23*(8), 1859–1873. https://doi.org/10.1093/cercor/bhs170

Bemis, D. K., & Pylkkänen, L. (2013c). Flexible composition: MEG evidence for the deployment of basic combinatorial linguistic mechanisms in response to task demands. *PLoS One*, *8*(9), e73949. https://doi.org/10.1371/journal.pone.0073949

Bever, T. (1970). The cognitive basis for linguistic structures. In J. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). Wiley. https://doi.org/10.1093/acprof:oso/9780199677139.003.0001

Blanco-Elorrieta, E., Ferreira, V. S., Del Prato, P., & Pylkkänen, L. (2018). The priming of basic combinatory responses in MEG. *Cognition*, *170*, 49–63. https://doi.org/10.1016/j.cognition.2017.09.010

Blank, I. A., & Fedorenko, E. (2020). No evidence for differences among language regions in their temporal receptive windows. *NeuroImage*, *219*, 116925. https://doi.org/10.1016/j.neuroimage.2020.116925

Brennan, J. (2016). Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass*, *10*, 299–313. https://doi.org/10.1111/lnc3.12198

Brennan, J., & Pylkkanen, L. (2012). The time-course and spatial distribution of brain activity associated with sentence processing. *Neuroimgae*, *60*, 1139–1148. https://doi.org/10.1016/j.neuroimage.2012.01.030

Brennan, J., Stabler, E., Van Wagenen, S., Luh, W., & Hale, J. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, *157–158*, 81–94. https://doi.org/10.1016/j.bandl.2016.04.008

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 134. https://doi.org/10.1038/s42003-022-03036-1

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179–194. https://doi.org/10.1006/nimg.1998.0395

Dobs, K., Isik, L., Pantazis, D., & Kanwisher, N. (2019). How face perception unfolds over time. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-09239-1

Flick, G., Oseki, Y., Kaczmarek, A. R., Al Kaabi, M., Marantz, A., & Pylkkänen, L. (2018). Building words and phrases in the left temporal lobe. *Cortex*, *106*, 213–236. https://doi.org/10.1016/j.cortex.2018.06.004

Flick, G., & Pylkkänen, L. (2020). Isolating syntax in natural language: MEG evidence for an early contribution of left posterior temporal cortex. *Cortex*, *127*, 42–57. https://doi.org/10.1016/j.cortex.2020.01.025

Friston, K. J., Price, C. J., Fletcher, P., Moore, C., Frackowiak, R. S. J., & Dolan, R. J. (1996). The trouble with cognitive subtraction. *NeuroImage*, *4*(2), 97–104. https://doi.org/10.1006/nimg.1996.0033

Goncalves, S. I., de Munck, J. C., Verbunt, J. P. A., Bijma, F., Heethaar, R. M., & Lopes da Silva, F. (2003). In vivo measurement of the brain and skull resistivities using an EIT-based method and realistic models for the head. *IEEE Transactions on Biomedical Engineering*, *50*(6), 754–767.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), Article 3. https://doi.org/10.1038/s41593-022-01026-4

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., & Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, *86*, 446–460. https://doi.org/10.1016/j.neuroimage.2013.10.027

Hagoort, P. (2005). On Broca, brain, and binding: A new framework. *Trends in Cognitive Sciences*, *9*(9), 416–423. https://doi.org/10.1016/j.tics.2005.07.004

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Proceedings of NAACL*, *2*, 159–166. https://doi.org/10.3115/1073336.1073357

Hämäläinen, M. S., & Ilmoniemi, R. J. (1994). Interpreting magnetic fields of the brain: Minimum norm estimates. *Medical & Biological Engineering & Computing*, *32*(1), 35–42. https://doi.org/10.1007/bf02512476

Hasson, U., & Egidi, G. (2015). What are naturalistic comprehension paradigms teaching us about language?

In R. M. Willems (Ed.), *Cognitive Neuroscience of Natural Language Use* (pp. 228–255). Cambridge University Press. https://doi.org/10.1017/cbo9781107323667.011
Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, *28*(10), 2539–2550. https://doi.org/10.1523/jneurosci.5487-07.2008
Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences*, *4*(4), 131–138. https://doi.org/10.1016/s1364-6613(00)01463-7
Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. https://doi.org/10.1038/nature17637
Kandylaki, K. D., & Bornkessel-Schlesewsky, I. (2019). From story comprehension to the neurobiology of language. *Language, Cognition and Neuroscience*, *34*(4), 405–410. https://doi.org/10.1080/23273798.2019.1584679
King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210. https://doi.org/10.1016/j.tics.2014.01.002
King, J.-R., Gramfort, A., Schurger, A., Naccache, L., & Dehaene, S. (2014). Two distinct dynamic modes subtend the detection of unexpected sounds. *PLoS One*, *9*(1), e85791. https://doi.org/10.1371/journal.pone.0085791
Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*, 203–205. https://doi.org/10.1126/science.7350657
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13
Law, R., & Pylkkänen, L. (2021). Lists with and without syntax: A new approach to measuring the neural processing of syntax. *Journal of Neuroscience*, *41*(10), 2186–2196. https://doi.org/10.1523/jneurosci.1179-20.2021
Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, *31*(8), 2906–2915. https://doi.org/10.1523/jneurosci.3684-10.2011
Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006
Lew, S., Wolters, C. H., Anwander, A., Makeig, S., & MacLeod, R. S. (2009). Improved EEG source analysis using low-resolution conductivity estimation in a four-compartment finite element head model. *Human Brain Mapping*, *30*(9), 2862–2878.
Li, J., & Pylkkänen, L. (2021). Disentangling semantic composition and semantic association in the left temporal lobe. *Journal of Neuroscience*, *41*(30), 6526–6538. https://doi.org/10.1523/jneurosci.2317-20.2021
Lyu, B., Choi, H. S., Marslen-Wilson, W. D., Clarke, A., Randall, B., & Tyler, L. K. (2019). Neural dynamics of semantic composition. *Proceedings of the National Academy of Sciences*, *116*(42), 21318–21327. https://doi.org/10.1073/pnas.1903402116
Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024
Matchin, W., & Hickok, G. (2020). The cortical organization of syntax. *Cerebral Cortex*, *30*(3), 1481–1498. https://doi.org/10.1093/cercor/bhz180
Matchin, W., Liao, C.-H., Gaston, P., & Lau, E. (2019). Same words, different structures: An fMRI investigation of argument relations and the angular gyrus. *Neuropsychologia*, *125*, 116–128. https://doi.org/10.1016/j.neuropsychologia.2019.01.019
Meyers, E. M. (2018). Dynamic population coding and its relationship to working memory. *Journal of Neurophysiology*, *120*(5), 2260–2268. https://doi.org/10.1152/jn.00225.2018
Nelson, M. J., Karoui, I. E., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S. S., Naccache, L., Hale, J. T., Pallier, C., & Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, *114*(18), E3669–E3678.
Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4
Oostendorp, T. F., Delbeke, J., & Stegeman, D. F. (2000). The conductivity of the human skull: Results of in vivo and in vitro measurements. *IEEE Transactions on Biomedical Engineering*, *47*(11), 1487–1492.
Parrish, A., & Pylkkänen, L. (2022). Conceptual combination in the LATL with and without syntactic composition. *Neurobiology of Language*, *3*(1), 46–66. https://doi.org/10.1162/nol_a_00048
Price, A. R., Bonner, M. F., Peelle, J. E., & Grossman, M. (2015). Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. *The Journal of Neuroscience*, *35*(7), 3276–3284. https://doi.org/10.1523/jneurosci.3446-14.2015
Pylkkänen, L. (2019). The neural basis of combinatory syntax and semantics. *Science*, *366*(6461), 62–66. https://doi.org/10.1126/science.aax0050
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. *24*. https://api.semanticscholar.org/CorpusID:160025533
Ségonne, F., Dale, A. M., Busa, E., Glessner, M., Salat, D., Hahn, H. K., & Fischl, B. (2004). A hybrid approach to the skull stripping problem in MRI. *NeuroImage*, *22*(3), 1060–1075.
Schmitt, L.-M., Erb, J., Tune, S., Rysop, A. U., Hartwigsen, G., & Obleser, J. (2021). Predicting speech from a cortical hierarchy of event-based time scales. *Science Advances*, *7*(49), eabi6070. https://doi.org/10.1126/sciadv.abi6070
Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118. https://doi.org/10.1073/pnas.2105646118
Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*(1), 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061
Stromswold, K., Caplan, D., Alpert, N., & Rauch, S. (1996). Localization of syntactic comprehension by positron

emission tomography. *Brain and Language*, *52*, 452–473. https://doi.org/10.1006/brln.1996.0024

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, *1*, 173–180. https://doi.org/10.3115/1073445.1073478

Wallentin, M., Nielsen, A. H., Vuust, P., Dohn, A., Roepstorff, A., & Lund, T. E. (2011). Amygdala and heart rate variability responses from listening to emotionally intense parts of a story. *NeuroImage*, *58*(3), 963–973. https://doi.org/10.1016/j.neuroimage.2011.06.077

Wehbe, L., Blank, I. A., Shain, C., Futrell, R., Levy, R., von der Malsburg, T., Smith, N., Gibson, E., & Fedorenko, E. (2021). Incremental language comprehension difficulty predicts activity in the language network but not the multiple demand network. *Cerebral Cortex*, *31*(9), 4006–4023. https://doi.org/10.1093/cercor/bhab065

Westerlund, M., & Pylkkänen, L. (2014). The role of the left anterior temporal lobe in semantic composition vs. semantic memory. *Neuropsychologia*, *57*, 59–70. https://doi.org/10.1016/j.neuropsychologia.2014.03.001

Zhang, L., & Pylkkänen, L. (2015). The interplay of composition and concept specificity in the left anterior temporal lobe: An MEG study. *NeuroImage*, *111*, 228–240. https://doi.org/10.1016/j.neuroimage.2015.02.028