Behavioral correlates of cortical semantic representations modeled by word vectors

Satoshi Nishida[1,2,*], Antione Blanc[1], Naoya Maeda[4], Masataka Kado[4], and Shinji Nishimoto[1,2,3]

[1]Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology (NICT), Suita, Osaka, Japan
[2]Graduate School of Frontier Biosciences,
[3]Graduate School of Medicine,
Osaka University, Suita, Osaka, Japan
[4]NTT DATA Corporation, Tokyo, Japan

[*]**Corresponding author**: Satoshi Nishida
1-4 Yamadaoka, Suita, Osaka 565-0871, Japan
s-nishida@nict.go.jp

Abstract

The quantitative modeling of semantic representations in the brain plays a key role in understanding the neural basis of semantic processing. Previous studies have demonstrated that word vectors, which were originally developed for use in the field of natural language processing, provide a powerful tool for such quantitative modeling. However, whether semantic representations revealed by the word vector-based models actually capture our perception of semantic information remains unclear, as there has been no study explicitly examining the behavioral correlates of the modeled semantic representations. To address this issue, we compared the semantic structure of nouns and adjectives estimated from word vector-based brain models with that evaluated from human behavior. The brain models were constructed using voxelwise modeling to predict the functional magnetic resonance imaging (fMRI) response to natural movies from semantic contents in each movie scene through a word vector space. The semantic dissimilarity of word representations was then evaluated using the brain models. Meanwhile, data on human behavior reflecting the perception of semantic dissimilarity between words were collected in psychological experiments. We found a significant correlation between brain model- and behavior-derived semantic dissimilarities of words. This finding suggests that semantic representations in the brain modeled via word vectors appropriately capture our perception of word meanings.

Keywords

Semantic processing, Voxelwise modeling, Natural language processing, FastText, Human, fMRI

Introduction

Natural language processing is a branch of machine learning that aims to develop machines that understand the meanings of words. In the field of natural language processing, a number of algorithms have been developed to capture the semantic representations of words from word statistics in large-scale text data as word vectors (e.g., Bojanowski et al., 2017; Deerwester et al., 1990; Lund and Burgess, 1996; Mikolov et al., 2013; Pennington et al., 2014). The word vectors obtained using these algorithms have effectively captured the latent semantic structure of words and further performed various types of natural language tasks, such as word similarity judgment (Bojanowski et al., 2017; Mikolov et al., 2013b; Pennington et al., 2014), sentiment analysis (Li et al., 2017; Tang et al., 2014), and question answering (Wang and Ittycheriah, 2015).

Furthermore, word vectors can be also used in neuroimaging studies to model semantic representations in the brain (Güçlü and van Gerven, 2015; Nishida et al., 2020a, 2015; Nishida and Nishimoto, 2018; Pereira et al., 2018; Ruan et al., 2016; Vodrahalli et al., 2018; Wehbe et al., 2014). These studies have reported that word vector-based models have the ability to predict the brain response evoked by semantic perceptual experiences (Güçlü and van Gerven, 2015; Nishida et al., 2020a, 2015; Ruan et al., 2016; Vodrahalli et al., 2018; Wehbe et al., 2014). These models are also able to recover perceived semantic contents from brain response (Nishida and Nishimoto, 2018; Pereira et al., 2018; Wang et al., 2020). These findings suggest that word vectors capture at least some aspects of the semantic representations in the brain. However, whether the semantic representations modeled by word vectors accurately reflect the semantic perception of humans is yet to be determined. In other words, no study has yet identified the behavioral correlates of the modeled semantic representations. This clarification is important in order to establish the brain modeling with word vectors as an accurate methodology for investigating human semantic processing.

To examine the behavioral correlates of the modeled semantic representations, we compared the semantic representational structure estimated by word vector-based models with that evaluated from human behavior. The modeled semantic representational structure was obtained using voxelwise modeling with a word vector space of fastText (Bojanowski et al., 2017; Fig. 1A). This voxelwise model predicts functional magnetic resonance imaging (fMRI) signals evoked by natural movies from the semantic contents in individual movie scenes in the fastText vector space. We then transformed fastText word vectors into brain representations using the model weights, and a brain-derived word dissimilarity matrix from

these brain representations was obtained (Fig. 1B). Meanwhile, the behavior-derived semantic representational structure was obtained from behavioral data in a psychological task, in which participants separately arranged tens of words (nouns and adjectives) in a two-dimensional space as per their semantic relationship (Fig. 2). This task was a modified version of a psychological task introduced previously (Kriegeskorte and Mur, 2012). A behavior-derived word dissimilarity matrix was then estimated using these behavioral data. Finally, we examined the correlation between the brain- and behavior-derived word dissimilarity matrices separately for both nouns and adjectives.

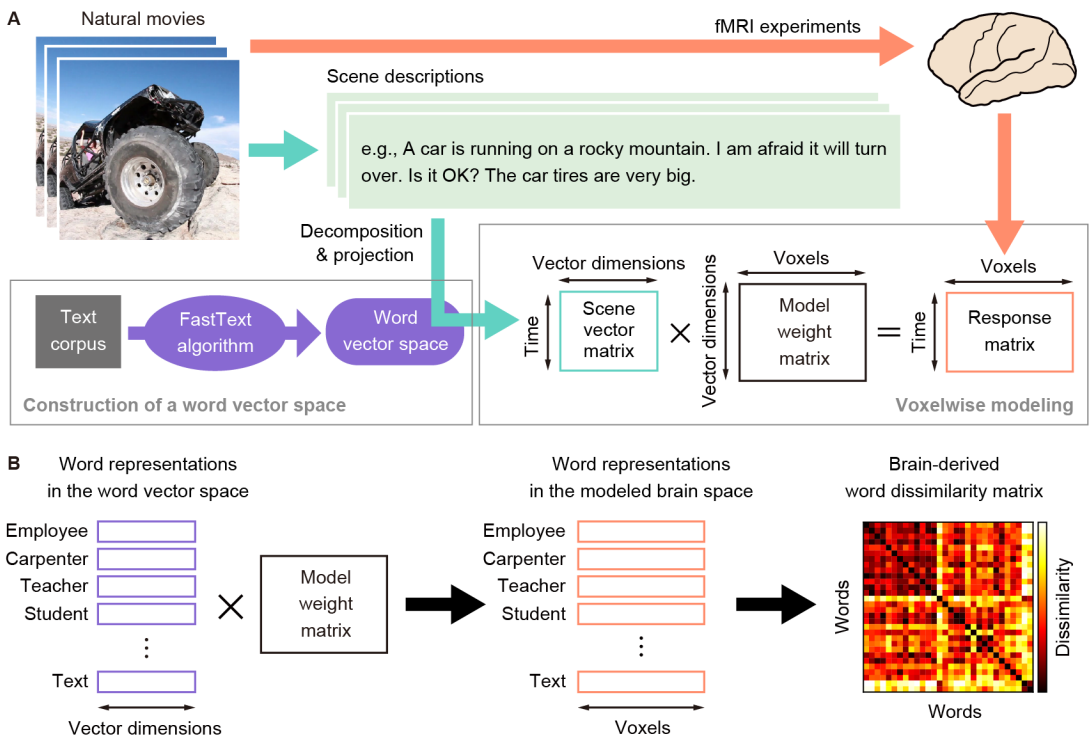

Fig. 1. Voxelwise modeling and brain-derived word dissimilarity. (A) Voxelwise modeling based on a fastText vector space. The model predicts fMRI signals evoked by natural movie scenes via a weighted linear summation of the vector representations of semantic descriptions of each scene (scene vectors). Scene vectors were obtained by transforming manual descriptions of each movie scene through a fastText vector space pretrained using a text corpus. The weights of the linear prediction model were trained using the corresponding time series of movies and fMRI signals of each brain. (B) Estimation of brain-derived word dissimilarity. Word representations in the fastText vector space were transformed into word representations in the modeled brain space by multiplying fastText word vectors by the model weights. Then, the correlation distance of the modeled word representations between all possible pairs of words was calculated,

producing a brain-derived word dissimilarity matrix.

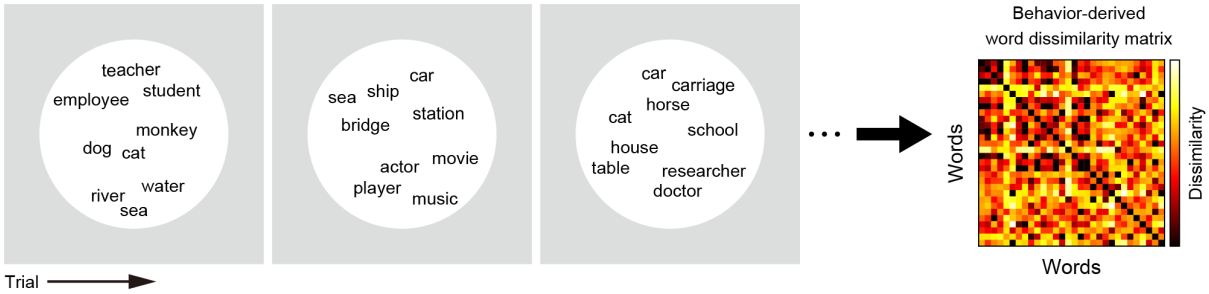

Fig. 2. Word-arrangement task and behavior-derived word dissimilarity. To evaluate the word dissimilarity structure derived from human behavior, we conducted psychological experiments in which each participant performed a word-arrangement task. On each trial of this task, participants were required to arrange multiple (≤60) words in a two-dimensional space according to the semantic relationship of those words. After each participant completed ≤1 h of this task, a behavior-derived word dissimilarity matrix was established using inverse multidimensional scaling (see Material and Methods for more details).

1. Material and methods

1.1. Participants

In total, 52 healthy Japanese participants (21 females and 31 males; age 20–61, mean ± SD = 26.8 ± 8.9 years) were recruited for the 2 sets of fMRI experiments. Among the recruits, 36 participated in one or the other of these fMRI experiments, and 16 participated in both. In addition, another 36 healthy Japanese participants (20 females and 16 males; age 18–58, mean ± SD = 25.9 ± 10.1 years) were recruited for psychological experiments. Of these, 6 were also participants of the fMRI experiments and 30 were unique participants. All participants had normal or corrected-to-normal vision. Written informed consent was obtained from all of the participants. The experimental protocol was approved by the ethics and safety committees of the National Institute of Information and Communications Technology. The fMRI data, but not the psychological data, used here were also utilized in a previous publication (Nishida et al., 2020b).

1.2. MRI experiments

Functional and anatomical MRI data were collected via a 3T Siemens MAGNETOM Prisma scanner (Siemens, Germany), with a 64-channel Siemens volume coil. Functional data were

collected using a multiband gradient echo EPI sequence (Moeller et al., 2010; TR = 1000 ms; TE = 30 ms; flip angle = 60 °; voxel size = 2 × 2 × 2 mm; matrix size = 96 × 96; FOV = 192 × 192 mm; the number of slices = 72; multiband factor = 6). Anatomical data were also gathered using a T1-weighted MPRAGE sequence (TR = 2530 ms; TE = 3.26 ms; flip angle = 9 °; voxel size = 1 × 1 × 1 mm; matrix size = 256 × 256; FOV = 256 × 256 mm; the number of slices = 208) on the same 3T scanner.

In the two sets of fMRI experiments, participants were asked to view movie stimuli on a projector screen inside the scanner (27.9 × 15.5 of visual angle at 30Hz) and used MR-compatible headphones for the sounds. The participants were given no explicit task. The fMRI data for each participant upon viewing of the movies were collected in three separate recording sessions over 3 days for each set of fMRI experiments.

The movie stimuli consisted of Japanese television advertisements for one set of experiments (movie set 1) and Japanese web advertisements for the other set of experiments (movie set 2; see also Nishida et al., 2020b). Movie set 1 has included 420 ads broadcasted on Japanese TV between 2011 and 2017; meanwhile, movie set 2 included 368 ads broadcasted on the Internet between 2015 and 2018. The ad movies were all unique, include a wide variety of product categories (see Table S1 for more details). The length of each movie was 15 or 30 s. To create the movie stimuli for each experiment, the original movies in each of the movie sets 1 and 2 were sequentially concatenated in a pseudo-random order. For each movie set, 14 non-overlapping movie clips of 610 s in length were obtained.

Individual movie clips were then displayed in separate scans. The initial 10 s of each clip served as a dummy in order to discard hemodynamic transient signals caused by clip onset. fMRI responses collected during the 10 s dummy part were not used for modeling. Twelve clips from each movie set were only presented once. The fMRI responses to these clips were used for the training of the voxelwise models (training dataset; 7200 s in total). The other two clips for each movie set were presented four times each in four separate scans. The fMRI responses to these clips were then averaged across four scans, which were further used for the test of the voxelwise models (test dataset; 1200 s in total).

1.3. MRI data pre-processing

Motion correction in each functional scan was carried out via the statistical parameter mapping toolbox (SPM8, http://www.fil.ion.ucl.ac.uk/spm/software/ spm8/). For each

subject, all volumes were aligned to the first image from the first functional run. Low-frequency fMRI response drift was then eliminated by subtracting median-filtered signals (within a 120-s window) from raw signals. Then, the response for each voxel was normalized by subtracting the mean response and scaling to the unit variance. FreeSurfer (Dale et al., 1999; Fischl et al., 1999) was used to identify cortical surfaces from anatomical data and register these surfaces to the voxels of functional data. Then, each voxel was assigned to one cortical region based on the atlas for cortical segmentation (Destrieux et al., 2010).

1.4. Word vector spaces

This study used the skip-gram algorithm of fastText (Bojanowski et al., 2017) to construct a word vector space. This algorithm has been originally developed to learn a word vector space based on nearby word co-occurrence statistics in natural language texts. A fastText vector space was constructed from a text corpus of the Japanese Wikipedia dump on April 1, 2020. All Japanese texts in the corpus were segmented into words and lemmatized using MeCab (http://taku910.github.io/mecab). Used were only nouns, verbs, and adjectives. In an attempt to improve the reliability of fastText learning, the vocabulary size was restricted to ~100,000 words by excluding words that appeared fewer than 262 times in the corpus. The learning parameters of fastText used were as follows: window size = 10; the number of negative samples = 5; downsampling rate = 0.001; and the number of learning epochs = 10. The fastText vector spaces were constructed using five different numbers of vector dimensions (100, 300, 500, 1000, and 2000). Of these, the 1000-dimension vector space was used for the main analysis (Figs. 3–6), and the other dimensions were used to test the effect of vector dimensionality on modeling performance and behavioral correlates. To eliminate the effects of trial variations in the learning quality of word vector spaces on the performance of voxelwise modeling, five different fastText spaces were learned independently using the same text corpus and the same parameters. All results were obtained from the average over five learned spaces for each parameter set.

1.5. Movie scene descriptions

Manual scene descriptions using natural Japanese language were provided for every 1-s scene of each movie in movie sets 1 and 2, in a manner similar to that described previously (Nishida et al., 2020b, 2020a; Nishida and Nishimoto, 2018). The annotators were native Japanese speakers (movie set 1: 68 females and 28 males, age 19 –62 years; movie set 2: 11 females and 2 males, age 20–56 years), who were not the fMRI participants. They were instructed to describe each scene (the middle frame of each 1-s clip) using more than 50

Japanese characters. Multiple annotators (movie set 1, 12–14 annotators; movie set 2, 5 annotators) were randomly assigned for each scene to reduce the potential effect of personal bias. The descriptions contain a variety of expressions reflecting not only objective perceptions but also the subjective perceptions of the annotator (e.g., impression, feeling, association with ideas; for more details, see the Supplementary Methods in Nishida et al., 2020b, https://osf.io/3hkwd).

Each description for a given scene was also segmented, lemmatized, and decomposed into nouns, verbs, and adjectives via MeCab as described above; then, they were transformed into fastText vectors. The word vectors were then averaged within each description. For each scene, all vectors obtained from the different descriptions were averaged. Through this procedure, a single vector (scene vector) was obtained for each 1-s scene, which was later used for modeling.

1.6. Voxelwise modeling

The procedure of voxelwise modeling was similar to that described previously (Nishida et al., 2020a). A series of fMRI signals evoked in each voxel by a series of movie scenes was modeled as a weighted linear combination of a scene-vector matrix plus isotropic Gaussian noise. A set of linear temporal filters was then used to model the slow hemodynamic response and its coupling with the brain response (Nishimoto et al., 2011). In an attempt to capture the hemodynamic delay in the responses, the scene-vector matrix was constructed by concatenating three sets of scene vectors with delays of 3, 4, 5, and 6 s. The weighted matrix was estimated using an L2-regularized linear least-squares regression, which can obtain good estimates even for models containing a large number of regressors (Huth et al., 2012).

In order to estimate the regularization parameters, the training dataset was randomly divided into two subsets containing 80 % and 20 % of the samples, for model fitting and validation, respectively. This random resampling procedure was repeated 10 times. Regularization parameters were optimized, according to the mean Pearson's correlation coefficient between the predicted and measured fMRI signals for the 20 % validation samples. An optimal parameter was obtained separately for each model.

The brain-response prediction performance of the model was evaluated using the test dataset, which was not used for model fitting or parameter estimation. The prediction accuracy of the model was quantified as the Pearson's correlation coefficient between the predicted and

average measured fMRI signals in the test dataset.

1.7. Psychological experiments

Participants were asked to perform a word-arrangement task on a PC. This task is considered to be a modified version of the psychological task introduced previously (Kriegeskorte and Mur, 2012). Kriegeskorte and colleagues originally used this task to examine the behavioral correlates of cortical object representations, which were quantified using the representational similarity of fMRI signals evoked by object stimuli (Charest et al., 2014; Mur et al., 2013). The present study used this task to study the methodological validity of word vector-based brain modeling by testing the behavioral correlates of modeled cortical semantic representations. The psychological data are available online (https://osf.io/um3qg/).

In each trial of this task, the participants were required to arrange ≤60 words (nouns or adjectives) in a two-dimensional space according to their semantic relationship on a computer screen by mouse drag-and-drop operations (Fig. 2). This paradigm has allowed for the efficient collection of the perceptual semantic dissimilarity between words from participants (Kriegeskorte and Mur, 2012).

The words used in this task included 60 nouns and 60 adjectives in Japanese (Table 1). These words were selected from the vocabulary of the fastText space (i.e., top 100,000 most frequently used words in the Japanese Wikipedia corpus). Nouns were selected in terms of the following six semantic categories: humans, non-human animals, non-animal natural things, constructs, vehicles, and other artifacts. Ten words were selected for each category. For adjectives, such category-based selection was deemed difficult due to the small number of adjectives in the vocabulary (only 473 words). Nouns and adjectives were separately used in two distinct sessions of the task, which was performed over 2 days.

The words were arranged in a designated circular area on the computer screen ("arena"). The words were initially displayed outside the arena. The participants used mouse drag-and-drop operations in moving each word item into the arena and arranging them. The arrangement of words was according to the participants' own judgment of semantic similarity and dissimilarity between words. More similar or dissimilar word pairs should be closer or further apart in the arena, respectively. The participants were allowed to move any words within the arena as many times as they wished. Once all words were moved into the arena, the participants could click a button marked "Next" anytime to move on to the next trial.

On the initial trial of each session, the participants have arranged the entire set of 60 words (nouns or adjectives). On subsequent trials, they arranged subsets of those 60 words. After the end of each trial, the rough estimate of the word dissimilarity matrix and the evidence (0–1) for each word pairwise dissimilarity were computed as described previously (Kriegeskorte and Mur, 2012). Words in the subsets chosen for each trial were determined so as to increase evidence for pairwise dissimilarities of the words whose evidence estimated on the previous trial was the weakest. The session continued until the evidence of every word pair dissimilarity was below a threshold (0.75) or until the total duration of the session approached 1 h.

Table 1. Words used in psychological experiments and for calculating word dissimilarity matrices. Original words were in Japanese, and, for display purposes, here they were translated into English.

Nouns	Adjectives
Employee, carpenter, teacher, student, driver, player actor, writer, doctor, researcher, monkey, dog, cat, horse, cow, sheep, bird, frog, fish, bug, star, sky, mountain, sea, river, water, flower, tree, dirt, vegetable, house, school, factory, hospital, station, airport, bridge, hotel, stadium, park, car, motorbike, train, ship, boat, airplane, helicopter, rocket, bicycle, carriage, movie, book, music, phone, PC, TV, cloth, table, cup, text	Many, high, strong, large, absent, good, near, long, new, few, deep, broad, low, young, bad, bold, small, short, red, early, white, beautiful, black, old, weak, intense, far, detailed, childish, tender, bright, quick, correct, strict, blue, slow, sweet, thick, narrow, difficult, dim, rare, thin, heavy, light, pleasant, slender, wide, dark, hot, scary, shallow, sharp, circular, interesting, pale, sad, equal, remarkable, terrific

1.8. Word dissimilarity matrix

The brain-derived word dissimilarity matrix was estimated using the following procedure: first, the voxelwise model weights for each participant were averaged across the three sets of hemodynamic delay terms. Second, all or top N voxels with the highest prediction accuracy were then selected from the model weights. All voxels were used in the main analysis, and selected voxels (N = 2000, 5000, 10000, 30000, and 50000) were used in an additional analysis that tested the effect of voxel selection on brain–behavior correlation. Third, the

fastText vectors for the words used in the psychological experiments (Table 1) were multiplied by the weight matrix, which yielded the word representations in the modeled brain space. Finally, a word dissimilarity matrix was calculated by the correlation distance (1 – Pearson's r coefficient) between these word representations, which was averaged across participants separately for nouns and adjectives.

The behavior-derived word dissimilarity matrix was evaluated from multiple word-subset arrangements on the word-arrangement task. The dissimilarity for a given word pair was estimated as a weighted average of the scale-adjusted dissimilarity estimates from individual arrangements as has been described previously (Kriegeskorte and Mur, 2012). These estimates produced a word dissimilarity matrix for the entire set of words. In this way, the word dissimilarity matrix was estimated separately for nouns and adjectives and was averaged across participants.

In addition, a word vector-derived word dissimilarity matrix was also examined in order to test whether the behavioral correlates of semantic dissimilarity structures changed through the transformation from original word vector representations to brain word representations. The word vector-derived matrix was computed by the correlation distance between fastText vectors of the same word sets (Table 1) separately for nouns and adjectives.

The behavioral correlates of brain- or word vector-derived data were evaluated using Pearson's correlation of the brain- or word vector-derived word dissimilarity matrix with the behavior-derived matrix. The lower triangular portion of each matrix was used for calculating correlation coefficients. A bootstrapping test was performed to statistically compare correlation coefficients. In each repetition of this test, word pairs were randomly resampled from each word dissimilarity matrix with replacement. Then, the difference of correlation coefficients between different pairs of word dissimilarity matrices was calculated. This procedure was repeated 10,000 times to obtain the distribution of correlation coefficient differences between different matrix pairs and estimate the p values.

2. Results

2.1. Performance of voxelwise models based on word vectors

To determine whether the voxelwise models based on a fastText vector space appropriately capture semantic representations in the brain, we first tested the performance of the models in terms of brain-response prediction. The performance was evaluated using the following two

measures: (1) prediction accuracy calculated as the correlation coefficients between predicted and measured brain responses in the test dataset, and (2) the fraction of significant voxels that reach their prediction accuracy to a threshold value ($P < 0.05$ after the correction for multiple comparisons using the false discovery rate [FDR]). The prediction accuracy averaged across all cortical voxels and across participants was 0.0928 for movie set 1 (28 participants) and 0.0865 for movie set 2 (40 participants); no participant showed mean prediction accuracy less than 0 (Fig. 3A and C). This accuracy was sufficiently high because it was averaged over all cortical voxels, and previous studies on voxelwise modeling have reported a similar tendency (Huth et al., 2016; Naselaris et al., 2015; Nishida et al., 2020b). The fraction of significant voxels averaged across participants was 0.575 for movie set 1 and 0.526 for movie set 2 (Fig. 3B and D). These results indicate that the voxelwise models trained showed sufficient performance in the modeling of semantic representations consistently across different stimulus sets.

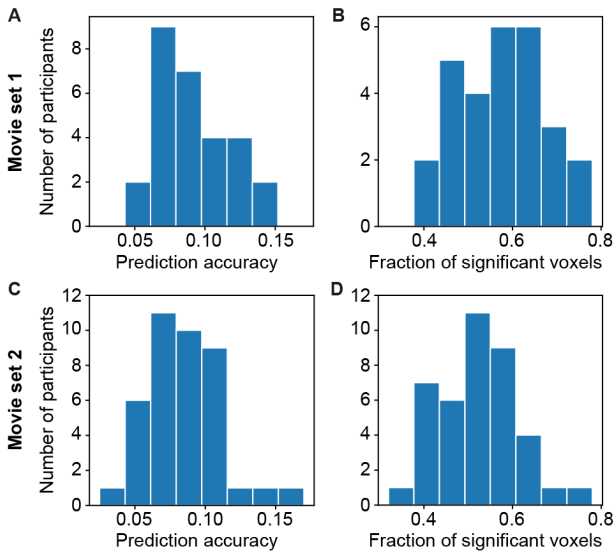

Fig. 3. Performance of voxelwise models in brain-response prediction. The performance of voxelwise models (vector dimension = 1000) was evaluated in terms of predicting brain responses in the test dataset. For this purpose, prediction accuracy and the fraction of significant voxels were calculated for each brain. The distribution of prediction accuracy (A and C) and that of the fraction of significant voxels (B and D) were separately shown for movie sets 1 (A and B) and 2 (C and D).

Although the vector dimensionality of the fastText vector space had little effect on the model performance, there was no clear tendency toward vector-dimensionality dependency (Fig.

S1). The vector space of 100 dimensions has showed the highest performance for movie set 1, whereas the vector space of 2000 dimensions showed the highest performance for movie set 2. In addition, the change of model performance was not monotonic against the change of vector dimensions.

2.2. Localization of cortical regions highly predicted by voxelwise models

We next identified which cortical regions were predictable by the word vector-based models in order to determine that the models could capture the semantic representations in appropriate cortical regions. For this purpose, the prediction accuracy averaged across participants was mapped onto the cortical surface of a reference brain (Fig. 4A–B). The highly predictable regions were localized over widespread cortical regions, including the occipital, superior and inferior temporal, and posterior parietal regions, for both movie sets. This localization was consistent with previous reports, in which the semantic representations in the brain were modeled with word features (Huth et al., 2012; Nishida et al., 2020a). There was a strong correlation of mean prediction accuracy over 150 cortical regions between movie sets 1 and 2 (Pearson's $r = 0.974$); this indicates the consistency of model predictability of these cortical regions across stimulus sets. In addition, there were similar tendencies across vector dimensionality (Fig. S2 and Table S2). These results indicate that the voxelwise models trained reliably showed the localization of highly predictable regions.

A similar localization was also observed when the fraction of significant voxels in each cortical region was mapped onto the cortical surface (Fig. 4C–D). The fraction was relatively large in the occipital, superior and inferior temporal, and posterior parietal regions compared with the other regions. This localization pattern was observed to be highly consistent across both movie sets and vector dimensions (Fig. S3 and Table S3). Taken together, these results suggest that the word vector-based models capture the semantic representations in appropriate cortical regions.

It should be noted that, however, the fraction of significant voxels was not small even in the other cortical regions, including the prefrontal cortex. The minimum fraction within any individual region was 0.237, which is significantly higher than 0.05 (corresponding to the significance level of 0.05; Wilcoxson test, $P < 0.00001$). This result suggests that although high prediction accuracy was observed in specific regions (Figs. 4A–B and S2), the word vector-based model can potentially capture semantic information even from other regions across the cortex.

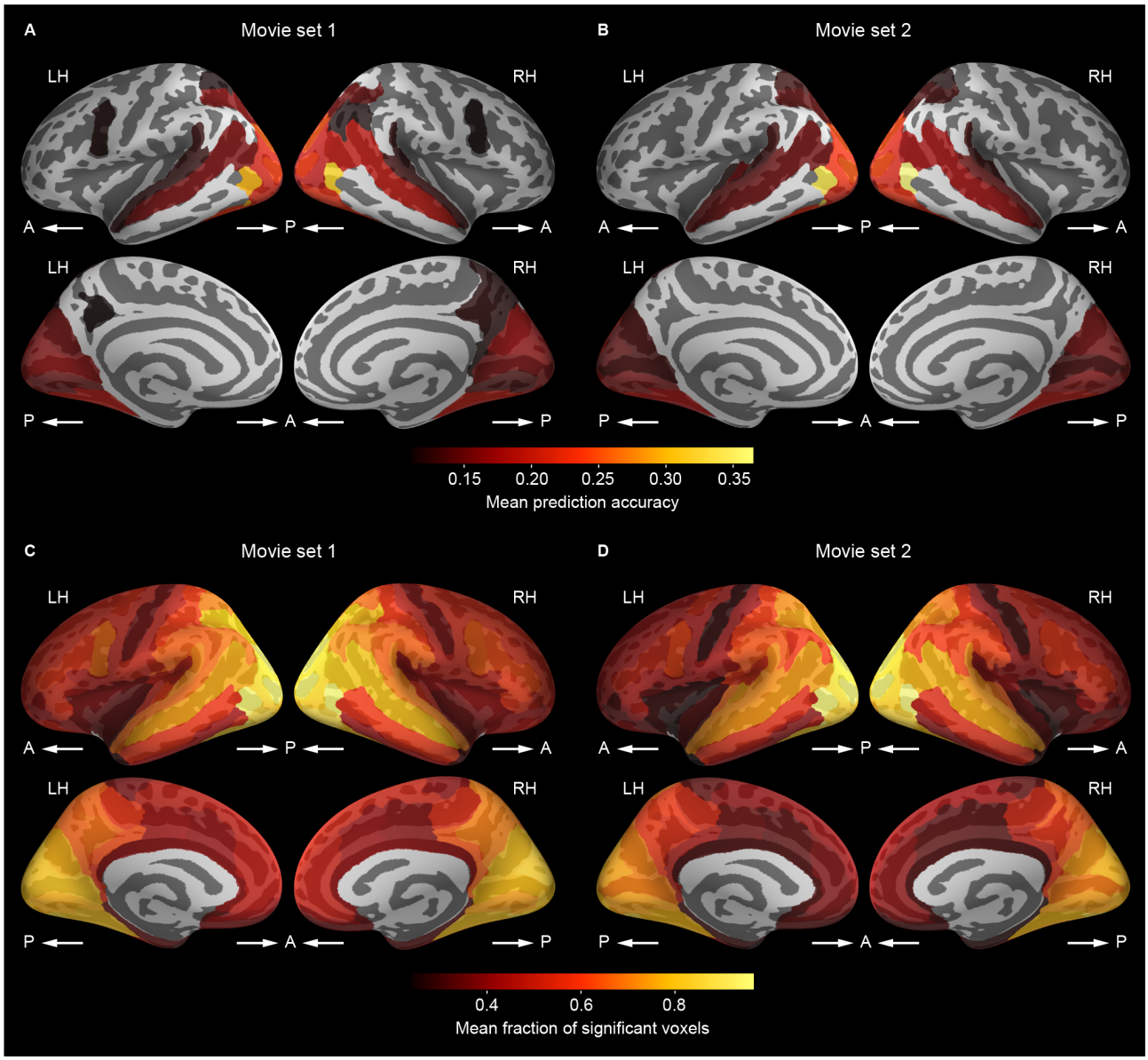

Fig. 4. Cortical mapping of model performance. (A and B) Participant-averaged prediction accuracy of voxelwise models (vector dimension = 1000) mapped onto the cortical surface of a reference brain for each of the movie sets 1 (A) and 2 (B). The prediction accuracy was averaged within each of the cortical regions that were anatomically segmented. Brighter colors in the surface maps indicate cortical regions that have higher prediction accuracy. Only regions with mean prediction accuracy above 0.11 (corresponding to P = 0.0001 ~ 0.05/150 regions) are shown. (C and D) Fraction of significant voxels of the same models mapped onto the same cortical surface separately for movie sets 1 (C) and 2 (D). The fraction was then computed within each cortical region. Brighter colors indicate regions that have larger fraction. LH, left hemisphere; RH, right hemisphere; A, anterior; P, posterior.

2.3. Correlation between brain- and behavior-derived word dissimilarities

Finally, we tested the correlation between the word dissimilarity matrix derived from the voxelwise models and that derived from behavioral data to clarify the behavioral correlates of modeled semantic representations. Word representations in the modeled brain space were calculated by multiplying the fastText word vectors by model weights, and the brain-derived word dissimilarity matrix was obtained from the correlation distance between all possible pairs of word representations (Fig. 1B). Meanwhile, the behavior-derived word dissimilarity matrix was measured from the behavioral data from the word-arrangement task (Fig. 2), in which 36 participants completed 18.8 (SD = 10.7) trials on average for each session.

These brain- and behavior-derived word dissimilarity matrices were constructed by averaging the matrices over all model or behavioral data separately for nouns (Fig. 5A) and adjectives (Fig. 5B). We found significant correlations between these two types of word dissimilarity matrices. For nouns, the correlation coefficients were consistently high across movie sets ($r = 0.612$ and 0.635 for movie sets 1 and 2, respectively; $P < 0.00001$). For adjectives, the correlation coefficients were deemed smaller but significant for both movie sets ($r = 0.342$ and 0.358 for movie sets 1 and 2; $P < 0.00001$). In these cases, the populations from which we obtained these dissimilarity matrices were different between brain- and behavior-derived data. However, we observed a significant correlation for both nouns and adjectives and for both movie sets even when these matrices were collected from the same population ($P < 0.00001$; Fig. S4). In addition, significant correlations were consistently observed regardless of vector dimensionality (Fig. S5A and C); although the correlation coefficients differed across vector dimensions, no clear tendency of changes in correlation coefficients across vector dimensions was determined (Fig. S5B and D). These results indicate that the brain-derived dissimilarity structure of semantic representations correlates with the behavior-derived dissimilarity structure.

To determine whether the behavioral correlates of word dissimilarity change through the transformation from original fastText word vector representations to brain representations, we evaluated the correlation of a word vector-derived word dissimilarity matrix with the behavior-derived matrix. We found that correlation coefficients between the word vector- and behavior-derived matrices were significantly lower than those between the brain- and behavior-derived matrices (Fig. 5; $r = 0.318$ and 0.250 for nouns and adjectives, respectively; bootstrapping test, $P < 0.0001$). This observation was consistent across vector dimensions

(except for the vector dimension = 100 for adjectives; Fig. S5A and C). These results suggest that the semantic representations of word vectors become more similar to human perception as they are transformed into brain semantic representations.

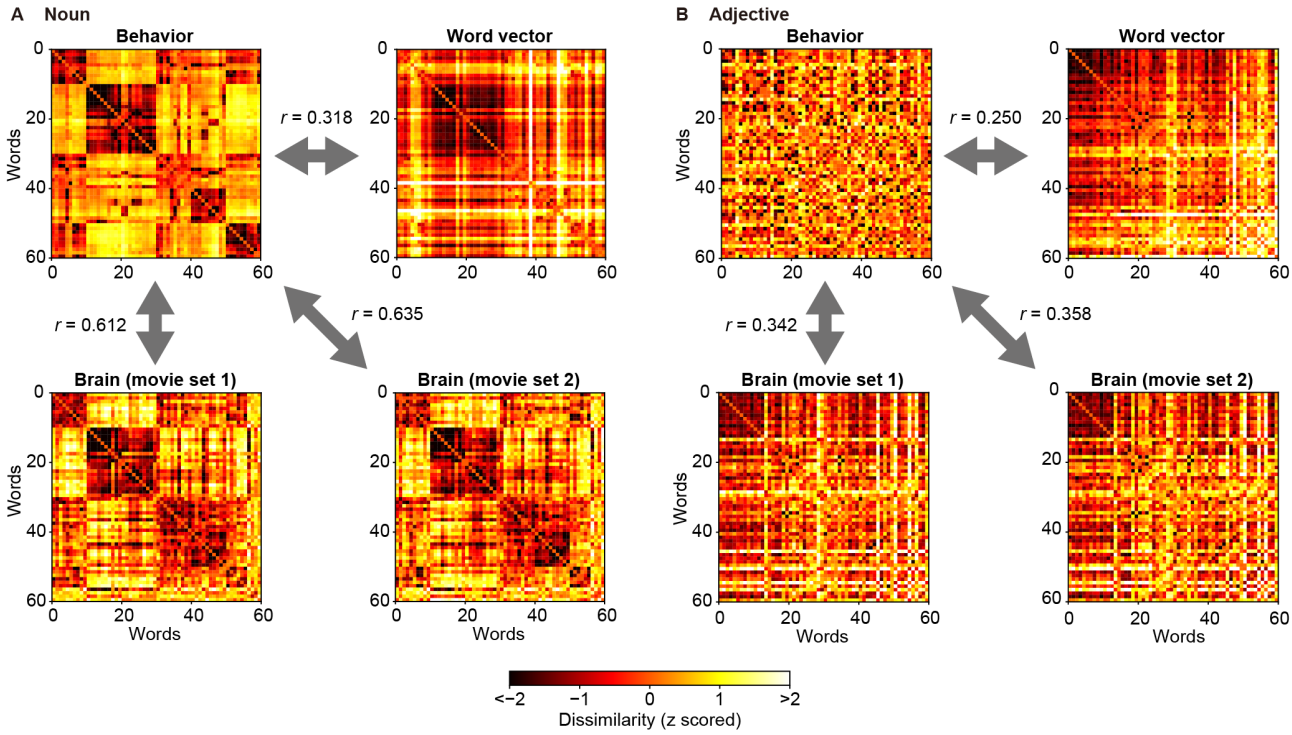

Fig. 5. Correlations between dissimilarity matrices. We have constructed word dissimilarity matrices separately for nouns (A) and adjectives (B). Each color map shows the behavior-derived matrices obtained from behavioral data (top left in A and B), the word vector-derived matrices obtained directly from the fastText vector space (top right), or the brain-derived matrices obtained from voxelwise models (vector dimension = 1000) for each movie set (bottom). Brighter colors in each map indicate higher dissimilarity of word pairs. Significant correlations were identified between behavior- and word vector-derived matrices and between behavior- and brain-derived matrices ($P < 0.0001$). Pearson's correlation coefficients (r) are indicated.

The brain-derived dissimilarity matrices in Figs. 5, S4, and S5 were calculated from the model weights of all cortical voxels (the number of voxels was 54914–73301 [mean ± SD = 62245 ± 5018] for movie set 1 and 52832–73301 [mean ± SD = 62150 ± 5190] for movie set 2). However, the high performance of the voxelwise models was observed in localized cortical regions (Figs. 4, S2, and S3). Hence, the correlation of brain- and behavior-derived dissimilarity matrices may be stronger when only the cortical regions with high model

performance were used. To test this possibility, we constructed brain-derived dissimilarity matrices using only those voxels with the highest prediction accuracy (top 2000, 5000, 10000, 30000, or 50000 voxels). Then, their correlations with behavior-derived matrices were compared with the original correlations calculated using all cortical voxels. However, we found the strongest correlation when using all cortical voxels (Figs. 6 and S6). This result suggests that semantic information correlated with behavior are distributed broadly across the cortex and can be captured using word vector-based voxelwise models.

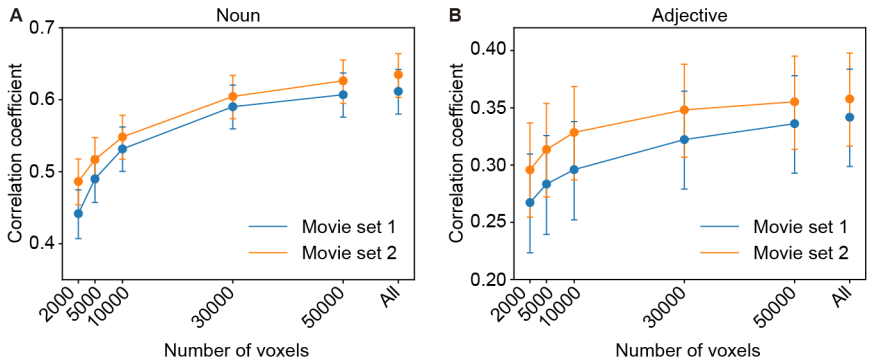

Fig. 6. Effects of voxel selection on brain–behavior correlation. We have evaluated the correlations between brain- and behavior-derived word dissimilarity matrices when the brain-derived matrix was obtained from the selected voxels with the highest model accuracy. The correlation coefficients are shown separately for nouns (A) and adjectives (B), while the numbers of selected voxels are changed (2000, 5000, 10000, 30000, 50000, and all voxels). Error bars indicate 95 % confidence interval (CI) estimated by bootstrapping.

3. Discussion

We have examined the behavioral correlates of semantic representations estimated by word vector-based brain models. We constructed a voxelwise model that has the ability to predict movie-evoked fMRI signals in individual brains through a fastText vector space. The voxelwise models showed substantial response-prediction performance in reasonable cortical regions consistently across stimulus and parameter sets. There were significant correlations between the word dissimilarity structure derived from the voxelwise models and that derived from behavioral data. These results suggest that the semantic representations estimated from word vector-based brain models can appropriately capture the semantic relationship of words in human perception.

Word vectors have been extensively leveraged for the modeling of semantic information in the human brain. Some studies have successfully predicted brain responses, through word vector spaces, evoked by audiovisual stimuli, such as movies (Nishida et al., 2020b, 2015; Vodrahalli et al., 2018), pictures (Güçlü and van Gerven, 2015; Ruan et al., 2016), and sentences (Wehbe et al., 2014). In addition, one of these studies has visualized the semantic representational space in the brain through word vector-based modeling (Nishida et al., 2015). Another line of studies has successfully recovered perceived semantic contents from brain responses by using word vector spaces (Nishida and Nishimoto, 2018; Pereira et al., 2018; Wang et al., 2020). In this way, word vectors will be a useful tool to investigate comprehensively the semantic processing in the human brain. However, no study has explicitly clarified the behavioral correlates of the brain semantic representations revealed by word vector-based models. Although previous studies in which movie-evoked brain responses were predicted from manual descriptions for the movie scenes (Nishida and Nishimoto, 2018; Vodrahalli et al., 2018) employed human behavior implicitly, they did not reveal the association between brain- and behavior-derived semantic representations. To the best of our knowledge, this is the first study demonstrating the behavioral correlates of semantic representations modeled by word vectors.

One of the advantages of word vector-based modeling is that semantic representations can be quantified using much more vocabularies (Nishida and Nishimoto, 2018) than modeling based on discrete word features (e.g., Huth et al., 2012). This allows the quantification of different aspects of semantic information in the brain by using different types of words. In particular, adjectives have a potential for the visualization of impression information (Nishida and Nishimoto, 2018). We found the behavioral correlates of modeled semantic representations on nouns and even adjectives, which were consistently observed across stimulus and parameter sets (Figs. 5, S4, and S5). These findings indicate that word vectors are deemed suitable for modeling the cortical representations of impression information. Thus, word vector-based modeling provides a useful framework for extensive investigation of semantic representations in the brain.

In this study, we found localized cortical regions in which brain responses were accurately predicted by our models (Figs. 2, 4, S2, and S3). Although these brain regions, primarily the posterior and temporal cortices, are consistent with previous reports on the modeling of semantic representations (Huth et al., 2012; Nishida et al., 2020a), it may be argued that these regions do not encompass the entire language network, including the left inferior frontal and

superior temporal cortices (Friederici and Gierhan, 2013). However, a recent study has clearly demonstrated that language network is not necessarily involved in semantic processing that requires no explicit linguistic processing (Ivanova et al., 2019). They reported that aphasia patients with physical damage to the language network manifested severely declined performance in semantic tasks with linguistic information, but performance equivalent to healthy controls in semantic tasks with nonlinguistic information. This notion is also supported by evidence that semantic models using word features can predict brain response in the language network when the response is evoked by linguistic stimuli, e.g., sentences and speech (Deniz et al., 2019; Huth et al., 2016; Wehbe et al., 2014); however, it was not applicable when the response is evoked by nonlinguistic stimuli, e.g., movies (Huth et al., 2012; Nishida et al., 2020a). Thus, the semantic representations modeled using the brain response evoked by nonlinguistic stimuli may be biased toward specific modalities, such as vision and audition.

In the field of natural language processing, recently developed algorithms based on deep learning have exhibited state-of-the-art performance in many types of natural language tasks (e.g., Devlin et al., 2018; Peters et al., 2018; Yang et al., 2019). The feature representations obtained by these algorithms are also beginning to be used for the modeling of semantic representations (Jain and Huth, 2018; Jat et al., 2019), which show state-of-the-art performance even in brain-response prediction (Schrimpf et al., 2020). Hence, it is possible that the correlation between brain- and behavior-derived semantic representations improves by using such deep learning features instead of word vectors. However, some advantages are noted using word vectors compared with deep learning features: First, word vectors are easily trained with much smaller sets of parameters than deep learning features. Second, the internal processing of word vector algorithms is more explainable than that of deep learning-based algorithms. Third, there are many existing methods to improve word-vector spaces for specific types of tasks (Barrett et al., 2018; Faruqui et al., 2014; Fyshe et al., 2014; Ling et al., 2015; Liu et al., 2015; Ruan et al., 2016). These advantages may be also beneficial for the modeling of semantic representations. Therefore, it is important to use these two types of algorithms separately depending on the aims of specific studies on semantic processing.

4. Conclusions

This study provides the first evidence on the behavioral correlates of semantic representations modeled by word vectors. The word vector-based models capture semantic information distributed across the cortex, and the dissimilarity structure of modeled semantic

representations appropriately reflects human perception of noun and adjective meanings. These findings contribute to the establishment of word vector-based brain modeling as a powerful tool in investigating human semantic processing.

Acknowledgments: The work was supported by JSPS KAKENHI Grant-in-Aid for Early-Career Scientists (18K18141) and for Young Scientists A (15H05311), and JST ERATO (JPMJER1801) from MEXT. We thank Ms. Hitomi Koyama, Mr. Yusuke Nakano, Mr. Koji Takashima, Mr. Takeshi Matsuda, Mr. Susumu Minamiyama, Ms. Mami Yamashita, Mr. Ryo Yano, Ms. Risa Matsumoto, Mr. Masato Okino, and Mr. Akira Nagaoka for their analytical and experimental support.

Author contributions: S. Nishida designed the research. S. Nishida and A.B. designed and performed the experiments. S. Nishida and A.B. performed the analysis. S. Nishida, A.B., N.M., M.K., and S. Nishimoto wrote the manuscript.

Competing Interests: This study was funded by NTT Data Corp. NM and MK are employees of NTT Data Corp.

References

Barrett, M., Bingel, J., Hollenstein, N., Rei, M., Søgaard, A., 2018. Sequence classification with human attention. CoNLL 2018 - 22nd Conf. Comput. Nat. Lang. Learn. Proc. 302–312. https://doi.org/10.18653/v1/k18-1030

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. 5, 135–146.

Charest, I., Kievit, R.A., Schmitz, T.W., Deca, D., Kriegeskorte, N., 2014. Unique semantic space in the brain of each beholder predicts perceived similarity. Proc. Natl. Acad. Sci. U. S. A. 111, 14545–14570. https://doi.org/10.1073/pnas.1402594111

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

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., 1990. Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. 41, 391–407.

Deniz, F., Nunez-Elizalde, A.O., Huth, A.G., Gallant, J.L., 2019. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. J. Neurosci. 39, 7722–7736. https://doi.org/10.1523/JNEUROSCI.0675-19.2019

Destrieux, C., Fischl, B., Dale, A., Halgren, E., 2010. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage 53, 1–15. https://doi.org/10.1016/j.neuroimage.2010.06.010

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv 1810.04805.

Faruqui, M., Dodge, J., Jauhar, S.K., Dyer, C., Hovy, E., Smith, N.A., 2014. Retrofitting word vectors to semantic lexicons. arXiv 1411.4166.

Fischl, B., Sereno, M.I., Dale, A.M., 1999. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. Neuroimage 9, 195–207. https://doi.org/10.1006/nimg.1998.0396

Friederici, A.D., Gierhan, S.M., 2013. The language network. Curr. Opin. Neurobiol. 23, 250–254. https://doi.org/10.1016/j.conb.2012.10.002

Fyshe, A., Talukdar, P.P., Murphy, B., Mitchell, T.M., 2014. Interpretable semantic vectors from a joint model of brain- and text-based meaning, in: ACL. pp. 489–499.

Güçlü, U., van Gerven, M.A.J., 2015. Semantic vector space models predict neural responses to complex visual stimuli. arXiv 1510.04738.

Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, Jack, L., 2016. Natural

speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458. https://doi.org/10.1038/nature17637

Huth, A.G., Nishimoto, S., Vu, A.T., Gallant, J.L., 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76, 1210–1224. https://doi.org/10.1016/j.neuron.2012.10.014

Ivanova, A.A., Mineroff, Z., Zimmerer, V., Kanwisher, N., Varley, R., Fedorenko, E., 2019. The language network is recruited but not required for non-verbal semantic processing. bioRxiv 696484. https://doi.org/10.1101/696484

Jain, S., Huth, A.G., 2018. Incorporating context into language encoding models for fMRI, in: Proceedings of the 32Nd International Conference on Neural Information Processing Systems, NIPS'18. Curran Associates Inc., USA, pp. 6629–6638.

Jat, S., Tang, H., Talukdar, P., Mitchell, T., 2019. Relating simple sentence representations in deep neural networks and the brain, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 5137–5154. https://doi.org/10.18653/v1/P19-1507

Kriegeskorte, N., Mur, M., 2012. Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. Front. Psychol. 3, 1–13. https://doi.org/10.3389/fpsyg.2012.00245

Li, Y., Pan, Q., Yang, T., Wang, S., Tang, J., Cambria, E., 2017. Learning word representations for sentiment analysis. Cognit. Comput. 9, 843–851.

Ling, W., Dyer, C., Black, A.W., Trancoso, I., 2015. Two/too simple adaptations of Word2Vec for syntax problems, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1299–1304. https://doi.org/10.3115/v1/N15-1142

Liu, Q., Jiang, H., Wei, S., Ling, Z.-H., Hu, Y., 2015. Learning semantic word embeddings based on ordinal knowledge constraints, in: Proceedings of ACL. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1501–1511. https://doi.org/10.3115/v1/P15-1145

Lund, K., Burgess, C., 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. Behav. Res. methods, instruments, Comput. 28, 203–208.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013a. Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. 26, 3111–3119.

Mikolov, T., Yih, W., Zweig, G., 2013b. Linguistic regularities in continuous space word representations, in: HLT-NAACL. pp. 746–751.

Moeller, S., Yacoub, E., Olman, C.A., Auerbach, E., Strupp, J., Harel, N., Uğurbil, K., 2010. Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. Magn. Reson. Med. 63, 1144–1153. https://doi.org/10.1002/mrm.22361

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P.A., Kriegeskorte, N., 2013. Human object-similarity judgments reflect and transcend the primate-IT object representation. Front. Psychol. 4, 1–22. https://doi.org/10.3389/fpsyg.2013.00128

Naselaris, T., Olman, C.A., Stansbury, D.E., Ugurbil, K., Gallant, J.L., 2015. A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. Neuroimage 105, 215–228. https://doi.org/10.1016/j.neuroimage.2014.10.018

Nishida, S., Huth, A.G., Gallant, J.L., Nishimoto, S., 2015. Word statistics in large-scale texts explain the human cortical semantic representation of objects, actions, and impressions. Soc. Neurosci. Abstr. 45, 333.13.

Nishida, S., Matsumoto, Y., Yoshikawa, N., Son, S., Murakami, A., Hayashi, R., Nishimoto, S., Takahashi, H., 2020a. Reduced intra- and inter-individual diversity of semantic representations in the brains of schizophrenia patients. bioRxiv 2020.06.03.132928. https://doi.org/10.1101/2020.06.03.132928

Nishida, S., Nakano, Y., Blanc, A., Maeda, N., Kado, M., Nishimoto, S., 2020b. Brain-mediated transfer learning of convolutional neural networks, in: Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence. pp. 5281–5288. https://doi.org/10.1609/aaai.v34i04.5974

Nishida, S., Nishimoto, S., 2018. Decoding naturalistic experiences from human brain activity via distributed representations of words. Neuroimage 180, 232–242. https://doi.org/10.1016/j.neuroimage.2017.08.017

Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L., 2011. Reconstructing visual experiences from brain activity evoked by natural movies. Curr. Biol. 21, 1641–1646. https://doi.org/10.1016/j.cub.2011.08.031

Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation, in: Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014). pp. 1532–1543.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S.J., Kanwisher, N., Botvinick, M., Fedorenko, E., 2018. Toward a universal decoder of linguistic meaning from brain activation. Nat. Commun. 9, 963. https://doi.org/10.1038/s41467-018-03068-4

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations, in: Proceedings of the 2018 Conference of

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 2227–2237. https://doi.org/10.18653/v1/N18-1202

Ruan, Y.-P., Ling, Z.-H., Hu, Y., 2016. Exploring semantic representation in brain activity using word embeddings, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 669–679.

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E.A., Kanwisher, N., Tenenbaum, J., Fedorenko, E., 2020. Artificial neural networks accurately predict language processing in the brain. bioRxiv 2020.06.26.174482. https://doi.org/10.1101/2020.06.26.174482

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B., 2014. Learning sentiment-specific word embedding, in: ACL. pp. 1555–1565.

Vodrahalli, K., Chen, P.-H., Liang, Y., Baldassano, C., Chen, J., Yong, E., Honey, C., Hasson, U., Ramadge, P., Norman, K.A., Arora, S., 2018. Mapping between fMRI responses to movies and their natural language annotations. Neuroimage 180, 223–231. https://doi.org/10.1016/j.neuroimage.2017.06.042

Wang, S., Zhang, J., Wang, H., Lin, N., Zong, C., 2020. Fine-grained neural decoding with distributed word representations. Inf. Sci. (Ny). 507, 256–272. https://doi.org/10.1016/j.ins.2019.08.043

Wang, Z., Ittycheriah, A., 2015. FAQ-based question answering via word alignment. arXiv 1507.02628.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., Mitchell, T., 2014. Simultaneously uncovering the patterns of brain regions involved in different story reading Subprocesses. PLoS One 9, 1–19. https://doi.org/10.1371/journal.pone.0112575

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q. V, 2019. XLNet: Generalized autoregressive pretraining for language understanding, in: Advances in Neural Information Processing Systems. pp. 5753–5763.

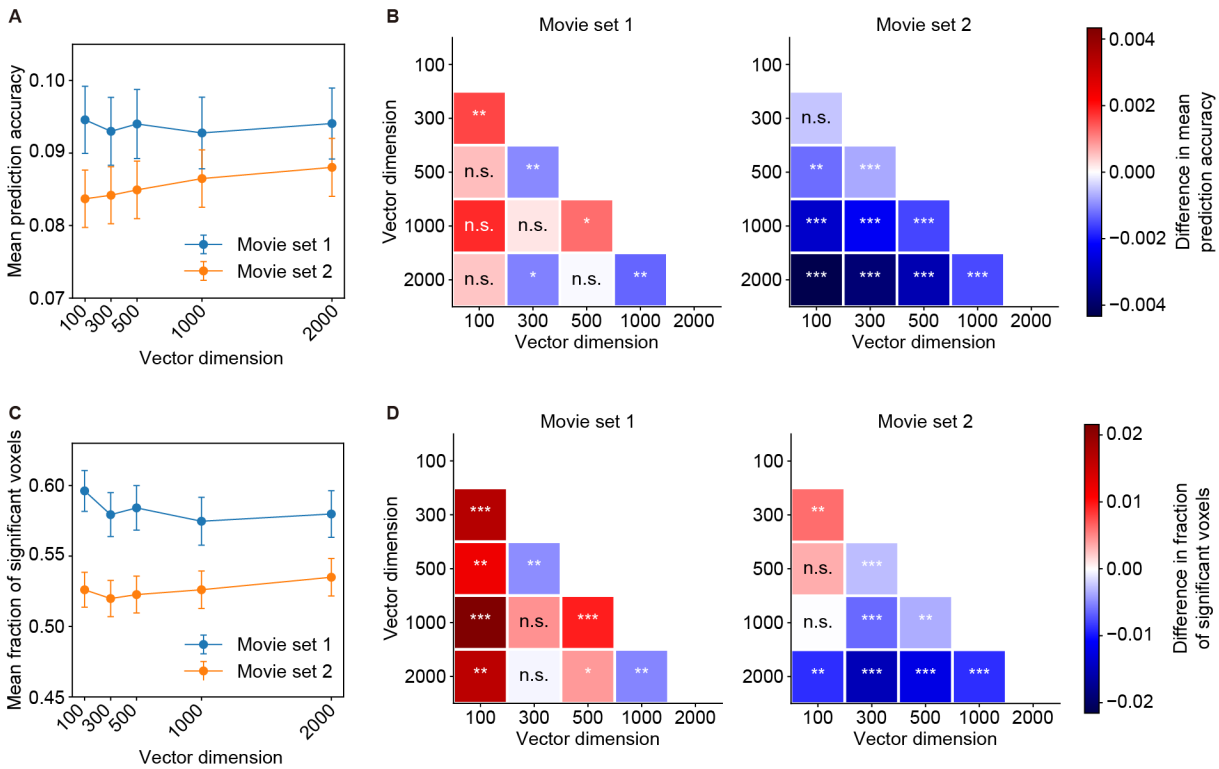

Fig. S1. Model performance comparison across vector dimensions. (A) Mean prediction accuracy of voxelwise models with different vector dimensions. Error bars indicate standard error of the mean (SEM). (B) Difference of mean prediction accuracies between different vector dimensions. The difference was evaluated separately for movie sets 1 (left) and 2 (right). The color of each cell represents the accuracy difference of the dimension on the x-axis minus the dimension on y-axis (red, positive values; blue, negative values). The mark in each cell indicates the statistical significance of the difference (Wilcoxon test, ***P < 0.0001, **P < 0.01, *P < 0.05, n.s., P > 0.05, FDR corrected). (C) Fractions of significant voxels for voxelwise models with different vector dimensions. The same conventions were used as in A. (D) Difference in fractions of significant voxels between different vector dimensions. The same conventions were used as in B.

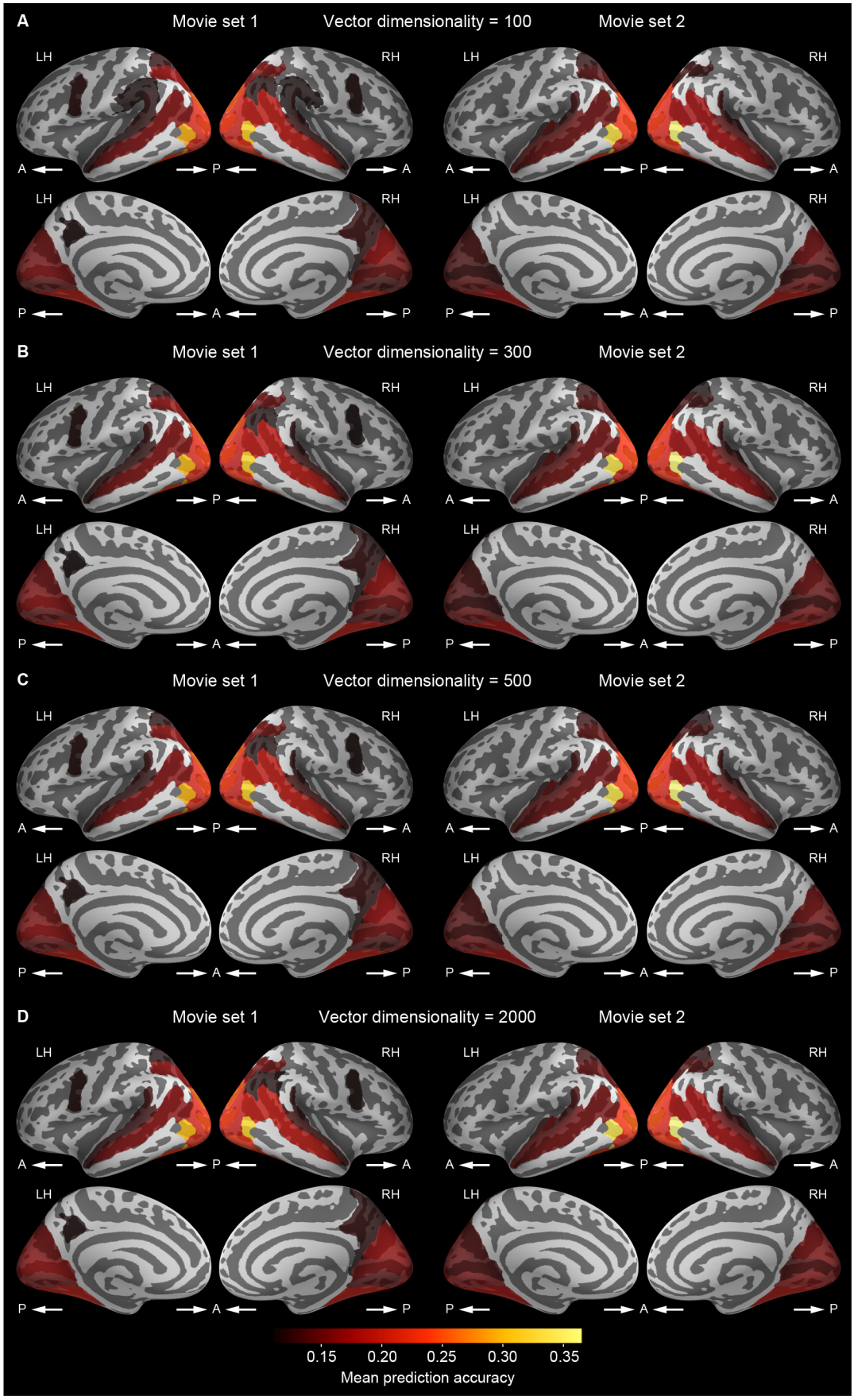

Fig. S2. Cortical mapping of prediction accuracy for other vector dimensions. The mean prediction accuracy of voxelwise models in each brain region was mapped onto the cortical surface (A, vector dimension = 100; B, 300; C, 500; D, 2000). The same conventions were used as in Fig. 4A–B.

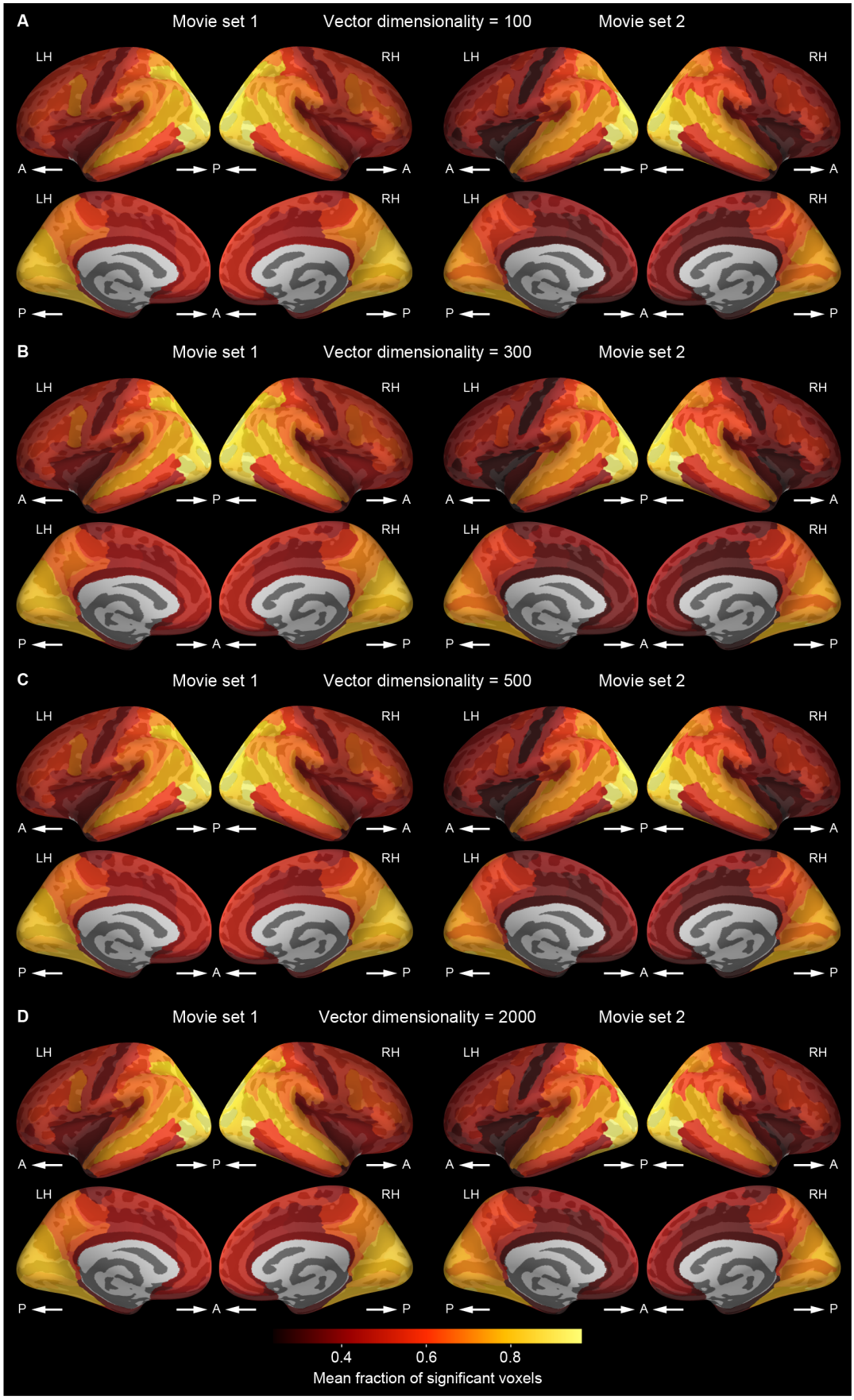

Fig. S3. Cortical mapping of the fraction of significant voxels for other vector dimensions. The fraction of significant voxels in each brain region was mapped onto the cortical surface (A, vector dimension = 100; B, 300; C, 500; D, 2000). The same conventions were used as in Fig. 4C–D.

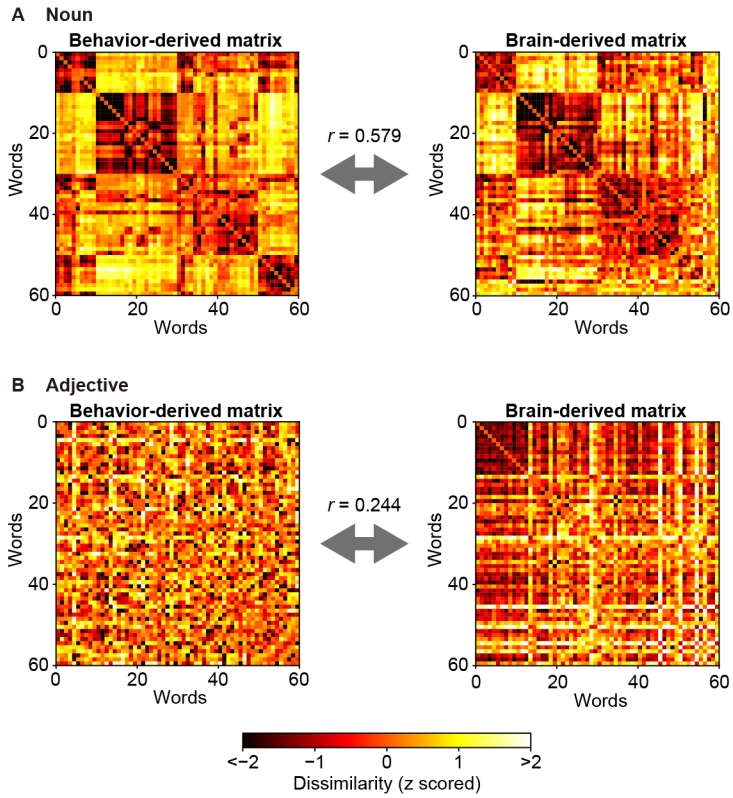

Fig. S4. Correlations between brain- and behavior-derived dissimilarity matrices from the same participant population. We constructed word dissimilarity matrices separately for nouns (A) and adjective (B) using behavioral and voxelwise-model data obtained from the same population of 6 participants (word-vector dimension = 1000; movie set 1). The same conventions were used as in Fig. 5. There were significant correlations between behavior- and brain-derived word dissimilarity matrices for both nouns and adjectives ($P < 0.00001$).

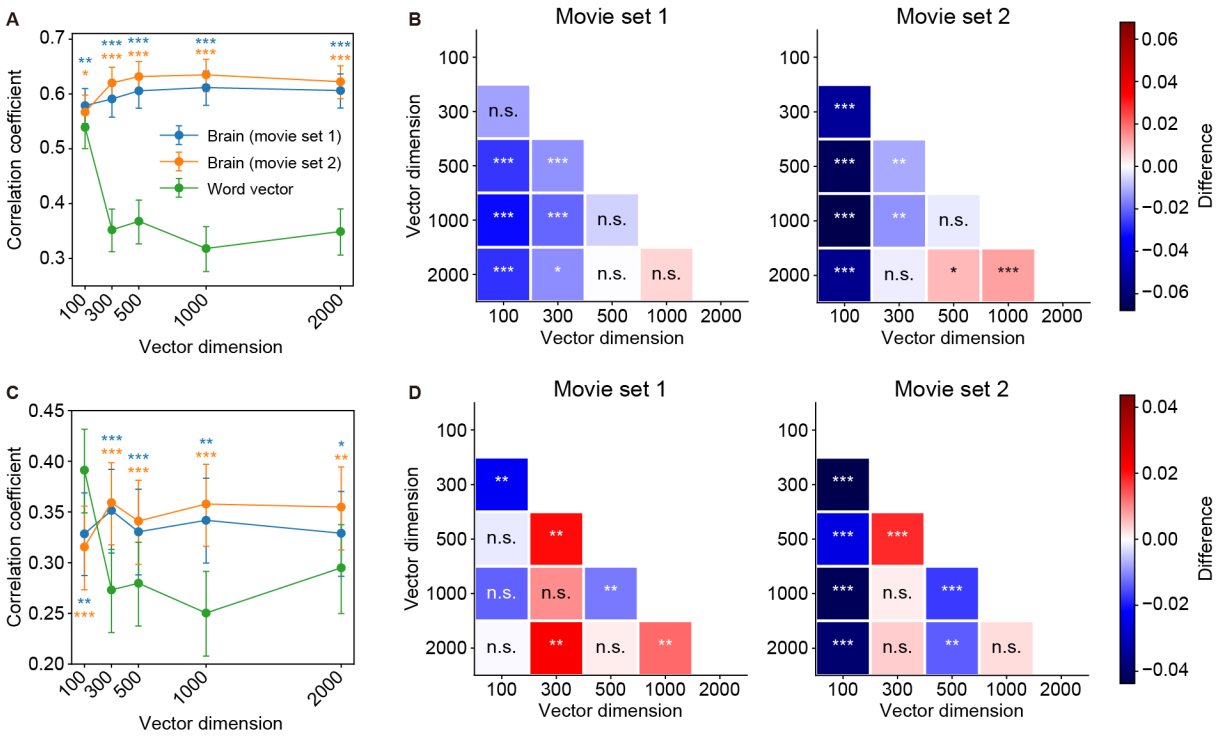

Fig. S5. Correlations between dissimilarity matrices for all vector dimensions. (A) Brain–behavior and word vector–behavior correlations of noun dissimilarity matrices for different vector dimensions. Error bars indicate 95% confidence interval (CI) estimated by bootstrapping. Marks above or below each plot indicate the statistical significance of the difference between brain–behavior and word vector–behavior correlation coefficients for each movie set (bootstrapping test, ***P < 0.001, **P < 0.01, *P < 0.05, FDR corrected) (B) Difference between brain–behavior correlation coefficients of different vector dimensions for nouns. The color of each cell represents the coefficient difference of the dimension on the x-axis minus the dimension on the y-axis (red, positive values; blue, negative values). The mark in each cell indicates the statistical significance of the difference (bootstrapping test, ***P < 0.0001, **P < 0.01, *P < 0.05, n.s., P > 0.05, FDR corrected). (C and D) The same analyses were performed as in A and B using adjective dissimilarity matrices.

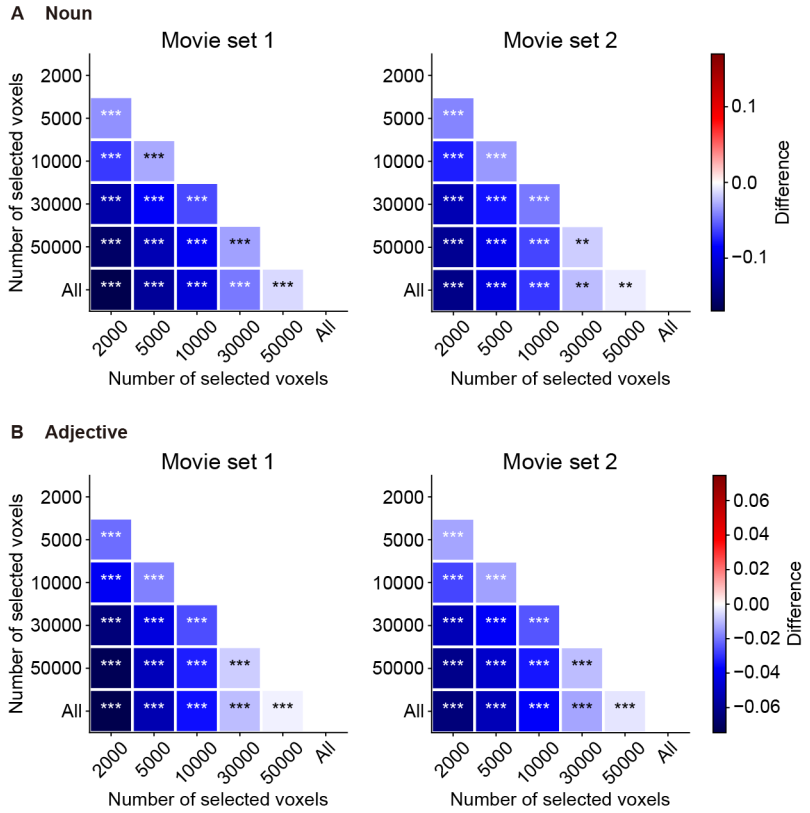

Fig. S6. Difference of brain–behavior correlation coefficients between each pair of different voxel selections. The color of each cell represents the difference between each pair of the number of selected voxels; the difference is the brain–behavior correlation coefficient for the dimension on the x-axis minus the coefficient for the dimension on the y-axis (red, positive values; blue, negative values). The mark in each cell indicates the statistical significance of the difference (bootstrapping test, ***P < 0.0001, FDR corrected).

Table S1. Number of stimulus movie clips in individual product/service categories.

Categories	Movie set 1	Movie set 2
Electronic & Precision	9	4
Audiovisual	1	5
Appliance	5	16
Car	33	31
Food & Confectionery	69	7
Beverage & Alcoholic drink	38	20
Medical & Health	21	35
Cosmetics	6	49
Sundries & Home equipment	45	10
Garment/apparel	11	9
Entertainment	43	42
Media & Education	16	41
Distribution & Retailer	20	12
Communication & Service	51	35
House & Construction	12	9
Finance	20	9
Enterprise, Public service, & Others	20	34

Table S2. Consistency of inter-regional patterns of prediction accuracy across vector dimensionality and across movie sets. We calculated the mean prediction accuracy within each cortical region (Figs. 4A–B and S2) and compared the inter-regional patterns of prediction accuracy between arbitrary pairs of vector dimensionality and movie sets. The value in each cell denotes the Pearson's correlation coefficient of the inter-regional patterns between each pair.

			Vector dimensionality									
			Movie set 1					Movie set 2				
			100	300	500	1000	2000	100	300	500	1000	2000
Vector dimensionality	Movie set 1	100	-	-	-	-	-	-	-	-	-	-
		300	0.9992	-	-	-	-	-	-	-	-	-
		500	0.9994	0.9999	-	-	-	-	-	-	-	-
		1000	0.9988	0.9999	0.9998	-	-	-	-	-	-	-
		2000	0.9985	0.9998	0.9997	0.9999	-	-	-	-	-	-
	Movie set 2	100	0.9687	0.9726	0.9708	0.9730	0.9740	-	-	-	-	-
		300	0.9683	0.9726	0.9707	0.9730	0.9741	0.9997	-	-	-	-
		500	0.9677	0.9721	0.9701	0.9725	0.9737	0.9996	0.9999	-	-	-
		1000	0.9677	0.9724	0.9704	0.9729	0.9740	0.9992	0.9998	0.9999	-	-
		2000	0.9676	0.9722	0.9703	0.9728	0.9740	0.9990	0.9997	0.9998	0.9999	-

Table S3. Consistency of inter-regional patterns of significant-voxel fraction across vector dimensionality and across movie sets. As in the case of the mean prediction accuracy (Table S2), the inter-regional patterns of significant-voxel fraction (Figs. 4C–D and S3) were compared between arbitrary pairs of vector dimensionality and movie sets. The value in each cell denotes the Pearson's correlation coefficient of the inter-regional patterns between each pair.

			Vector dimensionality									
			Movie set 1					Movie set 2				
			100	300	500	1000	2000	100	300	500	1000	2000
Vector dimensionality	Movie set 1	100	-	-	-	-	-	-	-	-	-	-
		300	0.9991	-	-	-	-	-	-	-	-	-
		500	0.9993	0.9999	-	-	-	-	-	-	-	-
		1000	0.9985	0.9997	0.9997	-	-	-	-	-	-	-
		2000	0.9983	0.9996	0.9995	0.9998	-	-	-	-	-	-
	Movie set 2	100	0.9576	0.9637	0.9612	0.9650	0.9665	-	-	-	-	-
		300	0.9579	0.9644	0.9619	0.9659	0.9673	0.9995	-	-	-	-
		500	0.9582	0.9648	0.9624	0.9663	0.9677	0.9994	0.9999	-	-	-
		1000	0.9590	0.9655	0.9631	0.9670	0.9684	0.9991	0.9998	0.9999	-	-
		2000	0.9600	0.9663	0.9640	0.9677	0.9691	0.9989	0.9996	0.9997	0.9999	-