Decoding the Neural Representation of Story Meanings across Languages

Morteza Dehghani,[1]* Reihane Boghrati,[1] Kingson Man,[1] Joe Hoover,[1] Sarah I. Gimbel,[1] Ashish Vaswani,[2] Jason D. Zevin,[1] Mary Helen Immordino-Yang,[1] Andrew S. Gordon,[1] Antonio Damasio,[1] and Jonas T. Kaplan[1]

[1]*University of Southern California, Los Angeles, CA*
[2]*Google Brain, Mountain View, California*

Abstract: Drawing from a common lexicon of semantic units, humans fashion narratives whose meaning transcends that of their individual utterances. However, while brain regions that represent lower-level semantic units, such as words and sentences, have been identified, questions remain about the neural representation of narrative comprehension, which involves inferring cumulative meaning. To address these questions, we exposed English, Mandarin, and Farsi native speakers to native language translations of the same stories during fMRI scanning. Using a new technique in natural language processing, we calculated the distributed representations of these stories (capturing the meaning of the stories in high-dimensional semantic space), and demonstrate that using these representations we can identify the specific story a participant was reading from the neural data. Notably, this was possible even when the distributed representations were calculated using stories in a different language than the participant was reading. Our results reveal that identification relied on a collection of brain regions most prominently located in the default mode network. These results demonstrate that neuro-semantic encoding of narratives happens at levels higher than individual semantic units and that this encoding is systematic across both individuals and languages. *Hum Brain Mapp 38:6096–6106, 2017.* **© 2017 Wiley Periodicals, Inc.**

Key words: language neuroscience; semantics; natural language processing; machine learning; knowledge representation

INTRODUCTION

One of the defining characteristics of human language is its capacity for semantic extensibility. Drawing from a common lexicon of morphemes and words, humans generate and comprehend sophisticated, higher-level utterances that transcend the sum of their individual units. This is perhaps best exemplified in stories, in which sequences of events invite inferences about the intentions and motivations of characters, about cause and effect, and about theme and message. The kind of meaning that emerges over time as one listens to a story is not easily captured by analysis at the word level alone. Further, a necessary condition for generating higher-level semantic constructs

Additional Supporting Information may be found in the online version of this article.

Contract grant sponsor: Defense Advanced Research Projects Agency

*Correspondence to: Morteza Dehghani, University of Southern California, Los Angeles, CA 90089-1061. E-mail: mdehghan@usc.edu

Received for publication 9 June 2017; Revised 1 September 2017; Accepted 6 September 2017.

DOI: 10.1002/hbm.23814

Published online 20 September 2017 in Wiley Online Library (wileyonlinelibrary.com).

is that speakers of the same language infer similar meanings from expressions of both lower and higher level semantic units. For example, it can be assumed that when speakers of the same language listen to stories, the perceived meanings of these stories have much in common. Does this imply that individuals' brain activity shows systematic patterns related to the meaning of the stories? If so, can these neural representations of meaning be associated post-hoc with particular stories? Also, when a story is translated from one language to another, the particulars of sounds and words are left behind, but a core of semantic content survives the translation. Does it then follow that the neural representations related to this semantic content are also comparable across the brains of speakers of different languages?

Understanding how conceptual knowledge is represented and organized in the human brain is one of the central problems of cognitive science and many studies have aimed at exploring and understanding the neural representations of concepts [Damasio et al., 2004]. Relying on multiple approaches and methods, this body of work has identified a collection of separable brain regions, constituting a large network, within which different conceptual categories and the corresponding words are represented [e.g., Binder et al., 2009; Damasio et al., 2004; Grabowski et al., 2001; Tranel et al., 1997]. More recently a word-level *semantic map* has been proposed [Huth et al., 2016].

In this work, our aim is to move beyond word-level semantics to investigate neuro-semantic representations at the story-level across three different languages. Specifically, we set out to determine if there are systematic patterns in the neuro-semantic representations of stories beyond those corresponding to word-level stimuli. Our aim is motivated by the long-standing understanding that discourse representations are different from the sum of all of their lexical or clausal parts. Most psycholinguistic models of discourse processing are concerned with the construction and representation of structures such as propositions and arguments [Kintsch, 1974], and situation models [Kintsch, 1988] that are, of course, derived from specific words and sentences, but are more abstract. Indeed, many interesting effects of discourse representations can be demonstrated by comparing the comprehension of the same sentence under conditions that differ in the plausibility of the most common interpretation of the sentence in isolation [Sanford and Garrod, 1998] or, conversely, comprehension of the same passage [Johnson-Laird, 1983] under conditions that differ in the availability of a situation model to begin with. Hence, we aim to leverage recent advances in natural language processing for representing aggregate meaning in high-dimensional semantic space to decode neural activations (i.e., map back to the semantic space) during story reading. This will allow us to investigate the networks involved in representing higher-level semantics, and to explore whether activation patterns related to higher-level meaning are common across people and languages.

One approach that has been proven to be fruitful in exploring neuro-semantic representations is to investigate the relationship between various co-occurrence patterns of words within large textual corpora, and to relate these patterns to neural activity recorded during exposure to those words. Examining the representations of concrete nouns, Mitchell et al. [2008] proposed a computational model that predicts voxel activation using a weighted sum of various semantic features of nouns, calculated based on word co-occurrence in a large English corpus. Using fMRI data from 9 participants who viewed 60 word-picture pairs, they demonstrated that their model captures aspects of neuro-semantic representations of these concepts; it could correctly predict which word/picture participants were viewing with a higher-than-chance accuracy. The most predictive voxels were distributed around the cerebral cortex, and included sensory regions in the temporal and occipital lobes, motor related regions in the frontal and parietal lobes, and other regions of the orbitofrontal and inferior frontal cortices.

Other researchers have proposed different representational [Fyshe et al., 2014; Yogatama et al., 2014], computational [Huth et al., 2012; Just et al., 2010; Shinkareva et al., 2011], and methodological [Guimaraes et al., 2007; Sudre et al., 2012] modifications to the general Mitchell et al. (2008) approach in order to better capture and explore the neural architectures involved in representing concrete nouns. Notably, Huth et al. (2016) use word-embeddings, a quantitative representation of meaning calculated based on co-occurrence patterns of 985 common English words, along with a generative model of areas covering the cortex, to create a detailed semantic-map reported to be consistent across individuals. Generally, in word-level analysis applied in the above approaches, each word is treated as an isolated symbol, independent of any relation to context or to the overall meaning that the story conveys. Therefore, even though these approaches have been very useful in mapping the meaning of individual words, they do not reveal how the higher-level meaning that emerges through the course of a story is represented.

Discourse has typically been operationalized for fMRI analysis by reference to theory-dependent identification of key discourse features [Whitney et al., 2009], general linear model [Yarkoni et al., 2008] or cross-correlation [Ames et al., 2015; Lerner et al., 2011] analyses comparing texts differing in coherence. More recently researchers have started to investigate the neural architectures involved in the semantic representation of sentences and story segments. Wehbe et al. (2014) present a dynamic model for studying how different regions of the brain encode various types of information while reading a story. Passages from a Harry Potter story were presented to 9 participants using Rapid Serial Visual Presentation (each word presented for 0.5 s). The story segments (16 words) were then represented using a combination of various syntactic features, semantic features of individual words, and low-level features such as number of letters, and used in a predictive model that was able to predict which one of the two novel passages was being read. They showed that different aspects of story processing were encoded in different brain networks. For example, activity in temporo-parietal regions in both hemispheres was related to processing sentence length and complexity, while semantic classification was most accurate in voxels from the middle and superior temporal gyri and the left inferior frontal gyrus. In general, brain regions on the lateral surfaces of both hemispheres in the temporal, occipital, and parietal regions appear to have been most predictive of linguistic content in the stories.

While previous methodologies have found semantic representations to be widely distributed throughout the brain, there now seems to be growing evidence for a special involvement of the default mode network (DMN) in representation of high-level meaning and language comprehension. The DMN was originally identified as a "resting state" network that shows high baseline activity when people are asked to rest without engaging in any specific externally-focused task [Raichle and Snyder, 2007; Raichle, 2015]. This bilateral network includes midline cortical structures (medial prefrontal cortex, precuneus, and posterior parietal cortex) as well as lateral structures (inferior parietal lobe and anterior temporal lobes). Activity in these nodes is highly correlated during resting fMRI [Buckner et al., 2008] and has been thought to reflect such cognitive processes as mind-wandering [Smallwood and Schooler, 2015], thinking about one's self [Qin and Northoff, 2011], remembering the past and imagining the future [Østby et al., 2012], and in general to support stimulus-independent thought [Smallwood et al., 2013].

Yet, the specific function of this network is not well understood. A growing body of work has implicated the DMN in more active forms of cognition [Spreng, 2012], and specifically in semantic processing [Binder et al., 2009]. A series of studies have shown that the DMN seems to be involved in representing the global meaning of passages, rather than meaning at the word or sentence-level [Ferstl et al., 2008; Lerner et al., 2011]. Further, the activity in the DMN is consistent when a story is presented in different modalities (spoken vs. written), or in different languages (Russian vs. English to native speakers of these languages), indicating highly abstracted representations of the stimuli in this network [Chen et al., 2017; Honey et al., 2012; Regev et al., 2013; Zadbood et al., 2016]. The word-level semantic map produced by Huth et al. [2016] also demonstrates significant overlap with the DMN. More recently, Simony et al. [2016] show that the DMN reconfigures consistently across subjects when processing narrative stimuli. Also, a recent study demonstrates that patterns of activity in the DMN when people are describing a narrative are highly consistent across individuals and specific to events in the narrative [Chen et al., 2017]. Our own recent finding shows that activity in some DMN nodes increases throughout the course of a story, and is greatest when reading stories containing strong moral values [Kaplan et al., 2016].

Parallel to this line of work, multi-voxel pattern analysis (MVPA) has also been used to explore common representations across languages by predicting the neural response to a noun in speakers of one language, based on the noun's neural representation in speakers of a different language [Buchweitz et al., 2012; Correia et al., 2014; Yang et al., 2017; Zinszer et al., 2015]. Supporting evidence from lesion studies, Correia et al. [2014] find the left anterior temporal lobe to be an area with high predictive accuracy independent of language. The ATL is part of the DMN, confirming this region as part of a network in which the abstract representations of words are coded.

In this paper, we use *story-level* embeddings to examine the neuro-semantic representations of stories in people from three different linguistic-cultural backgrounds, across their three native languages. We demonstrate that story-level representations can capture story content across languages, and can be used to localize the neuro-semantic content of stories to specific regions of the brain. We mapped the brain regions in which language-independent semantic story content was detected, revealing a set of regions in the DMN that have been implicated in the generation of rich internal experiences. While previous work has demonstrated that the DMN responds to abstract meaning of stories, we show that the patterns of activation in the DMN correspond to the unique encoding of the meaning of narratives, and that these patterns can be used to decode the meaning of narratives across three different languages.

METHOD

Stimuli

For our stimuli, we used real-world personal narratives written by people describing their experiences. In order to find such narratives, we started with a corpus of over 20 million weblog story posts that were compiled from Spinn3r.com [Sagae et al., 2013]. Next, the corpus was queried on different topics (e.g., telling a lie, getting a divorce) via a text retrieval system (Apache Lucene). Forty stories were chosen and each was condensed to a paragraph of 145–155 words. Professional translators were used to translate these English stories into Mandarin Chinese and Farsi. Lastly, these translations were back-translated by native speakers into English, and the back-translated versions were checked for any inconsistencies with the original stories. Any minor inconsistencies were resolved by the translators.

Distributed Representation of Stories

Distributed word representations refer to n-dimensional numeric vector representations that capture some semantic and syntactic aspects of words. Prior to the recent widespread upsurge of deep neural networks, methods such as Latent Semantic Analysis [Deerwester et al., 1990] and Latent Dirichlet Analysis [LDA; Blei et al., 2003] were used to build such distributed word representation.

In recent years, neural network approaches for language processing have gained considerable attention. One notable example is the *word2vec* [Mikolov et al., 2013] method which generates distributed word vectors from a text corpus. It efficiently trains a model that predicts a word conditioned on its surrounding context (e.g., the five words before and after the target word) by maximizing vector similarity of words that appear together and minimizing the similarity of words that do not.

Moving beyond representations at the word level, several related methods have been proposed for distributed representation of higher-level textual units (i.e., sentences, paragraphs, or documents). Paragraph vector [Le and Mikolov, 2014] is a technique that simultaneously learns vector representations for words and texts, and captures different aspects of the semantics of the text in the training process.

In this work, we used a version of paragraph vector derived from word2vec called *doc2vec* [Le and Mikolov, 2014]. At a high level, doc2vec uses a technique called the *Distributed Memory* Model which aims to maximize the probability of a word w_i, given the paragraph D_i that the word is drawn from and the words adjacent to it (i.e., its context) $w_{i-j}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+j}$. This is treated as a multi-class classification problem, and paragraphs and context words are represented as n-dimensional vectors. During training, these vectors are optimized via stochastic gradient descent with backpropogation [Rumelhart et al., 1986] to maximize the probability of the target word. That is, at each iteration of gradient descent, a context for a given target word is sampled from a random paragraph. Then the representation of this paragraph D_i and the representations of the context words are used to predict the target word. The gradient from this prediction is then used to update the paragraph vector D_i and context vectors.

An important component of this model is the paragraph matrix D, which both provides context for the interpretation of individual words, and merges the semantics at the lexical level. The columns of matrix D contain "paragraph vectors" that are each optimized to represent a paragraph or other units of higher level text drawn from the training corpus. These paragraph vectors are learned via a sliding window function that predicts each word in the paragraph, given the paragraph vector and vectors representing the words contained in the sliding window. Importantly, paragraph vectors are constant across windows sampled from the same paragraph; thus, they can be thought of as representing the component of the paragraph that transcends and unifies each of the windows sampled from the paragraph. Accordingly, doc2vec represents paragraphs as unique higher-level units and not simply as aggregations of individual words. The relationship between paragraph vectors and the contexts contained in a paragraph can also be understood as hierarchical, such that words and contexts are nested in paragraphs. From this view, a portion of the variance associated with words can be attributed to their immediate context (e.g., local semantic and syntactic phenomena), but an additional portion of this variance is explained by the paragraph from which the word is drawn. Accordingly, doc2vec's paragraph vectors represent the component of a paragraph that cannot be reduced to word-level semantics, they represent the higher-level meaning expressed by the paragraph—or, as in our case, narrative.

We would like to emphasize that our goal in this paper is not to compare the performance of doc2vec against any other NLP technique (namely word2vec), but to show that aggregate story-level representations of text can be used to successfully decode narrative-processing fMRI data. Specifically, one could use word2vec, or any other distributed representation at the word-level, to construct story-level representations by aggregating (or concatenating) individual word-level vectors [Garten et al., 2015]. However, given that doc2vec does not process individual words in isolation but takes context into account, it has been shown that it captures the overall meaning of larger pieces of text significantly better than other related techniques [Dai et al., 2015; Lau and Baldwin, 2016; Le and Mikolov, 2014]. These findings motivated the use of doc2vec over aggregated word2vec vectors. We also performed a series of behavioral experiments demonstrating the effectiveness of doc2vec compared to word-level operations in capturing the overall gist of stories (see Supporting Information). We further compared the performance of two different word-level representations of the stories to the doc2vec representations in our whole-brain analysis, and show that doc2vec representations outperformed the other methods (see Supporting Information). Further, doc2vec is only one method for modeling narrative-level semantics, and other techniques for capturing sentence-level and/or document-level semantics have recently been proposed [e.g., Conneau et al., 2017; Ma et al., 2015; Palangi et al., 2016]. To reiterate, our goal is to demonstrate the effectiveness of aggregate story-level representations, rather than representation of words in isolation as it has been done before [e.g., Huth et al., 2016; Mitchell et al., 2008], in decoding fMRI data, and using these representations to further advance our understanding of neuro-semantic encoding at the narratives level.

All model training was performed using the doc2vec c library, now available through *gensim* [Řehůřek and Sojka, 2010] in python.[1] To train the paragraph vector model for

[1] Detailed instruction on how to use the doc2vec framework in python is available at https://radimrehurek.com/gensim/models/doc2vec.html

English stories, we used an English corpus of weblogs of personal stories [~2 billion words; See Gordon and Swanson, 2009, for more details about how this corpus was compiled]. Similarly, Chinese story vectors were built using a Chinese blog corpus of personal stories [~2 billion words; Gordon et al., 2013]. A similar process was applied to the Farsi stories with a minor difference on the type of corpora used for training. Since Farsi blog posts are mostly written in informal Farsi, and the stories we used are in formal Farsi, we used a combination of Farsi news corpora [AleAhmad et al., 2009] along with a large weblog corpus (~2 billion words) of personal stories to have a better representation of stories in formal Farsi. Model training took between 20–24 hours for each language. After completion of the training process, each language model was then used to represent each of the stories as a 100-dimensional vector. This resulted in 40 story-vectors per language.

Participants

Ninety-five healthy participants with no history of psychological or neurological disorders were recruited from the University of Southern California community and the surrounding Los Angeles Area. We recruited participants from three groups: Americans, Chinese, and Iranians. All American participants were born in the United States and were native English speakers who grew up in exclusively English-speaking households. All Chinese and Iranian participants were born and raised in their native country, and had been in the United States for fewer than five years. Chinese and Iranian participants were all fluent in English in addition to their native languages.

One American participant, two Chinese participants, and two Iranian participants were excluded either for excessive motion or for not completing the scans. This left 30 American participants (mean age: 23.83 ± 0.92, 15 male), 30 Chinese participants (mean age: 23.47 ± 0.39, 16 male), and 30 Iranian participants (mean age: 26.66 ± 0.60, 16 male). Subjects were paid [dollar]20 per hour for their participation and gave informed consent approved by the Institutional Review Board of the University of Southern California.

fMRI Experiment

Participants read study instructions in their native language. If they had any questions about the study, they had the opportunity to ask them before entering the fMRI scanner. For each participant there were five story-reading scans (556 s), one resting state scan not presented here, and one additional task run not presented here. During the story scans, each story was preceded by a context slide (2s) identifying the protagonist of the story (e.g., American Mother). Following a 1.5 s delay, the story was presented over the course of 3 slides of text, each displayed for 12 s. After a variable delay (1.25–4.75s) a question appeared asking the participant a question about the values of the protagonist. In this paper we only report analysis of the 36-second story reading period.

fMRI Parameters

Imaging was performed using a 3T Siemens MAGNETON Trio System with a 12-channel matrix head coil at the Dana and David Dornsife Neuroscience Institute at the University of Southern California. Functional images were acquired using a gradient-echo, echo-planar, T2*-weighted pulse sequence (TR = 2000 msec, one shot per repetition, TE = 25 msec, flip angle = 90°, 64 x 64 matrix). 40 slices covering the entire brain were acquired with a voxel resolution of 3.0 × 3.0 × 3.0 mm. Functional data were continuously acquired for each run, with a short break between runs. A T1-weighted high-resolution (1 × 1 × 1 mm) image was acquired using a three-dimensional magnetization-prepared rapid acquisition gradient (MPRAGE) sequence (TR = 2530 msec, TE = 3.09 msec, flip angle = 10°, 256 × 256 matrix). Two hundred and eight coronal slices covering the entire brain were acquired. Total scan time for each participant was approximately 80 minutes.

fMRI Data Preprocessing

Univariate data analysis was performed with FSL (FMRIB's Software Library http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/). Data were preprocessed using standard steps: motion correction (Jenkinson et al., 2002), 8mm FWHM spatial smoothing, highpass temporal filtering using Gaussian-weighted least-squares straight line fitting with a sigma of 60s (corresponding to a period of 120 s) and slice timing correction. Data were also corrected for magnetic field inhomogeneities using field maps acquired for each subject.

Data were then analyzed using the General Linear Model. Each component of the task (context, the combined three screens of the story, and the question) was modeled by convolving the task design with a double-gamma hemodynamic response function. The model for each functional scan included one regressor for all of the context periods, one regressor for all of the question periods, and a set of regressors for each 12-second story period in the scan. Also included in the model were the temporal derivative of each story regressor, and six motion correction parameters. This analysis resulted in a statistical map of standardized *z* scores for each story for each participant, representing the brain activity for that story relative to resting baseline. A contrast of interest for each story combined across the 3 story slides to yield a single z-map representing the average activity for the entire story. These storywise z-maps formed the input to our classifier.

To register the data to a common space for cross-subjects analysis, we used FSL's FLIRT tool in two stages [Jenkinson et al., 2002; Jenkinson and Smith, 2001]. First, functional images from each participant were aligned with their own T1-weighted MPRAGE using a 6 degrees of freedom rigid-body warp. Next, the MPRAGE was registered to the standard 2mm MNI atlas with a 12 degrees of freedom affine transformation, and then this transformation was refined using FNIRT nonlinear registration [Andersson et al., 2010].

Analysis

As discussed earlier, 90 participants (30 from each culture) were scanned while reading 40 different stories. The fMRI data were preprocessed and the z-score activation map for each story was converted to a vector of 212,018 dimensions (voxels) for each participant-story pair, resulting in a matrix of 40*212,018 for each participant. Next, as discussed previously, the text of each of the 40 stories was converted to a 100 dimensional semantic vector using the word2vec method described earlier, resulting in a matrix of dimension 40*100 for each language.

Searchlight-based multi-voxel pattern analysis was performed in order to investigate specific brain networks that encode neuro-semantic representations of stories. Specifically, a ridge regression model was fitted on the neural activity recorded in a sphere (of four-voxel radius), successively centered around every voxel for each story (as predictor variables), with its accompanied 100 dimensional semantic representation of that story (as observation variables). The fitted model was evaluated using *k*-fold cross-validation: the ridge regression model was trained on every possible pair of 38 stories and tested on the two remaining stories, resulting in $\binom{40}{2}$ analyses per voxel. In each fold, using the trained model on the 38 stories, the story vectors were predicted for the two left-out stories, $Story_A$ and $Story_B$. A similar approach to Mitchell et al. [2008] was then used to evaluate the predictions: if the predicted vectors were more similar (as assessed by cosine similarity) to the vectors of their target stories, as compared to the vectors of their non-target stories, then the classification was counted as correct. More formally, this can be described as:

$$\cos\left(\overrightarrow{\text{predicted}_A}, \overrightarrow{\text{actual}_A}\right) + \cos\left(\overrightarrow{\text{predicted}_B}, \overrightarrow{\text{actual}_B}\right) > \cos\left(\overrightarrow{\text{predicted}_A}, \overrightarrow{\text{actual}_B}\right) + \cos\left(\overrightarrow{\text{predicted}_B}, \overrightarrow{\text{actual}_A}\right) \quad (1)$$

The accuracy of the classification for the voxel was then averaged over the 780 folds. For each participant, this analysis was then repeated at all the voxels in the brain, resulting in 780 (cross-validation) * 212,018 (voxels) * 90 (participants) classifications. For the inter-language analyses, the fMRI vectors of one cultural group were modeled with story vectors generated from a different language. To be more specific, we used the American subjects' fMRI data, which were recorded while reading stories in English, to map to Farsi and Mandarin story vectors; the Iranian subjects' fMRI data, recorded while reading stories in Farsi, to map to English and Mandarin story vectors; and the Chinese subjects' fMRI data, recorded while reading stories in Mandarin, to map to English and Farsi story vectors. The inter-language analysis added a factor of three to the number of analyses, resulting in total of 44,650,990,800 classifications. The analysis was implemented using the *scikit—learn* library in python [Pedregosa et al., 2011], and was ran on 30 Google Computing Engine machines,[2] each with four CPUs and 26GB of RAM, and it took about six weeks to complete.

The result of this analysis is a map of story-prediction accuracy. In other words, the above method scores each voxel neighborhood based on how accurately it can predict the story vectors. To assess the statistical significance of these accuracy values, non-parametric one-sample *t*-tests were performed with the FSL Randomise program [Winkler et al., 2014]. FSL's Randomise algorithm works as follows. For a given classification, the accuracy maps from 30 participants were concatenated and expressed as 30 positive or negative values indicating above- or below-chance decoding accuracy, respectively. Under the null hypothesis, positive and negative signs are equally likely to occur and may be randomly flipped with no effect on their distribution. Random sign-flipping occurs for the set of 30 values at each voxel, followed by the calculation of a *t*-statistic. We performed 5mm HWHM spatial smoothing of the variance (over subjects), a commonly included step for increasing study power. This is performed for all voxels in the map, creating a permuted *t*-map. The *t*-map then undergoes threshold-free cluster enhancement (TFCE), a procedure that further improves power by taking into account the underlying extent of spatial support. TFCE enhances values within cluster-like structures, while preserving voxel-level inference [Smith and Nichols, 2009]. The TFCE map ultimately contributes a single value, its maximum, to an empirical null distribution of maximum TFCE values. This distribution was populated by repeating the permutation procedure 5000 times. The preceding steps were all implemented in a relatively standard one-line command to FSL Randomise (see Supporting Information). The value at the 95th percentile of the distribution was used to threshold the original, unpermuted data, thereby controlling the family-wise error rate at the 0.05 level. To facilitate comparison across different classification types, we conservatively restricted visualization to the set of voxels that survived thresholding across all of the classification types. For example, we created a unified intra-language decoding map, selecting only voxels present in all three of the English, Chinese, and Farsi classification maps.

RESULTS

The searchlight analyses yielded maps of brain regions that contained information regarding the content of stories.

[2] https://cloud.google.com/compute/

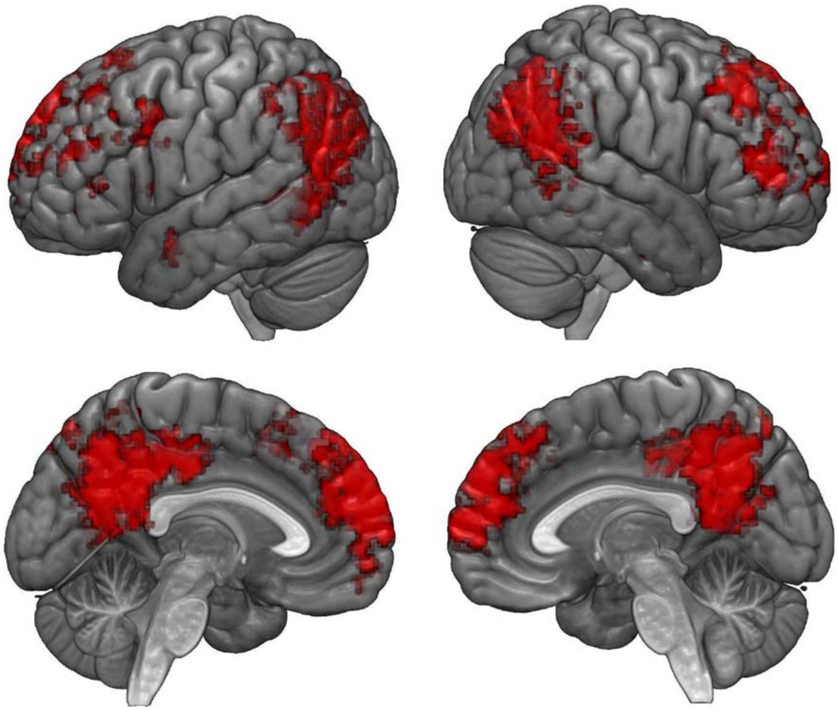

Figure 1.

Intra-language searchlight maps. Colored voxels indicate regions in which English, Farsi, and Chinese stories, presented to participants in their native languages, were all successfully decoded. The most prominent cluster was found in the posteromedial cortices, bilaterally. Other clusters on the medial surface included the superior frontal gyrus, paracingulate gyrus, and frontal pole. On the lateral surface, a prominent cluster was centered on the angular gyrus. [Color figure can be viewed at wileyonlinelibrary.com]

Decoding of stories presented in the participants' native language was successful in a set of spatially-defined brain regions (Fig. 1). Maps for all three within-language classifications were similar (for individual intra-language classification maps see Supporting Information Fig. S3). The most prominent cluster was in the bilateral posteromedial cortices, including the precuneus and posterior cingulate. Also on the medial surface was a bilateral cluster including the paracingulate gyrus, superior frontal gyrus and frontal pole. On the lateral surfaces, the junction of the temporo-parieto-occipital lobes was implicated, specifically, a cluster centered on the angular gyrus bilaterally and extending into the posterior supramarginal gyrus bilaterally and superior lateral occipital cortex. Rostrally, somewhat lateralized to the right but also present on the left, were the superior frontal sulcus and middle frontal gyrus.

The brain activations of participants that were evoked by native-language stories were also successfully decoded using the representations of the same stories, translated into different, unfamiliar languages. We mapped this semantic information to a similar set of regions as found for native-language story decoding, namely the cortical midline structures and fronto-parietal regions (Fig. 2) (for individual inter-language maps see Supporting Information Fig. S4, and for the overlaying inter-language and intra-language maps see Supporting Information Fig. S5). It should be noted that the inter-language story vectors are not linearly correlated ($r^2_{\text{Mandarin}-\text{Farsi}}=0.0008, r^2_{\text{Mandarin}-\text{English}}=0.0003, r^2_{\text{English}-\text{Farsi}}=0.0183$), while there is much higher correlation between the story vectors in the same language (average $r^2=0.308$).

Across all combinations of cultural backgrounds and story languages, the overall accuracies of each searchlight map were broadly similar (Table I). The mean accuracies of significant voxels in each map were generally above 0.55 and the maximum accuracies on each map were all above 0.6 (For whole-brain analysis accuracies, please see Supporting Information Figs. S1 and S2).

DISCUSSION

In this study, we investigated neural representations that transcend narratives' language-specific words and syntactic features, and the degree to which these representations are systematic across people. Using distributed representations of narrative texts, we were able to decode

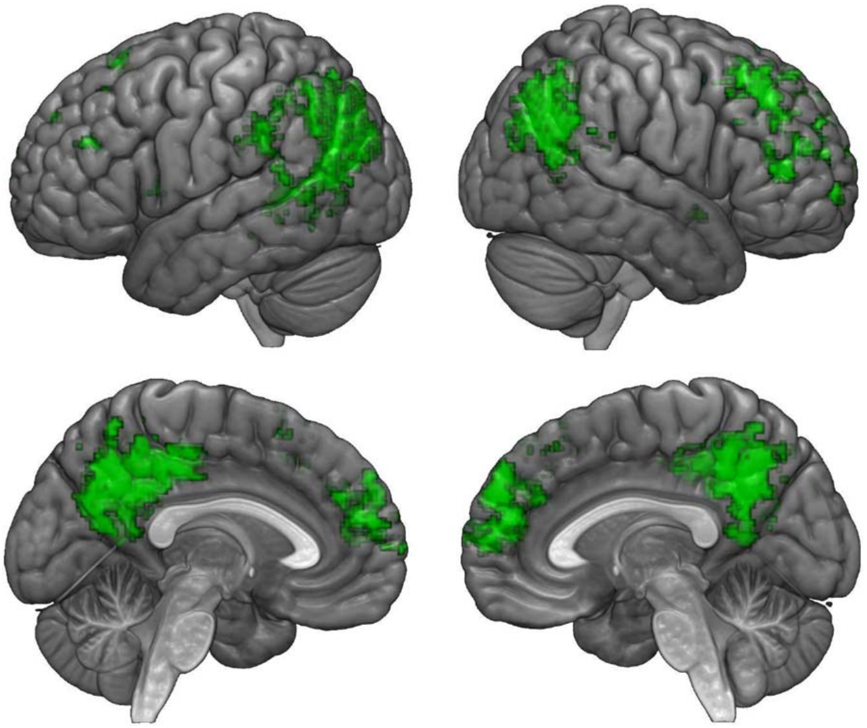

Figure 2.

Inter-language searchlight maps. Colored voxels indicate regions in which story representations, which were generated from stories in a language unfamiliar to the participant, were successfully decoded in English, Farsi, and Chinese speakers. A similar set of regions were found as for intra-language decoding, including the cortical midline structures and fronto-parietal regions. [Color figure can be viewed at wileyonlinelibrary.com]

neural activity corresponding to the reading of particular stories irrespective of the language being used to convey the story. Within- and between-language narrative decoding highlighted similar patterns of neural activity, with the highest classification performance in the DMN. We saw above-chance classification accuracy in all of the major nodes of the DMN, including the posterior medial cortices, the medial prefrontal cortex, and the lateral parietal cortex (See Supporting Information Fig. S6 for relationship between story classification accuracy and the DMN).

The spatial distribution of classification performance implicates the DMN in these high-level semantic representations. One of the distinguishing features of narrative comprehension is that it requires the integration of

TABLE I. Decoding accuracy of story vector models for three languages, based on the fMRI activations of three cultural groups

Culture	Story language	Mean accuracy of Significant Voxels	Maximum accuracy among Searchlight Spheres	Coordinates of center voxel of Maximum Searchlight Sphere
Americans	English	0.566	0.622	48, 40, 50
	Farsi	0.575	0.624	48, 39, 50
	Mandarin	0.579	0.637	44, 40, 54
Iranians	English	0.565	0.623	68, 26, 47
	Farsi	0.549	0.605	48, 33, 48
	Mandarin	0.553	0.607	67, 26, 47
Chinese	English	0.562	0.622	71, 29, 51
	Farsi	0.55	0.608	70, 26, 55
	Mandarin	0.552	0.611	67, 22, 53

Coordinates are in X,Y,Z MNI space.

information over time. To grasp the meaning of a story we must continually stitch together individual words, link events with their causes, and remember what came before to build a holistic understanding. While some aspects of semantic integration appear to happen very quickly [Christiansen and Chater, 2016; Hagoort et al., 2004], there is evidence that the DMN is involved in a process that unfolds over longer time scales [Hasson et al., 2015; Honey et al., 2016; Lerner et al., 2011]. The relatively slow fluctuations found in the DMN [Zou et al., 2008] may therefore relate to a semantic integration process that extends over long time windows such as those required to understand a story. This is consistent with fMRI data showing that shared stimulus-locked activity in the DMN increases as longer segments of a story remain intact [Lerner et al., 2011; Simony et al., 2016], and with evidence that there are shared patterns of activity in the DMN across participants when describing scenes from an audiovisual story [Chen et al., 2017]. The DMN has a profile of anatomical and functional connectivity [Heuvel and van den Sporns, 2013] that situates it to participate in the process of abstracting across information from a diversity of neural systems, and it may function as a very high-level convergence and divergence zone [Fernandino et al., 2015; Kaplan et al., 2016]. In this hierarchical model of neural architecture, the DMN can be activated by, and in turn is able to reactivate, multiple lower-level brain systems that represent information in more explicit, mapped formats [Meyer and Damasio, 2009]. Such properties may be instrumental to the DMN's apparent role in constructing meaning that transcends the specifics of the language in which information is presented.

While previous work has investigated cross-linguistic decoding, it has focused on translating semantic representations of individual words [Buchweitz et al., 2012; Correia et al., 2014, 2015; Zinszer et al., 2016] rather than higher-level semantic constructs. Most recently, Zinszer et al. [2016] showed that neuro-semantic representations of words are preserved across both individuals and languages by translating English words into their Chinese counterparts using brain activation of independent groups of native English and Chinese speakers. Notably, other work has examined modal [Regev et al., 2013; Wilson et al., 2008] and linguistic [Honey et al., 2012] representational invariance in narrative processing, however these studies have not investigated whether neural representations of story-level features can be decoded. Thus, while Honey et al. [2012] found that a set of regions showed similar activation patterns among both English and Russian-speaking participants who listened to time-matched recordings of the same narrative, the invariance of story-level representations has not previously been addressed.

Accordingly, while this research collectively makes a strong case for the cross-linguistic invariance of word-level semantic representation and suggests that invariance might exist for higher-level narrative constructs, the fidelity with which this higher-level representational invariance holds for higher-level semantic constructs has been less certain. In contrast, by recording participant responses to multiple narratives, we were able to explore not only cross-linguistic invariance, but also test whether fMRI data during story reading contained sufficient information for identifying which narrative the activity corresponded to. Once we learned the relationship between the neural maps and the story representations, we were able to predict the representation of the left-out stories based just on fMRI activity data. Our inter-language results do not appear to be easily explained by relationships between the story vectors across languages, given that vectors for a given story were not correlated with each other across languages. Even though the inter-language vectors are very likely correlated in a higher order, given that we are using *linear* regression, the higher order correlations cannot be picked up by our model. Therefore, our inter-language decoding results cannot be due to the fact that the story vectors across the different languages are highly similar, but that there is considerable overlap in the neuro-semantic representation of the stories across the languages.

To our knowledge, this research is the first to demonstrate within and cross-linguistic neural decoding of entire narrative sequences. By conducting cross-language decoding, we were able to functionally control for low-level, language-specific semantic constructs. If the accuracies of the within-language story predictions were contingent on the variation of specific lexical or syntactic features captured in the story representations, a substantial drop in accuracy should have been observed in the cross-language predictions. However, the accuracies of the within- and between-language predictions were comparable, which indicates that these predictions exploited systematic encodings of story-level elements that transcended the idiosyncratic lexical and syntactic features of individual languages.

Overall, our research is an attempt to demonstrate that abstracted beyond the level of independent concepts and language units, the brain seems to systematically encode high-level narrative elements. Further, despite the remarkable variety of human language, the combination of a shared cognitive architecture and overlapping sociocultural experiences produces a remarkable cross-language consistency not only in the words and concepts that we use, but also in the narratives that we construct. Our results indicate that this similarity is echoed in the neuro-semantic representation of narrative-level information.

AUTHOR CONTRIBUTION

MD and JTK developed the concept and were involved in all phases of the study. Data analysis was performed by RB. KM, JH, and JDZ were involved in interpretation of the results and writing of the manuscript. SIG contributed to the study design and data collection. AV was consulted on the computational analysis of the data. ASG contributed to the study design and assisted with NLP model training. AD was involved in the study design and contributed to the interpretation of the study. All authors approved the final version of the manuscript for submission.

ACKNOWLEDGMENTS

This research was sponsored by Institute for New Economic Thinking and the Defense Advanced Research Projects Agency, Grant number D12AP00069. The content of this publication does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. Part of the fMRI data used in this paper has been previously published in Kaplan et al. [2016]. The research focus and data analysis of the two papers, however, are mutually exclusive. A.V. worked on this project while at USC.

REFERENCES

AleAhmad A, Amiri H, Darrudi E, Rahgozar M, Oroumchian F (2009): Hamshahri: A standard persian text collection. Knowl Based Syst 22:382–387.

Ames DL, Honey CJ, Chow MA, Todorov A, Hasson U (2015): Contextual alignment of cognitive and neural dynamics. J Cogn Neurosci 27:655–664.

Andersson JLR, Jenkinson M, Smith S (2010): Non-linear registration, aka spatial normalisation. FMRIB technical report TR07JA2.

Binder JR, Desai RH, Graves WW, Conant LL (2009): Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. Cerebr Cortex 19:2767–2796.

Blei DM, Ng AY, Jordan MI (2003): Latent dirichlet allocation. J Mach Learn Res 3:993–1022.

Buchweitz A, Shinkareva SV, Mason RA, Mitchell TM, Just MA (2012): Identifying bilingual semantic neural representations across languages. Brain Lang 120:282–289.

Buckner RL, Andrews-Hanna JR, Schacter DL (2008): The brain's default network. Ann N Y Acad Sci 1124:1–38.

Chen J, Leong YC, Honey CJ, Yong CH, Norman KA, Hasson U (2017): Shared memories reveal shared structure in neural activity across individuals. Nat Neurosci 20:115–125.

Christiansen MH, Chater N (2016): The now-or-never bottleneck: A fundamental constraint on language. Behav Brain Sci 39:e62.

Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A (2017): Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv 1705: 02364.

Correia J, Formisano E, Valente G, Hausfeld L, Jansma B, Bonte M (2014): Brain-based translation: fmri decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe. J Neurosci 34:332–338.

Correia J, Jansma B, Hausfeld L, Kikkert S, Bonte M (2015): Eeg decoding of spoken words in bilingual listeners: from words to language invariant semantic-conceptual representations. Front Psychol 6:71.

Dai AM, Olah C, Le QV (2015): Document embedding with paragraph vectors. arXiv preprint arXiv 1507:07998.

Damasio H, Tranel D, Grabowski T, Adolphs R, Damasio A (2004): Neural systems behind word and concept retrieval. Cognition 92:179–229.

Deerwester SC, Dumais ST, Landauer TK, Furnas GW, Harshman RA (1990): Indexing by latent semantic analysis. JAsIs 41:391–407.

Fernandino L, Binder JR, Desai RH, Pendl SL, Humphries CJ, Gross WL, Conant LL, Seidenberg MS (2015): Concept representation reflects multimodal abstraction: A framework for embodied semantics. Cerebr Cortex 26:2018–2034.

Ferstl EC, Neumann J, Bogler C, Von Cramon DY (2008): The extended language network: a meta-analysis of neuroimaging studies on text comprehension. Hum Brain Mapp 29:581–593.

Fyshe A, Talukdar PP, Murphy B, Mitchell TM (2014): Interpretable semantic vectors from a joint model of brain-and text-based meaning. In *Proc. of ACL*.

Garten J, Sagae K, Ustun V, Dehghani M (2015): Combining distributed vector representations for words. In *Proceedings of NAACL-HLT* (pp. 95–101).

Gordon A, Huangfu L, Sagae K, Mao W, Chen W (2013): Identifying personal narratives in Chinese weblog posts. In *Intelligent narrative technologies workshop, Boston, MA*.

Gordon A, Swanson, R (2009): Identifying personal stories in millions of weblog entries. In *Third international conference on weblogs and social media, data challenge workshop, San Jose, CA* (Vol. 46).

Grabowski TJ, Damasio H, Tranel D, Ponto LLB, Hichwa RD, Damasio AR (2001): A role for left temporal pole in the retrieval of words for unique entities. Hum Brain Mapp 13:199–212.

Guimaraes MP, Wong DK, Uy ET, Grosenick L, Suppes P (2007): Single-trial classification of meg recordings. IEEE Trans Biomed Eng 54:436–443.

Hagoort P, Hald L, Bastiaansen M, Petersson KM (2004): Integration of word meaning and world knowledge in language comprehension. Science 304:438–441.

Hasson U, Chen J, Honey CJ (2015): Hierarchical process memory: memory as an integral component of information processing. Trends Cognit Sci 19:304–313.

Heuvel MP, van den Sporns O (2013): Network hubs in the human brain. Trends Cognit Sci 17:683–696.

Honey CJ, Chen J, Müsch K, Hasson U (2016, Commentary): How long is now? the multiple timescales of language processing. Commentary on: Christiansen & Chater "The Now-or-Never Bottleneck: A Fundamental Constraint on Language." Behav Brain Sci 39.

Honey CJ, Thompson CR, Lerner Y, Hasson U (2012): Not lost in translation: neural responses shared across languages. J Neurosci 32:15277–15283.

Huth AG, Heer WA, de, Griffiths TL, Theunissen FE, Gallant JL (2016): Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532:453–458.

Huth AG, Nishimoto S, Vu AT, Gallant JL (2012): A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76:1210–1224.

Jenkinson M, Bannister P, Brady M, Smith S (2002): Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17:825–841.

Johnson-Laird PN (1983): Mental models: Towards a cognitive science of language, inference, and consciousness, Vol. 6. Cambridge, MA: Harvard University Press.

Just MA, Cherkassky VL, Aryal S, Mitchell TM (2010): A neurosemantic theory of concrete noun representation based on the underlying brain codes. PloS One 5:e8622.

Kaplan JT, Gimbel SI, Dehghani M, Immordino-Yang MH, Sagae K, Wong JD, et al. (2016): Processing narratives concerning protected values: A cross-cultural investigation of neural correlates. Cerebr Cortex bhv325.

Kintsch W (1974): The Representation of Meaning in Memory. Lawrence: Erlbaum.

Kintsch W (1988): The role of knowledge in discourse comprehension: a construction-integration model. Psychol Rev 95:163.

Lau JH, Baldwin T (2016): An empirical evaluation of doc2vec with practical insights into document embedding generation. arXiv preprint arXiv 1607–05368. https://arxiv.org/abs/1607.05368.

Le QV, Mikolov T (2014): Distributed representations of sentences and documents. In *ICML* (Vol. 14, pp. 1188–1196).

Lerner Y, Honey CJ, Silbert LJ, Hasson U (2011): Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J Neurosci 31:2906–2915.

Ma M, Huang L, Xiang B, Zhou B (2015): Dependency-based convolutional neural networks for sentence embedding. arXiv preprint arXiv 1507:01839.

Meyer K, Damasio A (2009): Convergence and divergence in a neural architecture for recognition and memory. Trends Neurosci 32:376–382.

Mikolov T, Chen K, Corrado G, Dean J (2013): Efficient estimation of word representations in vector space. arXiv preprint arXiv 1301:3781.

Mitchell TM, Shinkareva SV, Carlson A, Chang K-M, Malave VL, Mason RA, Just MA (2008): Predicting human brain activity associated with the meanings of nouns. Science 320:1191–1195.

Østby Y, Walhovd KB, Tamnes CK, Grydeland H, Westlye LT, Fjell AM (2012): Mental time travel and default-mode network functional connectivity in the developing brain. Proc Natl Acad Sci 109:16800–16804.

Palangi H, Deng L, Shen Y, Gao J, He X, Chen J, Song X, Ward R (2016): Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. IEEE/ACM Trans Audio, Speech Lang Process (TASLP) 24:694–707.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J (2011): Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830.

Qin P, Northoff G (2011): How is our self related to midline regions and the default-mode network? Neuroimage 57:1221–1233.

Raichle ME (2015): The brain's default mode network. Ann Rev Neurosci 38:433–447.

Raichle ME, Snyder AZ (2007): A default mode of brain function: a brief history of an evolving idea. Neuroimage 37:1083–1090.

Regev M, Honey CJ, Simony E, Hasson U (2013): Selective and invariant neural responses to spoken and written narratives. J Neurosci 33:15978–15988.

Řehůřek R, Sojka, P (2010, May 22): Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA. (http://is.muni.cz/publication/884893/en)

Rumelhart DE, Hinton GE, Williams RJ (1986): Learning representations by back-propagating errors. Nature 323:533–536.

Sagae K, Gordon AS, Dehghani M, Metke M, Kim JS, Gimbel SI, Tipper C, Kaplan J, Immordino-Yang M (2013). A data-driven approach for classification of subjectivity in personal narratives. Hamburg, Germany: 2013 Workshop on Computational Models of Narrative (pp. 198–213).

Sanford AJ, Garrod SC (1998): The role of scenario mapping in text comprehension. Discourse Processes 26:159–190.

Shinkareva SV, Malave VL, Mason RA, Mitchell TM, Just MA (2011): Commonality of neural representations of words and pictures. Neuroimage 54:2418–2425.

Simony E, Honey CJ, Chen J, Lositsky O, Yeshurun Y, Wiesel A, Hasson U (2016): Dynamic reconfiguration of the default mode network during narrative comprehension. Nature Commun 7: 12141.

Smallwood J, Schooler JW (2015): The science of mind wandering: empirically navigating the stream of consciousness. Annu Rev Psychol 66:487–518.

Smallwood J, Tipper C, Brown K, Baird B, Engen H, Michaels JR, Grafton S, Schooler JW (2013): Escaping the here and now: evidence for a role of the default mode network in perceptually decoupled thought. Neuroimage 69:120–125.

Smith SM, Nichols TE (2009): Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44:83–98.

Spreng RN (2012): The fallacy of a "task-negative" network. Front Psychol 3:145.

Sudre G, Pomerleau D, Palatucci M, Wehbe L, Fyshe A, Salmelin R, Mitchell T (2012): Tracking neural coding of perceptual and semantic features of concrete nouns. NeuroImage 62:451–463.

Tranel D, Damasio H, Damasio AR (1997): A neural basis for the retrieval of conceptual knowledge. Neuropsychologia 35: 1319–1327.

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:e112575.

Whitney C, Huber W, Klann J, Weis S, Krach S, Kircher T (2009): Neural correlates of narrative shifts during auditory story comprehension. Neuroimage 47:360–366.

Wilson SM, Molnar-Szakacs I, Iacoboni M (2008): Beyond superior temporal cortex: intersubject correlations in narrative speech comprehension. Cerebr Cortex 18:230–242.

Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE (2014): Permutation inference for the general linear model. Neuroimage 92:381–397.

Yang Y, Wang J, Bailer C, Cherkassky V, Just MA (2017): Commonality of neural representations of sentences across languages: Predicting brain activation during portuguese sentence comprehension using an english-based model of brain function. NeuroImage 146:658–666.

Yarkoni T, Speer NK, Zacks JM (2008): Neural substrates of narrative comprehension and memory. Neuroimage 41:1408–1425.

Yogatama D, Faruqui M, Dyer C, Smith NA (2014): Learning word representations with hierarchical sparse coding. arXiv preprint arXiv 1406:2035.

Zadbood A, Chen J, Leong YC, Norman KA, Hasson U (2016): How we transmit memories to other brains: constructing shared neural representations via communication. bioRxiv 081208.

Zinszer BD, Anderson AJ, Kang O, Wheatley T, Raizada R (2015). You say potato, i say tǔdòu: How speakers of different languages share the same concept. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

Zinszer BD, Anderson AJ, Kang O, Wheatley T, Raizada RD (2016): Semantic structural alignment of neural representational spaces enables translation between english and chinese words. J Cognit Neurosci 28:1749–1759.

Zou Q-H, Zhu C-Z, Yang Y, Zuo X-N, Long X-Y, Cao Q-J, Wang YF, Zang YF (2008): An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: fractional ALFF. J Neurosci Methods 172:137–141.