Analysis of Story Vectors

As discussed in the main paper, the intention of this paper is not to compare representations produced by doc2vec to other story-level representations. Our aim is to show that aggregated story-level embeddings can be used to decode narrative-processing fMRI data. Nonetheless, we performed a behavioral experiment to demonstrate the representativeness of doc2vec vectors compared to aggregate word2vec vectors.

As discussed previously, doc2vec has been shown to be more effective in capturing the semantics of long pieces of text compared to other word-level techniques (Dai, Olah, & Le, 2015; Lau & Baldwin, 2016). Here we present a behavioral experiment demonstrating that doc2vec captures the overall meaning of stories more precisely than aggregated word-level operations (i.e. word2vec). Specifically, we compare the stories to their nearest neighbors in two semantic spaces, one constructed using doc2vec and the other using word-level operations, and show that the semantic representations at the story level are more indicative of the overall meaning of the stories. Further, we investigate how much overlap there is between the first three nearest neighbors constructed using the two techniques.

Method. First, doc2vec was used to represent each story in our corpus of 40 English stories in 100 dimensional semantic space. For each story, we then queried the closest nearest neighbor (in the semantic space) for that story from the rest of the 39 stories. Thus, for each story, we obtained a nearest neighbor prediction from doc2vec. Next, we used word2vec to represent each word in each story in a semantic space, and aggregated the word vectors to represent the full story. Summation of word vectors is an operation that is meaningful and used frequently in natural language processing (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). We then, queried the closest nearest neighbors of each word2vec represented story using this word-level summation technique. We then randomly chose 10 stories (out of 40) along with their nearest neighbors calculated using the two methods described above and used them as stimuli in the behavioral experiment described below.

One-hundred American participants were recruited through Amazon Mechanical

Turk. The task was described as follows to the participants:

> Thank you for taking time to participate in this experiment. In this task, you will be reading several different stories. For each story, you will be asked to judge how similar that story is to two other stories. You should base your judgment on the overall meaning of the stories, or what is often refer to as "the gist" of the stories, not on the similarity of the individual words included them. Please read each story separately, and think about the overall meaning of it before making any judgments. Again, we appreciate your help with this task.

Each participants was then presented with four sets of stories, along with the two nearest neighbors (calculated using the above techniques) for each story. We also presented two attention check questions to each participant. For each story, the participants were asked the following question: "Is the overall meaning of the above story more similar to the first or second option?". The order of the stories, and the two choices for each story, were fully randomized.

Results and Discussion. 38 participants failed one of the two attention test questions and were excluded from the analysis. Overall, a significantly larger number of participants chose the doc2vec nearest neighbor as the option that represented the "gist" of the original stories compared to the word-level method $\chi^2(1) = 11.121, p = 0.0008$. This result provides further evidence that the method we used in the paper represents the overall meaning of the stories more precisely that closely related word-level analysis techniques.

Further, we investigated how many of the first three closest neighbors of the stories match across the two representations. Knowing this detail is important as it provides insight into how much the semantic spaces constructed using the word-level technique and doc2vec are similar. Supporting our argument, this analysis revealed only a 21.67% overlap between the nearest neighbors of the same stories between the two representations. In other words, in 78% of the cases the closest neighbors did not match between the space constructed using doc2vec and the one constructed using word2vec.

Together, these studies indicate that doc2vec representations more closely

approximate human perceptions of narrative similarity, compared to word2vec representations, and that doc2vec predictions of similarity are substantially different from word2vec predictions of similarity. While these results do not *prove* that doc2vec actually represents higher-level meaning, rather than mere word-level meaning, in conjunction with the recent literature reviewed above, it is difficult to argue the converse.

Whole Brain Analysis

Whole brain analysis was also performed on the data discussed in the paper. The goal here was to predict the story vectors based on fMRI data from the whole brain, rather than within spatially restricted searchlight neighborhoods (as discussed the main text). For each participant, a ridge regression model was fitted on the 40*212,018 fMRI matrix, with its accompanied responses at 40*100 story matrix. 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 participant. In each fold, using the trained model on the 38 stories, the story vectors were predicted for the two left-out stories. The evaluation criteria was exactly the same as formula (1) in the paper.

Similar to the first experiment, the decoding accuracy for each subject was calculated by averaging the accuracies of the classification over the 780 folds. In order to do the cross-lingual analyses, the same process as described above was performed, with one difference: the fMRI vectors of one cultural group and story vectors of another language were used for decoding.

In order to establish a baseline chance performance, forty 100 dimensional random vectors with the same variance and mean as the actual story vectors were generated for each language, and were analyzed using the same procedure as actual story vectors.

We also tested two alternative numeric representations of the English stories: 1. Latent Dirichlet Allocation (LDA, Blei, Ng, & Jordan, 2003): we used LDA to generate 100 topics for a corpus of 1000 English personal stories (subset of the corpus discussed

in the Distributed Representation of Stories section). We then calculated topic weights for the stimuli stories. These weights were used to decode the fMRI data. 2. Linguistic Inquery and Word Count (LIWC, Tausczik & Pennebaker, 2010): We ran LIWC 2010 on the stimuli set and used the weights of the 64 LIWC categories as the representations of the stories. These representations were then used to decode the data.

Results and Discussion. The intra and inter-language results are illustrated in Figure 1. Overall, the intra-lingual decoding was performed with an accuracy of 59.1%. This performance is both higher than chance $t(29) = 14.366, p < 0.0001$, d[95% CI] = 3.7093[2.8608, 4.5454] and higher than the performance of the random story vectors $t(58) = 8.7492, p < 0.0001$, d[95% CI] = 2.2590[1.6016, 2.9044]. The inter-lingual decoding was performed with accuracy of 58.9% which is again higher than chance $t(29) = 9.9939, p < 0.0001$, d[95% CI] = 2.5804[1.8846, 3.2638] and higher than using random vectors $t(58) = 7.5023, p < 0.0001$, d[95% CI] = 1.9371[1.3151, 2.5476] for decoding. Figure 2 illustrates the results broken down by language and culture (See Table 1 for the complete statistics of this figure). A possible explanation for why the decoding performance on American participants is higher could be due to the fact that all these stories originated from a corpus of English stories from popular American blogposts.

The LDA representation of the stories did not perform better than the random vectors $t(58) = 1.6524, p = 0.1039$, nor chance $t(29) = 1.409, p = 0.1695$. The LIWC representations, however, performed better than the random vectors $t(58) = 3.6572, p = 0.0005$, d[95 %CI] = 0.9443[0.4063 , 1.4749] and chance $t(29) = 4.0067, p = 0.0004$, d[95 %CI] = 1.0345[0.4908, 1.5703]. LIWC representations perform significantly lower than doc2vec $t(58) = -2.9188, p = 0.005$, d[95 %CI] = -0.7536[-1.2748, -0.2264]. We would like to note that both (vanilla) LDA and LIWC are bag-of-words approaches, and therefore they do not capture sequential relation between words in passages. doc2vec, on the other hand, sequentially adds words to the model and hence, in theory, can capture these relations.

FSL Randomise program

These steps were all implemented by a single line of code to the FSL Randomise program:

```
randomise -i <inputData> -o <outputName> -1 -v 5 -T
```

The -i option identifies the input data, a single 4-D image (3 spatial dimensions plus a 4th concatenating data from 30 subjects). The -o option attaches the specified prefix to the output files. The -1 option indicates a one-sample t-test. The -v option selects variance smoothing at the default 5 mm value. The -T option selects the TFCE procedure.

Individual Language Searchlight Maps

Individual intra-language searchlight maps are presented in Figure 3, and inter-language maps in Figure 4.

Relationship to Default Mode Network

In order to visualize the relationship between our classification results and the spatial distribution of the Default Mode Network, we have overlayed the intra-language searchlight map on the seed-based correlation map of the DMN from (Kaplan et al., 2016). This map of the DMN was derived from an analysis of resting state data from the same participants, using a seed centered in the precuneus to identify functionally correlated brain regions. This relationship is visualized in Figure 6.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993–1022.

Dai, A. M., Olah, C., & Le, Q. V. (2015). Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998*.

Kaplan, J. T., Gimbel, S. I., Dehghani, M., Immordino-Yang, M. H., Sagae, K., Wong, J. D., ... Damasio, A. (2016). Processing narratives concerning protected values: A cross-cultural investigation of neural correlates. *Cerebral Cortex*, bhv325.

Lau, J. H., & Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24–54.

Table 1

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

Culture	Comparison (Language vs random vectors)	*d*[95% CI]	*df*	*t*	sig
Americans	English	1.7070[1.1081, 2.2950]	58	6.6112	<.001
	Farsi	1.6658[1.0709, 2.2500]	58	6.4517	<.001
	Mandarin	1.9265[1.3056, 2.5359]	58	7.4613	<.001
Iranians	English	1.3351[0.7692, 1.8915]	58	5.1708	<.001
	Farsi	1.0235[0.4805, 1.5587]	58	3.9639	<.001
	Mandarin	1.0576[0.5123, 1.5949]	58	4.0962	<.001
Chinese	English	1.0124[0.4702, 1.5470]	58	3.9212	<.001
	Farsi	0.6039[0.0837, 1.1192]	58	2.3391	0.0228
	Mandarin	0.7425[0.2158, 1.2632]	58	2.8759	0.0056

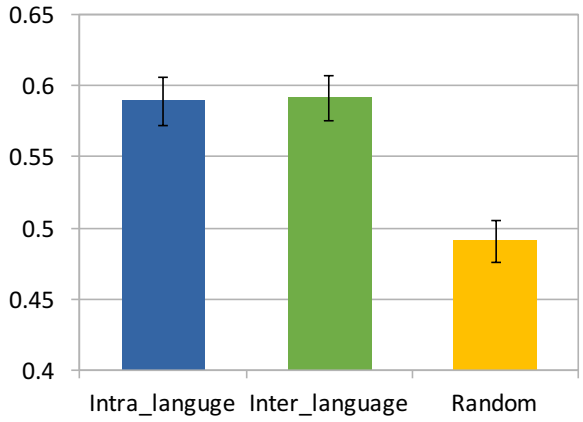

Figure 1. Intra and Inter language decoding performances

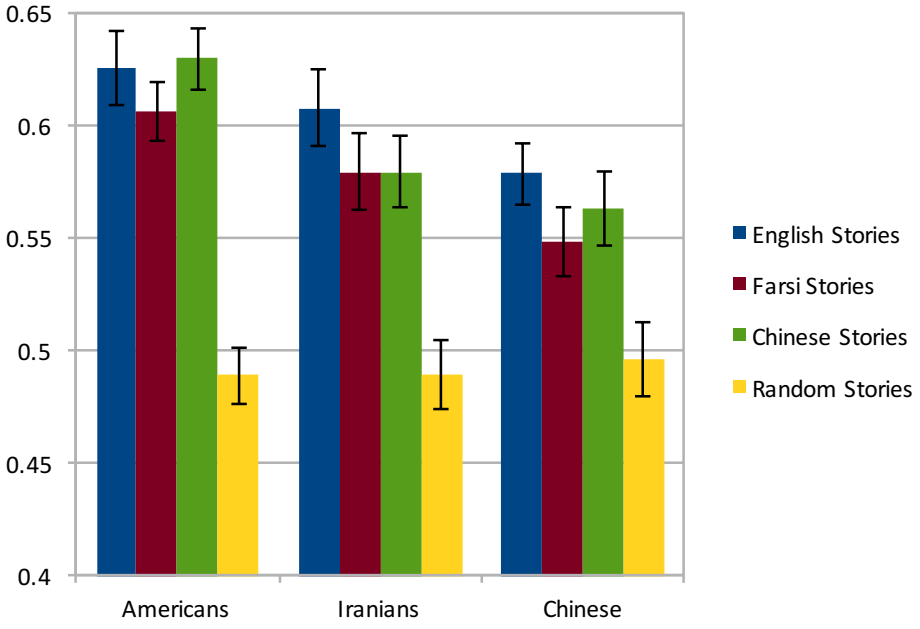

Figure 2. Decoding performances broken down by language and culture

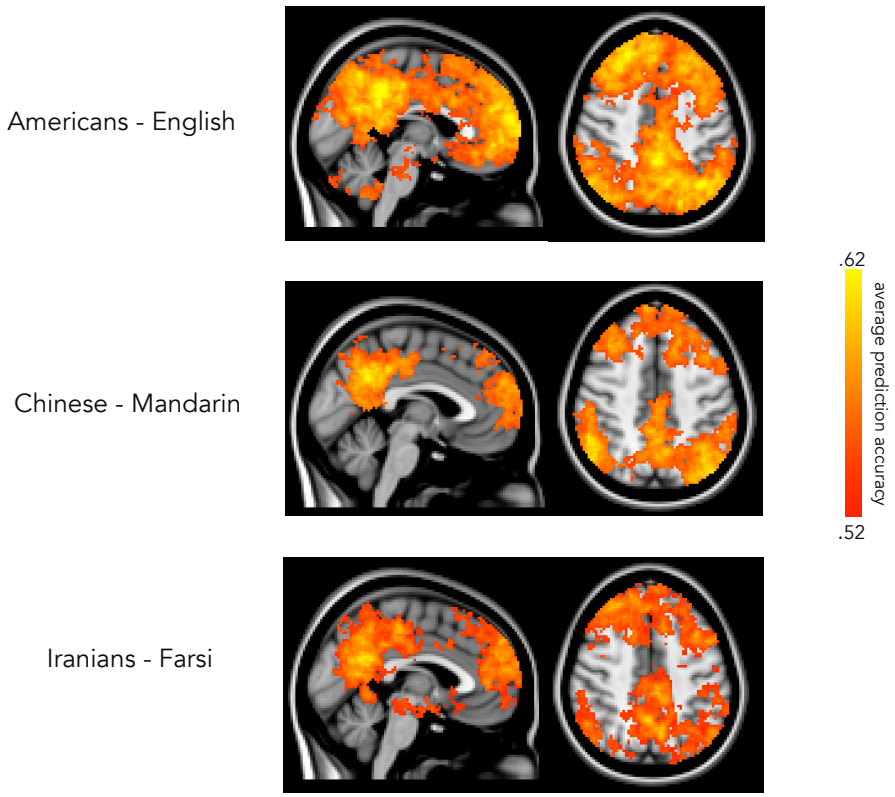

Figure 3. Intra-language searchlight maps

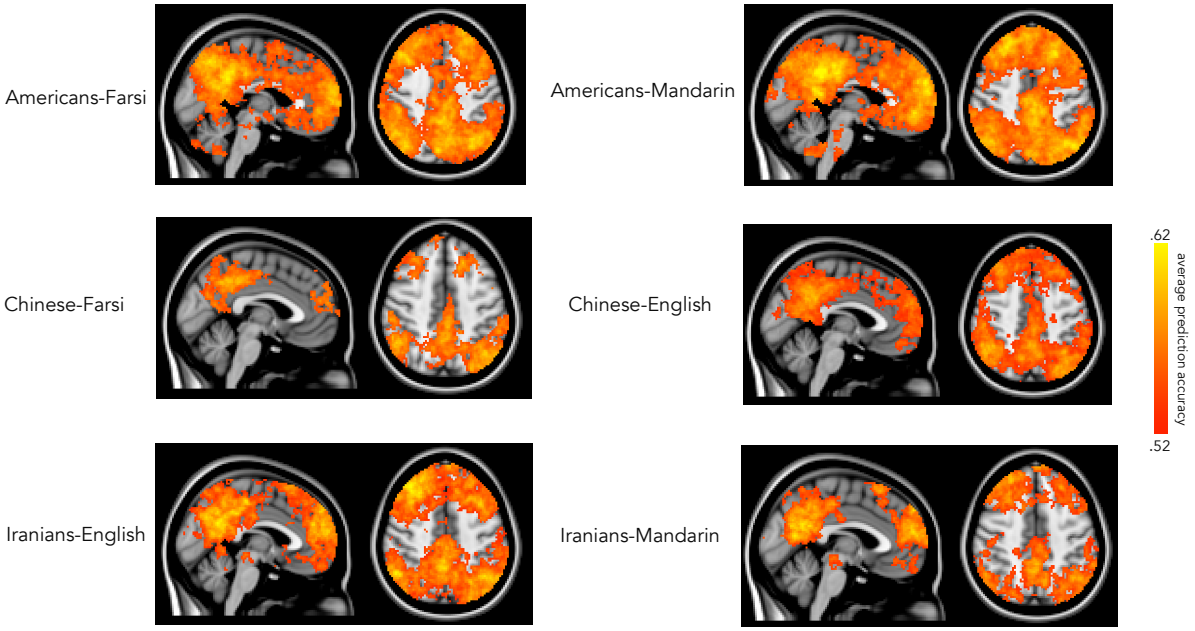

Figure 4. Inter-language searchlight maps

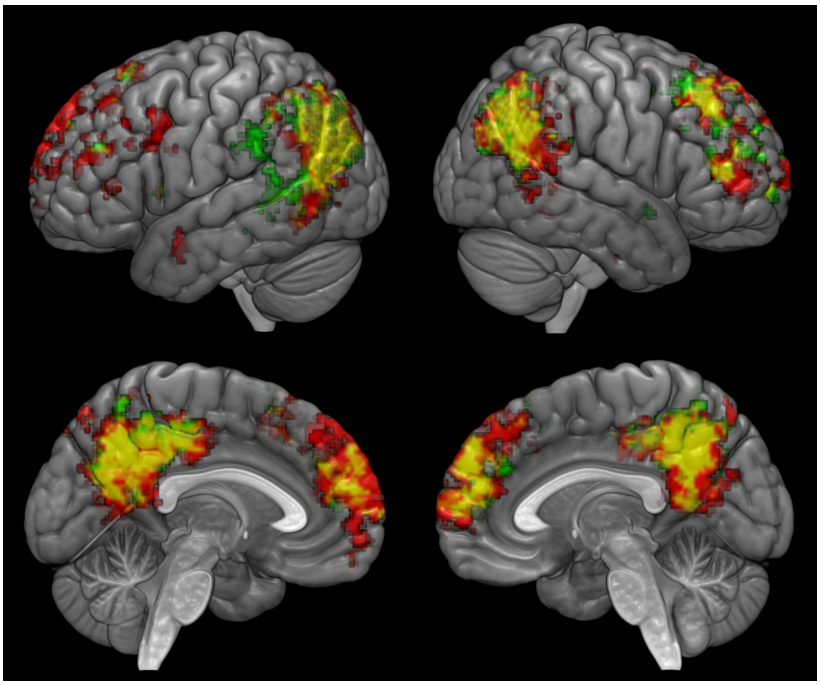

Figure 5. Overlap of intra- and inter-language searchlight maps. The maps from intra-language classification (red, see Fig. 1 in main text) and inter-language classification (green, see Fig. 2) are superimposed onto the same brain, with overlapping regions shown in yellow. There is large overlap between maps, with the notable exception of intra language-specific decoding observed in left superior and middle frontal gyrus.

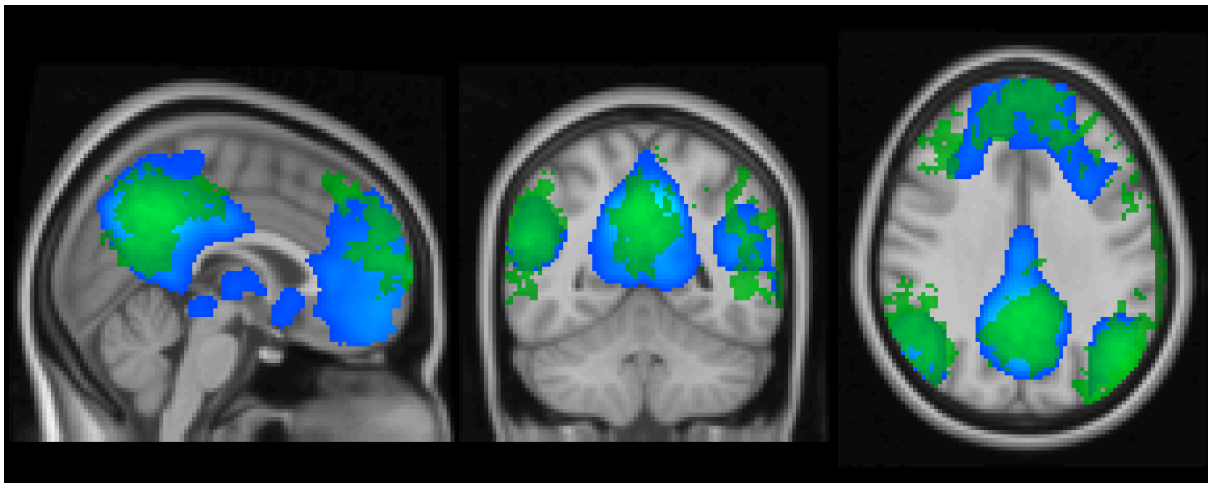

Figure 6. Relationship between story classification and Default Mode Network. A precuneus seed-based correlation map of the DMN is shown in blue, with intra-language classification overlayed in green.