### 5.2.2 Effect on Rows Without Brain Data

It is possible that some JNNSE(Brain+Text) dimensions are being used exclusively to fit brain activation data, and not the semantics represented in both brain and corpus data. If a particular dimension $j$ is solely used for brain data, the sparsity constraint will favor solutions that sets $A_{(i,j)} = 0$ for $i > w'$ (no brain data constraint), and $A_{(i,j)} > 0$ for some $0 \leq i \leq w'$ (brain data constrained). We found that there were no such dimensions in the JNNSE(Brain+Text). In fact for the $\ell = 1000$ JNNSE(Brain+Text), all latent dimensions had greater than $\sim 25\%$ non-zero entries, which implies that all dimensions are being shared between the two data inputs (corpus and brain activation), and are used to reconstruct both.

To test that the brain activation data is truly influencing rows of $A$ not constrained by brain activation data, we performed a *dropout* test. We split the original 60 words into two 30 word groups (as evenly as possible across word categories). We trained JNNSE(fMRI+Text) with 30 words, and tested word prediction with the remaining 8 subjects and the other 30 words. Thus, the training and testing word sets are disjoint. Because of the reduced size of the training data, we did see a drop in performance, but JNNSE(fMRI+Text) vectors still gave word prediction performance 7% higher than NNSE(Text) vectors. Full results appear in the supplementary material.

## 5.3 Predicting Corpus Data

Here we ask: can an accurate latent representation of a word be constructed using only brain activation data? This task simulates the scenario where there is no reliable corpus representation of a word, but brain data is available. This scenario may occur for seldom-used words that fall below the thresholds used for the compilation of corpus statistics. It could also be useful for acronym tokens (lol, omg) found in social media contexts where the meaning of the token is actually a full sentence.

We trained a JNNSE(fMRI+Text) with brain data for all 60 words, but withhold the corpus data for 30 of the 60 words (as evenly distributed as possible amongst the 12 word categories). The brain activation data for the 30 withheld words will allow us to create latent representations in $A$ for withheld words. Simultaneously, we will learn a mapping from the latent representation to the corpus data ($D^{(c)}$). This task cannot be performed with a NNSE(Text) model because one cannot learn a latent representation of a word without data of some kind. This further emphasizes the impact of brain imaging data, which will allow us to generalize to previously unseen words in corpus space.

Table 2: Mean rank accuracy over 30 words using corpus representations predicted by a JNNSE(MEG+Text) model trained with some rows of the corpus data withheld. Significance is calculated using Fisher's method to combine p-values for each of the subject-dependent models.

| Latent Dim. size | Rank Accuracy | p-value |
|---|---|---|
| 250 | 63.30 | $< 10^{-19}$ |
| 500 | 67.37 | $< 10^{-24}$ |
| 1000 | 63.47 | $< 10^{-15}$ |

We use the latent representations in $A$ for each of the words without corpus data and the mapping to corpus space $D^{(c)}$ to predict the withheld corpus data in $X$. We then rank the withheld rows of $X$ by their distance to the predicted row of $X$ and calculate the mean rank accuracy of the held out words. Results in Table 2 show that we can recreate the withheld corpus data using brain activation data. Peak mean rank accuracy (67.37) is attained at $\ell = 500$ latent dimensions. This result shows that neural semantic representations can create a latent representation that is faithful to unseen corpus statistics, providing further evidence that the two data sources share a strong common element.

How much power is the remaining corpus data supplying in scenarios where we withhold corpus data? To answer this question, we trained an NNSE(Brain) model on 30 words of brain activation, and then trained a regressor to predict corpus data from those latent brain-only representations. We use the trained regressor to predict the corpus data for the remaining 30 words. Peak performance is attained at $\ell = 10$ latent dimensions, giving mean rank accuracy of 62.37, significantly worse than the model that includes both corpus and brain activation data (67.37).

## 5.4 Mapping Semantics onto the Brain

Because our method incorporates brain data into an interpretable semantic model, we can directly map semantic concepts onto the brain. To do this, we examined the mappings from the latent space to the brain space via $D^{(b)}$. We found that the most interpretable mappings come from mod-