Tracking Story Reading in the Brain

Leila Wehbe[1,2,3]

[1] Machine Learning Department, Carnegie Mellon University, Pittsburgh
[2] Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh
[3] lwehbe@cs.cmu.edu

Abstract. Story comprehension is a rich and rapid phenomenon that requires multiple simultaneous processes (e.g. letter recognition, word understanding, sentence parsing...). Our goal is to study these complex processes in the brain by modeling the fMRI brain activity during story reading at a close to normal speed. This is a challenging goal, one reason being the coarse time-resolution of fMRI and another the lack of a comprehensive model of word meaning composition. Classically, fMRI has been used to localize brain areas that process specific elements of text processing (e.g. which areas are involved in syntactic processing) but not to model how the brain represents different instances of these elements (e.g. how do those areas represent different syntactic structures). We present here a generative model that predicts the fMRI activity created when subjects read a complex story where the words are presented in a serial manner, for 0.5 seconds each. Using this model, we performed an exploratory analysis in which we tested several types of story features (e.g. word length, syntax, semantics, story characters) to search for a good basis of features for story comprehension. We found different patterns of representation in the brain for different types of features. These patterns align with the predictions from the field. We tested the expressivity of our model using a classification task that decodes a passage of the story from a time segment of brain activity. We obtain a classification accuracy that is significantly higher than chance with $p < 10^{-6}$. We show that we can indeed study multiple components of reading simultaneously in fMRI at a close to normal speed. Our approach has the advantage of being flexible: any feature of language can be added to the model and tested, and features can range from simple perceptual features, to compositional semantics, to higher order reasoning about narrative structure and story comprehension.

1 Introduction

Story comprehension is a rich and rapid phenomenon that requires multiple simultaneous processes (e.g. letter recognition, word understanding, sentence parsing...). Our goal is to study these complex processes in the brain by modeling the fMRI brain activity during story reading at a close to normal speed. Although this is a challenging goal - one reason being the coarse time-resolution of fMRI - success in experiments that reconstruct video from fMRI activity encourages such an endeavor [22]. However, unlike the visual system, the mapping between elements of the language stimulus and the brain regions that represent them is poorly understood. Furthermore, it is not clear what kind of language features provide an adequate representation of the content of language and how to automate their extraction from the stimulus.

There has been multiple studies of story processing using fMRI. Some approaches use a story paradigm to analyze language processing in a natural setting. In [7], the authors use a story listening paradigm in which they parse all of the sentences and compute a measure of syntactic complexity for the integration of each word with the words that preceded it. They then identify the brain regions that have a time course of activity that correlates with syntactic complexity. Other approaches to studying story processing focus on narrative structure building [30, 32]. For instance, in [30], the authors identify a few narrative features such as the change of protagonist, a change in the protagonist's goals or a change in location. They then identify brain regions that are correlated with a change in a narrative feature. Both of these types of approaches therefore aim to identify regions that are involved in a given sub-process of story reading. We are interested in a more comprehensive problem: we want to model generatively the brain representation of the meanings and scenes of a story, and not only find the network of regions that are implicated in processing it. For instance, in addition to looking for correlates of semantic processing load, we want to understand how the brain represents different semantic entities. And instead of only looking for correlates of a change of character, goal or location, we want to find the representations of the different characters, goals and location.

FMRI has been used in language processing in this generative manner in several studies. In [20], the authors constructed a model that predicts the neural activity for concrete nouns. In order to achieve this, nouns are represented in a semantic space in which every dimension corresponds to their frequency of cooccurence with one of 25 verbs. This model learns the neural representation of every semantic dimension, and can predict the neural activity of any noun given its semantic features. However, this approach is limited to studying the representation of single words or a small group of words (via functions that combines semantic dimensions of consecutive words, as in [9]); while we would like

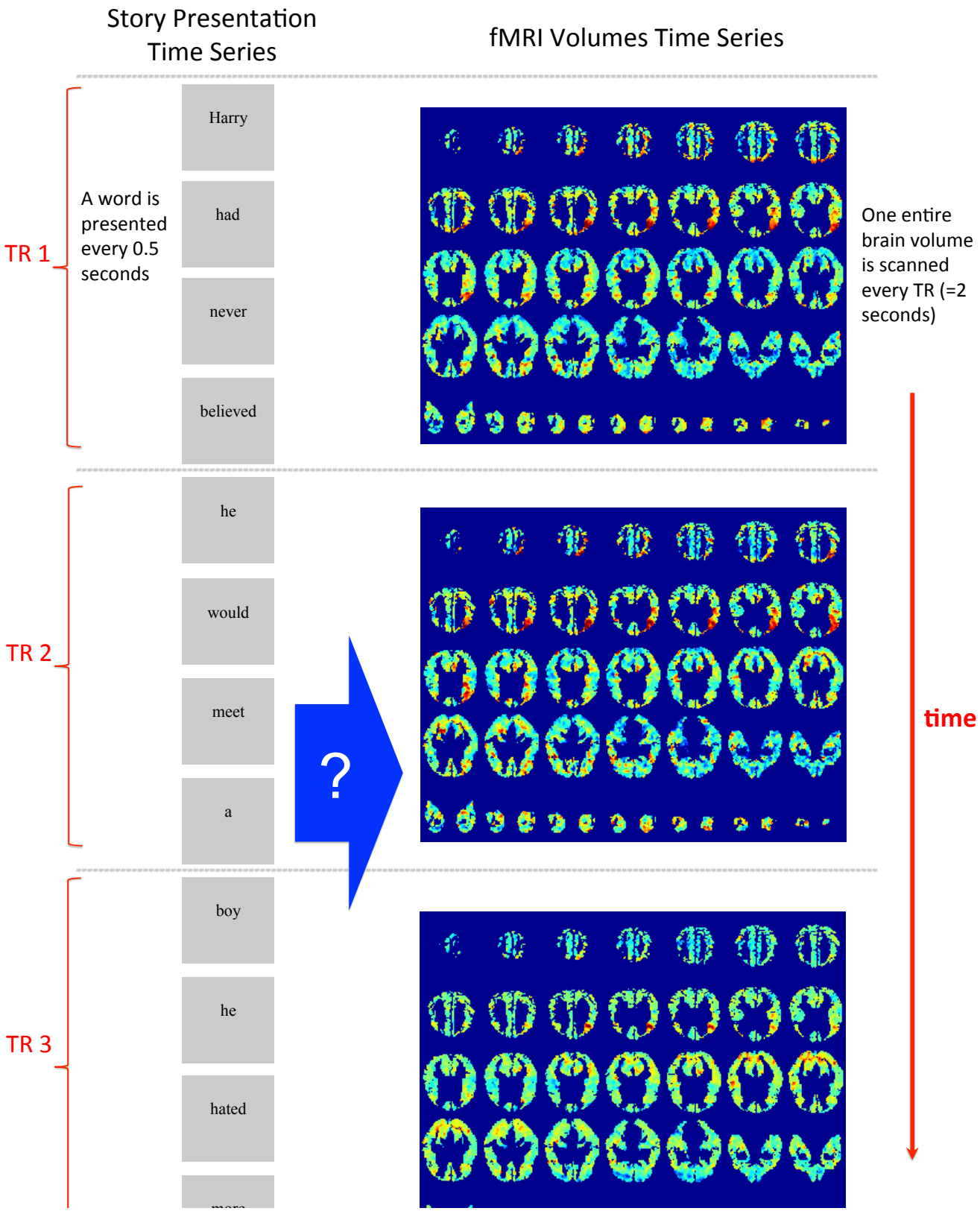


Fig. 1. Illustration of our method. We present the words of a story serially for 0.5 seconds each while recording brain activity with fMRI at a rate of one entire brain image by 2 seconds. Our goal is to model the perception and representation of this story by the brain via the fMRI data. Each fMRI activity volume is shown here in 35 horizontal slices: going right to left through the slices, then bottom-up, corresponds to looking at slices from the bottom of the brain up. On each slice, the top of the slice corresponds to the posterior of the brain, and the right side of the slice corresponds to the left side of the brain. The images are on a scale from blue to red where blue indicates negative deviation from baseline and red indicates positive deviations. A TR is the time needed to record one brain volume.

to model a more complex phenomenon that accounts for multiple levels of processing, meaning representation and reasoning.

Here, we propose a method that studies story comprehension when subjects are reading complex texts. We record fMRI data while presenting the words of a story to the subject in a serial manner, each for 0.5 seconds (see figure 1). We then perform an exploratory analysis in which we test several types of story features to search for a good basis of features for story comprehension. Our model can find the neural signatures of these different features in fMRI data while the subject is reading at a pace close to the pace of natural reading. Our model identifies where these types of features are represented in the brain, and what patterns of activity do different instances of these features have. To test the validity of this model, we perform a classification task that aims to decode what portion of the story a time segment of brain activity corresponds to. This classification task yields a classification accuracy that is significantly higher than chance ($p < 10^{-6}$), indicating that we can indeed study multiple components of reading simultaneously in fMRI at a close to normal pace. Our approach has the advantage of being flexible: any feature of language can be added to the model and tested, and features can range from simple perceptual features, to compositional semantics, to higher order reasoning about narrative structure and story comprehension.

2 Methods

2.1 Experimental design

Material Participants read chapter 9 of *Harry Potter and the Sorcerer's Stone* [26]. We chose this chapter because it involves many characters and spans multiple locations and scenes. We chose a famous book series because we hypothesized all subjects already had characteristic mental representations of the different characters and locations, and that at least a part of this representation would remain constant throughout the reading of chapter 9. This assumption allows us to use data from the entire chapter to look for the representation of the different characters, e.g. the protagonist Harry Potter. In contrast, had we chosen an unfamiliar story in which we learn about the protagonist's personality throughout the text, the mental representation of this protagonist will arguably change more than Harry's would.

Participants fMRI data was collected from 8 subjects (5 females and 3 males) recruited through Carnegie Mellon University, aged 18 to 35 years. The participants were all native English speakers and right handed. They were chosen to be familiar with the material: we made sure they had read the Harry Potter books or seen the movies series and were familiar with the characters and the story. All the participant were screened for safety, signed the consent form and were compensated for their participation. Data from one of the subjects was excluded from the analysis because of artifact that was not removed by our preprocessing procedure.

Design

The words of the story were presented in rapid serial visual format [8]. Words were presented one by one at the center of the screen for 0.5 seconds each (see figure 1). The background was gray and the font was black. We used MATLAB and the Psychophysics Toolbox extensions [5, 24, 18].

The chapter was divided into four runs, of approximately 11 minutes each. Subjects had short breaks between runs. Each run started with a fixation period of 20 seconds in which the subjects stared at a cross in the middle of the screen. The words presentation started after the fixation period. The total length of the runs was 45 minutes, during which about 5200 words were presented. Chapter 9 was presented in it's entirety without modifications and each subject read the chapter only once.

Before the experiment, we supplied the subjects with a summary of the events preceding chapter 9 and a summary of the main characters and concepts in *Harry Potter and the Sorcerer's Stone* to refresh their memory. We also instructed them to practice rapid serial presentation by viewing a video that replicated the parameters of our design, but with another story (*The Tale of Peter Rabbit* [29]). On the day of the experiment, the subjects were instructed to lay in the scanner and read the chapter as naturally as possible while remaining alert.

fMRI procedure Functional images were acquired on a Siemens Verio 3.0T scanner (Siemens, Erlangen, Germany) at the Scientific Imaging & Brain Imaging Center at Carnegie Mellon University, using a T2* sensitive echo planar imaging pulse sequence with repetition time (TR)= 2s, echo time=29 ms, flip angle=79°, 36 slices and $3 \times 3 \times 3$mm voxels. Anatomical volumes were acquired with a T1-weighted 3D-MPRAGE pulse sequence.

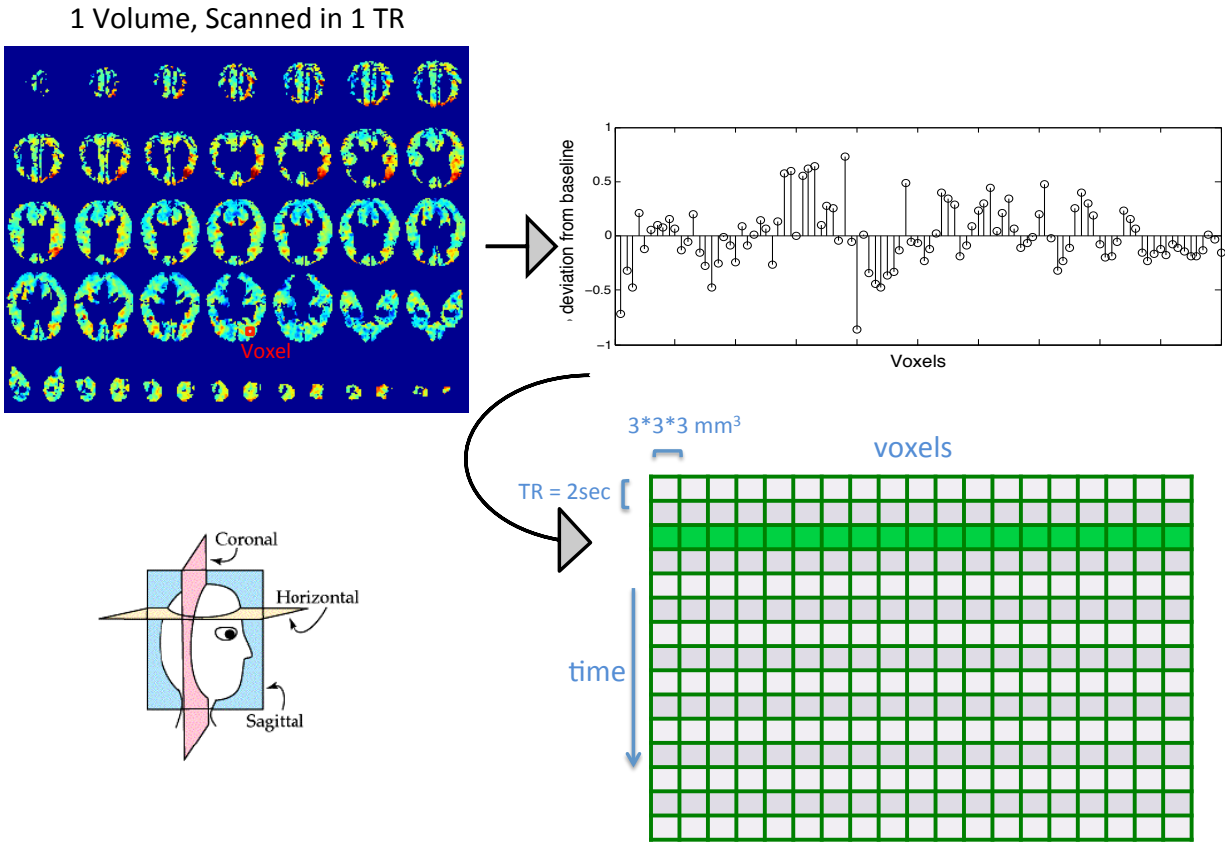


Fig. 2. Structure of fMRI data. One fMRI activity volume is shown in horizontal slices. A diagram of brain slices is provided for illustration, from [1]. A TR is the time needed to record one brain volume. Each 3D pixel is called a voxel and measures $3 \times 3 \times 3$mm. In each voxel, we measure a change from rest activity that has been normalized by the variance of that voxel during rest. The entire brain volume corresponds to one data point. After collapsing all the voxel from the brain in a 1D vector, we can represent the data from the entire experiment as a matrix where every row is one brain volume acquisition.

Data preprocessing We used the MATLAB suite SPM8 [3] to preprocess the data. Each subject's functional data underwent realignment, slice timing correction and co-registration with the subject's anatomical scan, which was segmented into grey and white matter and cerebro-spinal fluid. The subject's scans were normalized to the Montreal Neurological Institute (MNI) space and smoothed with a $6 \times 6 \times 6$mm Gaussian kernel smoother.

Using the Python toolbox PyMVPA [14], we masked the functional data using the segmented anatomical mask, discarding cerebrospinal-fluid voxels. Each voxel's activity was then normalized by removing the mean activity in that voxel during the fixation periods and dividing by the standard deviation during those periods.

Finally, we selected voxels from each subject, keeping only voxels in 78 cortical Regions Of Interest (ROIs), defined using the AAL brain atlas [31]. We ended up with an average of 29227 voxels per subject. The anatomical union (number of MNI voxel locations for which at least one subject had a voxel) of these 8 subject's brains was a set of 41073 voxel locations. The intersection of the 8 subjects voxel locations was a set of 17403 voxel locations.

2.2 Data

Because we use a TR of 2 seconds, and we present each word for 0.5 seconds, the subject sees 4 words at every TR (see figure 1). In figure 1, we displayed each volume of data (equivalent to one whole brain data acquisition, or one data point) in horizontal slices, along with the time series of words presented. We will use 1 TR as a time unit in our analysis. We can represent the data from the entire experiment in one matrix in which every row is the activity in the brain volume at a given TR: all the voxels in the brain are collapsed into a 1D vector (see figure 2).

3 Generative Model

3.1 Representing stories in a feature space

Our goal is to understand how the brain processes stories. Our approach is to train a computer program to learn the mapping between features of the story (e.g. its words, the appearance of specific characters in the story) and the observed fMRI activity when a person reads that story. To measure the validity of this trained model, we define the objective of decoding what passage of the story is being read during a particular time segment **D** of brain activity. We use the mapping learned by the model to predict brain activity for any passage of the story. We then use these predictions to identify the correct passage that created **D**.

We represented our story features as a multivariate discrete time series. We used one TR as a unit of time. This enables us to have the same time scale for the features and the data time series. We compute the value of a feature at any TR by aggregating the features of the four words that were read during that TR (see Figure 3).

We extracted the story features at multiple levels of representation. Specifically, we obtained simple perceptual features such as the average word-length in a TR, as well as semantic features of individual words and sentence level features such as syntactic dependency relationship. We also included discourse level features such as the presence of different story characters (Figure 3).

Story Words	TR Number	Average Word Length	Story Characters			Syntax			Semantics		
			Harry	Hermione	Ron	Verb	Sbj	Obj	PC1	PC2	PC3
Harry	1	5.25	1	0	0	1	1	0	0.35	0.23	0.65
had											
never											
believed											
he	2	3	1	0	0	1	1	0	0.42	-0.54	0.87
would											
meet											
a											
...	...	...	...	...	...	...	...	...	...	...	...

Fig. 3. Example of the time course of the different types of story features. Stories have to be represented in a feature space that allows for learning the brain response to individual features. The neural response to a novel part of the story can then be predicted as the combination of the responses associated with its features.

Average Word Length

We compute the average word length in every TR.

Story Characters

We resolve all pronouns to the character to whom they refer, and make binary features to signal which of the 10 characters are mentioned.

Semantic features

An approximation of the meaning of a word can be obtained by the pattern of its occurrence with other words over a large text corpus. For example, "apple" is likely to occur with other food items or the verb "eat", but not so likely to occur with building materials or power tools.

We used a semantic feature set derived by [21]. It is a flat word-form based model with some similarties to HAL [19]: we obtain the co-occurrence statistics of all words or punctuation found up to 4 positions on either side of every word out of 50 million documents. We reduce the resulting sparse matrix of words by words and punctuation to 300 principal dimensions.

For every word in our story, we therefore obtain 300 features. We sum the features of the four words within each TR.

Syntactic features

Using an automated parser [23] we determined the part of speech of every word in the story and obtained the dependency role of every word from the parse tree of the sentences.

We obtained a set of 28 unique parts of speech and 17 unique dependency relationships, for a total of 45 syntactic binary features that indicate if a given part of speech or a dependency relationship occurred within a TR. We also included an additional feature that records the position of a word in the sentence, i.e. its number starting from the beginning of the sentence. This value is averaged for the four words in a TR. Finally, we added a feature that tracks if the current word being read is part of a dialog between the story characters.

3.2 Modeling the time dynamics of the neural activity

We aim to find the mapping between the different types of features we presented above and the neural activity v_i of a voxel i. We want to learn the response of this voxel i to every feature j.

We first assume that each feature j has a signature activity in voxel i that is consistently repeated every time the brain encounters this feature (for the regions that do not encode this feature, we will ideally learn a signature activity equal to 0). Figure 4(a) shows a hypothetical pattern of activation elicited by the semantic feature j in a given voxel. Due to the TR = 2 seconds we use in our experiment, and the typical latency of the hemodynamic response, we are only interested in the points of the signature response that are sampled 2, 4, 6 and 8 seconds after the onset of feature j (w_1^{ij}, w_2^{ij}, w_3^{ij} and w_4^{ij}). It is important to note that we do not constrain the shape of the signature response to be learned.

The second assumption is that the signature activity is scaled by the value of feature j at the time the feature is presented. See figure 4(b).

Therefore, if we assume that the responses created by successive occurrences of a feature are additive then the activity at time t in voxel v_i is:

$$v_i(t) = \sum_{k=1}^{4} f_j(t-k) \times w_k^{ij} \tag{1}$$

where $f_j(t)$ is the value of feature j at time t. Another way to think about this is that the activity created by the feature is the convolution of the signature response with the time course of the feature. Above we considered the brain activity to be created by one story feature. Now we include the responses created by all of the features we have defined above, again assuming they are additive. This gives the model:

$$v_i(t) = \sum_{j=1}^{F} \sum_{k=1}^{4} f_j(t-k) \times w_k^{ij} \tag{2}$$

We therefore model the voxel's activity $v_i(t)$ as a linear combination of the values of all the features at times $t-4 \rightarrow t-1$. We know time courses of the feature values and the voxel's activity, and we need to predict the set of signature responses.

Our approach is similar to Hidden Process Models [15, 16] that also use a multiple regression setup. The neural activity is assumed to be generated by linearly additive processes and all instantiations of the same process share the same response, but unlike our model, the delay in the onset of the response is variable.

3.3 Learning the Signature Responses

In this document, we compare two methods for learning the signature responses from equation 2. In that equation, we did not consider different subjects, and only considered a hypothetical voxel i. However, in reality, we have S subjects, and $V_T^{(s)}$ voxels for each subject. The regression in equation 2 can therefore be rewritten as:

$$\mathbf{v}_i^{(s)} = \mathbf{F} \times \mathbf{w}_i^{(s)} + \boldsymbol{\epsilon}_i^{(s)} \tag{3}$$

where:

- s is the index of a given subject ($1 \leq s \leq S$)
- n is the number of TRs (or time points)
- $\mathbf{v}_i^{(s)}$ is the $n \times 1$ vector of activity of voxel i of subject s
- $\mathbf{F}$ is the $n \times K$ matrix of time shifted features (every row contains the features of the 4 previous TR, i.e. $K = 4 \times F$)
- $\mathbf{w}_i^{(s)}$ is the $K \times 1$ vector of signature responses in voxel i of subject s
- $\boldsymbol{\epsilon}_i^{(s)} \sim N(0, \sigma_i^2)$ is the $n \times 1$ vector of errors (n is the number of TRs) caused by error in voxel voxel i of subject s

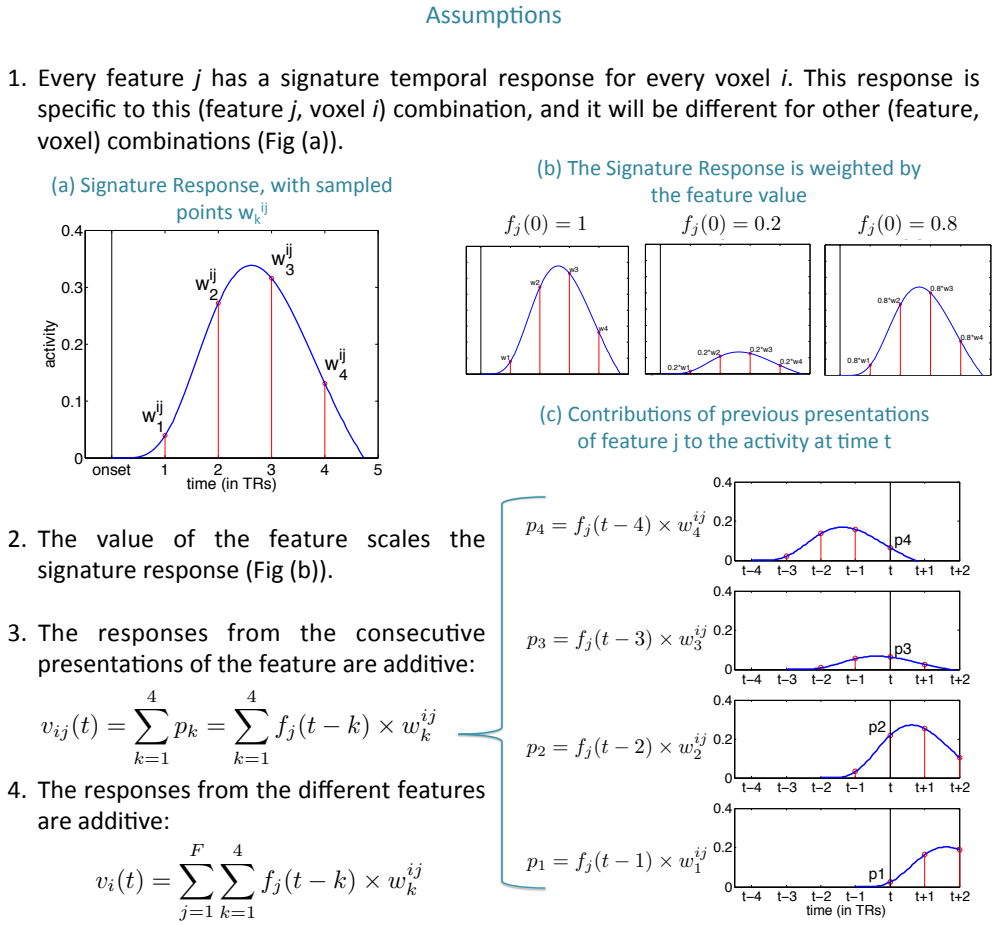


Fig. 4. Time model of a voxel's response to the consecutive occurrences of the features of a story. Because of the hemodynamic response latency, the occurrence of a feature at time t will affect the activity of the voxel for several TRs after time t. This latency is accounted for by considering occurrences of features at previous TRs when modeling a voxel's activity at time t.

Individual Subject Regression

If we consider a small number of features K, such that n is greater than K, we can learn the responses $\mathbf{w}_i^{(s)}$ via the ordinary least squares solution. If K exceeds n, we solve the following L2 regularized regression:

$$\min_{\mathbf{w_i}} ||\mathbf{v_i^{(s)}} - \mathbf{F} \times \mathbf{w_i^{(s)}}||_2^2 + \lambda ||\mathbf{w_i^{(s)}}||_2 \tag{4}$$

independently, for each voxel i and each subject s. This equation has a closed form solution

$$\hat{\mathbf{w}}_\mathbf{i} = (\mathbf{F}^T\mathbf{F} + \lambda \mathbf{I}_K)^{-1}\mathbf{F}\mathbf{v}_\mathbf{i}^{(\mathbf{s})} \tag{5}$$

where $\mathbf{I}_K$ is the $K \times K$ identity matrix, and we choose λ by cross-validation [13].

Hierarchical Linear Modeling

Hierarchical Linear Models (also know as mixed models) [27, 12] are common in fMRI studies for group analysis. These models assume the brains of different subjects respond similarly to the same stimulus and tries to learn better and more reliable weights by learning jointly from all subjects. To use a Hierarchical Linear Model, we will only pick the intersection of voxels that are shared among all the subjects. Therefore, $V_T^{(s)} = V_T = 17403$.

For every subject s, the activation in a particular voxel i is modeled as:

$$\mathbf{v}_i^{(s)} = \mathbf{F}^{(s)} \times \mathbf{w}_i^{(s)} + \boldsymbol{\epsilon}_i^{(s)} \tag{6}$$

where

$$\mathbf{w}_i^{(s)} = \boldsymbol{\gamma}_i + \mathbf{u}_i^{(s)} \tag{7}$$

Here we assumed the coefficients $\mathbf{w}_i^{(s)}$ are sampled from a multivariate normal centered around a group-wide vector $\boldsymbol{\gamma}_i$ (representing a "universal" brain response) with a covariance matrix $\mathbf{u}_i^{(s)}$.

Since we have $\mathbf{F}^{(s)} = \mathbf{F}$ in our setting, this is a particular case of the Hierarchical Linear Model, and it has a closed form solution that can be derived from the log-likelihood of the data for all subjects under this model: [27, 12]

$$l = -\frac{1}{2}\Big\{ Mn \log \sigma^2 + \sum_{s=1}^{S} \big(\log |\mathbf{I}_n + \mathbf{F}^T \mathbf{T}^* \mathbf{F}| + \frac{1}{\sigma^2} (\mathbf{v}_i^{(s)} - \mathbf{F}\boldsymbol{\gamma}_i)^T (\mathbf{I}_n + \mathbf{F}^T \mathbf{T}^* \mathbf{F})^{-1} (\mathbf{v}_i^{(s)} - \mathbf{F}\boldsymbol{\gamma}_i) \big) \Big\} \tag{8}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix and $\mathbf{T}^* = \frac{1}{\sigma^2}\mathbf{T}$.

We maximize the log-likelihood by setting [27, 12]:

$$\hat{\boldsymbol{\gamma}}_i = (S\mathbf{F}^T\mathbf{F})^{-1} \sum_{s=1}^{S} \mathbf{F}^T \mathbf{v}_i^{(s)} \tag{9}$$

$$\hat{\sigma}_i^2 = \frac{1}{M \times (n-K)} \sum_{s=1}^{S} (\mathbf{v}_i^{(s)})^T (\mathbf{I}_n - \mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T) \mathbf{v}_i^{(s)} \tag{10}$$

$$\hat{\mathbf{E}}\hat{\mathbf{E}}^T = \sum_{s=1}^{S} (\mathbf{v}_i^{(s)} - \mathbf{F}\hat{\boldsymbol{\gamma}}_i)(\mathbf{v}_i^{(s)} - \mathbf{F}\hat{\boldsymbol{\gamma}}_i)^T \tag{11}$$

$$\hat{\mathbf{T}}^* = \frac{1}{M\hat{\sigma}_i^2} (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\hat{\mathbf{E}}\hat{\mathbf{E}}^T\mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1} - (\mathbf{F}^T\mathbf{F})^{-1} \tag{12}$$

After we obtain these solutions, we can estimate the subject specific $\mathbf{w}_i^{(s)}$ as: [27]

$$\hat{\mathbf{w}}_i^{(s)} = \boldsymbol{\Lambda}_i^{(s)} \hat{\mathbf{w}}_{i,OLS}^{(s)} + (\mathbf{I}_K - \boldsymbol{\Lambda}_i^{(s)})\boldsymbol{\gamma}_i \tag{13}$$

where:

$$\boldsymbol{\Lambda}_i^{(s)} = \mathbf{T}(\mathbf{T} + \hat{\sigma}_i^2 (\mathbf{F}^T\mathbf{F})^{-1})^1 \tag{14}$$

$$\hat{\mathbf{w}}_{i,OLS}^{(s)} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{v}_i^{(s)} \tag{15}$$

4 Classification

4.1 Cross-Validation Procedure

To learn the signature responses we time-shift the story feature matrix: we make matrix $\mathbf{F}$ in which **every row** t **contains the values of all the features at times** $t-4$**,** $t-3$**,** $t-2$ **and** $t-1$ (Figure 5(a)). We also create an fMRI data matrix containing in **each row t the concatenation of the entire brain images for all subjects, at TR** t.

We introduce here the matrix $\mathbf{W}$, which is the concatenation of all the vectors $\mathbf{w}_i^{(s)}$ from part 3.3. i.e.

$$\mathbf{W} = [\mathbf{W^{(1)}}, \mathbf{W^{(2)}} ... \mathbf{W^{(S)}}] \tag{16}$$

where

$$\mathbf{W^{(s)}} = [\mathbf{w}_1^{(s)}, \mathbf{w}_2^{(s)}, ..., \mathbf{w}_{V_T^s}^{(s)}] \tag{17}$$

To test the validity of the learned signature responses, we constructed a binary classifier that decodes which passage of the story is being read from a given fMRI data frame. We start by partitioning the timeline into non-overlapping time windows, each of length 20 TRs. Then, for every cross validation fold (i.e. for every pair of 20 TRs segments), the steps are:

1. Divide the data and the corresponding feature matrix into test data (the two 20 TRs segments) and training data (see Figure 5(a)).

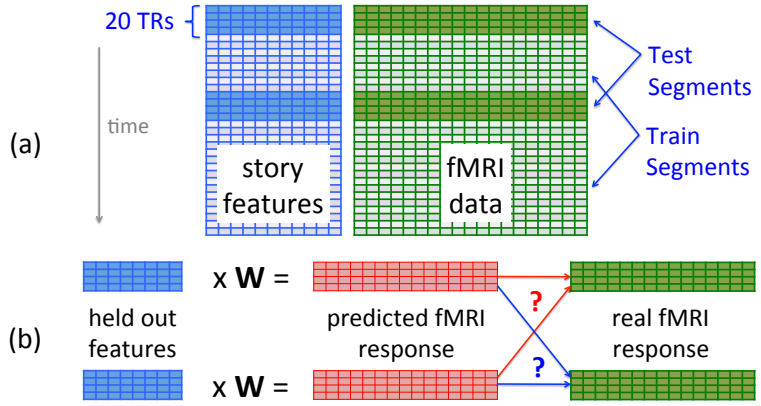


Fig. 5. Diagram of the classification task. The training data (light green) is used to learn the responses of the brain to the different story features. The test data (dark green) consist of two 20 TR segments of brain activity. The task is to assign to each segment the 20 seconds portion of the story to which it corresponds (one of the the two dark blue segments).

2. Use the training data to estimate the signature responses of all features in all voxels and all subjects ($\mathbf{W}$), using the methods in part 3.3.
3. Take the two test story-frames and predict the corresponding brain activity using the learned responses $\mathbf{W}$, as shown in Figure 5(b).
4. Use the two predictions to classify each of the two test data-frames independently: i.e. assign to each data-frame the story-frame with the closest prediction, using a distance function explained in part 4.2 (Figure 5(b)) .

We average the results of all the cross-validation folds and obtain an overall classification accuracy.

4.2 Selecting Voxels for Testing

Here we describe how the distances between a test segment $\mathbf{T}$ and the two predicted segments $\mathbf{P_1}$ and $\mathbf{P_2}$ that we compare it to are computed (see figure 5(b)). We use three methods:

- **Whole-Brain** classification:
  This method uses all the voxels from all the subjects in order to determine the distance between the predicted segments and the true segment. Because we are working with single trial data, concatenating the voxels from different subjects in a row acts as a substitute for multiple repetitions. We compute the Euclidean distance between the two images: $||\mathbf{T}-\mathbf{P_1}||_\mathbf{2}$ and $||\mathbf{T}-\mathbf{P_2}||_\mathbf{2}$. We refer to the classification accuracy that we obtain with this method as "Whole-Brain accuracy".

- **Local** classification:
  Whole-Brain accuracies do not tell us about which parts of the brain are contributing to the classification accuracy. In order to assess this, we perform the classification "locally", looking in one region of the brain at a time. Regions are defined as $k \times k \times k$-voxel boxes centered around one MNI voxel location, k being an odd integer (see figure 6). This method is similar to the Searchlight approach commonly used in neuroimaging [25], however we expand it to include data from multiple subjects:
  - We pick a box size k: for example a $3 \times 3 \times 3$ voxels box (to look at one voxel at a time we take a $1 \times 1 \times 1$ voxel box)
  - For every voxel location (x_i, y_i, z_i), we select the set of voxels whose coordinates fall in the $k \times k \times k$ voxels box centered around that location. This can be done for each subject independently, in the case where we are interested to look for regions with high accuracy on a single subject basis. It can also be done by selecting the union of voxels from all subjects that fall in this box.
    Because we are working with single trial data, concatenating the corresponding voxels from different subjects in a row acts as a substitute for multiple repetitions. Additionally, since the alignment of the subjects to the same anatomical space is not perfect, taking a $k \times k \times k$ voxel box with $k > 1$, allows us to circumvent small variations in the anatomical configuration of the subjects brains.
  - For each of these sets L_i of voxels, we compute the Euclidean distances:

$$||\mathbf{T}(\text{all rows, voxels in } L_i) - \mathbf{P_1}(\text{all rows, voxels in } L_i)||_2$$

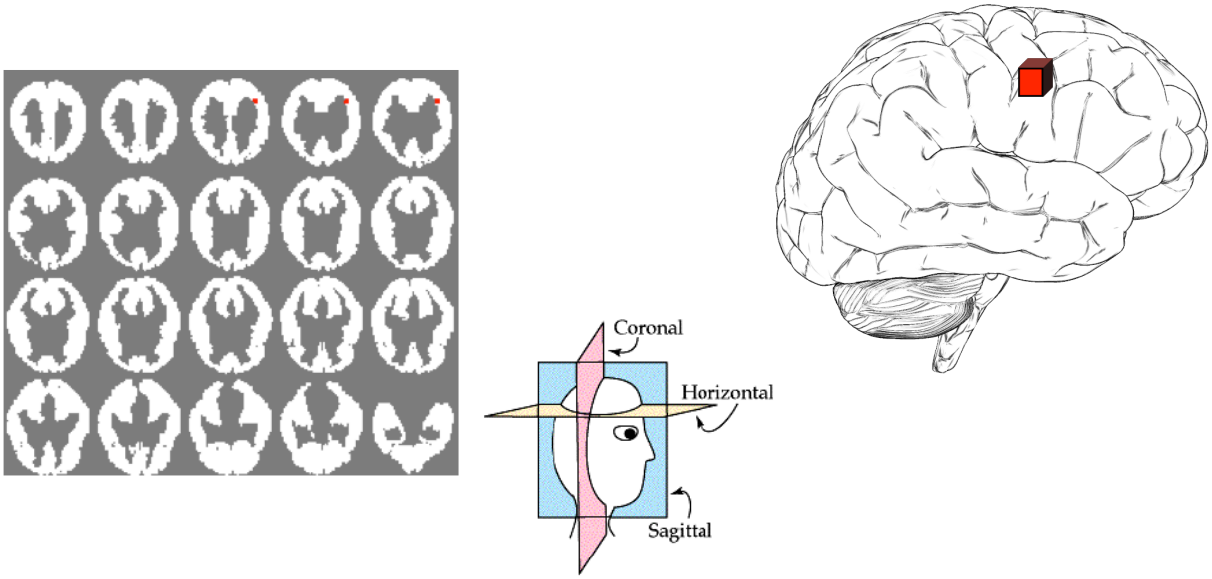


Fig. 6. Searchlight Approach to Find Region Specific Accuracies. Voxels from the fMRI volume are shown in horizontal slices, see diagram for illustration, from [1] . The red voxels correspond to the voxel locations used by one local classifier, that takes into account voxels inside a $3 \times 3 \times 3$ coordinate box. The box is represented on top of a brain as an illustration (adapted from [2]). We slide the box over to every location and repeat the classification there. When using data from multiple subjects, we concatenate the voxels in that location from the multiple subjects.

$$||\mathbf{T}(\text{all rows, voxels in } L_i) - \mathbf{P_2}(\text{all rows, voxels in } L_i)||_2$$

Note: we are performing this computation at every voxel, so we are actually performing V_L classifications.

We refer to the classification accuracies we obtain with this method as "local accuracies".

- **Best-Voxels** classification:
  We identify accurate voxels for each subject. For this purpose, we perform the following:
  - We take out the data from the first block of the experiment (which correspond to the first 11 minutes). On this data, we perform training and validation testing in the same way as described in part 4.1. However, for testing, we use the local distance method explained above with a box size of $1 \times 1 \times 1$, for each subject independently. For each subject s, we therefore can identify the set $\mathbf{B_s}$ of the top 1% performing voxels. We can therefore take the union of these sets of good performing voxels: $\mathbf{B}$.
  - We now discard the data from the first block. We perform the classification in part 4.1 using the rest of the block, and at testing, we use the Euclidean distances:

$$||\mathbf{T}(\text{all rows, voxels in } \mathbf{B}) - \mathbf{P_1}(\text{all rows, voxels in } \mathbf{B})||_2$$

$$||\mathbf{T}(\text{all rows, voxels in } \mathbf{B}) - \mathbf{P_2}(\text{all rows, voxels in } \mathbf{B})||_2$$

We call the results of this method "Best Voxels accuracy". We will also discard the first block of the experiment when reporting the results of the first two methods.

4.3 Whole-Brain Classification Accuracy

To show that Whole-Brain classification accuracy is higher than chance accuracy, which is 50% in this balanced binary classification task, we compute an empirical null distribution. The null hypothesis that story features cannot predict neural activity is approximated by iteratively permuting the time series of story features (before the time-shifting step), then running the same train/test procedure.

4.4 Comparing Multi-Subject Learning and Single-Subject Learning

We predict that the Multi-Subject Hierarchical Linear Model is useful when the size of the training data is small, but that when there is enough data it no longer has an advantage on the single-subject regression model. To test this assumption, we perform the Whole-Brain accuracy experiment detailed above with varying sizes of training data. We only use Word Length features to test this hypothesis because it is a low level perceptual feature and we can expect the brain response to be more universal among subjects, making the hierarchical learning more useful. Another reason is that the Hierarchical model is less computationally expensive with a small number of features.

For every block of the story, we repeat the experiment for training sets of time-length $k * 20$ TRs with $k = 1...11$, where each 20 TRs segment is contiguous in time, but the set of k segment doesn't have to be. For every size k, we sample a training set H times (H =39) and we perform testing on all the possible pairs of remaining segments. By averaging these H values we obtain an average accuracy for a set of size k. We estimate the standard error of this set via the bootstrap method: we sample many sets out of the H values with replacement, compute their means and the standard deviation of their means, which is our estimate of the standard error.

We perform this analysis for the Multi-Subject and Single-Subject method. For both methods, we use only the intersection of voxels that are present for all subjects. For each method we can estimate a confidence interval for each point using the standard error from the bootstrap method, and a Bonferroni correction for multiple comparisons.

4.5 Assessing Replicability of the Learned Weights and their Distribution in the Brain

To assess how reliable and how informative the weights learned for a feature in a given voxel are, one way is to estimate the empirical distribution of weights under the null hypothesis. We do that via permutation testing: by randomly permuting the rows of the feature matrix (before the time shifting step) we break the relationship between the time series of features and the time series of fMRI data. The weights are then learned using the randomized feature matrix. This is repeated 50 times, allowing us to obtain the null empirical distribution. Then we find the weights for which the null hypothesis is rejected with a rate of $\frac{0.05}{N}$ where N is the number of weights per subject (this corresponds to doing a Bonferroni correction).

We group the voxels by anatomical ROI. For every one of the 4 types of features, and for every ROI, we find the frequency of non-null weights (by averaging across subjects, the features in each type and their 4 weights, and the voxels in that ROI). For every type of feature, we can therefore determine the 10 ROIs with the highest frequency of non-null weights. This allows us to formulate hypotheses about the representations of different types of features in different regions.

4.6 Identifying Brain Regions Correlated with Different Feature Types:

To find out where in the brain each type of feature is useful, we followed a similar training approach as in section 4.1, except that (1) only one type of feature (Word Length, Story Characters, Syntax or Semantics) was used at a time and (2) we used a Local Accuracy procedure at test time with $k = 3$ and using data from all subjects (part 4.2). Precisely, for every voxel location i, we took the box of $3 \times 3 \times 3$ voxel coordinates centered around that location. The union of voxels from all subjects that have coordinates included in this box were selected. Therefore, for every location, we performed the classification of 2 segments of size $20 \times |\mathbf{V}_i|$, where $\mathbf{V}_i$ is the set of voxels $\{v|coords(v) \in box_i\}$.

For every one of these combination of type of feature/subset of data, we obtain a local classification accuracy. We measure significance by computing an empirical null distribution in the same way as for Whole-Brain Accuracies, then correcting for multiple comparisons using the Benjamini-Hochberg False Discovery Rate procedure [4]. We therefore obtain a set of local accuracies that are significantly higher than chance. For a given type of feature, we call the regions that achieve such an accuracy "successful regions".

5 Results

5.1 Whole Brain Classification Accuracy

Using the Whole-Brain distance, we obtained an average accuracy of **67%** when using all features, which is significantly better than chance with $\mathbf{p < 10^{-6}}$. Interestingly, using the Semantic features separately performs better than using all the features combined.

Using the Best-Voxels distance, we obtained an average accuracy of **78%** when using all features, which is significantly better than chance with $\mathbf{p < 10^{-8}}$. Using Semantic features and Word Length individually, the accuracy jumps to more than 90%. This might be due to the fact that the selection of the top 1% voxels for Semantic features and Word Length are more replicable across blocks of the experiment, while the top 1% voxels for all features are a more varied set and their activity is less replicable across blocks.

Distance	All Features	Word Length	Syntax	Semantics	Characters
M1	67% $p < 10^{-6}$	52% $p < 0.3$	64% $p < 10^{-4}$	85% $p < 10^{-8}$	53% $p < 0.2$
M3	78% $p < 10^{-8}$	92% $p < 10^{-8}$	74% $p < 10^{-8}$	91% $p < 10^{-8}$	55% $p < 0.03$

5.2 Multi-Subject versus Single-Subject Learning

We plot in figure 7 the change in the mean accuracy as the size of the training sample increases for the Multi-Subject Hierachical Linear Model and the Single-Subject Regression. We are using here Word Length as input feature and Best-Voxels as distance function (Whole-Brain is not powerful enough to yield significant result for Word Length as shown in part 5.1. We also plot confidence intervals with rate $\frac{0.05}{11*2}$ (Bonferroni correction).

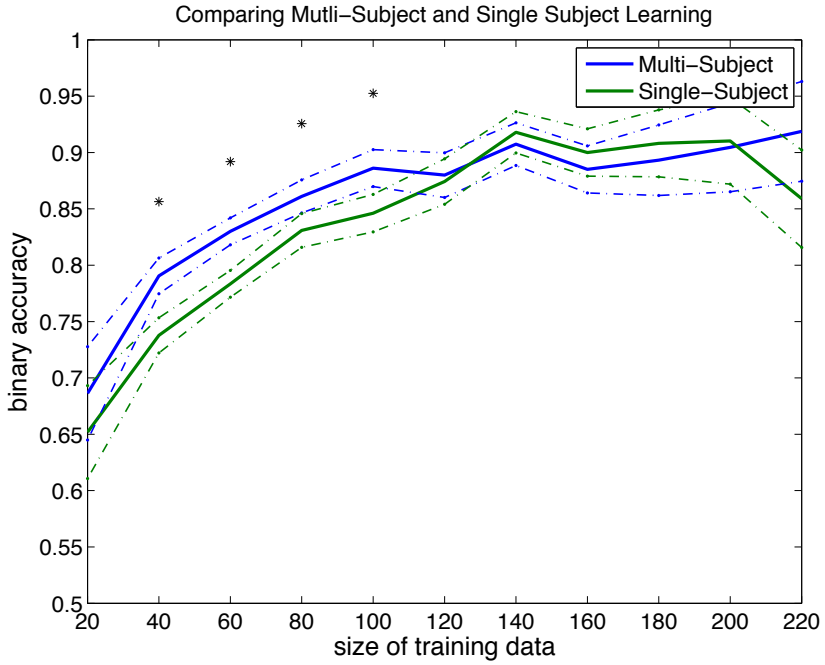


Fig. 7. Difference in the Multi-Subject Hierachical Linear Model and the Single-Subject Regression as training set size varies. The multi-subject model outperforms the single-subject model for small set sizes, while the performance of both models becomes indistinguishable for both test sizes. Input features are Word Length.

Our observations align with our prediction: the multi-subject model outperforms the single-subject model for a small sample size (there is a significant difference between the two models for a set size of 40 to 100 TRs. Furthermore, when the training set size becomes larger, the performance of the two methods becomes indistinguishable.

5.3 Identifying Brain Regions Correlated with Different Feature Types:

Different Patterns of Weights We ran the experiment described in section 4.5 using all the features simultaneously as input. We also ran it with each type of feature as input. For each case, and for each feature type, we display in the table below the ROIs that have the highest frequency of replicable weights.

Learning Types of Features Simultaneously			
Word Length	Syntax	Semantics	Characters
Lingual L	Heschl L	Precuneus L	Cingulum Ant L
Calcarine L	Frontal Mid Orb L	Precuneus R	Frontal Med Orb L
Calcarine R	Insula R	Cingulum Mid R	Frontal Inf Oper L
Lingual R	Rolandic Oper L	Lingual L	Lingual L
Occipital Inf L	Frontal Inf Tri R	Occipital Sup L	Cingulum Ant R
Occipital Mid L	Insula L	Frontal Inf Oper L	Calcarine R
Occipital Sup R	ParaHippocampal R	Cuneus R	Fusiform L
Fusiform L	Olfactory R	Cingulum Ant R	Calcarine L
Cingulum Ant R	Temporal Sup L	Cingulum Post R	Cingulum Post R
Occipital Mid R	Temporal Sup R	Supp Motor Area R	Frontal Sup Medial L

Learning Types of Features Individually			
Word Length	Syntax	Semantics	Characters
Occipital Inf L	Insula R	Lingual L	Occipital Inf R
Lingual R	Rolandic Oper L	Occipital Inf L	Occipital Inf L
Lingual L	Insula L	Calcarine L	ParaHippocampal R
Occipital Mid L	Occipital Inf R	SupraMarginal L	Temporal Pole Mid R
Calcarine L	Temporal Inf L	Temporal Mid R	Frontal Mid Orb R
Fusiform L	Fusiform R	Temporal Mid L	Temporal Sup R
Calcarine R	Frontal Inf Oper L	Lingual R	Calcarine R
Occipital Inf R	Frontal Inf Tri L	Calcarine R	Occipital Mid L
Fusiform R	Cingulum Post R	Temporal Sup R	Cingulum Post L
Cingulum Post L	Temporal Sup R	Temporal Sup L	Frontal Med Orb L

Different Patterns of Successful Regions Using a False Discovery Rate of **0.005**, we obtained different patterns of successful regions for different types of features. We plot them in Figure 8(a).

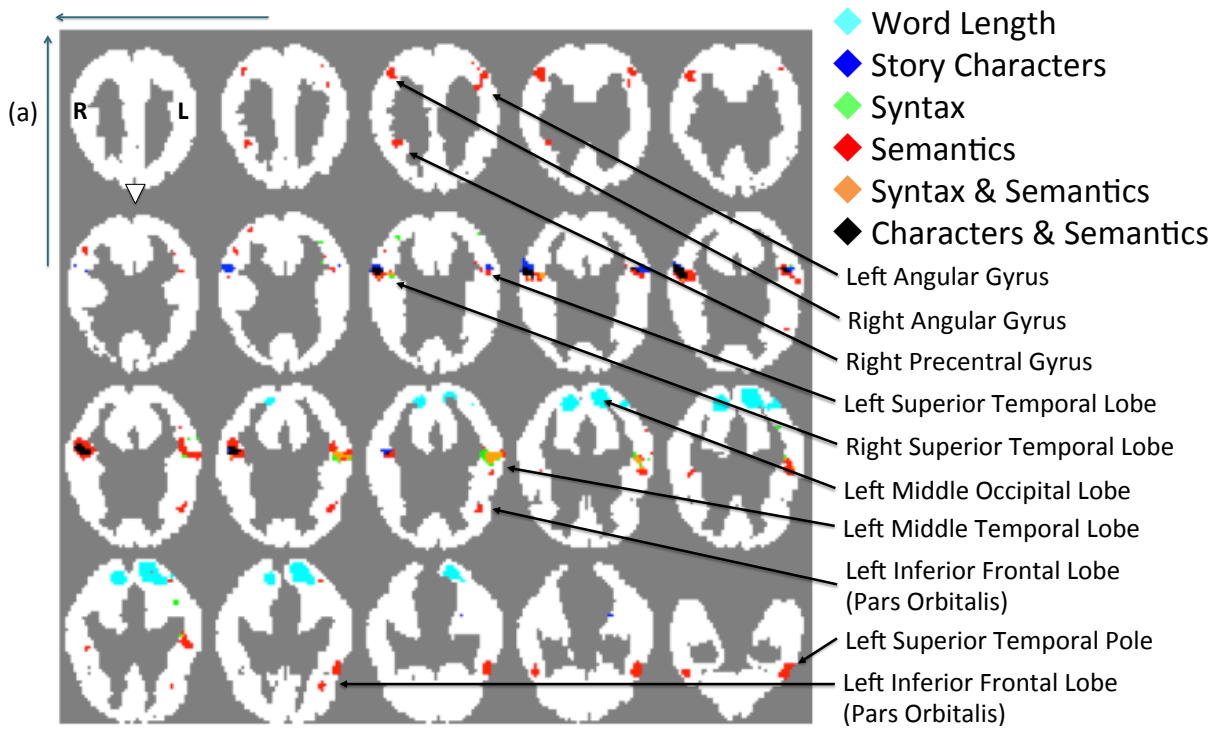
6 Discussion

The language processing literature has identified key areas in the brain that are associated with language. These main areas are the Left Middle Temporal Gyrus (MTG), the Left Superior Temporal Gyrus (STG) and the Left Inferior Frontal Gyrus (IFG)[10]. The Right homologues of these regions are also activated when the parsed stimulus increases in difficulty [17]. There has been a long debate about what areas process semantic versus syntactic information. In [6], authors found that syntactic structure building during natural story listening correlated with the activity in the Left Anterior Temporal Regions (Left MTG and STG). Earlier experiments have shown a differential implication of Pars Opercularis of Left IFG in processing syntax, and of Pars Orbitalis in processing semantics [11]. Since the input is visual, the Occipital cortex is expected to be activated [10]. These predictions are summarized in figure 8(b).

The patterns of successful regions we obtained align with these findings (figure 8(a)). The Occipital Lobe was successful with Word Length. Most successful regions with the Syntax and Semantic features are in the Left Temporal Lobe. In the Left MTG, one cluster was successful with both Syntax and Semantics, and a small cluster of voxels was successful with Syntax only. A region of Pars Orbitalis of the Left IFS was also successful with Semantic features, as well as the bilateral Angular Gyrus and the Left Superior Temporal Pole.

We found the Bilateral Posterior Superior Temporal (STC) cortex to be successful for Agents features, with more successful voxels in the Right hemisphere. In [30], the Bilateral Posterior STC was shown to increase in activity with the change of characters in a story and their change of goal. Moreover, the Posterior Right Superior Temporal sulcus has been linked to the representation of other people's actions in terms of goals [28].

Our approach is flexible. The story feature basis we presented here is simple, we plan to use more complicated features in the future. For example, we can have features that correspond to narrative structure, such as locations or characters' goals [30]. We also plan on using models of meaning composition for successive words instead of only summing the semantic features of the words as we currently do. These future goals outline an important contribution of our model: it can be used to assess different models of reading or of word meaning composition. To test multiple



(b)

What we already knew		
Word Length (Visual features)	Syntax and Semantics (Language)	Characters
Occipital Cortex	Bilateral Temporal and Inferior frontal (but mainly in the Left Hemisphere) Separation between Syntax and Semantics not clear	Right Posterior Superior Temporal (interpreting other people's actions)

Fig. 8. (a) Voxels with significant classification accuracy when using different types of story elements as features, shown in different colors. The slices are drawn such that they increase in the Z MNI-coordinate when going right to left, then bottom-up. On each slice, the top of the slice corresponds to the posterior of the brain, and the right side of the slice corresponds to the left side of the brain (only slices from Z=-22 to Z=35 had significant accuracies and are plotted here). Each voxel location represents the result of using a box of $3 \times 3 \times 3$ voxel coordinates, centered at that location, such that the union of voxels from all subjects whose coordinates are in that box are used. (b)Predictions from the literature on the location of regions with significant classification.

models of semantic composition for example, one could create a set of features for the words of a story using each model, and then compare the performance and the pattern of successful regions for each set of features. This give a measure of how close each model is to representing the information processed in the brain.

Other future goals we have is to (1) test the model's linearity assumptions and (2) perform a more adapted learning. Instead of the Hierarchical Linear Model, we plan to have a multi-task learning setting in which spatial smoothness of the learned weights, as well as between-subject similarity, is encouraged.

7 Conclusion

We presented a generative model that predicts the brain activity when subjects read a story, by learning the mapping between the different story features and the fMRI data. Our model is successfully used to decode a passage of a story from a segment of fMRI data. This approach also recovers patterns of representation for different types of features that align well with the findings in the field. These patterns are found simultaneously for all features, using a reading paradigm with a close to normal reading rate and a real sample of text, without specifically controlling for each feature separately. These results encourage the use of fMRI to model reading in realistic settings. Furthermore, this approach could be used to compare multiple hypotheses of story reading and meaning composition: sets of story features can be extracted according to each hypothesis and their performance and pattern of representation can be compared.

References

1. Coronal and horizontal sections. `http://www.bioon.com/bioline/neurosci/course/corhor.html`. Accessed: 11/28/2012.
2. Study guide for the parts of the brain. `http://www.brighthubeducation.com/science-homework-help/61901-brain-anatomy-study-guide/`. Accessed: 11/28/2012.
3. J. Ashburner, CC Chen, G. Flandin, R. Henson, S. Kiebel, J. Kilner, V. Litvak, R. Moran, W. Penny, K. Stephan, et al. Spm8 manual. *Functional Imaging Laboratory, Institute of Neurology*, 2008.
4. Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.
5. D.H. Brainard. The psychophysics toolbox. *Spatial vision*, 10(4):433–436, 1997.
6. J. Brennan, Y. Nir, U. Hasson, R. Malach, D.J. Heeger, and L. Pylkkänen. Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and language*, 2010.
7. J.R. Brennan. *Incrementally Dissociating Syntax and Semantics*. PhD thesis, New York University, 2010.
8. A. Buchweitz, R.A. Mason, L. Tomitch, and M.A. Just. Brain activation for reading and listening comprehension: An fmri study of modality effects and individual differences in language comprehension. *Psychology & Neuroscience*, 2(2):111–123, 2009.
9. K.K. Chang, V.L. Cherkassky, T.M. Mitchell, and M.A. Just. Quantitative modeling of the neural representation of adjective-noun phrases to account for fmri activation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 638–646. Association for Computational Linguistics, 2009.
10. R.T. Constable, K.R. Pugh, E. Berroya, W.E. Mencl, M. Westerveld, W. Ni, and D. Shankweiler. Sentence complexity and input modality effects in sentence comprehension: an fmri study. *Neuroimage*, 22(1):11–21, 2004.
11. M. Dapretto and S.Y. Bookheimer. Form and content: dissociating syntax and semantics in sentence comprehension. *Neuron*, 24(2):427–432, 1999.
12. E. Demidenko. *Mixed models: theory and applications*, volume 518. Wiley-Interscience, 2004.
13. J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer Series in Statistics, 2001.
14. M. Hanke, Y.O. Halchenko, P.B. Sederberg, S.J. Hanson, J.V. Haxby, and S. Pollmann. Pymvpa: A python toolbox for multivariate pattern analysis of fmri data. *Neuroinformatics*, 7(1):37–53, 2009.
15. R.A. Hutchinson, T.M. Mitchell, and I. Rustandi. Hidden process models. In *Proceedings of the 23rd international conference on Machine learning*, pages 433–440. ACM, 2006.
16. R.A. Hutchinson, R.S. Niculescu, T.A. Keller, I. Rustandi, T.M. Mitchell, et al. Modeling fmri data generated by overlapping cognitive processes with unknown onsets using hidden process models. *NeuroImage*, 46(1):87–104, 2009.
17. M.A. Just, P.A. Carpenter, T.A. Keller, W.F. Eddy, and K.R. Thulborn. Brain activation modulated by sentence comprehension. *Science*, 274:114–116, 1996.
18. M. Kleiner, D. Brainard, D. Pelli, A. Ingling, R. Murray, and C. Broussard. Whats new in psychtoolbox-3. *Perception*, 36(14):1–1, 2007.
19. K. Lund and C. Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, 28(2):203–208, 1996.
20. T.M. Mitchell, S.V. Shinkareva, A. Carlson, K.M. Chang, V.L. Malave, R.A. Mason, and M.A. Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, 2008.
21. B. Murphy, P. Talukdar, and T. Mitchell. Selecting corpus-semantic models for neurolinguistic decoding. In *First Joint Conference on Lexical and Computational Semantics (* SEM)*, pages 114–123, 2012.
22. S. Nishimoto, A.T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J.L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 2011.
23. J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kubler, S. Marinov, and E. Marsi. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95, 2007.
24. D.G. Pelli. The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision*, 10(4):437–442, 1997.
25. F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and fmri: a tutorial overview. *Neuroimage*, 45(1 Suppl):S199, 2009.
26. J.K. Rowling. *Harry Potter and the Sorcerer's Stone*. Harry Potter US. Pottermore Limited, 2012.
27. I. Rustandi. Predictive fmri analysis for multiple subjects and multiple studies. 2010.
28. R. Saxe et al. Uniquely human social cognition. *Current opinion in neurobiology*, 16(2):235–239, 2006.
29. C. Scott. The tale of peter rabbit. *Beatrix Potter's Peter Rabbit: A Children's Classic at 100*, 1:19, 2002.
30. N.K. Speer, J.R. Reynolds, K.M. Swallow, and J.M. Zacks. Reading stories activates neural representations of visual and motor experiences. *Psychological Science*, 20(8):989–999, 2009.
31. N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, M. Joliot, et al. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289, 2002.
32. C. Whitney, W. Huber, J. Klann, S. Weis, S. Krach, and T. Kircher. Neural correlates of narrative shifts during auditory story comprehension. *Neuroimage*, 47(1):360–366, 2009.