Probing Word Syntactic Representations in the Brain by a Feature Elimination Method

Xiaohan Zhang[1,2], **Shaonan Wang**[1,2], **Nan Lin**[3,4], **Jiajun Zhang**[1,2], **Chengqing Zong**[1,2,5]
[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] CAS Key Laboratory of Behavioural Sciences, Institute of Psychology
[4] Department of Psychology, University of Chinese Academy of Sciences
[5] CAS Center for Excellence in Brain Science and Intelligence Technology
{xiaohan.zhang, shaonan.wang, jjzhang, cqzong}@nlpr.ia.ac.cn; linn@psych.ac.cn

Abstract

Neuroimaging studies have identified multiple brain regions that are associated with semantic and syntactic processing when comprehending language. However, existing methods cannot explore the neural correlates of fine-grained word syntactic features, such as part-of-speech and dependency relations. This paper proposes an alternative framework to study how different word syntactic features are represented in the brain. To separate each syntactic feature, we propose a feature elimination method, called Mean Vector Null space Projection (MVNP). This method can remove a specific feature from word representations, resulting in one-feature-removed representations. Then we respectively associate one-feature-removed and the original word vectors with brain imaging data to explore how the brain represents the removed feature. This paper for the first time studies the cortical representations of multiple fine-grained syntactic features simultaneously and suggests some possible contributions of several brain regions to the complex division of syntactic processing. These findings indicate that the brain foundations of syntactic information processing might be broader than those suggested by classical studies.

Introduction

Word semantics and its syntactic features form the whole picture of word representations and enable the flexibility of human language. Therefore, investigating how the brain encodes semantic and syntactic features of words is crucial to studying brain language-comprehension mechanisms.

The mainstream theory of lexical semantics assumes that words can be represented by sets of features, but it is still an open question as to what constitutes a primitive word feature. Brain imaging studies have accumulated evidence supporting that word semantic representations are at least partly "embodied" in the modal neural systems through which concepts are experienced. For example, the word "cat" is composed of primitive features such as furry (vision), fast (motion), mew (sound), and so on (Binder et al. 2016). They have found that several semantic features are associated with activation in corresponding sensory-motor regions and convergence regions that integrate multi-modal features (Fernandino et al. 2016).

Due to the complexity of extracting and representing syntactic features, there are still hot debates on whether and how syntax is represented in the brain (Pylkkänen 2019). Previous work mainly studied the overall syntactic processing demands of sentences (Hagoort and Indefrey 2014) and the specific syntactic structures of sentences, such as relative clauses (Chen et al. 2006). These studies showed that the processing of syntactic features is associated with brain regions located within the inferior frontal cortex, the lateral temporal cortex, and the inferior parietal cortex. However, it remains unclear how fine-grained syntactic features are represented and whether the neural correlates of different syntactic features overlap or dissociate from each other.

The main difficulty of studying the above problem is to separate a specific syntactic feature from the others. Different from the mainstream neuroimaging studies that employ manually designed disassociate-stimuli, this paper proposes an alternative framework to study the brain representations of word syntactic features. Specifically, we propose a feature elimination method, called Mean Vector Null space Projection (MVNP), to separate different features. This method can remove a target feature from word embeddings by projecting them into a subspace in which contains minimum information of the target feature. Based on the original and one-feature-removed word embeddings, we explore how the brain encodes syntactic features by associating these vectors with brain imaging data. The motivation of removing one feature from representations is that if a specific feature is removed from the original word embeddings and if this feature is represented in the brain, the predictability of the brain areas associated with this feature will be severely damaged.

Our results show that word syntactic features are distributively represented across the temporal, frontal, and parietal lobe and their brain networks are largely overlapped with the classic semantic networks. Furthermore, for the first time, this paper studies the relations of various syntactic features represented in the brain, suggests a hierarchical organization of brain areas in syntactic processing, and illustrates a more detailed division in the classic syntactic network.

To summarize, our main contributions include:

- We propose a new framework that employs the latest computational models to study word syntactical representations in the brain.

- We propose a feature elimination method that can remove a specific feature from word embeddings while retaining other features.
- Our results provide new evidence for the brain representations of several word syntactical features, hopefully helping promote related neuroscience studies.

Related Work

Word representations and interpretation methods Contextual word representations such as ELMo (Peters et al. 2018) and BERT (Devlin et al. 2018) have achieved impressive results on downstream tasks. Previous work investigated the contents encoded in these word representations and found that various types of syntactic information can be effectively decoded from them (Linzen, Dupoux, and Goldberg 2016; Conneau and Kiela 2018; Hewitt and Manning 2019; Tenney et al. 2019). Another line of work, which can be categorized into two groups: adversarial and projection methods, studies how to control or separate different kinds of information.

Adversarial methods mainly rely on adding adversarial components to the main task objective (Mathieu et al. 2016; Xie et al. 2017; Zhang, Lemoine, and Mitchell 2018; Bao et al. 2019). For instance, to learn separated syntactic and semantic representations, Bao et al. (2019) used adversarial components to minimize the syntactic information encoded in semantic vectors and vice versa. Although widely used in various tasks, adversarial methods have two main drawbacks: they get unstable results and rely on an additional loss, making them less suitable for a feature elimination task.

Projection methods aim to project representations onto a space in which only the target feature is retained or is removed (Bolukbasi et al. 2016; Xu et al. 2017; Ravfogel et al. 2020). The most similar work to our method is the Iterative Null space Projection (INLP) proposed by Ravfogel et al. (2020), which repeatedly trains linear classifiers on the unwanted feature and then projects the representations onto the null space of these classifiers. However, this method aims to remove all separable information for the target feature, which might remove excessive information and severely affect the word representation space, especially when the target feature is complex.

Word representations in the brain Mainstream neuroimaging studies employ the hypothesis-based method to study the neural correlates of syntactic features, including brain representations of grammatical categories (Yu et al. 2011), argument structures (Thompson et al. 2007), relative clauses (Chen et al. 2006), and the overall syntactic processing demands of sentences (Hagoort and Indefrey 2014). These studies usually design artificial stimuli with different conditions that differ in one syntactic feature and identify brain regions correlated with a specific feature by contrasting the brain activation between these different conditions.

The main problem with this method is that the syntactic changes of a language expression usually alter its semantics, making it difficult to distinguish whether the change of brain activity is caused by syntactic or semantic change (Pylkkänen 2019). Moreover, this method that relies on manually designed stimuli can only explore one specific feature in one experiment, thus it cannot study the relations between different word syntactical features.

With the surge of language computational approaches, neuroimaging studies begin to explore brain representations in a data-driven way. Mitchell et al. (2008) firstly used explainable word representations to predict brain activations elicited by words. A series of the following work used different types of representations to explore the language processing mechanism in the brain (Huth et al. 2016; Gauthier and Levy 2019; Sun et al. 2019; Toneva and Wehbe 2019; Jain et al. 2020). To probe the sentence-level semantic and syntactic brain activation patterns, Wang et al. (2020) proposed a two-channel variational autoencoder model to dissociate sentences into semantic and syntactic representations and separately associate them with brain imaging data to find feature-correlated brain regions. Wehbe et al. (2014) presented an integrated computational model that incorporates multiple reading sub-processes to predict the detailed neural representation of diverse story features. Reddy and Wehbe (2021) extended this work to syntactic structure features and showed their effectiveness in predicting brain activity.

Different from these previous works, this paper focuses on exploring detailed cortical representation and the relationships between multiple fine-grained word syntactic features.

The Framework of Probing Brain Representations

As shown in Figure 1, the framework contains three modules: 1) the feature removal module, which removes a specific feature from the original word representations, resulting in one-feature-removed word representations. 2) the brain encoding module that predicts brain activation from the original and one-feature-removed word representations. 3) the significance test module, which verifies whether one-feature-removed representations (compared with the original word representations) cause a significant prediction accuracy drop for every brain voxel.

Feature Elimination

Inspired by the INLP (Ravfogel et al. 2020), our feature elimination method MVNP projects word embeddings onto a subspace in which only contains minimum information about the target feature. By doing so, the target feature is removed from the representations. The core concept used in both INLP and our MVNP is called **null space projection**.

Null Space Projection Given a set of vectors $x_i \in \mathbf{R}^d$, and corresponding discrete attributes $Z, z_i \in \{1, 2, ..., k\}$ (e.g. Z can be part-of-speech), we aim to learn a transformation $g : \mathbf{R}^d \rightarrow \mathbf{R}^d$, such that z_i cannot be predicted from $g(x_i)$, and meanwhile g should have a low impact to the vector space. A feature is linearly eliminated if no linear classifier $c(\cdot)$ can predict z_i from $g(x_i)$ with an accuracy greater than the proportion of the majority class in Z.

Let c be a trained linear classifier and $W \in \mathbf{R}^{k \times d}$ be its weight matrix. Null space projection is the operation of projecting word embeddings onto the null space of W. Since the null space of a matrix W is defined as the space

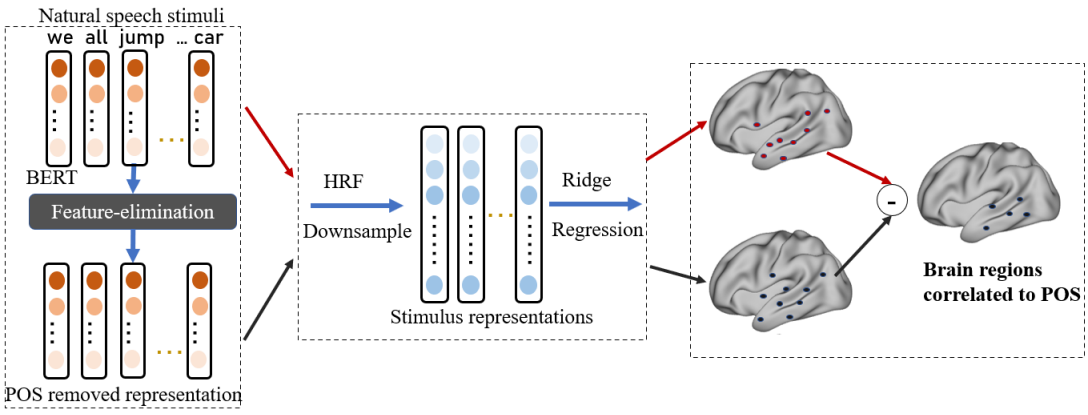

Figure 1: The framework of probing syntactic word representations in the brain, including (1) feature elimination module, (2) brain encoding module, (3) significance test module.

$N(W) = \{x|Wx = 0\}$, projecting word embedding x_i onto $N(W)$ makes W useless for predicting z_i from x_i. Thus the transformation $g(x_i)$ can be defined as $g(x_i) = Px_i$, P is the orthogonal projection matrix onto $N(W)$.

Since word embeddings are high-dimensional vectors that result in multiple linear directions on which feature Z is separable, a single classifier may not learn all these directions or capture the most effective directions. Therefore, the INLP method conducts iterative projection. That is, after obtaining the first classifier's weight matrix W_0 and the corresponding projection matrix $P_{N(W_0)}$, the INLP trains classifier W_1 on $P_{N(W_0)}X$, obtains a projection matrix $P_{N(W_1)}$, trains W_1 on $P_{N(W_1)}P_{N(W_0)}X$ and so on, until no classifier W_{m+1} can be trained. However, for the effect to X in each iteration is accumulated, INLP can severely affect the feature space of representations as iteration time increases.

To alleviate the effect to feature space caused by projection, we aim to find the most effective weight matrix W which makes one projection enough to remove the target feature.

Mean Vector Null Space Projection Let $W^T = [w_1^T, w_2^T, ..., w_k^T]$ be a linear classifier's weight matrix. The geometric interpretation of Wx is that: x is projected onto the subspace spanned by W's rows, and is classified according to the dot product between x and W's rows. Ideally, W's rows should have such a property: among all rows of W, w_i should have the largest dot product with x which belongs to class z_i: $w_i x > w_j x, j \neq i$. Therefore, the process to learn w_i is to solve to the following optimization problem:

$$\max \sum_{x \in z_i} \sum_{j \neq i} w_i x - w_j x$$
$$s.t. \ \|w_l\| = 1, l = 1, 2, ..., k$$

which can be reorganized as:

$$\max \ (\sum_{x \in z_i} x)(\sum_{j \neq i} (w_i - w_j))$$
$$s.t. \ \|w_l\| = 1, l = 1, 2, ..., k$$

The analytic solution for this problem would be very difficult to compute. So we seek to find a approximate solution. The inner product of two vectors is decided by the length and direction of both vectors. To simplify the above problem, we assume that the length and direction of $(\sum_{x \in z_i} x)$ and $(\sum_{j \neq i}(w_i - w_j))$ are independent. Then, the optimization problem reaches its optimal solution when $(\sum_{x \in z_i} x)$ and $(\sum_{j \neq i}(w_i - w_j))$ have the same direction. Therefore $(\sum_{x \in z_i} x)$ can be written as the linear combination of $\{w_1, ..., w_k\}$.

Let $\bar{X}_i$ be the mean vector of all word vectors that belong to the i_{th} class, $\bar{X}_i$ has the same direction as $\sum_{x \in z_i} x$. Then, the subspace spanned by W's rows is equal to or is a subspace of the space spanned by the class mean vectors:

$$L(w_1, w_2, ..., w_k) \subseteq L(\bar{X}_1, \bar{X}_2, ..., \bar{X}_k)$$

Therefore, the null space of the class mean vector matrix $\bar{X}^T = [\bar{X}_1, \bar{X}_2, ..., \bar{X}_k]$ is the same or a subspace of the null space as the matrix W. Since we only need W's null space projection matrix, we use $\bar{X}$'s null space projection matrix as its approximation.

An intuitive explanation of MVNP is that the mean vector of one class is the most representative direction for this class. Thus the subspace spanned by mean vectors encodes rich information of this feature. And this information can be removed by projecting word vectors onto its orthogonal complement.

For clarity, we use P_Z to denote the final projection matrix to remove feature Z from word embeddings, then we have

$$P_Z = P_{N(\bar{X})}$$

Brain Encoding

Let S be the stimuli and R be the brain activity elicited by S. In this paper, S is a sequence of words $w_1, w_2, ..., w_n$ and R is functional Magnetic Resonance Imaging (fMRI) signals elicited by S. The brain encoding models learn to map from

Sentence	When somebody wrote a story in the Washington Post on Friday morning ...
POS	When [WRB] ... Washington [NNP] Post [NNP] on [IN] Friday [NNP] morning [NN]
NE	When [*] ... Washington [ORG] Post [ORG] on [*] Friday [TIME] morning [TIME]
SR	(wrote when) [ARGM-TMP] (wrote somebody) [ARG0]... (wrote story) [ARG2] ...
DEP	(wrote when) [advmod] (wrote somebody) [nsubj] (story a) [det] (wrote story) [dobj] ...

Table 1: A example sentence and labelled features. In NE, * means this word is not an entity.

stimuli S to the elicited brain response R for every voxel. Let $g_Z(S)$ be the feature elimination function which would remove the feature Z from S, then we have

$$R = B \times g_Z(S) + b$$

$g_Z(S) \in \mathbf{R}^d$, $R \in \mathbf{R}^v$. $B \in \mathbf{R}^{v \times d}$ is the regression weight, and b is the bias of the regression model. v is the number of voxels in the brain, and d is the dimension of stimuli feature, which is the same as word embeddings here.

In practice, fMRI measures the blood-oxygen-level-dependent (BOLD) signal, which is a slow changing process compared to neural activity. To align with the temporal delay of the BOLD signal, all feature vectors are convolved with a canonical hemodynamic response function (HRF)[1] and then downsampled to the sampling rate of fMRI.

$$\hat{R} = B \times downsample(conv(g_Z(S), hrf)) + b$$

After the training process is finished, we calculate the Pearson correlation $Corr(\hat{R}, R)$ between the real response R and estimated signals $\hat{R}$ to evaluate the model performance.

Significance Test

We run a block-wise permutation test (Adolf et al. 2014) with 10,000 permutations for each brain voxel to compute one-sided p-value and conduct False Discovery Rate (FDR) correction to get significant brain voxels (FDR $q < 0.05$) corresponding to input stimuli. On these significant voxels, we compute whether removing a specific feature causes a significant drop in encoding results.

To achieve that, we run a block-bootstrap test (Reddy and Wehbe 2021) to compute the significant difference between the original and one-feature-removed word representations. Specifically, we divide brain signals and stimuli representations into blocks, with approximately 20 seconds for each block. Then we randomly select ninety percent of the blocks to train regression models for all kinds of representations simultaneously and test on the remaining ten percent of blocks. By repeating this 10,000 times, we get an empirical distribution of each representation's prediction results (the Pearson correlation between real and predicted voxel activities), from which we can compute the difference between the results of the original and one-feature-removed representations. A voxel is correlated to a specific feature if removing this feature causes its encoding results significantly lower (FDR $q < 0.01$) than the original ones.

[1]The canonical HRF is a mathematical model that describes what the BOLD signal would theoretically be in response to a neural impulse.

	POS	NE	SR	DEP
Training	173322	173322	140517	203150
Validation	29962	29962	23104	24956
Test	35952	35952	30254	24949

Table 2: Details of the syntactic feature datasets.

Experiments

Syntactic features This paper adopts four syntactic features, including two word features, i.e., part-of-speech (POS) and name entity (NE) and two word-relation features, i.e., word dependency (DEP) and semantic role (SR)[2]. See an example of the four features in Table 1.

Syntactic feature datasets Since there is no annotation corpus that includes all the above four features, we use the Ontonotes 5.0 corpus[3] for POS, NE and SR features, and English Web Treebank of the Universal Dependencies 2.5 release[4] for DEP feature[5]. See Table 2 for more details. For word-level features (POS and NE), each word with its label in the corpus is taken as a sample. While for word relation features (DEP and SR), the concatenated representation of each word (or a predicate) x_i and its syntactic head (or its argument) x_j is taken as one sample, and the label of this sample is the relation between these two words. We train the MVNP on these two datasets and apply it on the stimuli text used in fMRI collection.

Brain activation data The fMRI dataset we use comes from (Zhang et al. 2020) which is publicly available at https://osf.io/eq2ba/. The dataset contains fMRI signals collected from 19 human subjects. While being scanned for fMRI, each subject listened to several audio stories collected from the Moth Radio Hour[6].

Experimental setup We adopt two types of pre-trained language models—the pre-trained BERT-base (Devlin et al. 2018) model[7] and the pre-trained ELMo (Peters et al. 2018) model[8]—to learn original word embeddings for two syntac-

[2]This paper follows the cognitive-linguistics literature and regard SR as a syntactic feature.

[3]https://catalog.ldc.upenn.edu/

[4]https://universaldependencies.org/

[5]To alleviate the class-imbalance existed in these features, similar classes are merged into one class and labels with less than 500 samples are discarded, resulting in 22 POS, 19 NE, 20 SR, and 35 DEP tags.

[6]https://themoth.org/radio-hour

[7]https://huggingface.co/bert-base-cased

[8]https://allennlp.org/

			POS	NE	SR	DEP
Random			11.77	19.54	42.43	12.30
ELMo	Word		97.65 ± 0.05	77.81 ± 0.99	77.49 ± 2.18	92.71 ± 0.15
	INLP	Null POS	$\mathbf{11.12 \pm 0.61}$	71.13 ± 0.49	64.34 ± 0.11	79.23 ± 0.14
		Null NE	97.19 ± 0.08	$\mathbf{12.09 \pm 0.71}$	74.91 ± 0.14	92.35 ± 0.10
		Null SR	83.99 ± 0.07	71.94 ± 0.32	$\mathbf{17.21 \pm 1.85}$	81.30 ± 0.18
		Null DEP	62.17 ± 0.09	57.16 ± 0.50	52.86 ± 0.22	$\mathbf{17.76 \pm 0.28}$
	MVNP	Null POS	$\mathbf{21.70 \pm 3.35}$	78.98 ± 1.47	71.09 ± 0.28	87.93 ± 0.10
		Null NE	94.69 ± 0.10	$\mathbf{13.79 \pm 8.77}$	76.70 ± 0.15	92.24 ± 0.09
		Null SR	96.79 ± 0.05	80.23 ± 0.60	$\mathbf{22.03 \pm 1.52}$	88.28 ± 0.17
		Null DEP	77.79 ± 0.10	67.38 ± 0.67	62.15 ± 0.20	$\mathbf{15.98 \pm 1.64}$
BERT	Word		97.89 ± 0.06	78.89 ± 2.66	78.04 ± 0.37	93.21 ± 0.20
	INLP	Null POS	$\mathbf{16.55 \pm 2.03}$	48.00 ± 0.76	66.89 ± 0.21	84.72 ± 0.17
		Null NE	96.71 ± 0.07	$\mathbf{5.02 \pm 0.89}$	75.53 ± 0.33	92.98 ± 0.14
		Null SR	70.93 ± 0.11	59.30 ± 1.19	$\mathbf{25.20 \pm 0.70}$	72.13 ± 0.10
		Null DEP	57.22 ± 0.13	37.45 ± 1.25	53.86 ± 0.36	$\mathbf{13.39 \pm 0.38}$
	MVNP	Null POS	$\mathbf{21.76 \pm 2.68}$	77.45 ± 0.64	72.80 ± 0.35	86.53 ± 0.15
		Null NE	95.23 ± 0.06	$\mathbf{12.30 \pm 6.38}$	77.23 ± 0.27	93.12 ± 0.18
		Null SR	96.86 ± 0.07	79.07 ± 0.84	$\mathbf{23.40 \pm 1.93}$	88.11 ± 0.14
		Null DEP	76.33 ± 0.11	63.10 ± 0.77	62.69 ± 0.40	$\mathbf{13.96 \pm 1.80}$

Table 3: Classification accuracy and its confidence interval of the INLP and MVNP methods. Rows are different types of representation models and "Null $*$" means $*$-removed representations. Columns are four syntactic features. We train each classifier for 10 times to compute the confidence intervals.

tic feature datasets and story stimuli used in fMRI collection. To evaluate which layer of BERT-base and ELMo best encodes the four syntactic features, we train a classifier for each layer of these two models and sum up the classification accuracy across all four features. Among all the 12 layers of BERT-base, layer 7 has the highest accuracy sum. And for ELMo, layer 1 has the highest accuracy sum. Therefore, we adopt BERT-base layer 7 and ELMo layer 1 activation as word embeddings.

For each syntactic feature, we compute the mean vectors of each class to get the mean vector matrix and its null space projection matrix on two syntactic feature datasets. To evaluate the elimination results of each projection matrix, we train linear classifiers for each feature on the word embeddings after projection and test the trained classifier on the test datasets. The results are compared with two baselines. One is the random results, which are calculated as the proportion of the majority class in a specific feature. Another baseline is the classification results of the original BERT word embeddings. Then the projection matrix of each feature is used to remove the feature in the word embeddings of story stimuli [9].

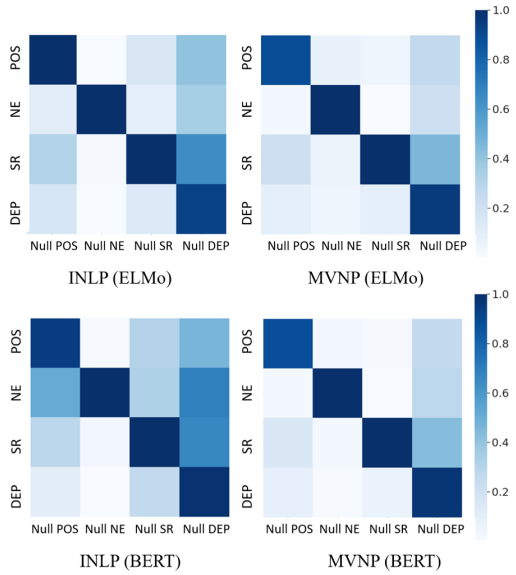

Figure 2: The accuracy drop percentage by the feature-elimination methods.

Results and Analysis

Feature-Elimination Results

To test models' feature elimination ability, we show the feature classification results for both ELMo and BERT embeddings by the INLP and MVNP models in Table 3. As shown, the elimination results on ELMo and BERT embeddings are similar. In general, both INLP and MVNP methods can remove a specific feature from two types of word embeddings effectively. Specifically, as the bold numbers in Table 3 shows, the INLP method can achieve a slightly better (lower) classification accuracy on three features than our MVNP method. However, compared with our method, the INLP method has a larger influence on other features after eliminating one feature, especially when removing POS, SR and DEP features respectively from the original embeddings.

To intuitively compare the results of the two feature-elimination methods, we compute the accuracy drop per-

[9]The code used in this paper is available at https://github.com/xhzhang-1/probing-syntactic-representation

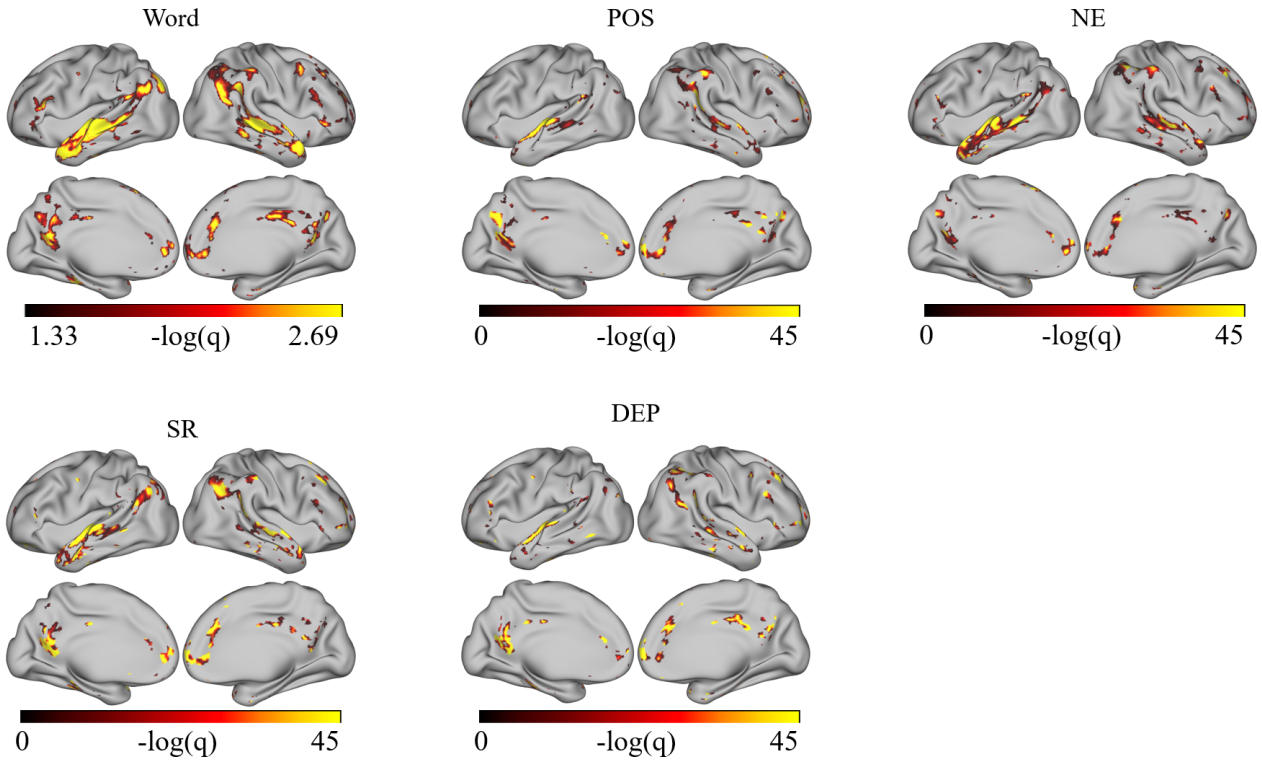

Figure 3: Cortical representations of words and syntactic features. The color-highlighted areas for each feature include voxels with significant difference (one-sided z-test, FDR $q < 0.01$) between the original and one-feature-removed embeddings.

centage after eliminating each feature:

$$Acc_{drop} = \frac{acc_{BERT/ELMo} - acc_{Null*}}{acc_{BERT/ELMo} - acc_{random}}$$

As shown in Figure 2, compared with the INLP method, the MVNP has a smaller influence on other features when removing one feature from ELMo and BERT embeddings. But the original BERT embeddings have slightly higher classification accuracy. Hence, we choose BERT embeddings and the MVNP method in the subsequent brain encoding experiment.

Furthermore, we see that the relationship between the four features is not symmetric. That is, removing SR or NE causes a slight effect on other features. But removing POS or DEP has a more significant influence on other features. This is reasonable because the four features are correlated to a certain degree. For example, named entity words are mostly nouns, but a noun can be any entity category. Therefore, it is inevitable to cause some disturbance to others when removing one specific feature. To eliminate this impact to some extent, we conduct a significant test analysis in the subsequent brain encoding experiments.

BERT	Null POS	Null NE	Null SR	Null DEP
0.6993	0.6870	0.6820	0.6775	0.6875

Table 4: The Pearson correlation between representational cosine-similarity scores and human judgments.

To further illustrate that our MVNP method can successfully remove a specific feature but retain other information, including semantics, we use Stanford's Contextual Word Similarities (SCWS) (Huang et al. 2012) to quantitatively evaluate the influence of MVNP on word semantic similarity. As shown in Table 4, our proposed MVNP model can successfully hold semantic information while eliminating one specific syntactic feature.

In sum, the purpose of this paper is to investigate how word syntactic features are represented in the brain, thus we need only one specific feature removed from the original word embeddings at one time and the others remain unchanged. Experimental results have shown the effectiveness of the proposed MVNP method.

Cortical Representations of Syntactic Features

Based on the original BERT and one-feature-removed word embeddings, we use the brain encoding method and significance test analysis to find brain voxels that are responsible for the four syntactic features. Figure 3 shows the brain encoding results of word embeddings generated by BERT, and POS, NE, SR, DEP removed representations computed by MVNP.

Syntactic brain networks are largely overlapped with semantic brain networks In general, word representa-

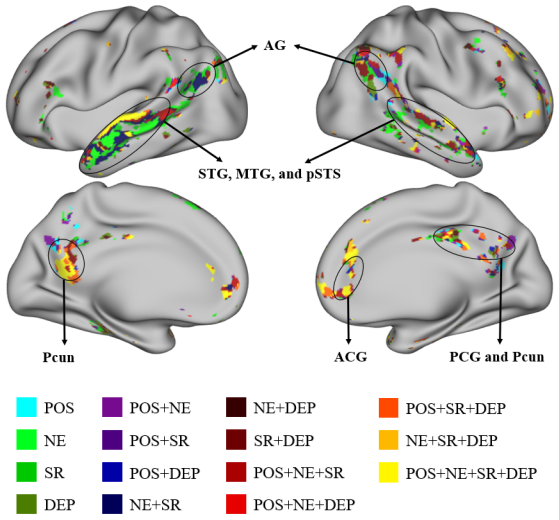

Figure 4: Overlapped brain regions for words' syntactic features.

tions, including all semantic and syntactic components, are encoded in the inferior frontal gyrus (IFG), superior temporal gyrus (STG), posterior superior temporal sulcus (pSTS), middle temporal gyrus (MTG), inferior temporal gyrus (IFG), precuneus (Pcun), cingulate gyrus (CG) and angular gyrus (AG). Most of these brain areas agree with previous findings on the brain networks of language processing (Fedorenko and Kanwisher 2011; Yang et al. 2019), supporting the validity of our proposed framework that uses computational representations to study brain language comprehension.

Compared with word representations, four word syntactic features (POS, NE, SR, DEP) have much less activated brain regions. These brain regions (mainly including the lateral temporal cortex and angular gyrus) largely replicate the results of previous work in brain representations of syntax (Blank et al. 2016; Hagoort and Indefrey 2014). These brain networks also overlap with classic semantic brain networks (Hagoort and Indefrey 2014; Huth et al. 2016) to a great extent, at least under the spatial resolution of the current fMRI technique. This finding supports a popular research view that language networks as a whole are sensitive to both semantic and syntactic information (Blank et al. 2016), and the brain networks activated by semantics and syntax are largely overlapped (Wang et al. 2020).

In addition, we also find two brain regions activated by syntactic features that have not been found in previous work, i.e., the precuneus and the cingulate gyrus. Both of them are sensitive to several syntactic features, suggesting that they may play an important role in brain language processing and the brain foundations of syntactic information processing might be broader than those suggested by classical studies.

Syntactic features are distributively represented in the brain As shown in Figure 3, four syntactic features are distributively represented across the brain cortex, indicating that word syntactic features highly rely on distributed brain networks instead of a local brain region. There are two possible explanations. One is that our brain uses distributed representation to encode everything even primitive features. Another reason may be that the four word syntactic features defined by linguists are not the primitive components used by our brain.

There are also recent works that found a distributive representation network as ours (Yang et al. 2017). Different from previous work that precisely controls one variable at a time and investigates one specific feature representation in the brain, we adopt the data-driven method and analyze various syntactic features in the same experiment. These are all cutting-edge research questions and different exploratory methods have their advantages. Together, using the classic hypothesis method and the data-driven method, we may draw a big picture of human language understanding.

Different syntactic features are represented and integrated in a hierarchical brain system To further show the functional cortical division of different syntactic features, we put four encoding results together. As shown in Figure 4, there are some brain areas (especially the STG, pSTS, Pcun, and AG) that encode several syntactic features, while other areas are only sensitive to one feature. And each feature has slightly different brain networks. For instance, the rostral area of STG and anterior superior temporal sulcus (aSTS) is only correlated to NE. In precuneus areas, there are scattered small regions that only correlate to POS. These findings suggest a hierarchical organization for the neural representations of syntactic features.

Conclusions

The classic syntactic-related brain areas mainly include IFG, MFG, pSTS, and AG. Through a data-driven method, this paper discovered the contribution of these brain regions to syntactic representation in the brain and additionally suggested some possible contributions of other brain regions. This work corroborates and extends previous findings, highlighting the value of introducing the latest language processing models in studying brain language comprehension, and suggests that the brain foundations of syntactic information may be broader than those suggested by classical studies.

However, our purposed framework has several limitations. First, the MVNP method is limited to categorical features and cannot be directly used on numerical features such as word frequency or word surprisal. Second, both the MVNP method and the voxel-wise brain encoding model are under linear assumptions. Although the linear assumption is widely used in fMRI researches, future work can explore more complex transforms with the increase of data and computational resources. Finally, the voxel-wise encoding models used in our framework can only study the activation of single voxels. Future work can combine MVNP with other methods such as multivariate pattern analysis to explore the spatial activation patterns of multiple voxels in the brain.

Acknowledgments

This work is supported by the Natural Science Foundation of China under Grant 62036001 and 61906189. This work is also supported by the independent research project of National Laboratory of Pattern Recognition.

References

Adolf, D.; Weston, S.; Baecke, S.; Luchtmann, M.; Bernarding, J.; and Kropf, S. 2014. Increasing the reliability of data analysis of functional magnetic resonance imaging by applying a new blockwise permutation method. *Frontiers in Neuroinformatics*, 8: 72–72.

Bao, Y.; Zhou, H.; Huang, S.; Li, L.; Mou, L.; Vechtomova, O.; yu Dai, X.; and Chen, J. 2019. Generating Sentences from Disentangled Syntactic and Semantic Spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6008–6019.

Binder, J. R.; Conant, L. L.; Humphries, C. J.; Fernandino, L.; Simons, S. B.; Aguilar, M.; and Desai, R. H. 2016. Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, 33: 130–174.

Blank, I.; Balewski, Z.; Mahowald, K.; and Fedorenko, E. 2016. Syntactic processing is distributed across the language system. *NeuroImage*, 127: 307–323.

Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; and Kalai, A. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS'16 Proceedings of the 30th International Conference on Neural Information Processing Systems*, volume 29, 4356–4364.

Chen, E.; West, W. C.; Waters, G.; and Caplan, D. 2006. Determinants of bold signal correlates of processing object-extracted relative clauses. *Cortex*, 42(4): 591–604.

Conneau, A.; and Kiela, D. 2018. SentEval: An Evaluation Toolkit for Universal Sentence Representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. N. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Fedorenko, E.; and Kanwisher, N. 2011. Some regions within broca's area do respond more strongly to sentences than to linguistically degraded stimuli: A comment on rogalsky and hickok (2011). *Journal of Cognitive Neuroscience*, 23(10): 2632–2635.

Fernandino, L.; Binder, J. R.; Desai, R. H.; Pendl, S. L.; Humphries, C. J.; Gross, W. L.; Conant, L. L.; and Seidenberg, M. S. 2016. Concept Representation Reflects Multimodal Abstraction: A Framework for Embodied Semantics. *Cerebral Cortex*, 26(5): 2018–2034.

Gauthier, J.; and Levy, R. 2019. Linking artificial and human neural representations of language. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 529–539.

Hagoort, P.; and Indefrey, P. 2014. The Neurobiology of Language beyond Single Words. *Annual Review of Neuroscience*, 37(1): 347–362.

Hewitt, J.; and Manning, C. D. 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4129–4138.

Huang, E.; Socher, R.; Manning, C.; and Ng, A. 2012. Improving Word Representations via Global Context and Multiple Word Prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 873–882.

Huth, A. G.; de Heer, W. A.; Griffiths, T. L.; Theunissen, F. E.; and Gallant, J. L. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600): 453–458.

Jain, S.; Vo, V. A.; Mahto, S.; LeBel, A.; Turek, J. S.; and Huth, A. 2020. Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech. In *Advances in Neural Information Processing Systems*, volume 33.

Linzen, T.; Dupoux, E.; and Goldberg, Y. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4(1): 521–535.

Mathieu, M.; Zhao, J.; Sprechmann, P.; Ramesh, A.; and LeCun, Y. 2016. Disentangling factors of variation in deep representations using adversarial training. In *NIPS'16 Proceedings of the 30th International Conference on Neural Information Processing Systems*, volume 29, 5047–5055.

Mitchell, T. M.; Shinkareva, S. V.; Carlson, A.; Chang, K.-M.; Malave, V. L.; Mason, R. A.; and Just, M. A. 2008. Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science*, 320(5880): 1191–1195.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, 2227–2237.

Pylkkänen, L. 2019. The neural basis of combinatory syntax and semantics. *Science*, 366(6461): 62–66.

Ravfogel, S.; Elazar, Y.; Gonen, H.; Twiton, M.; and Goldberg, Y. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7237–7256.

Reddy, A. J.; and Wehbe, L. 2021. Syntactic representations in the human brain: beyond effort-based metrics. In *bioRxiv*.

Sun, J.; Wang, S.; Zhang, J.; and Zong, C. 2019. Towards Sentence-Level Brain Decoding with Distributed Representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1): 7047–7054.

Tenney, I.; Xia, P.; Chen, B.; Wang, A.; Poliak, A.; McCoy, R. T.; Kim, N.; Durme, B. V.; Bowman, S. R.; Das, D.; and Pavlick, E. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019*.

Thompson, C. K.; Bonakdarpour, B.; Fix, S. C.; Blumenfeld, H. K.; Parrish, T. B.; Gitelman, D. R.; and Mesulam, M. M. 2007. Neural correlates of verb argument structure processing. *Journal of Cognitive Neuroscience*, 19(11): 1753–1767.

Toneva, M.; and Wehbe, L. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, volume 32, 14954–14964.

Wang, S.; Zhang, J.; Lin, N.; and Zong, C. 2020. Probing Brain Activation Patterns by Dissociating Semantics and Syntax in Sentences. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(5): 9201–9208.

Wehbe, L.; Murphy, B.; Talukdar, P.; Fyshe, A.; Ramdas, A.; and Mitchell, T. 2014. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLOS ONE*, 9(11).

Xie, Q.; Dai, Z.; Du, Y.; Hovy, E. H.; and Neubig, G. 2017. Controllable Invariance through Adversarial Feature Learning. In *Advances in Neural Information Processing Systems*, volume 30, 585–596.

Xu, K.; Cao, T.; Shah, S.; Maung, C.; and Schweitzer, H. 2017. Cleaning the Null Space: A Privacy Mechanism for Predictors. In *AAAI*, 2789–2795.

Yang, H.; Lin, Q.; Han, Z.; Li, H.; Song, L.; Chen, L.; He, Y.; and Bi, Y. 2017. Dissociable intrinsic functional networks support noun-object and verb-action processing. *Brain and Language*, 175: 29–41.

Yang, X.; Li, H.; Lin, N.; Zhang, X.; Wang, Y.; Zhang, Y.; Zhang, Q.; Zuo, X.; and Yang, Y. 2019. Uncovering cortical activations of discourse comprehension and their overlaps with common large-scale neural networks. *NeuroImage*, 203: 116200.

Yu, X.; Law, S. P.; Han, Z.; Zhu, C.; and Bi, Y. 2011. Dissociative neural correlates of semantic processing of nouns and verbs in Chinese–a language with minimal inflectional morphology. *NeuroImage*, 58(3): 912–922.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

Zhang, Y.; Han, K.; Worth, R. M.; and Liu, Z. 2020. Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature Communications*, 11(1): 1877.