

# Cross-Lingual Word Alignment Grounded on Brain Semantics

Nikitas Theodoropoulos

December 20, 2021

## 1 Introduction

Most methods for learning dense word representations (i.e., embeddings), rely on the distributional hypothesis: words with similar meanings will appear in similar contexts. Word representations learned based on this assumption have been immensely useful for a vast number of NLP tasks. Interestingly, it has also been observed that continuous word embeddings learned in this manner exhibit similar properties across languages, relating to geometric and distributional features of their vector spaces. This similarity can be used to learn a *mapping* from two (or more) languages to a shared space, creating cross-lingual embeddings. The shared embedding space can then be used to approximate a *dictionary* between the two languages, effectively translating from one language to another by finding nearest neighbors.

This alignment, however, most commonly relies on the assumption that the embedding spaces are approximately isomorphic, and a simple linear mapping is used. So far, linear mapping methods with sophisticated refinement techniques have achieved good performance both in the supervised and unsupervised settings. However, current alignment methods suffer in low-resource settings, occasionally failing to converge, and are challenged by etymologically and morphologically distant language pairs.

To address these issues, we propose a novel approach to the alignment problem, relying on the most powerful language processing system that we know of: the human brain. Intuitively, the representation of meaning in the brain should be independent of the exact form of words. Neuroscientific studies have validated this claim, by showing that a common signal exists *across languages*, despite cultural or individual speaker differences. The brain semantic space can be captured by functional MRI (fMRI) activations, taken when people are processing a word or a sentence during a lexical task. We hypothesize that by learning to map traditional embedding spaces of two languages to a neural, brain-derived embedding space, we can perform an unsupervised alignment of the two spaces, grounded in brain semantics.

To achieve this, we learn a mapping from distributional embeddings to fMRI-based word representations. This follows a recent line of work, where the semantic information present in fMRI brain scans from word-stimulus experiments has been used to improve state-of-the-art language models. We combine these approaches with methods for unsupervised embedding space alignment, for inducing a bilingual dictionary. We hypothesize that the brain responses are to a degree naturally aligned across languages, and this bilingual signal can be used to increase performance of current translation methods, effectively addressing cases of low-resource and etymologically diverse languages. Our main research questions are: *Does the shared semantic signal in brain semantics lead to better performance in unsupervised dictionary learning?*

Importantly, apart from improving current state-of-the-art techniques in embedding space alignment, this work will also provide empirical and experimental proof for the existence of shared semantic representations across languages.

We investigate the processing and semantic analysis of fMRI in literature in §3. Recent work on cross-lingual alignment is summarized in §4. Neuroscience based evidence on the shared brain response for different languages is reviewed in §2. Finally, we state our proposed approach in §5.

## 2 Neuroscientific Background

We review key neuroscience literature for evidence of language-independent representation of meaning in the human brain. Across these methods, most commonly, a classifier mapping brain encodings to word IDs is trained in one language. The resulting classifier is then tested in word identification in the other language. Above chance accuracy suggests the existence of a shared meaning space across languages. Finally, in a more fine-grained analysis, many methods also investigate the correlation of activations in specific brain areas.

Early work by [Buchweitz et al. \(2012\)](#) focused on semantic neural responses in late bilinguals. They used English fMRI activity to predict nouns in Portuguese. Despite the limited amount of stimuli in two categories: tools (e.g., hammer, screwdriver) and dwellings (e.g., castle, apartment), there was a significant and above chance prediction accuracy with cross-language training and testing. Following this line of work, [Correia et al. \(2014\)](#) demonstrated that brain-based decoding at the level of within a semantic category (e.g., animals) is possible both across languages (English, Dutch) and within a language, reporting also observations of language-invariant word representations.

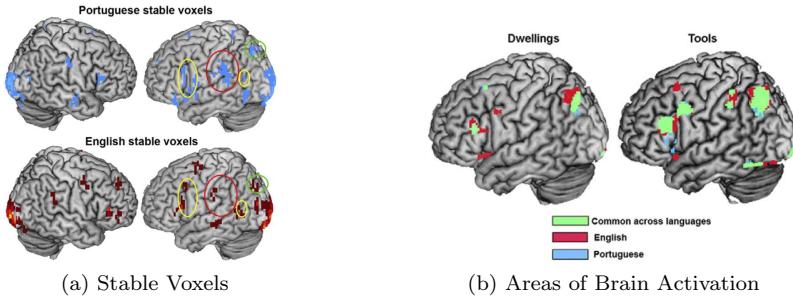


Figure 1: Brain activation patterns reported by [Buchweitz et al. \(2012\)](#). (a) Stable voxels in English and Portuguese: voxel clusters show similarities in activated brain areas across languages, with some locations previously associated with tool manipulation and dwellings. (b) Areas of brain activations in the two languages when thinking about the tested objects. Objects in the “tools” category show significant activation overlap, potentially resulting from their high concreteness.

Recently, [Hu et al. \(2019a\)](#) expanded on previous works by moving beyond the word-level, to investigate the semantic processing of sentences across languages, in proficient bilinguals. The stimuli consisted of 48 tuples of images, Chinese and Japanese sentences describing different activities. Participants had to classify a pair as either coherent or incoherent. A binary SVM classifier was trained to distinguish fMRI activations with the same two labels. Significant results were achieved in across-language classification, with accuracies however in the lower range ([50,55]). Finally, the authors suggest the existence of a common neural system across languages in sentence processing, and name possible brain locations.

In a related approach, [Yang et al. \(2017\)](#) investigated differences in the neural representation in three languages, with a novel experimental paradigm. They showed that training on two languages and testing with a prediction task on a third, can result in a richer representation, with accuracy increasing and getting closer to single language performance. The authors determined that some categories reliably benefited from the joint training: (Person, Communication, Social, Knowledge, Natural). The results suggest that using two languages in training is always beneficial, especially with respect to abstract and social concepts.

Closer to our line of work, [Zinszer et al. \(2016\)](#) investigated the underlying semantic representation of stimuli by cross-translating English and Chinese. English native speakers, and Chinese bilinguals were shown translation pairs of concrete monosyllabic nouns. Participants completed a semantic relatedness task. For each Region of Interest (ROI) and for each word a voxel activations vector was computed and afterward, the correlation matrix across vectors was calculated. The authors then relied on the isomorphic assumption: the matrices for two languages should be equal up to a permutation of rows and columns. By trying all

possible permutations, the authors translated between languages. Six ROIs in total produced 7/7 correct translations, and an additional 11 ROIs correctly translated 5/7 words.

In conclusion, the above review suggests that there is some evidence that meaning representation in the brain is to a degree unaffected by stimulus type (language or modality) and activations can encode also semantic information, rather than purely syntactic or other surface properties. Furthermore, studies point out to common semantics in the brain should at both the individual word and the sentence level. We note however that a limited set of stimuli were tested in each experiment, consisting mainly of concrete nouns with a more or less universally accepted meaning. Finally, semantic differences related to the cultural, social or personal understanding of words are present in brain responses, and can be leveraged for a richer representation.

### 3 Brain data in NLP

The seminal work of [Mitchell et al. \(2008\)](#) demonstrated that fMRI signals encode meaningful semantic information, which can be effectively used to map between distributed semantic representations and voxel activations. This was the first computational model to predict brain patterns associated with unknown words, and has sparked an active line of research in using brain representations to improve model architectures, task performance, or derive new neuroscientific insights. In the following sections, we summarize related work in terms of employed datasets, voxel selection, alignment between subjects, and connecting brain semantics to a distributional embedding space.

#### 3.1 Datasets and Stimuli

Generally, the most common neuroimaging modality is functional MRI (fMRI), which records the blood-oxygen response in the whole brain, while subjects are participating in a (lexical) task. In comparison to other methods, fMRI offers a good balance between spatial resolution ( $\sim 1\text{-}3$  mm) and temporal resolution ( $\sim 1\text{-}2$  sec), however the signal is inherently noisy and pre-processing is a crucial step for extracting useful semantic information. Here we focus mostly on the text modality, although visual and auditory stimuli have also been used. The examples can be single words, displayed one at a time, or in context as a word cloud or a series of sentences with a common theme. Beyond isolated input, narratives have been increasingly employed, in experimental designs where participants are reading a book chapter. These allow studying longer contexts and naturalistic settings, but come with significant drawbacks in disambiguating the effect of each sentence, or word. In support of our project proposal, we focus on datasets that expand to other languages beyond English.

**Sentences** [Pereira et al. \(2018\)](#) moved beyond isolated words, to evaluate abstract concepts using sentences in three experiments. In the first experiment, subjects were shown 180 concept words selected to cover a pre-defined semantic space, representing a cluster of related words based on GloVe vectors ([Pennington et al., 2014](#)). The stimuli were shown in three paradigms with multiple repetitions, in a sentence, as an image, or in a word cloud. In the second and third experiments, the stimuli consisted of a collection of sentences for different topics. One fMRI image was captured for each sentence. [Schoffelen et al. \(2019\)](#) have released a massive 204 participant study with both visual and auditory stimuli. The participants were native Dutch speakers, with the total stimuli consisting of 360 sentences in **Dutch**. The visual subjects read words one at a time in a sentence, in the correct and in a scrambled order, each subject viewing 60 sentences in total, alternating sentences and word lists. The dataset by [Hu et al. \(2019b\)](#) contains fMRI scans from 29 Chinese–Japanese bilingual speakers who were asked to assess the coherence of 48 pairs of images, 48 pairs of corresponding captions, one est in **Chinese** and one in **Japanese**. The images depicted one or two people performing common daily activities, and each pair was a sequence of coherent or incoherent events.

**Narratives** The dataset by [Wehbe et al. \(2014\)](#) includes fMRI data recorded while participants read a chapter from “Harry Potter”. The chapter contains 5176 words and was recorded from nine participants.

The dataset in [Bhattasali et al. \(2020\)](#) includes fMRI recordings of participants listening to the first chapter of Alice’s Adventure in Wonderland, which comprises 2,129 words in 84 sentences, demonstrating reasonable syntactic diversity. For the fMRI data, there are anatomical and functional scans, and the dataset is annotated with predictors for prosody, morphology, and syntax. The dataset collection by ([Hanke et al., 2014](#); [Liu et al., 2019](#); [Hanke et al., 2016](#); [Sengupta et al., 2016](#)) contains high-resolution fMRI data from 20 participants in response to prolonged auditory stimulation with the feature film “Forrest Gump” in **German**. The dataset by [Dehghani et al. \(2017\)](#) includes fMRI scans from 90 participants reading in their native language (**Farsi**, **Chinese**, **English**), 30 in each language, where subjects read 40 short personal stories that had been collected from weblogs, each roughly 150 words. However, the data is not publicly available. In the recently released dataset by ([Stehwien et al., 2020](#); [Li et al., 2021](#)), 49 **English**, 35 **Chinese** and 28 **French** speakers listened to the same audiobook of “The Little Prince” in their native language while fMRI images were collected. The audio spans around 100 minutes and contains approximately, 15000 words. The data has been pre-processed and aligned to a common brain template for all subjects and languages, providing also the audio stimuli, word and fMRI acquisition times, and linguistic annotations. In its full release, it plans to include translations of the original children’s story in 26 languages.

This review demonstrates that there is a plethora of multilingual fMRI datasets, with speakers being presented the same stimulus in different languages. We hypothesize that these brain scans paired with their lexical stimuli can be used for unsupervised cross-lingual alignment.

### 3.2 Preprocessing

The fMRI brain scan consists of voxels (3-d cubes), corresponding to different regions of aggregated signal in the brain. Their number varies with respect to the voxel size and the shape of an individual’s brain. The activity measured in many of these voxels is most likely not related to language processing, and might change due to physical processes like the noise perception in the scanner. In these cases, learning a mapping model from the stimulus representation to the voxel activation will not succeed because the stimulus has no influence on the variance of the semantic signal. For this reason, effective voxel selection is crucial for extracting semantic information from brain data. Importantly, as [Artemova et al. \(2020\)](#) mention, each area, represented by a voxel, responds largely independently of the other areas, thus a separate model is needed to fit responses in each cortical voxel. We highlight these approaches in this section.

Restricting the brain response to voxels that fall within a pre-selected set of regions of interests can be considered as a **theory-driven analysis**. [Toneva and Wehbe \(2019\)](#) reduced the voxels by using previous knowledge about groups of regions of interests. Past experiments have found that a set of regions in the temporo-parietal and frontal cortices are activated in language processing and are collectively referred to as the language network. [Brennan et al. \(2016\)](#) select regions related to sentence comprehension.

A more **information-driven** approach is proposed by [Kriegeskorte et al. \(2006\)](#). Searchlight analysis moves a sphere through the brain to select voxels (comparable to sliding a context window over text) and analyze the predictive power of the voxel signal within the sphere. [Pereira et al. \(2018\)](#) in a decoding experiment, selected 5000 most informative voxels by their power to predict embedding vectors (max correlation with true values). [Mitchell et al. \(2008\)](#) for each participant analyze all six brain responses for the same stimulus and select 500 voxels that exhibit a consistent variation in activity across all stimuli. Voxel stability can be calculated as the average Pearson coefficient  $r$  for all trial pair combinations.

As noted by [Beinborn et al. \(2019\)](#) for datasets where trials are not present (i.e., only one stimulus presentation per participant), a **prediction-driven** metric can be used to select informative voxels. Notably, [Jain and Huth \(2018\)](#) estimated a separate encoding model for each voxel and calculated model performance for a single voxel as the Pearson correlation coefficient between real and predicted responses. [Gauthier and Ivanova \(2018\)](#) recommend evaluating voxels based on explained variance. Lastly, [Bingel et al. \(2016\)](#) use 10-PCA for low-dimensional representations.

### 3.3 Mapping from Voxel Space to Lexical Embeddings

Due to the lack of large fMRI datasets, the most common method that is employed for obtaining lexical cognitive embeddings from fMRI data is linear or ridge regression. However, neural networks have also been used. A model mapping from a lexical space to a voxel space is referred to as an encoder, and respectively a decoder in the opposite direction.

In Mitchell et al. (2008) the activation of a voxel  $v$  for word a  $w$  is given by:

$$y_v(w) = \sum_{i=1}^m c_{v,i} f_i(w), \quad \forall v = 1 \cdots V, \quad (1)$$

For stimuli representation the authors use the similarity with 25 seed verbs manually selected with respect to psycholinguistic criteria. The similarity  $f_i(w)$  between seed word  $i$  and word  $w$  is calculated from co-occurrence statistics in a large corpus.  $V$  is the total number of voxels and  $c_{v,i}$  are learned weights that are estimated via regression by using fMRI data for known words.

Athanasiou et al. (2018) follows the same approach, deriving cognitive embeddings and evaluating their performance in NLP downstream tasks (MEN, ESSLLI, Sensicon, SNLI). Some authors Anderson et al. (2016) have also used similarity encoding, where the activation for an unknown word  $w$  is computed as a sum of activations of known words  $u_i$ , weighted by the similarity  $\text{sim}(u_i, w)$ .

Several works evaluated the initial mapping by Mitchell, Devereux et al. (2010) report that the automatically choosing the set of verbs leads to equally good results. Jelodar et al. (2010); António Rodrigues et al. (2018) use WordNet based features for the 25 seed words, achieving comparable results. Abnar et al. (2018) conclude that no input representation is better overall at predicting brain activations, although morphological and dependency based models can potentially perform better. Anderson et al. (2016) used a 65 experiential attribute that span different aspects of experience, ratings for each semantic dimension were crowd-sourced. Bulat et al. (2017) review many semantic models for input representation, including dependency, association and image based. They conclude that visual, rather than linguistic, information is a stronger predictor of brain activity for concrete nouns. Huth et al. (2016) use low dimensional co-occurrence vectors and sentence fMRI data to map words to cortical areas. They use a generative model, with a probability distribution for semantic category clusters in the brain, and emission probabilities modeled as Gaussians.

Pereira et al. (2018) use a ridge regression to predict GloVe vectors from voxel activations. They show that a decoder learned in the isolated word setting, can accurately classify sentences from their fMRI with different levels of granularity. In contrast with earlier works the stimuli include abstract nouns. Recent works Cao and Zhang (2019); Hollenstein et al. (2019) attempt to map conventional word embeddings to brain-derived embeddings, using a neural network with one hidden layer. Specifically, Cao and Zhang (2019) report that by using neural networks, both encoding and decoding accuracy is improved compared to a linear regression model on the same input.

Anderson et al. (2016) proposes similarity based decoding/encoding. Given  $b_i$  the known neural activations, and  $u_i$  the stimulus representation. We predict a brain image  $b'$  for an unseen word  $u'$  as a weighted sum of known brain activations  $b_i$ , where the weights are given by  $\text{corr}(u_i, u')$ .

**Natural Narratives** There has been growing interest in using more natural stimuli in fMRI experiments, instead of isolated words in a controlled setting. For many datasets the stimuli consist of sentences, often from large narratives. Due to the low temporal resolution of fMRI words are not clearly separated in neural images, and an fMRI Image  $N$  corresponds to a set of words  $\{w_1, w_2, \dots, w_n\}$  which makes the mapping harder. Wehbe et al. (2014) use an input representation of 195 features for each word  $w_i$  and try to predict  $N$  from  $\sum_{i=1}^4 w_i$  ( $n=4$ ), summing the representations of all corresponding words. Bingel et al. (2016) address the low temporal resolution problem by assigning the same fMRI to all corresponding words, and then sliding a Gaussian window across tokens. They use the resulting representations with an HMM to

improve performance in POS induction. [Schwartz et al. \(2019\)](#) use BERT to predict the fMRI image from the corresponding sentence.

Due to the sequential nature of the data, an LSTM can be used to map between word embeddings and neural activations. The common neural image for a set of words can be predicted as a function of the corresponding token hidden states. [Abnar et al. \(2019\)](#); [Qian et al. \(2016\)](#), conclude that LSTM sentence representations correlate well with brain data. [Jain and Huth \(2018\)](#) use a ridge regression on top of an LSTM pretrained for language modeling to map sentence stimuli to fMRI responses. They notice that LSTMs encode context and are better at predicting activations of individual words in a sentence.

### 3.4 Application to NLP tasks

Following the usual approach for transfer learning in NLP, brain-derived representations could be used to augment and increase performance of task-based models. The simplest approach is feature-based, where these embeddings are used as input with or without fusion with traditional embeddings. A pretrained model mapping words to cognitive space could also be fine-tuned end-to-end for a specific task. [Bingel et al. \(2016\)](#) use fMRI features derived from sentences combined with text features to increase performance in a POS tagging task.

The common approach for interpreting language model representations is by using specific NLP tasks, word annotations or behavioral measures. Some researchers used fine-tuned language models to predict brain activity and evaluate the brain representations. Such fine-tuning is a new paradigm in learning about human language processing and it relies on encoding information from targets of a prediction task (e.g., the brain representations in our case) into the model parameters. The goal is to optimize these models to take advantage of multiple sources of information about language processing in the brain.

There is little prior work that evaluates or improves NLP models through brain data. [Søgaard \(2016\)](#) investigate whether a word embedding contains cognition-relevant semantics by measuring how well it can predict eye tracking data and fMRI recordings. Similarly [Hollenstein et al. \(2019\)](#) proposed a framework for intrinsic word embedding evaluation based on how much they reflect brain semantics. Six types of word embeddings were evaluated by regressing on fMRI, EEG and eye tracking data. They report correlation between the cognitive evaluation and performance in NER and Question Answering tasks.

[Jain and Huth \(2018\)](#) aligned layers from a Long Short-Term Memory (LSTM) model to predict fMRI recordings of subjects listening to stories to differentiate between the amount of context maintained by each brain region. [Toneva and Wehbe \(2019\)](#) used brain activity recordings to show that different network representation encode information relevant to language processing at different context lengths. Both [Toneva and Wehbe \(2019\)](#); [Schwartz et al. \(2019\)](#) observed that by modifying the pretrained BERT model to better capture brain-relevant language information, they achieved higher accuracy results at NLP tasks. This finding suggests that altering an NLP model to better align with brain recordings of people processing language may lead to better language understanding by the NLP model.

[Affolter et al. \(2020\)](#) create an fMRI decoder for direct classification. achieving up to 15% top-5, 5.2% top-1 accuracy. They find that predicting word embeddings degrades model performance. The decoder is trained as an autoencoder using sentence data. Surprisingly, they use data from different subjects from Pereira, with no previous alignment, incorporating instead a subject separation term in the loss function. Finally, using GPT-2 and the decoder they achieve a neurally guided word generation.

## 4 Cross-Lingual space alignment

Cross lingual alignment consists in learning a shared word vector space, where words with similar meanings across two languages obtain close-distanced vectors. Given two monolingual embedding spaces  $X, Y$  this is usually formulated as a mapping  $W$  where  $WX \approx Y$ . Word pairs can be formed by selecting the nearest

neighbor  $y_t \in Y$  of  $x_s \in WX$  by a suitable metric (e.g., cosine similarity). Early work by Mikolov et al. (2013) showed that a simple linear mapping with a supervision of 5000 words, can successfully align embedding spaces. Xing et al. (2015) improved the approach by normalizing and enforcing the orthogonal constraint.

Initial methods were supervised, relying on different constraints and normalizations. Artetxe et al. (2018a) summarizes many of these methods as a series of Procrustes transformations  $\prod_i W_{(i)} X$ , with intermediate normalization steps. Related research attempted a semi-supervised mapping, relying only on small parallel corpora. Smith et al. (2017) relied on common strings and Artetxe et al. (2017) only on numerals and achieved comparable results to supervised methods. This opened the way for fully unsupervised approaches.

Unsupervised methods, without any parallel singal, differ from each other in how they build the initial bilingual dictionary, which is not given as input (like in supervised settings), but inferred from the monolingual data and later refined. The general method for unsupervised alignment is described in Figure 2, and a motivating example for the problem is given in Figure 3. We review these two key approaches that have remained competitive and serve as baselines for later works.

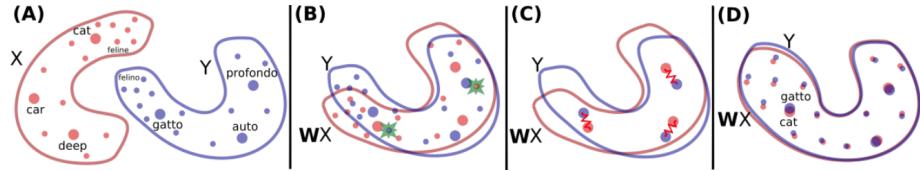


Figure 2: General steps for Unsupervised alignment: (A) The monolingual embedding spaces we want to align (B) An initial seed dictionary  $D$  or mapping  $W$  is constructed by means of distributional matching (C) The mapping  $W$  is further refined in a pseudo-Supervised (possibly iterative) step (D) We translate using the mapping  $W$  and by computing nearest neighbors in the shared space. Several metrics have been proposed to address the hubness problem. Image reproduced from Conneau et al. (2018).

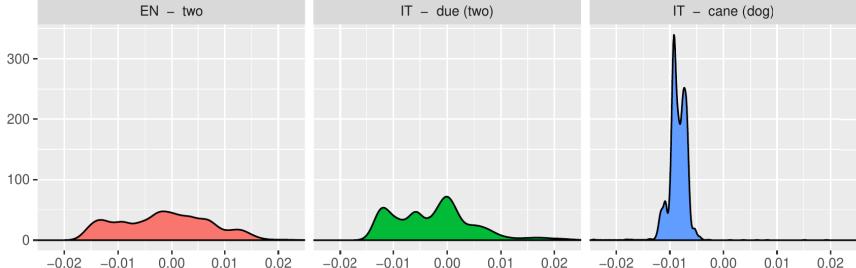


Figure 3: Motivating example for the unsupervised cross-lingual mapping problem, from Artetxe et al. (2018b). The similarity distributions between an English and two Italian words are depicted, demonstrating that equivalent translations, two and due (two), have more similar distributions than unrelated words, two and cane (dog). This property is exploited to build an initial solution, that is then iteratively refined.

In the GAN initialized alignment, realized by Conneau et al. (2018), the initial mapping  $W$  is learned in an generative adversarial game, where a generator learns the mapping  $W$ , and a discriminator has to differentiate between  $WX$  and  $Y$ . It is then refined with iterative Procrustes. Translating requires finding the nearest neighbor  $y_t \in Y$  of a source word  $x_s \in X$ . A simple cosine similarity metric however suffers from the hubness problem where some words are Nearest Neighbors (NNs) to many other words, resulting in poor translation pairs. To address this, different metrics have been proposed (Smith et al., 2017). Most commonly the CSLS metric is used, which intuitively selects a translation pair  $(x_s, y_t)$  if  $x_s, y_t$  are close to mutual NNs. Several other GAN initialized models exist, notably Mohiuddin and Joty (2019) first maps vector spaces  $X, Y$  to latent representations  $z_X, z_Y$ , and then adversarially aligns them enforcing cycle consistency.

The above methods rely on the idea that source and target spaces are approximately isomorphic and differ in a single rotation. This implies that a Procrustes transformation would perfectly align the spaces, which has the additional benefit of preserving monolingual quality of embeddings (e.g., dot product, and cosine similarity). It is easy to see that for many language pairs the assumption does not hold. Søgaard et al. (2018) concluded that unsupervised methods perform much worse for morphologically distant languages and are sensitive to similarity of corpora and embedding algorithms. As a future direction the isomorphic assumption is dropped and recent works attempt non-linear mappings. Moshtaghi (2019) transform spaces to possess the property, while Zhang et al. (2019) enforces unit length and zero mean for each language.

Closer to our work, Sigurdsson et al. (2020) attempt unsupervised word translation grounded on the Image modality, using large sets of instructional videos and their narration. The image modality functions as a naturally aligned space between languages. In low-resource settings the model outperformed text-based methods, and showed robustness with respect to known weaknesses: dissimilarity of text corpora, vocabulary size, and language relatedness.

## 5 Proposed Approach

We know that fMRIs encode semantic information related to concrete and visual objects. Despite individual brain and cultural differences, for equivalent words as stimuli there exist similar fMRI responses across languages. This shared signal can be used to learn an initial orthogonal mapping  $W$  between embedding spaces  $X$  and  $Y$  such that  $WY \approx X$ , which will be used as an initialization to unsupervised alignment methods. Generally, these methods learn the seed mapping by exploiting topological similarities in embedding spaces. Artetxe et al. (2018b) assumes that translation pairs have approximately the same similarity distributions. The mapping is then expanded and refined using heuristics and other methods, e.g., iterative Procrustes.

We hypothesize that the bilingual signal in fMRI is strong enough to offer a good initial mapping  $W$ . We then use this mapping with known refinement techniques, and test it on common benchmarks. Because the fMRI-derived cognitive embeddings are less affected by factors like the amount of training data or vocabulary size, they may perform better than distributional embeddings in a low resource setting.

### 5.1 Data and processing

The dataset recently released in Li et al. (2021) provides fMRI data for three languages (English, French, Chinese) where native speakers listened to the same audio recording translated in their language. The audio recording contains approximately 15000 words and 1500 sentences for each language; furthermore, the fMRI data is provided pre-processed and in the same brain template for all subjects. We propose to use this data that has been specifically created for multilingual analysis as combining different fMRI datasets with different conditions can in itself be very challenging.

Voxel selection should be uniform for all participants as we want the same future space for all the data. Due to the large size of the data a good way for selecting interesting voxels is to use specific ROIs in the brain (e.g., the language network or the visual cortex). The regions related to the shared signal in the brain, are reported in studies in section 2, and significant semantic voxels are also reported by Huth et al. (2016). We use atlases for segmenting fMRI data to brain areas available in Python libraries. An alternative approach would be a simple PCA for all the data, and selecting the first  $n$  components. Selecting voxels by predictive power (e.g., pearson correlation) of predictions and actual values is might be computationally infeasible.

After voxel selection we need to match the fMRI representations to the word stimuli. Using continuous text as input (e.g., subjects listening to a book recording) introduces some unique challenges. First of all, since an fMRI image was acquired every 2 seconds but the subjects listened continuously to the audio stimuli, fMRI images are not aligned with single words and a series of words may correspond to the same image. Furthermore the effect of the word stimulus in the brain is not instantly present and can take up to 10 seconds to show in fMRI activations. We thus need a method to create a correspondence between fMRI and

sentence stimuli. We can either use contextual representations (e.g., sentence embeddings) for the sentence preceding an fMRI image, or use some function (e.g., average) of the set of word embeddings for a time window before the fMRI image acquisition.

## 5.2 Model Overview

We present two model architectures for learning a matrix  $W$  that aligns two different embedding spaces  $X, Y$  corresponding to different languages, i.e.,  $WX \approx Y$ . We introduce the following notation:

$X \in \mathbb{R}^{K \times d}$ : an embedding space with  $K$  words (vocabulary size) and dimensionality  $d$ , for a language  $L^X$ .

$Y \in \mathbb{R}^{M \times d}$ : an embedding space with  $M$  words (vocabulary size) and dimensionality  $d$ , for a language  $L^Y$ .

For the language  $L^X$ : a set of fMRI images  $N^X$  and a set of collections of words  $S^X$  such that the word set  $s_j = [w_1^j, \dots, w_l^j] \in S^X$  correspond to the fMRI image  $n_j \in N^X$ .

For the language  $L^Y$ :  $S^Y, N^Y$  analogous to above.

**1. LSTM-based.** The first model uses contextual representations from an LSTM architecture as the literature suggests LSTM representations correlate well with brain activity. We are given an embedding space  $X$  and we also possess an fMRI dataset for the language  $L^X$  of the form  $(s_j, n_j) \in (S^X, N^X)$ , where the set of words  $s_j = [w_1^j, \dots, w_l^j]$  corresponds to an fMRI image  $n^j$ . We first use an LSTM to predict the fMRI activations  $n_j$  from the corresponding word embeddings  $f^X(w_i)$ ,  $w_i \in s_j$ . To do this we feed the word embeddings sequentially to the LSTM and use a combination of hidden states  $h_i$  to predict the fMRI image. After this initial training, for a second language  $Y$  with  $(s_j, n_j) \in (S^Y, N^Y)$  we *freeze* the LSTM architecture and use embeddings  $f^Y(w_i)$  to predict fMRI activations  $n_j$ . However, instead of feeding the raw embeddings to the LSTM we first apply a linear transformation  $W$  which is meant to align the two spaces i.e.,  $WX = Y$ . Given that the fMRI spaces are sufficiently aligned in order for the LSTM model to predict the images in  $N^Y$  it will need to learn a  $W$  that aligns the embedding spaces  $X, Y$ .

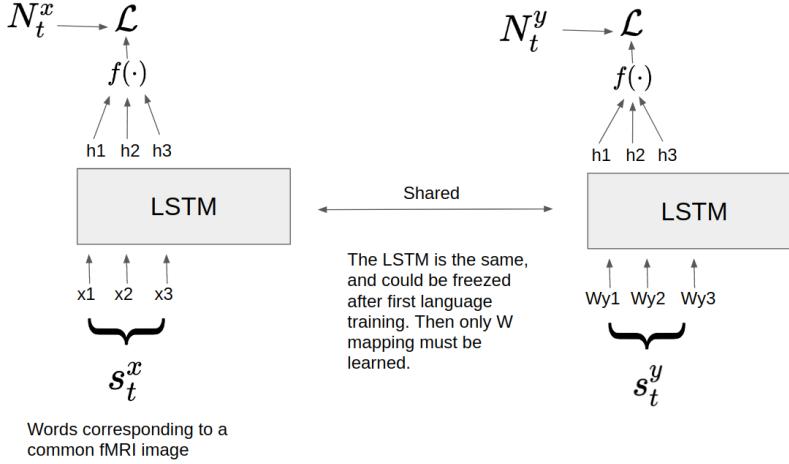


Figure 4: We learn a mapping from word embeddings  $x_i = f^X(w_i)$  to neural activations via a prediction task. The function  $f$  combines token hidden states to predict the corresponding fMRI. To align the different embedding spaces a matrix  $W$  is calculated while keeping the same encoder. If the neural spaces  $N_t^x, N_t^y$  are sufficiently aligned then the matrix  $W$  will map between spaces.

**2. Negative Sampling.** A different approach following [Sigurdsson et al. \(2020\)](#) is to use their architecture for word translation using video. Intuitively, if one possesses data of the form  $(X, N), (Y, N)$  where  $X, Y$

are embeddings spaces and  $N$  is a common space, then we can use  $N$  to find an alignment between  $X, Y$ . The authors use video as the common space, and they have data of different videos narrated in one of two languages  $L^X, L^Y$ . Instead for the common space we use fMRI activations.

For an fMRI image  $n_j$  corresponding to a set of words  $s_j$  we define the average embedding  $e_j = \sum_i f^X(w_i)/|s_j| : w_i \in s_j$ , and then create tuples  $(e_j, n_j)$ . We then learn the mapping  $e_j \rightarrow n_j$  using ridge regression. For the embeddings  $x_i$  in an embedding space  $X$  we perform *lexical expansion* to create a cognitive embedding for each word, and thus form a big dataset  $(x_i, \hat{n}_i^x)$  where  $\hat{n}_i^x$  is the predicted cognitive embeddings. We do the same for language  $L^Y$ . Now using a text encoder  $f$  and an fMRI encoder  $g$  we try to learn a space where  $f(x_i) \approx g(n_i^x) \approx g(n_j^y)$  where  $W$  is a matrix that aligns the two embedding spaces  $WX \approx Y$ . By keeping the encoders  $f$  and  $g$  the same for both languages during training the model will learn  $W$  to align the two embedding spaces. For training we sample equally from both languages  $X, Y$  and use a *negative sampling* loss: we sample matching pairs  $x_i, \hat{n}_i^x$  and random pairs  $x_i, \hat{n}_{i'}^x$  and demand that matching pairs are close in the learned space, while the random pairs are distant.

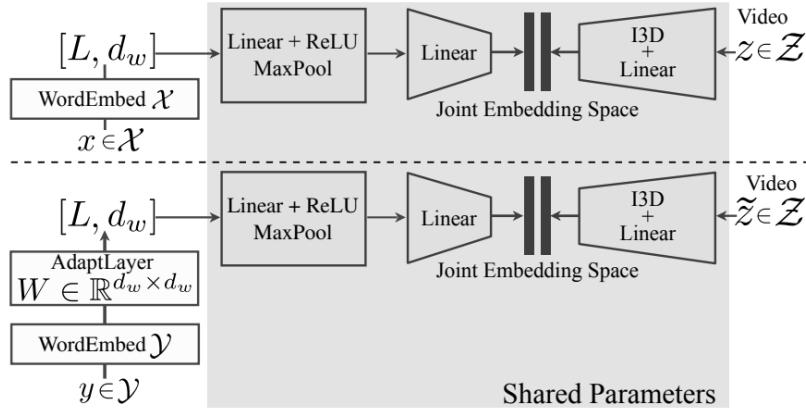


Figure 5: Matrix  $W$  aligns language  $Y$  with  $X$ . We map both languages and Neural data to the same space, using possibly non-linear layers. Then for training we use negative sampling, selecting a pair of matching word and fMRI  $(x, n)$  and a set of  $k$  non-matching pairs  $\{(x'_i, n'_i)\}_{i=1}^k$  randomly. The loss function demands that after mapping only the correct pair  $(x, n)$  is close in the shared space.

Having acquired an fMRI based mapping  $W$  we test for translation accuracy in MUSE<sup>1</sup> dictionaries of 5000 translation pairs. Afterwards we initialize known baselines Conneau et al. (2018); Artetxe et al. (2018b) with our  $W$  matrix, and continue refining with 100K most common word embeddings for each language as commonly done in literature. We compare results with baseline models, and try to demonstrate superiority of our mapping in low resource settings (reducing vocabulary and training data of embeddings). Finally qualitatively we show that predicted fMRIs for translation equivalent sentences are similar.

## References

- Samira Abnar, Rasyan Ahmed, Max Mijnheer, and Willem Zuidema. 2018. Experiential, Distributional and Dependency-based Word Embeddings have Complementary Roles in Decoding Brain Activity. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 57–66, Salt Lake City, Utah. Association for Computational Linguistics.

- Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. Blackbox meets blackbox: Rep-

<sup>1</sup><https://github.com/facebookresearch/MUSE>

resentational Similarity and Stability Analysis of Neural Language Models and Brains. *arXiv:1906.01539 [cs, q-bio]*. ArXiv: 1906.01539.

Nicolas Affolter, Beni Egressy, Damian Pascual, and Roger Wattenhofer. 2020. [Brain2word: Decoding brain activity for language generation](#).

Andrew James Anderson, Jeffrey R. Binder, Leonardo Fernandino, Colin J. Humphries, Lisa L. Conant, Mario Aguilar, Xixi Wang, Donias Doko, and Rajeev D. S. Raizada. 2016. [Predicting Neural Activity Patterns Associated with Sentences Using a Neurobiologically Motivated Model of Semantic Representation](#). *Cerebral Cortex*, 27(9):4379–4395.

João António Rodrigues, Ruben Branco, João Silva, Chakaveh Saedi, and António Branco. 2018. [Predicting Brain Activation with WordNet Embeddings](#). In *Proceedings of the Eight Workshop on Cognitive Aspects of Computational Language Learning and Processing*, pages 1–5, Melbourne. Association for Computational Linguistics.

Ekaterina Artemova, Amir Bakarov, Aleksey Artemov, Evgeny Burnaev, and Maxim Sharaev. 2020. [Data-driven models and computational tools for neurolinguistics: a language technology perspective](#).

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. [Learning bilingual word embeddings with \(almost\) no bilingual data](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *AAAI*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. [A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Nikos Athanasiou, Elias Iosif, and Alexandros Potamianos. 2018. Neural activation semantic models: Computational lexical semantic models of localized neural activations. In *Proceedings of the 27th International Conference on Computational Linguistics*.

Lisa Beinborn, Samira Abnar, and Rochelle Choenni. 2019. [Robust evaluation of language-brain encoding experiments](#). *CoRR*, abs/1904.02547.

Shohini Bhattacharji, Jonathan Brennan, Wen-Ming Luh, Berta Franzluebbers, and John Hale. 2020. [The alice datasets: fMRI & EEG observations of natural language comprehension](#). In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 120–125, Marseille, France. European Language Resources Association.

Joachim Bingel, Maria Barrett, and Anders Søgaard. 2016. [Extracting token-level signals of syntactic processing from fMRI - with an application to PoS induction](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 747–755, Berlin, Germany. Association for Computational Linguistics.

Jonathan Brennan, Edward Stabler, Sarah Wagenaen, Wen-Ming Luh, and John Hale. 2016. [Abstract linguistic structure correlates with temporal activity during naturalistic comprehension](#). *Brain and Language*, 157-158:81–94.

Augusto Buchweitz, Svetlana V. Shinkareva, Robert A. Mason, Tom M. Mitchell, and Marcel Adam Just. 2012. [Identifying bilingual semantic neural representations across languages](#). *Brain and language*, 21978845[pmid].

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. [Speaking, Seeing, Understanding: Correlating semantic models with conceptual representation in the brain](#). In *Proceedings of the 2017 Conference on*

*Empirical Methods in Natural Language Processing*, pages 1081–1091, Copenhagen, Denmark. Association for Computational Linguistics.

Lu Cao and Yue Zhang. 2019. *Investigating Lexical and Semantic Cognition by Using Neural Network to Encode and Decode Brain Imaging*.

Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. *Word translation without parallel data*. *CoRR*, abs/1710.04087.

João Correia, Elia Formisano, Giancarlo Valente, Lars Hausfeld, Bernadette Jansma, and Milene Bonte. 2014. *Brain-based translation: fmri decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe*. *The Journal of neuroscience : the official journal of the Society for Neuroscience*.

Morteza Dehghani, Reihane Boghrati, Kingson Man, Joe Hoover, Sarah I. Gimbel, Ashish Vaswani, Jason D. Zevin, Mary Helen Immordino-Yang, Andrew S. Gordon, Antonio Damasio, and Jonas T. Kaplan. 2017. *Decoding the neural representation of story meanings across languages: Decoding the neural representation*. 38(12):6096–6106.

Barry Devereux, Colin Kelly, and Anna Korhonen. 2010. Using fmri activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora. In *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, CN ’10, page 70–78, USA. Association for Computational Linguistics.

Jon Gauthier and Anna Ivanova. 2018. *Does the brain represent words? an evaluation of brain decoding studies of language understanding*.

Michael Hanke, Nico Adelhöfer, Daniel Kottke, Vittorio Iacobella, Ayan Sengupta, Falko R. Kaule, Roland Nigbur, Alexander Q. Waite, Florian Baumgartner, and Jörg Stadler. 2016. *A studyforrest extension, simultaneous fmri and eye gaze recordings during prolonged natural stimulation*. *Scientific Data*, 3(1):160092.

Michael Hanke, Florian J. Baumgartner, Pierre Ibe, Falko R. Kaule, Stefan Pollmann, Oliver Speck, Wolf Zinke, and Jörg Stadler. 2014. *A high-resolution 7-tesla fmri dataset from complex natural stimulation with an audio movie*. *Scientific Data*, 1(1):140003.

Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. 2019. *CogniVal: A Framework for Cognitive Word Embedding Evaluation*. *arXiv:1909.09001 [cs]*. ArXiv: 1909.09001.

Zhengfei Hu, Huixiang Yang, Yuxiang Yang, Shuhei Nishida, Carol Madden-Lombardi, Jocelyne Ventre-Dominey, Peter Ford Dominey, and Kenji Ogawa. 2019a. *Common neural system for sentence and picture comprehension across languages: A chinese-japanese bilingual study*.

Zhengfei Hu, Huixiang Yang, Yuxiang Yang, Shuhei Nishida, Carol Madden-Lombardi, Jocelyne Ventre-Dominey, Peter Ford Dominey, and Kenji Ogawa. 2019b. *Common neural system for sentence and picture comprehension across languages: A chinese-japanese bilingual study*. 13:380.

Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. 2016. *Natural speech reveals the semantic maps that tile human cerebral cortex*. *Nature*, 532:453–458.

Shailee Jain and Alexander G Huth. 2018. *Incorporating Context into Language Encoding Models for fMRI*. *Advances in Neural Information Processing Systems 31*, page 6628–6637.

Ahmad Jelodar, Mehrdad Alizadeh, and Shahram Khadivi. 2010. Wordnet based features for predicting brain activity associated with meanings of nouns.

Nikolaus Kriegeskorte, Rainer Goebel, and Peter Bandettini. 2006. *Information-based functional brain mapping*.

- Jixing Li, Shohini Bhattachari, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R. Nathan Spreng, Jonathan Brennan, Yiming Yang, Christophe Pallier, and John Hale. 2021. Le petit prince: A multi-lingual fmri corpus using ecological stimuli. *bioRxiv*.
- Xingyu Liu, Zonglei Zhen, Anmin Yang, Haohao Bai, and Jia Liu. 2019. A manually denoised audio-visual movie watching fmri dataset for the studyforrest project. *Scientific Data*, 6(1):295.
- Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting Similarities among Languages for Machine Translation. *arXiv:1309.4168 [cs]*. ArXiv: 1309.4168.
- T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.
- Tasnim Mohiuddin and Shafiq Joty. 2019. Revisiting adversarial autoencoder for unsupervised word translation with cycle consistency and improved training. 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)*, pages 3857–3867, Minneapolis, Minnesota. Association for Computational Linguistics.
- Masud Moshtaghi. 2019. Supervised and nonlinear alignment of two embedding spaces for dictionary induction in low resourced languages. 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)*, pages 823–832, Hong Kong, China. Association for Computational Linguistics.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J. Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications* 9.
- Peng Qian, Xipeng Qiu, and Xuanjing Huang. 2016. Bridging LSTM Architecture and the Neural Dynamics during Reading. *arXiv:1604.06635 [cs]*. ArXiv: 1604.06635.
- Jan-Mathijs Schoffelen, Robert Oostenveld, Nietzsche H. L. Lam, Julia Uddén, Annika Hultén, and Peter Hagoort. 2019. A 204-subject multimodal neuroimaging dataset to study language processing. *Scientific Data*.
- Dan Schwartz, Mariya Toneva, and Leila Wehbe. 2019. Inducing brain-relevant bias in natural language processing models. In *Advances in Neural Information Processing Systems 32*, page 14123–14133. Curran Associates, Inc.
- Ayan Sengupta, Falko R. Kaule, J. Swaroop Guntupalli, Michael B. Hoffmann, Christian Häusler, Jörg Stadler, and Michael Hanke. 2016. A studyforrest extension, retinotopic mapping and localization of higher visual areas. *Scientific Data*, 3(1):160093.
- Gunnar A. Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, João Carreira, Phil Blunsom, and Andrew Zisserman. 2020. Visual Grounding in Video for Unsupervised Word Translation. *arXiv:2003.05078 [cs]*. ArXiv: 2003.05078.
- Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *CoRR*, abs/1702.03859.
- Anders Søgaard. 2016. Evaluating word embeddings with fMRI and eye-tracking. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 116–121, Berlin, Germany. Association for Computational Linguistics.

- Sabrina Stehwien, Lena Henke, John Hale, Jonathan Brennan, and Lars Meyer. 2020. The little prince in 26 languages: Towards a multilingual neuro-cognitive corpus. page 7.
- Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. [On the Limitations of Unsupervised Bilingual Dictionary Induction](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.
- Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *NeurIPS*.
- Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014. [Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses](#). *PLoS ONE*, 9(11):e112575.
- Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. [Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation](#). In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.
- Ying Yang, Jing Wang, Cyntia Bailer, Vladimir Cherkassky, and Marcel Adam Just. 2017. [Commonalities and differences in the neural representations of english, portuguese, and mandarin sentences: When knowledge of the brain-language mappings for two languages is better than one](#). *Brain and Language*, 175:77–85.
- Mozhi Zhang, Keyulu Xu, Ken-ichi Kawarabayashi, Stefanie Jegelka, and Jordan Boyd-Graber. 2019. [Are Girls Neko or Shōjo? Cross-Lingual Alignment of Non-Isomorphic Embeddings with Iterative Normalization](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3180–3189, Florence, Italy. Association for Computational Linguistics.
- Benjamin D Zinszer, Andrew J , Olivia Kang, Thalia Wheatley, and Rajeev D S Raizada. 2016. [Semantic structural alignment of neural representational spaces enables translation between english and chinese words](#). *Journal of Cognitive Neuroscience*, 28(11):1749–1759.