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, (e.g., Pennington et al. (2014); Bojanowski et al. (2016)), 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 numerous 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 Artetxe et al. (2018b). This similarity can be used to learn a *mapping* from two (or more) languages to a shared space, with little to no supervision, 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 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 (Artetxe et al., 2019a). However, current alignment methods suffer in low-resource settings, occasionally failing to converge, and are challenged by etymologically and morphologically distant language pairs (Søgaard et al., 2018).

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 provided evidence for this claim by showing that a common signal exists across languages, despite cultural or individual speaker differences, e.g. (Buchweitz et al., 2012; Zinszer et al., 2016). This shared brain semantic space can be captured by Functional Magnetic Resonance Imaging (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. This, will help solve current alignment issues that stem from the language differences in form, morphology and other factors that are not purely semantic.

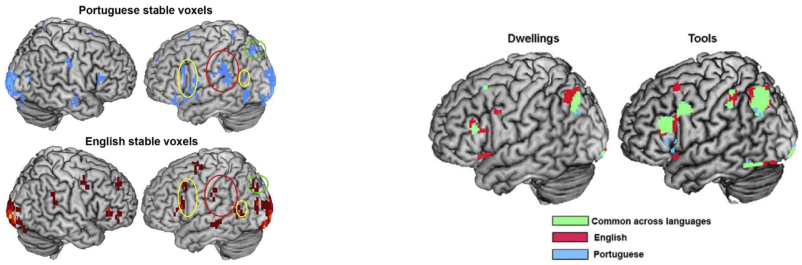
To achieve this, we propose to learn a mapping from distributional embeddings to fMRI-based word representations. This follows a recent line of work, where the semantic information present in brain scans, extracted from word-stimulus experiments, is used to improve state-of-the-art language models (Toneva and Wehbe, 2019). Combining this approach with methods for unsupervised embedding space alignment, we can then induce a bilingual dictionary using brain-derived representations. 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 methods, effectively addressing cases of low-resource and etymologically distant languages.

Apart from improving current state-of-the-art techniques in unsupervised embedding space alignment, the proposed approach could provide empirical proof for the existence of shared semantic representations across languages. Neuroscience evidence on shared brain signals is reviewed in §2; we investigate the use of brain data for NLP in §3; recent work on cross-lingual alignment is summarized in §4. Finally, we state possible approaches in §5 and summarize our conclusions in §6.

2 Neuroscientific Evidence

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 labels is trained in one language, and then tested in word identification in the other language. Above chance accuracy in this task suggests the existence of a shared meaning space across languages. Adopting a more theoretically motivated analysis, many works additionally 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.



(a) Stable Voxels (b) Areas of Brain Activation

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 due to their high concreteness and functionality or affordance.

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 in three "modalities": Images, Chinese sentences, 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 the resulting fMRI activations with the same two labels, as either coherent or incoherent. 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 representations of 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 compared to only training on a single language. Accuracy also increased, getting closer to inter-language performance (prediction with training and testing in the same language). 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.

In their study, 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, and then were asked to complete a semantic relatedness task. For each predefined Region of Interest (ROI) and for each word, they used the voxel activation vectors to compute the

correlation matrix across vectors. 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 and selecting the one with maximum similarity, the authors were able to translate between languages. In total, six ROIs 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 the existence of common semantics in the brain at both 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 insights about the function of the brain. In the following sections, we summarize related work in terms of employed datasets, voxel selection, alignment between subjects, connecting brain semantics to a distributional embedding space, and effect in NLP applications.

3.1 Datasets and Stimuli

For the use cases we are interested in, the most commonly employed neuroimaging modality is 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 high spatial resolution (e.g., 1-3 mm) and relatively low temporal resolution (e.g., 1-2 sec), requiring some care to adequately separate the time onset of different stimuli, and disentangle their effect on the brain activations. The signal is inherently noisy (e.g., due to scanning artifacts), and pre-processing is a crucial step for extracting useful semantic information.

Here we focus mostly on the text modality, although visual and auditory inputs have also been used. The stimuli can be single words, displayed one at a time or as a word cloud, or a series of sentences with a common theme. Beyond isolated input, narratives have been increasingly employed (e.g., participants reading a book). These allow studying longer contexts and naturalistic settings, but come with significant drawbacks, as it's harder to distinguish the effect of a single word or sentence. Lastly, in this review we focused on datasets that are not English-only, but involve other languages, and we categorize them depending on the stimulus type: words, sentences, and narratives.

Words The dataset introduced in Mitchell et al. (2008) contains fMRI scans from 9 participants, viewing line drawings and nouns of 60 concrete objects, corresponding to 12 semantic categories. Stimuli were shown 6 times, in a randomized order.

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 viewed 60 sentences in total, alternating between 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 and 48 pairs of corresponding captions, one 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 preprocessed 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 multiple speakers being presented the same stimulus in different languages. These brain scans, paired with their lexical stimuli, can be used for experiments in unsupervised cross-lingual alignment.

3.2 Data 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, and predominantly a gray matter mask is applied beforehand for an initial noise reduction and computational efficiency. We highlight the different approaches used for *voxel selection* below, driven by theoretical analysis or information theory measures, in a taxonomy introduced by Beinborn et al. (2019).

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*. Selecting voxels based on this prior knowledge can greatly reduce the computation needed in experiments. 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 the 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 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 low-dimensional representations.

Brain Alignment Here we discuss briefly a common preprocessing problem encountered when dealing with data from multiple subjects. Usually, data is combined across participants in order to reduce noise, and compute a more robust (and statistically significant) shared response representation. In practice this can be quite challenging, as each subject's activations belong to a separate voxel space (e.g., with different sizes), and an alignment of the different spaces must be performed. The simplest approach is *anatomical alignment*, which aligns voxel spaces to a common template using anatomical features from structural MRI. However, shape, size, and spatial location of functional areas might differ across subjects. This motivates *functional alignment*, which aligns brain spaces such that correlation of activations across participants for the same stimuli is maximized. Most commonly, Hyperalignment methods are used Haxby et al. (2011), featuring simple geometrical transformations that preserve shape. Lastly, some works opt for no alignment, using a simple average of activations across subjects, a selection of the best performing subjects, or a concatenation of low dimensional representations across subjects. While a detailed analysis is out of the scope of our work, we note that care and expertise are required to effectively combine representations across multiple subjects.

3.3 Mapping from Brain to Lexical Space

After selecting informative voxels based on theoretical priors or information-driven metrics, some mapping needs to be established between the voxel space, and a traditional semantic embedding space. Importantly, this mapping will allow the generalization of the brain-derived representations to new words that are not part of the original stimuli. Most commonly brain scan datasets tend to be small, due to experimental cost and privacy concerns. For this reason, most works employ simple linear mappings $y = Wx$ with some regularization, avoiding complex neural networks with big parameter counts. Importantly, as Artemova et al. (2020) mention, each area (represented by a voxel) responds largely independently of other areas, thus a separate model is needed to fit responses in each cortical voxel. We describe the mapping used by Mitchell et al. (2008), as a baseline that informs subsequent work. In this formulation, the activation y_v of a voxel v given a word stimulus w, is given by the following formula:

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

For stimuli representation, the authors use the similarity with 25 seed verbs manually selected with respect to psycholinguistic criteria, leading to a representation space $|F| = 25$. The similarity $f_i(w)$ between seed word i and word w is calculated from co-occurrence statistics in a large corpus. The total number of voxels are $|V|$, and the learned weights that are estimated via regression $c_{v,i}$ form a matrix $W \in \mathbb{R}^{|V| \times |F|}$.

These representations are used by Athanasiou et al. (2018) to measure the utility of neural versus traditional embeddings in downstream NLP tasks, e.g., natural language inference with the SNLI dataset Bowman et al. (2015). In a similar approach, some authors Anderson et al. (2016) have used a similarity encoding, where the activation for an unknown word w is computed as a sum of activations of known reference words u_i, weighted by their similarity $sim(u_i, w)$, e.g., as captured by a lexical embedding space.

Several subsequent works have analyzed the mapping methods introduced by Mitchell et al. (2008). Work by Devereux et al. (2010) reports that automatically choosing a set of verbs for feature representation 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, where ratings for each semantic dimension were crowdsourced. 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-occurence 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 Pennington et al. (2014) 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 also 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.

Natural Narratives There has been growing interest in using more natural stimuli in experiments, moving beyond isolated words in a controlled environment, to subjects reading a text (e.g., a book) in naturalistic settings while a machine records their brain patterns. Due to the low temporal resolution of fMRI, this makes it harder to distinguish the individual contributions of words in the brain images. To address this, Wehbe et al. (2014) use an input representation of 195 features for each word and try to predict the brain signal from a sum of the representations of all corresponding words in a 4-word window. Bingel et al. (2016) address the low temporal resolution problem by assigning the same brain representation to all corresponding words shown during an interval around the scan acquisition time, and then slide a Gaussian window across tokens. They use the resulting representations with a Hidden Markov Model, to improve performance in part of speech (POS) prediction. Schwartz et al. (2019) use BERT (Devlin et al., 2018), a large transformer model, to predict the neural image from the sentence the participant before brain scanning. Long Short-Term Memory (LSTM) models have also been used to map between word embeddings and neural activations. Abnar et al. (2019); Qian et al. (2016), conclude that LSTM hidden-state 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 brain responses. They notice that LSTMs encode context well and are good at predicting activations of individual words in a sentence.

3.4 Improving NLP Downstream Performance

Following common methods for transfer learning in NLP, brain-derived representations could be used to increase performance of task-specific models. Here we assume that for the set of words (or sentences) that are included in the evaluation task, we possess neural representation that might be derived from voxel activations as described in the previous sections. The simplest method of using these embeddings for increased performance is as **extra features**, that can be fused with features the model already uses (e.g., traditional word embeddings). A second approach involves learning task-specific brain representations. As seen previously, learning neural embeddings commonly involves finding a mapping W from a voxel to a dense word space. We can find this mapping by engaging in **end-to-end finetuning** at the evaluation task, selecting the voxels that are most beneficial for the specific lexical evaluation. Lastly, we can use the task of **predicting brain representations** to optimize a language model, or align its internals to reflect brain semantics. This has been shown to lead to general language task improvement.

There is little prior work that evaluates or improves NLP models through brain data. Bingel et al. (2016) use neural features, acquired from participants reading sentences, combined with text features to increase per-

formance in a POS tagging task. 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, eye tracking data, and electroencephalograms (EEG). They report correlation between the cognitive evaluation and performance in Named-Entity Recognition 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, in order to differentiate between the amount of context maintained by different brain regions. 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 at various language tasks. This finding suggests that altering a language model to better align its outputs with brain recordings of people processing language, may lead to better language understanding overall.

Affolter et al. (2020) create a decoder, mapping from fMRI activations to words, and evaluate it in direct word classification (predicting word indexes). The decoder is trained as an autoencoder using sentence data. With this setting, they achieve up to 15 (top-5) and 5.2 (top-1) accuracy. However, they find that predicting dense word embeddings, instead of word labels, degrades model performance. They use data from different subjects from the dataset of Pereira et al. (2018), incorporating a subject separation term in the loss function. Finally, using an autoregressive language model, together with their trained decoder, they attempt neurally-guided word generation.

4 Cross-Lingual Space Alignment

Cross lingual alignment consists in learning a shared word vector space, where for two languages, words with similar meanings obtain close-distanced vectors. Given two monolingual embedding spaces X, Y this is usually formulated as a mapping W where $WX \approx Y$. The X space corresponds to the *source* language (to be mapped), and the Y space to the target language. The mapping W serves to map the source to the target language. 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 an 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[1] transformations $\prod_i W_{(i)}X$, with intermediate normalization steps. Relaxing the strong assumptions of a fully supervised setting, related work attempts a semi-supervised mapping that relies only on small parallel corpora. Smith et al. (2017) relied on common strings and Artetxe et al. (2017) only on shared numeral representations, and achieved comparable results to supervised methods. This opened the way for fully unsupervised approaches.

Unsupervised methods, without any parallel signal, differ from each other in how they build an initial bilingual dictionary, which is subsequently refined through various transformations. This dictionary is not given as input like in supervised settings, but inferred from the monolingual data using geometric properties of their embedding spaces. 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.

In the GAN initialized alignment, proposed by Conneau et al. (2018), the initial mapping W is learned in a generative adversarial game. A generator learns a mapping W to map the source language to realistic vectors

[1]The Procrustes transformation is a geometric transformation involving translation, rotation, or uniform scaling, acting on the size, position and orientation of an embedding space, but leaving its shape unaffected.

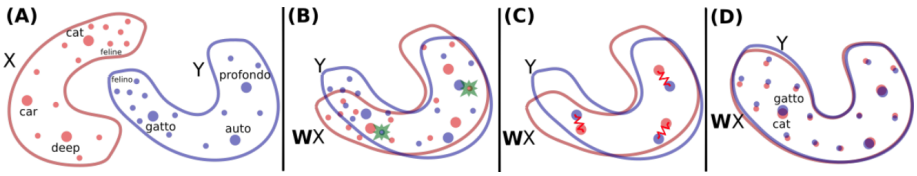


Figure 2: General steps for Unsupervised alignment: **(A)** The monolingual embedding spaces we want to align **(B)** An initial seed dictionary D or mapping W is constructed by matching words with similar context distributions **(C)** The mapping W is further refined in a pseudo-supervised (possibly iterative) step **(D)** We translate using the mapping W and by computing nearest neighbors in the shared aligned space. Image reproduced from Conneau et al. (2018).

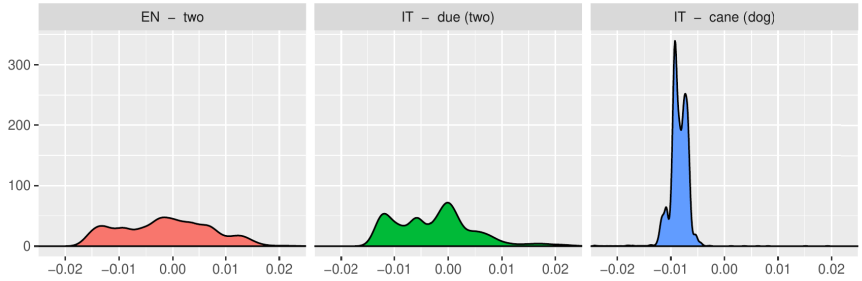


Figure 3: Motivating example for the unsupervised cross-lingual alignment problem, from Artetxe et al. (2018b). The similarity distributions between an English and two Italian words are depicted, demonstrating that equivalent translations, two and due (two), have more similar distributions than unrelated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

in the target language, and a discriminator has to differentiate between fake WX and real vectors Y. It is then refined with iterative Procrustes. Translating requires finding the nearest neighbor $y_t \in Y$ of a source word $x_s \in X$, using a suitable distance metric (e.g., cosine similarity). However, a simple cosine similarity metric suffers from the hubness proble, where some words are Nearest Neighbors (NNs) to many other words, resulting in poor translation pairs. To address this, many alternative metrics have been proposed. Most commonly, the CSLS metric (Smith et al., 2017) 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.

In later work by Artetxe et al. (2018b), the seed dictionary is derived by aligning word similarity matrices XX^T, YY^T assuming that if the embeddings spaces are isomorphic, then x_s and its translation pair y_t should have equal similarity distributions. The initial mapping is refined with a two-step iterative algorithm, and strong heuristics. Later authors Hartmann et al. (2019) have argued that the initialization is poor compared to GANs. Cai et al. (2021), introduce stochastic dictionary induction, where candidate word pairs are randomly dropped from the similarity matrix before dictionary induction to avoid poor local optima. However, the model by Artexte remains a competitive and robust baseline Vulić et al. (2019).

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 this property, while Zhang et al. (2019) enforces unit length and zero mean for each language.

The need for fully unsupervised methods is questioned by Vulić et al. (2019), as they usually fail to converge and are surpassed by low-supervised methods (e.g., with 1000 parallel words) using the same refinement procedures as a fair comparison. Fujinuma et al. (2019) propose a new intrinsic evaluation metric that is based on graph clustering and is language independent, while Glavaš et al. (2019) review notable caveats in the evaluation of models in literature. Other works (Ormazabal et al., 2019) compare monolingual space alignment, with directly learning multilingual representations by optimizing a monolingual prediction task and a cross-lingual objective (joint training). Wang et al. (2020) propose a two-step unified framework that uses unsupervised joint training as initialization and alignment as refinement, achieving competitive results.

Lastly, closer to our proposed approach, Sigurdsson et al. (2020) attempt unsupervised word translation grounded on images, using large sets of instructional videos and their narration. In this setting, the image modality functions as a naturally aligned space between languages; much like we hypothesize brain activations would. 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 Unsupervised Alignment with Brain Semantics

The previous sections have demonstrated evidence that fMRI signals encode many aspects of semantics, especially in relation to concrete and visual concepts. Additionally, despite differences in brain representations (related to individual or cultural factors), there exists some degree of shared neural response for participants across languages. In this section, we will demonstrate how this common signal can be used in cross-lingual alignment, to learn an initial orthogonal mapping W between embedding spaces X and Y for two languages, such that $WX \approx Y$. This mapping W can then function as an *initialization* to unsupervised embedding space alignment methods from the literature (§4), such as those by Conneau et al. (2018) and Artetxe et al. (2019b), to create a bilingual dictionary.

We hypothesize that the bilingual signal in neural responses is strong enough to result in an informative initial mapping W. Additionally, because the fMRI-derived neural word embeddings are less affected by factors like the amount of training data or vocabulary size, for low-resource setting this mapping may perform better than one relying solely on distributional embeddings. Since neural representation primarily hold *semantic* information largely unrelated to form and morphology, the use of brain signals in learning the mapping W may additionally result in improved performance for distant language pairs.

5.1 Data and processing

Li et al. (2021) recently introduced "The Little Prince" dataset, which provides fMRI data for three languages (English, French, Chinese) where native speakers listened to the same book audio recording translated in their language. The audio contains approximately 15000 words and 1500 sentences for each language. Furthermore, the neural data is provided pre-processed and is aligned to the same brain template for all subjects. We choose this data for our experimental approach, as it has been specifically created for *multilingual analysis*. In general, combining different fMRI datasets with diverse conditions and participants can in itself be very challenging, with the cross-lingual factor only adding to the difficulty.

Voxel selection should be uniform for all participants as we want the same feature space across languages and subjects. Due to the large data size, a good way for selecting informative 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 and the methods for extracting significant semantic voxels, are reported in studies in §2. In practice, we will use atlases for segmenting fMRI data to brain areas, readily available in Python libraries. An alternative approach is to apply PCA across data, and to then select the components with the most explained variance.

Lastly, selecting voxels using predictive power of a model, e.g., by the Pearson correlation between actual and predicted voxel values, is done by some works but might be computationally infeasible for our setting.

After voxel selection, we need to match the neural representations to the word stimuli. Using continuous text as input (e.g., subjects listening to a book recording) introduces some unique challenges. First, since in the dataset of Li et al. (2021) an fMRI image was acquired every two seconds while the subjects listened continuously to the audio stimuli, neural images are not aligned with single words or sentences, and a series of words *across sentences* may correspond to the same brain recording. Furthermore, the effect of the word stimulus in the brain is not instantly present and can take up to ten seconds to be recorded in the fMRI activations (known as, the *haemodynamic response*). We thus need a method to create a correspondence between brain representations and sentence stimuli. To achieve this, we can either use contextual representations (i.e., sentence embeddings) for a sentence preceding a brain image, or an average (or some other function) of the embeddings for words in a time window near the brain image acquisition.

5.2 Model Overview

Here, we present two different model architectures and corresponding experimental approaches, for learning an initial alignment between the vector spaces of two different languages X and Y. In particular, we learn a matrix W such that $WX \approx Y$. We introduce the following notation:

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

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

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

For language Y: $S^{\mathrm{Y}}, N^{\mathrm{Y}}$ analogous to above.

1. LSTM-based Approach Our first proposed model uses contextual representations from a Long Short-Term Memory (LSTM) architecture. This selection was motivated by the neural NLP literature (§3.3), which suggests that LSTM representations correlate well with brain activity. Importantly, the LSTM model could be first trained as a *language model*, in the source language X through traditional corpora, but with each word being an input feature (i.e., without subword tokenization). This way, the fMRI representations can augment the distributional information present in the embedding space X, already learned by the model.

We are given an embedding space X and we also possess an fMRI dataset for the language X of the form $(s_j, n_j) \in (S^X, N^X)$, where the set of words $s_j = [w_1^j, ..., w_l^j]$ corresponds to an fMRI image n^j. We first use an LSTM to predict the fMRI activations n_j from the corresponding word embeddings $x_i \in X : w_i \in s_j$. 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 $y_i \in Y : w_i \in s_j$ 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 the model will learn a mapping W that aligns the embedding spaces X, Y. This approach is illustrated in Figure 4.

2. Negative Sampling A different approach following Sigurdsson et al. (2020) is to adopt their existing methodology for word translation using videos, to translate using neural signals. 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 N can be used to find an alignment between X, Y. The authors use *video* as the share grounding space, and employ data of videos narrated in one of two languages X and Y. We adapt this method to our setting, by grounding alignment in a common brain activation (fMRI) space. The alignment is done via *contrastive learning*.

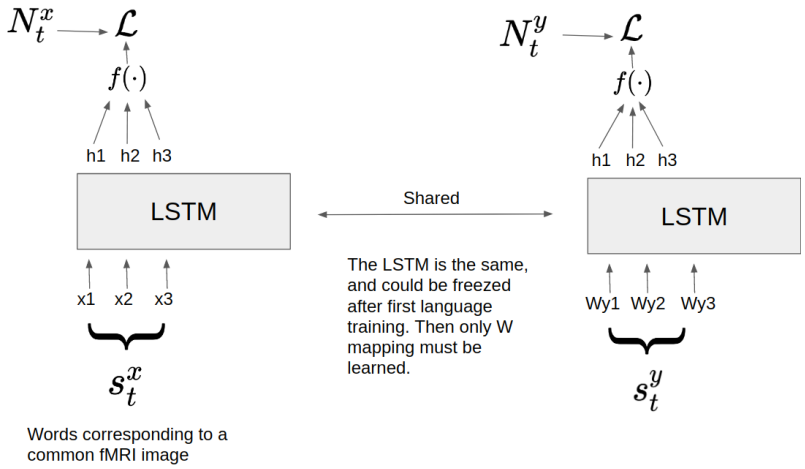


Figure 4: LSTM-based approach to neural word alignment. We learn a mapping from word embeddings $x_i \in X$ to neural activations via a prediction task. Function f combines word hidden states to predict the fMRI image. To align the different embedding spaces, a matrix W is learned, "freezing" other modules.

For an fMRI image n_j corresponding to a set of words s_j we define the average embedding $e_j = \sum_i f^X(w_i)/|s_j|$ for all $w_i \in s_j$, and then create tuples (e_j, n_j). We then learn the mapping $e_j \rightarrow n_j$ using simple regression. We use this mapping to compute neural embeddings for every word $x_i \in X$,v thus forming a big dataset $(x_i, \hat{n}_i^x)$ where $\hat{n}_i^x$ is the predicted neural embedding. We do the same for language Y. Now, using neural network encoders f for text and g for brain activations, we try to learn a space where the two modalities (text and brain activations) are aligned, i.e.: $f(x_i) \approx g(n_i^x)$ and $f(Wy_j) \approx g(n_j^y)$ 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. We train with *contrastive learning*, sampling 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 enforce that matching pairs are close in the learned space, while the random pairs are distant. This approach is illustrated in Figure 5.

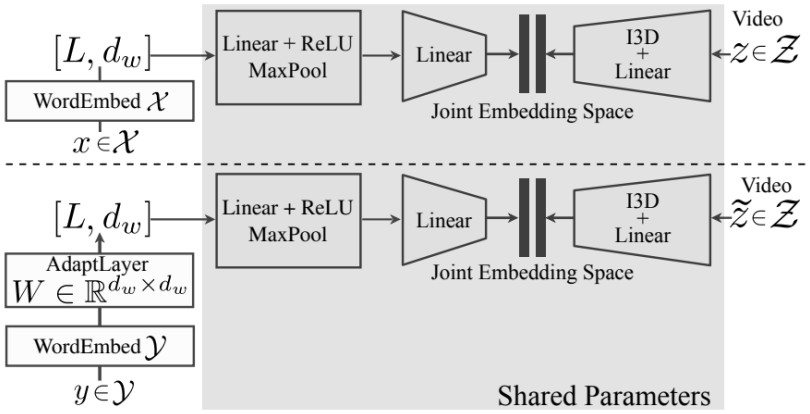


Figure 5: Negative sampling approach to neural word alignment. A matrix W is learned to align languages X and Y. We map text embeddings and neural representations to the same space using two encoders. For training, we employ negative sampling, using a pair of matching words and fMRI image (x, n) and a set of random non-matching pair. The loss function enforces that only the correct pair is close in the shared space.

Evaluation Having acquired an initial mapping W based on neural representations, we will measure the success of the alignment using translation performance in the standard evaluation benchmark MUSE[2], containing dictionaries of 5000 translation pairs for a variety of languages. Afterward, we will initialize known baselines Conneau et al. (2018); Artetxe et al. (2018b) with our W matrix, and continue refining with the 100K most common word embeddings for each language, as commonly done in literature. We then compare results with baseline models (defined above), in order to demonstrate the benefits of our mapping in low resource settings (reducing data needs), and for distant language pairs. Finally, we will qualitatively show that predicted fMRIs for equivalent sentences or words across languages are similar.

6 Conclusion

In this document, we have conducted a comprehensive literature review across two diverse research fields, Neuroscience and Natural Language Processing (NLP), that have been increasingly coming together in interdisciplinary work. With this work, we aim to provide a convincing view that (a) the neuroscientific literature suggests that similar brain activations exist across languages, and (b) this shared cross-lingual semantic space can be used to align two embedding space for different languages. Towards these goals, we have analyzed evidence on shared brain signals (§2), use of brain data for NLP (§3) and recent work on cross-lingual alignment (§4). Finally, we have concluded with two possible computational approaches using neural networks, one based on LSTM models and one on negative sampling, that might enable neural embeddings to be used for cross-lingual alignment, and subsequent bilingual dictionary induction. Our vision is that brain representations in this setting will improve prior methods for low-resource languages and distant language pairs, and additionally provide empirical evidence for the existence of shared semantic spaces in the brain.

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: Representational 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

[2] https://github.com/facebookresearch/MUSE

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.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019a. Bilingual lexicon induction through unsupervised machine translation. *CoRR*, abs/1907.10761.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019b. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy. 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 Bhattasali, 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.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Jonathan Brennan, Edward Stabler, Sarah Wagenen, 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.

Zhipeng Cai, Zuobin Xiong, Honghui Xu, Peng Wang, Wei Li, and Yi-Lun Pan. 2021. Generative adversarial networks. *ACM Computing Surveys (CSUR)*, 54:1 – 38.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Yoshinari Fujinuma, Jordan Boyd-Graber, and Michael J. Paul. 2019. A Resource-Free Evaluation Metric for Cross-Lingual Word Embeddings Based on Graph Modularity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4952–4962, Florence, Italy. Association for Computational Linguistics.

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

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy. Association for Computational Linguistics.

Michael Hanke, Nico Adelhöfer, Daniel Kottke, Vittorio Iacovella, 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.

Mareike Hartmann, Yova Kementchedjhieva, and Anders Sø gaard. 2019. Comparing unsupervised word translation methods step by step. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 6033–6043. Curran Associates, Inc.

James V. Haxby, J. Swaroop Guntupalli, Andrew C. Connolly, Yaroslav O. Halchenko, Bryan R. Conroy, M. Ida Gobbini, Michael Hanke, and Peter J. Ramadge. 2011. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416. PMC3201764[pmcid].

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 Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R. Nathan Spreng, Jonathan Brennan, Yiming Yang, Christophe Pallier, and John Hale. 2021. Le petit prince: A multilingual 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.

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the Limitations of Cross-lingual Word Embedding Mappings. *arXiv:1906.05407 [cs]*. ArXiv: 1906.05407.

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*.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do We Really Need Fully Unsupervised Cross-Lingual Embeddings? 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 4407–4418, Hong Kong, China. Association for Computational Linguistics.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime Carbonell. 2020. CROSS-LINGUAL ALIGNMENT VS JOINT TRAINING: A COMPARATIVE STUDY AND A SIMPLE UNIFIED FRAMEWORK.

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.