NeuroImage 146 (2017) 658–666

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

# Commonality of neural representations of sentences across languages: Predicting brain activation during Portuguese sentence comprehension using an English-based model of brain function

Ying Yang[a], Jing Wang[a], Cyntia Bailer[b], Vladimir Cherkassky[a], Marcel Adam Just[a,*]

[a] Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA
[b] Department of Foreign Language and Literature, Federal University of Santa Catarina, Brazil

ARTICLE INFO

*Keywords:*
Cross-language commonality
Sentence representations in bilinguals
Predictive modeling of sentence representations
Meta-language brain locations in sentence processing

ABSTRACT

The aim of the study was to test the cross-language generative capability of a model that predicts neural activation patterns evoked by sentence reading, based on a semantic characterization of the sentence. In a previous study on English monolingual speakers (Wang et al., submitted), a computational model performed a mapping from a set of 42 concept-level semantic features (Neurally Plausible Semantic Features, NPSFs) as well as 6 thematic role markers to neural activation patterns (assessed with fMRI), to predict activation levels in a network of brain locations. The model used two types of information gained from the English-based fMRI data to predict the activation for individual sentences in Portuguese. First, it used the mapping weights from NPSFs to voxel activation levels derived from the model for English reading. Second, the brain locations for which the activation levels were predicted were derived from a factor analysis of the brain activation patterns during English reading. These meta-language locations were defined by the clusters of voxels with high loadings on each of the four main dimensions (factors), namely *people*, *places*, *actions* and *feelings*, underlying the neural representations of the stimulus sentences.

This cross-language model succeeded in predicting the brain activation patterns associated with the reading of 60 individual Portuguese sentences that were entirely new to the model, attaining accuracies reliably above chance level. The prediction accuracy was not affected by whether the Portuguese speaker was monolingual or Portuguese-English bilingual. The model's confusion errors indicated an accurate capture of the events or states described in the sentence at a conceptual level. Overall, the cross-language predictive capability of the model demonstrates the neural commonality between speakers of different languages in the representations of everyday events and states, and provides an initial characterization of the common meta-language neural basis.

## 1. Introduction

### 1.1. Exploring the commonality of neural representations of sentences across languages

One of the new insights emerging about human brain function since the advent of fMRI is that individual concepts have identifiable neural signatures (Mitchell et al., 2008), and furthermore, that there is a high degree of commonality of such signatures across people (Just et al., 2010). Particularly germane to this study are previous investigations of the commonality of neural representations of concepts across different languages. For example, Buchweitz et al. (2012) demonstrated the commonality of the neural representations of 14 concrete objects (7 tools and 7 dwellings) across English and Portuguese. More recently Correia et al. (2014) demonstrated the commonality of the neural representations of 7 concrete objects (4 animals and 4 inanimate objects) across Dutch and English, while Zinszer et al. (2016) did so for 8 concrete objects across Mandarin Chinese and English. At least at the level of individual common concrete lexical items, the neural representations are to a large degree common across languages.

The goal of the current study was to assess the commonality across two languages of the neural representation of *sentences*, using a much larger vocabulary, and at the same time increasing the granularity of the scientific account of the phenomenon. The study developed a predictive model that learns the mediated mapping between semantic features of 96 word concepts (content words) and the resulting activation pattern of 60 sentences composed from these words in one language, and predicts the activation pattern of a new sentence

* Corresponding author.
*E-mail address:* just@cmu.edu (M.A. Just).

http://dx.doi.org/10.1016/j.neuroimage.2016.10.029
Received 13 May 2016; Accepted 18 October 2016
Available online 19 October 2016

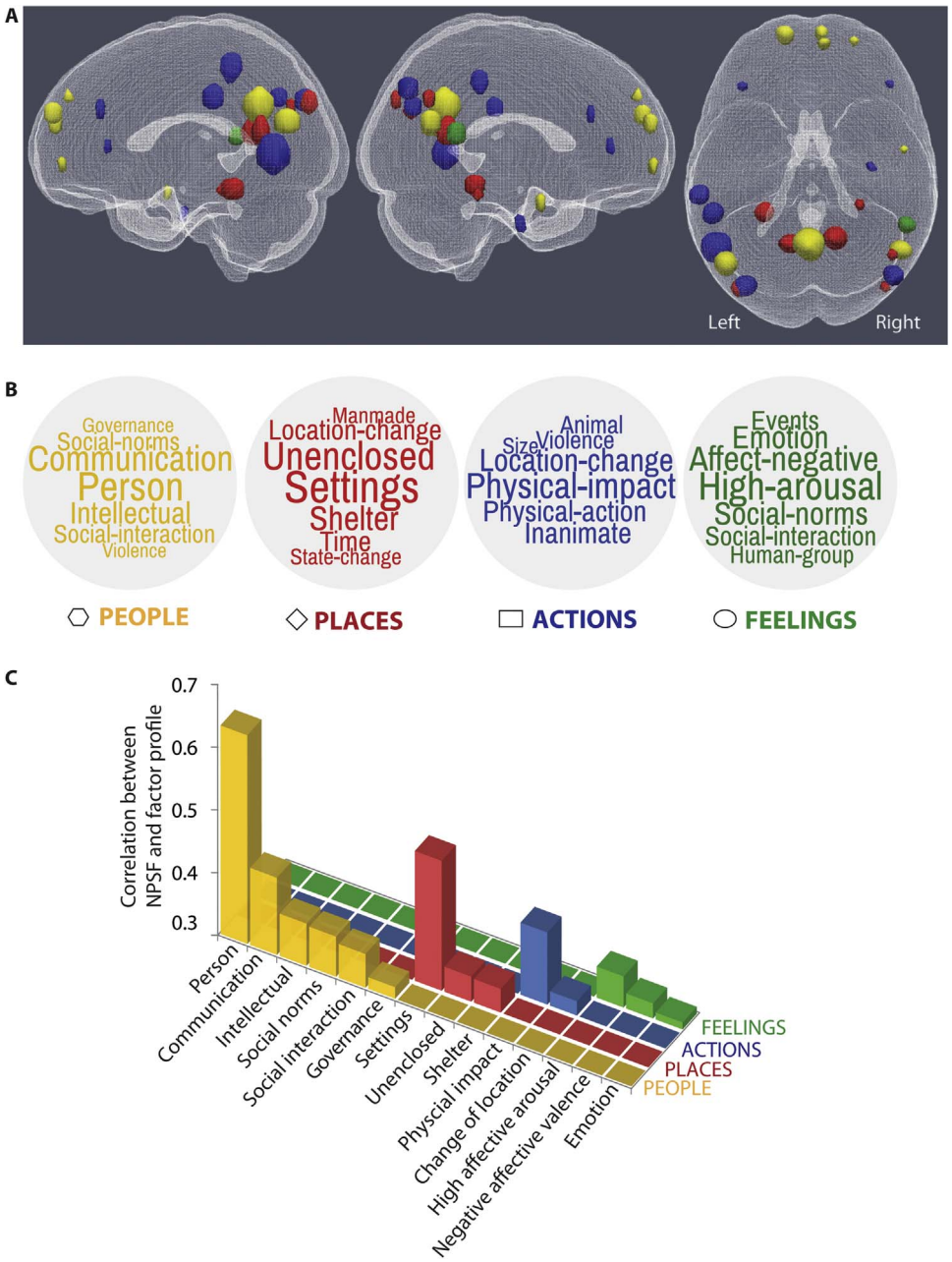


**Fig. 1.** Illustration of the mappings between neural activation patterns and semantic representations. (A) Brain regions associated with the four semantic factors: **people** (yellow), **places** (red), **actions** (blue) and **feelings** (green). (B) Word clouds associated with each factor. The clouds are formed using the 7 NPSFs most associated with each factor to illustrate the meaning components of each factor. (C) NPSFs that correlate with at least one factor with $r > 0.3$ ($p < 0.0001$). The pairwise correlations are computed between each NPSF's coding over the individual words and the factor scores over the same words. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

composed of new words (new to the model) in another language. (The 96 concepts consist of 58 nouns, 23 verbs, and 15 adjectives. The sentences each contain a mean of 3.2 content words).

Two very recent advances in the areas of neural modeling and brain reading provide the foundation for the current work. First, it has been possible to develop predictive models, rather than merely discriminative models, of the neural representations of concepts (Just et al., 2010; Mitchell et al., 2008). Discriminative models simply provide a mapping between stimulus items and brain activation patterns. Predictive or generative models, on the other hand, specify the principles or intervening variables that mediate this mapping, making it possible to predict the activation pattern for a new item. Thus the current study starts with a mapping between the semantic properties of word concepts and their neural representations developed from the data of English speakers reading English sentences, and then uses this mapping to predict the neural representation of a *new* word concept (new to the model) in Portuguese.

The second advance is that brain reading studies have progressed from decoding individual concepts from their fMRI signature to decoding entire sentences and narratives using predictive models (Huth et al., 2016; Wang et al., submitted). The neural representation of a sentence is construed here as the sum of the neural representations of its component content words, plus these words' thematic roles in the sentence. Taking the word concepts' thematic role in a given sentence into account characterizes some of the sentence-level meaning above the level of individual words. This construal is still an oversimplification of the nature of sentence meaning, which can additionally contain meaning elements that emerge from the contextual interaction of the

component words in a given syntactic configuration.

### 1.2. A generative mapping between word concepts and fMRI activation patterns

The type of generative mapping between word concepts and fMRI activation patterns used here requires a mediating layer of semantic elements that characterize word concepts. The semantic characterization of concepts was a set of 42 Neurally Plausible Semantic Features (NPSFs) that had been previously developed (Wang et al., submitted) to code the meanings of 242 content word concepts in 240 English stimulus sentences, 195 of which described an everyday event (e.g. *The woman left the restaurant after the storm*) and 45 of which described a state (e.g. *The flower was yellow*). These 240 sentences constitute a superset of the 60 sentences in the current study.

These NPSFs are hypothesized to encode semantic features that are common across word concepts, and have been shown in previous research to have neural bases that are common across people. For example, previous neuroimaging studies have found the neural bases of NPSFs such as *animals* (e.g. Martin et al., 1996), *concrete objects* (Just et al., 2010), *social interactions* (e.g. Just et al., 2014; Rilling et al., 2004; Schilbach, 2015; Schilbach et al., 2006; Van der Cruyssen et al., 2015), *shelter* (e.g. Huth et al., 2012; Just et al., 2010; Rustandi et al., 2009), *tools* (e.g. Johnson-Frey, 2004; Martin et al., 1996; Tranel et al., 2003), *eating/drinking* (e.g. Giuliani et al., 2014; Van der Laan et al., 2011), *emotions* (Kassam et al., 2013), and so on. (See Table S.1. in the Supplemental materials for a complete list of the NPSFs and the coding of some sample concepts).

Furthermore, the activation patterns corresponding to some of these NPSFs are largely similar across speakers of different languages (e.g. Zinszer et al., 2016) and among bilinguals and monolinguals (e.g. Kovelman et al., 2008; Palomar-García et al., 2015). Thus, NPSFs are hypothesized to be implemented at the level of the "language of thought" (Fodor and Pylyshyn, 1988; Marcus et al., 2014). We used these 42 NPSFs, developed to code the semantic properties of English words, testing their ability to generate accurate predictions concerning the neural representations of words in Portuguese. Notably, when the model generates the predicted activation pattern of a given Portuguese word, the model's training set from the English data excludes any information about the activation of the English translation equivalent of that Portuguese word. The prediction instead is based on the NPSFs of the Portuguese word, and how the NPSFs were related to activation patterns as they occurred in other words.

This modeling approach requires a specification of the brain areas where the mapping between NPSFs and activation patterns is implemented. These locations were derived from a factor analysis of the fMRI data of three English monolingual speakers in the previous study of 240 English sentences whose neural representations were particularly identifiable and similar to each other. More specifically, hierarchical factor analyses were applied to the datasets from these three English monolingual speakers to reduce the dimensionality of their data, uncovering the shared underlying semantic dimensions at a coarser level than NPSFs, and localizing each of these dimensions to a set of brain locations (implemented as voxel clusters, with locations shown in Table S.2. in the Supplemental materials). The factor analyses yielded four such dimensions and their associated brain locations, as illustrated in Fig. 1A and B. Specifically, the main underlying dimensions can be characterized as: (1) people; (2) places; (3) actions; and (4) feelings. These four labels each refer to a broad set of concepts, such as people referring to *social interactions*, *human knowledge*, *communication*, etc., some of which are indicated in the word clouds in Fig. 1B. A set of 2–15 brain locations was associated with each of the four underlying dimensions (clusters larger than 10 voxels associated with each factor are shown in Fig. 1A).

The correlation between NPSFs and these four basic dimensions can be assessed by relating the profile of a given factor's scores over individual stimulus words to the NPSF coding profile over these words. For example, the NPSF *communication* was associated with the factor people, as they both showed high scores for concepts such as *negotiate* and *speak*. Another example comes from the factor of places: the words *restaurant*, *hospital* and *car* all had high scores on this factor, and these words were coded with the NPSF *shelter*. Therefore, the NPSF *shelter* is correlated with the neural dimension of place, as indicated in Fig. 1C. Specifically, the 8 brain locations shown in yellow in Fig. 1A correspond to the people dimension in Fig. 1B, which is correlated with NPSFs like *communication* in 1C, and the 8 brain locations shown in red correspond to the *place* dimension in Fig. 1B, which is correlated with NPSFs like *shelter* in 1C. In sum, the factor analysis indicates the basic underlying dimensions, and the locations of voxel clusters with high loadings on these factors, and these locations are then used for mapping between NPSFs and activation levels in Portuguese.

### 1.3. A meta-language sentence prediction model

If the mapping between semantics and brain activation indeed has commonality across languages, then a predictive model should be able to learn a mapping between the semantic characterization and activation patterns in one language (English, in this case), and predict the activation patterns in another language, namely Portuguese.

To test this hypothesis, 60 arbitrarily selected sentences from the set of the 240 English sentences were translated into Brazilian Portuguese by two native speakers, to be used as stimuli for Portuguese speakers.

The mapping between NPSFs and activation in a given voxel location, expressed as model weights, were learned from the data of seven English monolingual speakers. The model weights computationally defined the mapping from NPSFs (and the thematic roles) of the content words in the sentences to the fMRI-measured neural activation in the factor-related locations (Wang et al., submitted). In the current study, these weights were used to predict the neural activation patterns of *new words* in *new* sentences as read by Portuguese speakers. Then, the predicted activation patterns of each of the individual content words of the sentence were added to produce predicted activation patterns of the entire sentence. This procedure has generated highly accurate predictions in the previous sentence decoding experiment on seven English monolingual speakers (mean rank sentence prediction accuracy=0.82, critical level at $p < 0.05$=0.54. obtained with random permutation testing).

The stimulus sentences in this study described everyday, concrete events and objects (as shown in Table S.3) making them unsuitable for addressing issues of cultural or environmental influences on neural representations of concepts and sentences. Cultural effects on neural activation patterns have been reported in several domains that intuitively seem sensitive to culture, such as self-representation (Zhu et al., 2007). Any conclusions regarding cross-language commonality based on the materials of the current study will be limited to sentences that describe relatively culture-free events and objects.

### 1.4. Hypotheses

Using the same brain locations, NPSFs, and trained model weights developed in the previous English sentence study (Wang et al., submitted), the following hypotheses were tested.

The main hypothesis is that the mapping between the sentence characterizing NPSFs/thematic roles and activation patterns in specific brain locations in English is above the level of an individual language and should predict the activation patterns associated with the reading of individual Portuguese sentences. This hypothesis also entails that there is a commonality across people, given that there is no overlap between the participants in the study of English reading and Portuguese reading.

Second, the cross-language prediction accuracy should be similar in

bilingual and monolingual participants, because the model is constructed at a conceptual level common between languages. Even though the model is based on data from English speakers, knowledge of English should not be relevant to prediction accuracy.

Additionally, the model should capture the mapping between activation and the gist of the sentence, rather than any superficial properties of the sentences. Thus the model's highly-ranked but incorrect sentence guesses should resemble the correct sentence in terms of the events or states they describe.

## 2. Material and methods

### 2.1. Participants

Fifteen native Brazilian Portuguese speakers participated and gave signed informed consent approved by the Carnegie Mellon University Institutional Review Board (IRB protocol HS14-474). Eight were Portuguese-English late bilinguals with high proficiency in L2 (English), all right-handed (5 females, 3 males), mean age 27.5 years (*SD*=2.3). Seven were Portuguese monolinguals, all right-handed (4 females, 3 males), mean age 28.7 years (*SD*=5.3).

The Portuguese-English bilingual participants were enrolled in or graduated from US universities as graduate or undergraduate students at the time of data collection. All the bilingual participants had been living in the U.S. for a mean of 2.02 years (SD=2.35), and all reported spending most of the day using English (6–12 h). The mean age of these bilinguals starting to learn English was 12.9 years (SD=4.7), and all had formal instruction in a school setting. To assess their reading proficiency, we adapted the reading section of a TOEFL test available online, and the Portuguese-English bilingual participants displayed high proficiency with a mean of 8.53 out of 10 (SD=1.16). (Note that the TOEFL reading comprehension test uses more complex sentences than our sentence stimuli.) A portion of the adapted TOEFL test administered to the Portuguese-English bilingual participants is shown in the Supplemental materials.

### 2.2. Experimental paradigm

Participants read 60 sentences in Portuguese while fMRI data were acquired. These Portuguese sentences were translation-equivalents of a subset of 240 English sentences (developed by Glasgow et al., 2016) used in a previous investigation (Wang et al., submitted). The fidelity of the translations was confirmed by back-translation and consultation among 4 advanced Portuguese/English Language scholars. The sentences obeyed subject-verb order, were in the active voice, and had a mean length of 3.2 content words. The sentences are shown in Supplemental materials (Table S.3). Of these 60 sentences, 49 described events (e.g. *O diplomata negociou na embaixada /The diplomat negotiated at the embassy*) and 11 described states (e.g. *A revista era amarela / The magazine was yellow*).

The sentences were presented one phrase at a time (e.g. *A família/ estava/feliz –The family/was/happy*) in a moving left-to-right window, as shown in Fig. 2. The duration of each phrase presentation was determined by an estimation formula: 300 ms×number of content words+16 ms×number of characters, where the number of characters includes all words except *the*). Phrases that contained adjectives that were followed by a noun remained on the screen until the noun disappeared. This display protocol (as opposed to presenting one word at a time or one sentence at a time) was adopted to approximate the type of encoding that is indicated by eye fixation studies of text reading (Just and Carpenter, 1980; Schuster et al., 2016). For example, during natural reading, it is rare for a reader to make separate eye fixations on the article *the*, so it was presented at the same time as the rest of its noun phrase. Furthermore, the presentation time for each phrase was also consistent with such studies that measured gaze durations on individual words in a text. In addition, because the hemodynamic

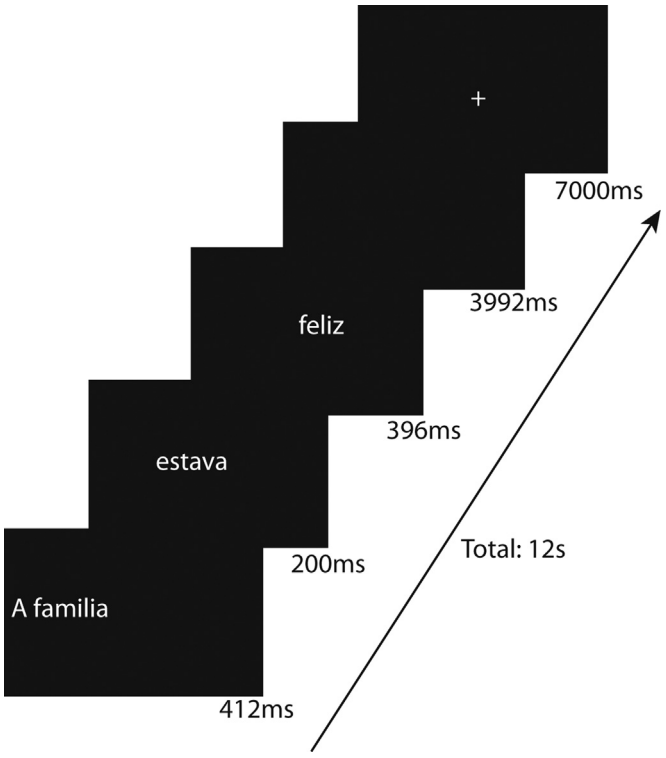


**Fig. 2.** Schematic representation of the experimental paradigm. Presentation of a sample sentence: *A familia estava feliz* (The family was happy). The duration of each phrase presentation was determined by a formula derived from eye movement studies of text reading (Just and Carpenter, 1980), namely 300 ms×number of content words +16 ms×number of characters, where the number of characters includes all words except *the*).

BOLD response in fMRI convolves the responses to temporally adjacent events, it is difficult to separate the responses to the article *the* and the noun it modifies. Thus, there is little loss of information in the fMRI signal if a simple phrase is presented in its entirety.

At the end of the sentence, a blank interval padded out the total presentation duration to 5 s. Participants were instructed to pay attention to the meaning of each phrase as it appeared by thinking about the properties of the concepts the phrase referred to. As each phrase of the sentence appeared, they were to integrate their conception of the phrase into their conception of the emerging sentence. During the blank interval, participants were instructed to continue thinking about the sentence, integrating the meaning of all the words. After the blank interval, a centered fixation cross appeared for 7 s during which participants were instructed to fixate and clear their minds.

The entire scanning session lasted one hour. Each of the 60 sentences was presented four times in four separate blocks in a randomized order. There were sixteen additional fixation or rest periods, 17 s each, distributed across the session, to provide a baseline measure of activation.

To assess the participants' attention to the task, they were given a sentence recognition test after the scan, consisting of 30 sentences that had been presented in the task (old) and 30 that were new. The resulting mean recognition accuracy was 95.4%.

### 2.3. fMRI acquisition and analysis

Functional images were acquired on a Siemens Verio 3.0T scanner at the Scientific Imaging & Brain Research Center (SIBR) of Carnegie Mellon University (gradient echo EPI pulse sequence; TR=1000 ms, TE=30 ms, and a 60° flip angle). Twenty 5-mm thick AC-PC aligned slices were imaged (1-mm gap between slices). The acquisition matrix

was 64×64 with 3.125×3.125×5-mm voxels.

The data were realigned and normalized to the Montreal Neurological Institute (MNI) template using SPM8 (http://www.fil.ion.ucl.ac.uk/spm/). For each presentation of a sentence, the percent signal change (PSC) was computed at each voxel, relative to the mean baseline activation level measured during fixation intervals.

Each sentence MPSC (mean PSC) image was measured as the mean of five PSC images, collected from 7 s to 11 s after sentence onset (one image per TR, each TR=1 s). This temporal window was determined by a preliminary investigation, which found that the most decodable neural signatures of all the content words in a simple sentence presented at normal reading speed occurred *after* the entire sentence had been read (Wang et al., submitted). The MPSC image was then normalized to mean of 0 and variance of 1 across sentences within each block of scans (presentations), to equate the overall intensities in each block. This procedure yielded a normalized MPSC image for each presentation of each sentence.

### 2.4. Neurally Plausible Semantic Features (NPSF) and thematic roles as word-level semantic features

The words of the sentences were characterized in terms of Neurally Plausible Semantic Features (NPSFs) and their thematic role in the sentence. These characterizations were identical for the English and Portuguese versions of the sentences. There were 42 binary NPSFs, such as *communication*, *shelter*, *impact*, and *emotion* as illustrated in Fig. 1C and described more completely in Table S.1. in Supplemental materials. The coding was performed by a group of raters guided by linguistic/semantic principles.

Six thematic role features for these simple sentence constructions were also coded and used in the model training: agent, main verb, predicate of copular sentence, patient, adjunct (most of which were propositional phrases), and modifier. These 6 thematic role features were conjoined with the 42 NPSFs to represent each word in a 48-long vector array. Words that had different thematic roles in different sentences were coded the same in the first 42 elements of this array (the NPSFs) but differently in the last 6 elements (thematic roles).

### 2.5. Specification of brain locations

The brain locations used in the modeling were derived from a hierarchical factor analyses (FA) of the fMRI sentence activation data from three participants in the English monolingual study whose activation patterns were most accurately predicted (Wang et al., submitted). The input to the first-level (i.e. individual participant level) of the hierarchical FA were the 600 voxels that were the most stable over all 240 sentences across four presentations in the data of each of these three participants. The first level FA produced seven factors for each participant. These first-level factors were then submitted to the second between-participant level FA, resulting in four semantic factors (plus a fifth perceptual factor pertaining to phrase length) common to the three participants, explaining 37% of the variation. The locations associated with the four semantic factors were obtained by clustering the voxels with the highest factor loadings into 38 brain locations. The clusters larger than 10 voxels are shown in Fig. 1. (All brain locations and their MNI coordinates are listed in the Table S.2. in Supplemental materials).

### 2.6. Sentence prediction model

The cross-language sentence model predicted the activation levels of the most stable voxels within the meta-language brain locations, shown in Fig. 1 and listed in Table S.2. Stable voxels were defined as voxels having consistent activation responses over the four presentations of the training sentences. The voxels with the highest stability within each brain location were selected. The numbers of voxels selected from each brain location were based on the brain location size: 5 for locations smaller than 50 voxels, 10 from locations of 50–150 voxels, and 20 from locations of 150–300 voxels.

In each cross-validation fold, the MPSC image (averaged over 5 PSC images from 7 s to 11 s post sentence onset) of one Portuguese sentence (averaged over four presentations) from one individual Portuguese speaker served as the test data. To obtain the weights from the English fMRI dataset, the model was trained on the MPSC images from four repetitions of a sentence (each averaged over 5 PSC images from 7 s to 11 s post sentence onset) of seven English monolingual speakers reading 240 sentences containing 242 words (Hastie et al., 2005), but excluding any sentence that contained any English translation equivalents of the component words of the test Portuguese sentence. These weights mapped between the 48 features (42 NPSFs +6 thematic roles) and the activation levels of the most stable voxels in the factor-related brain locations. Then these weights were applied the set of features (NPSFs+thematic roles) for each of the words of the Portuguese test sentence (that were new to the model), to predict their activation patterns. The predicted word images were then added together to compose the predicted sentence activation image for the Portuguese test sentence. The predicted activation patterns for the other 59 Portuguese sentences were similarly generated. To assess the accuracy of the sentence activation predictions, the similarity (cosine distance) between the actual left-out sentence image and all sixty predicted images was computed, and these predictions were rank ordered by their similarity to the actual left-out test sentence image. The rank accuracy of the prediction was computed as the normalized rank of its similarity to the actual target sentence in the list of 60 guesses.

## 3. Results

When the model weights and brain locations obtained from English monolingual participants' data were applied to the data of the 15 Portuguese participants, the mean rank accuracy of predicting the activation pattern of each of the 60 Portuguese sentences was .67 (SD=.07), and reliably above chance ($p < 0.001$, p value estimated by a 5000-iteration random permutation). Furthermore, the rank accuracies were significantly above chance for all but one of the participants (rank accuracies≥.56, $p < 0.05$). The mean prediction accuracies for sentences describing events and states were very similar (.68 and .67 respectively), and no significant difference was found between these two types of sentences. Thus the mapping developed in English is predictive of the activation evoked during the reading of Portuguese sentences, indicating both cross-language and cross – participant commonality of neural representations.

The detail of the model's functioning can be illustrated with an example, the sentence *O eleitor foi ao protesto* (*The voter went to the protest*). The actual observed image (the MPSC image averaged over participants) for this sentence is shown in the top row of Fig. 3, and the predicted image for the same sentence (using English-based weights) is shown in the bottom row. The three content words of this sentence (*voter*, *went*, *protest*) were coded with NPSFs such as "Governance", "Person", "Change of Location", "Social Interaction", "High affective Arousal", etc. These NPSFs are associated with high activation levels in areas such as left superior and middle frontal gyri, left middle temporal area, precuneus, and right temporoparietal junction. The observed and predicted images show this generally similar pattern, indicating that the generative model indeed captured the mapping between the semantics of the sentence and a specific neural activation pattern.

### 3.1. Confusion errors of the model reflect its semantic integrity

An additional way of assessing the degree to which the model was capturing the sentences' gist was to examine the types of events described by the model's highly ranked but incorrect sentences. We

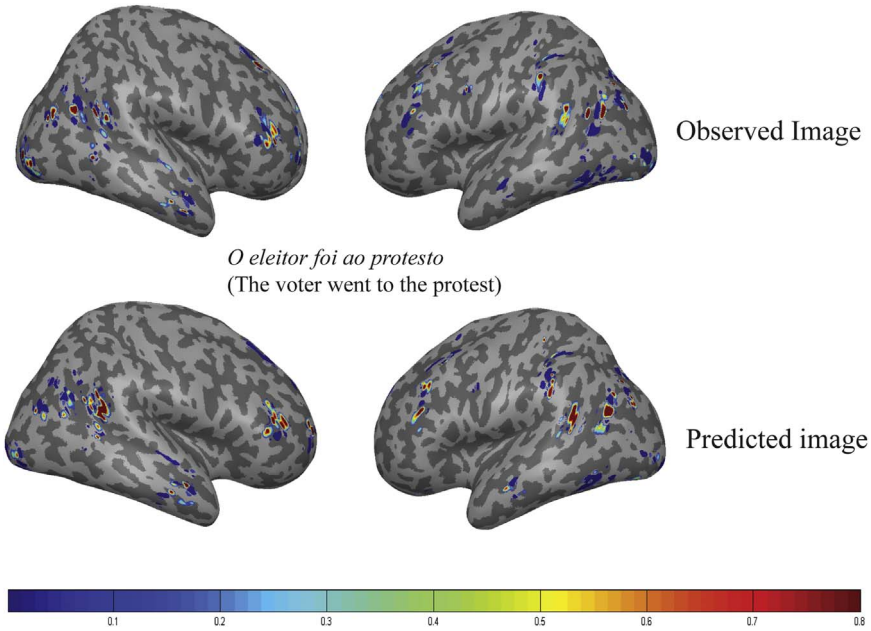


**Fig. 3.** Comparison of the neural activation patterns (in only the selected voxels) between the observed image evoked by the sentence *O eleitor foi ao protesto* (*The voter went to the protest*), and the predicted image of the sentence using weights from the English-based fMRI data. Both images are the MPSC images averaged across participants and normalized to values of 0–1.

focused on those items for which the predictive model's highest ranked sentence was the correct one, which represent cases of modeling success. The goal was to informally assess how similar the next few highest ranking sentences were to the correct sentence.

Some examples of such top ranked sentences are shown in Table 1. In these cases the mapping also assigned high ranks to other sentences that described similar events or situations. These runners-up came from the same semantic cohort as the target. That is, the runner-up sentences are similar in terms of the type of events they describe. This systematic semantic similarity of the runners-up indicates that the generative model captures the gist of the target stimulus sentence. This observation in turn suggests that events of a similar type have similar neural representations.

### 3.2. Thematic roles code sentence-level meaning

The predicted activation pattern of a sentence was generated by adding together the predicted activations of its component words, as well as taking those words' thematic roles into account (i.e. the same word with different thematic roles was modeled as different entities). The thematic roles encode a level of sentence meaning above individual words. Whether this level of sentence meaning contributed to the prediction accuracy was assessed by comparing models with and without thematic roles. The model that included thematic roles resulted in reliably higher accuracy (.67 vs 0.62) than the model without thematic roles $t$ (13)=3.14, ($p < 0.01$), indicating that inclusion of thematic roles captures a significant portion of the activation variance associated with sentence-level meaning representations. This approach is still an oversimplification of the nature of sentence meaning. There are likely to be additional sentence-level elements of meaning that emerge from the sentence as a whole, which are beyond the scope of the current model. Nevertheless, even without such postulated higher levels of meaning, the current model captures a significant amount of the systematicity in the fMRI data for the 60 sentences under investigation.

**Table 1**
The top five ranked sentences from the generative model in a subset of perfectly predicted sentences. The top five guesses are shown in English translation. In addition to the correct guess of the stimulus sentence at the top, the following four runners-up were also semantically similar to the target. This systematic runners-up cohort pattern indicates that the generative model captured the gist of the target stimulus sentence.

| | | | |
|---|---|---|---|
| Target sentence | *Os pais visitaram a escola* | Target sentence | *A flor era amarela.* |
| English translation | *The parent visited the school* | English translation | *The flower was yellow* |
| Top 5 predicted sentences | *The parent visited the school.*<br>*The politician visited the family*<br>*The happy couple visited the embassy*<br>*The parent bought the magazine*<br>*The family was happy* | Top 5 predicted sentences | *The flower was yellow*<br>*The magazine was yellow*<br>*The street was dark*<br>*The street was empty at night*<br>*The yellow bird flew over the field* |

## 4. Additional models and analyses

### 4.1. Bilingual versus monolingual participants

To determine whether knowledge of a second language (English) impacted the prediction accuracy of the cross-language model, this accuracy was computed separately for the bilingual and monolingual participants. The mean rank accuracy for the two groups was very similar: for bilingual participants, it was .66 ($SD$=.07) and for monolingual participants, it was .67 ($SD$=.05) ($t$ (13)=0.34, $n.s$). This result indicates that knowledge of English was neither essential nor helpful for producing accurate predictions. Thus the ability to predict sentence activation patterns in bilingual participants was apparently not due to them internally translating the sentences into English and thus conforming to the English-based weights and locations.

### 4.2. Activation prediction for Portuguese test sentences using data from both English and Portuguese sentences

The main model was trained only on data from English sentences. It is interesting to consider whether a model would make reliably more

accurate predictions if it were additionally trained on data from Portuguese sentences (excluding sentences that contain any component word of the test sentence). To assess this conjecture, an additional model was trained on activation data from both languages. This additional two-language model was trained on the averaged fMRI images of seven English monolingual speakers and on the images of the Portuguese speaker whose activation was being predicted (excluding the test sentence and any other sentence containing any words from the test sentence in both languages). The mean rank accuracy across 15 participants was .72, reliably higher ($t$ (14)=3.28, $p < 0.01$) than the .67 accuracy of the main model. (Note that the two-language model was given less information about the English activation (images for ~58–59 English sentences) than the main model (weights obtained from ~236 English sentences), with the same brain locations used in both cases, and the two-language model nevertheless provided higher accuracy). This analysis indicates that a predictive model is more accurate when trained on the data from not just another language but also on data from the target language. Nevertheless, the main conclusion of this paper is that it is possible to train a model exclusively on data from one language and make accurate activation predictions for another language.

### 4.3. Activation prediction for Portuguese sentences using data only from Portuguese sentences to derive model weights while still using the brain locations derived from the English data

When the model weights were obtained by training on the data of 14 of the participants on sentences that did not contain any component word from the one held-out test sentence, the mean rank accuracy for predicting the left-out test sentence in the left-out 15th participant, was .67 ($SD$=.05, $p < 0.001$). All the participants' accuracies reached significance (rank accuracy≥.56, $p < 0.05$). This mean accuracy is the same as that of the main model that was based exclusively on English data (but the main model was based on a larger amount of training data). Thus training a model on the same language as the target language does not substantially increase the prediction accuracy compared to a cross-language model with more training data.

### 4.4. Activation prediction using random sets of brain locations

All of the models and analyses above made their predictions for the brain locations derived from the factor analyses of the activation data evoked by English sentences. A model using random locations was developed to assess the contribution of using the factor-based brain locations. The random locations model used 1000 sets of 38 random brain volumes of the same sizes as the original set of brain locations. This model used no data whatsoever from the English study, but instead used data only from Portuguese sentences to derive model weights. The model was trained and tested on the averaged MPSC images averaged over the Portuguese participants.

In each of the 1000 permutation steps per sentence, the locations (centroid coordinates) of each of the 38 volumes were randomly selected (while retaining each volume's shape, and keeping them all within the boundary of the brain, and disallowing overlaps between volumes). In each permutation, the most stable voxels were selected within these new volumes, and the numbers of voxels per volume were determined as in the main model. This procedure was applied to predict the activation of each of the 60 Portuguese sentences (based on training data from the other 59 sentences).

The resulting mean prediction accuracy of the random locations model was .53, reliably lower than the accuracy of .67 obtained by the main model and using the original set of meta-language brain locations ($p < .001$). The random locations model's accuracy was also reliably lower than the accuracy of .67 obtained by the model above that used data only from Portuguese sentences to derive model weights while still using the brain locations derived from the English data. These results indicate that the brain locations obtained from the English monolingual speakers indeed encoded critical semantic information that is common across languages.

### 4.5. Prediction accuracy for different parts of speech

The word prediction accuracies for three parts of speech (noun, verb, adjective) were computed separately, producing means of .76, .73, and .76, respectively, with no reliable differences among them (F (2,108)=0.26, n.s.).

## 5. Discussion

### 5.1. Configurations of concept representations in sentences across languages

The findings clearly showed that it is possible to predict the fMRI activation patterns evoked by the reading of a sentence using a model developed entirely in another language. Furthermore, the model's confusion errors indicated that the prediction accuracy stemmed from capturing the gist of the event or the state that the sentence described, rather than any superficial properties.

Several assumptions were made in the cross-language sentence prediction model: that NPSFs provide a basis for estimating the activation patterns evoked by word-level semantic processing, and that sentence activation is composed of both the word-level activation patterns of the sentence's component words as well as activation patterns corresponding to these words' thematic roles.

Specifically, the NPSFs were construed as a set of implicit hypotheses concerning some key modulators of neural activation patterns. By using NPSFs as independent variables in the predictive model, we hypothesized them to be critical elements in the neural processing of word-based meaning representation. The key advantage of directly implementing NPSFs in the prediction model lies in the direct translation from basic neurolinguistic research to model development. By contrast, semantic vector representations of concepts (e.g. Latent Semantic Analysis, (Dumais, 2004)) have limited linguistic interpretability, although the use of such representations has been shown to be predictive of neural activation (e.g., Mitchell et al., 2008; Murphy et al., 2012; Schloss and Li, 2016).

NPSFs enable an intuitive assessment of the mapping from words of a given semantic class to activation patterns in a particular spatial distribution. Collectively, the results indicate that the mapping between NPSFs and neural activation patterns constitute a conceptually-based model, superseding the language differences at the lexical and syntactical levels.

To the best of our knowledge, all the cross-language neural decoding and prediction studies to date (Buchweitz et al., 2012; Correia et al., 2014; Zinszer et al., 2012), including the current study, used stimuli that only covered a small semantic space, which is selected to be simple, concrete and culturally similar. Whether the predictive ability of this concept-based model can generalize to the other uncovered domains of the vast human conceptual space requires further research and model development.

### 5.2. The meta-language brain locations

Using the 38 brain locations derived from the fMRI factor analyses of English monolingual speakers, the model succeeded in predicting the activation patterns evoked by Portuguese sentences, using the weights from either English speakers or other Portuguese speakers. These locations collectively constitute a meta-language conceptual brain network. Part of the universality of neural meaning representation can be attributed to the commonality of the neural infrastructure. Our findings lead to the testable prediction that the model may be extensible to concept prediction in other languages using similar brain

locations in the meta-language conceptual brain network with the ability to predict the activation of *new sentences composed of new words* of the same general type.

The knowledge of this common neural network can (1) advance the theories of the "language of thought", and (2) open new doors for cross-language and cross-modality decoding technologies for complex conceptual constructs.

These 38 brain locations belong to a number of networks that are relevant to semantic and conceptual processing. Some of them are related to basic neural dimensions, such as those related to the *people* dimension, including bilateral precuneus, bilateral middle temporal gyrus, and bilateral superior frontal areas, and those related to the *place* dimension, involving bilateral parahippocampal areas, bilateral precuneus, and bilateral middle occipital areas. There are also areas involved in sentence context processing, including semantic integration (angular gyrus, middle temporal gyrus) (Moseley and Pulvermüller, 2014; Price et al., 2016), theory of mind (precuneus, posterior cingulate cortex, bilateral prefrontal cortex) (Rilling et al., 2004), and concept binding (left middle and superior temporal area) (Frankland and Greene, 2015).

This study supports the hypothesis that the NPSFs or semantic concepts central to human experience are encoded in common neural areas across languages. The prediction model can bypass a language difference by characterizing neural activation patterns within these areas, and relating them to the corresponding elementary meaning units (NPSFs) at the "language of thought" level.

### 5.3. Commonalities between bilinguals and monolinguals in semantic representation

A few previous studies have demonstrated that bilinguals and monolinguals recruit similar neural areas for language processing (Kovelman et al., 2008; Proverbio et al., 2002). Although differences in neural activation levels between monolinguals and bilinguals have been demonstrated (Jones et al., 2011; Kovelman et al., 2008) due to factors such as exposure, age, and proficiency (Bloch et al., 2009), a common shared neural network has nonetheless repeatedly emerged (Buchweitz and Prat, 2013; Buchweitz et al., 2012; Correia et al., 2014; Kovelman et al., 2008).

This study further shows that similar neural activation patterns are evoked by the processing of English and Portuguese translation-equivalent sentences in both bilinguals and monolinguals. In other words, the neural codes for representing concepts in simple literal sentences are similar regardless of whether a person knows one language or two. This finding is consistent with a number of studies that compared activation patterns between monolinguals and bilinguals and found many commonalities and few differences between the two groups (Isel et al., 2010; Palomar-García et al., 2015; Parker Jones et al., 2012). At the same time, the findings do not imply that the neural processing of concepts in the bilingual brain is *identical* to the processing in the monolingual brain. For instance, previous studies have shown that several language-related areas show higher activation levels in bilinguals than monolinguals (Costa and Sebastián-Gallés, 2014; Parker Jones, et al. 2012), possibly due to the increased neural processing demands in the bilingual brain due to the need to control two languages (Costa and Sebastián-Gallés, 2014). Our study did not compare activation levels per se across languages, but instead compared *patterns* of activation levels across a set of voxels associated with individual concepts and sentences.

Although the bilingual participants in this study were all late bilinguals, we expect our sentence prediction algorithms to be comparably accurate for early bilinguals, since it is reasonable to assume that the concept representation pattern in L1 of early bilinguals is similar to late bilinguals and monolinguals.

### 5.4. Prediction direction

The modeling above mapped from semantics to neural activation, but the inverse mapping, from neural activation to NPSFs, is also possible. In a previous investigation of English monolingual speakers, both directions yielded similar accuracies in the mapping between the activation patterns and the NPSFs of sentences (Wang et al., submitted). Neural activation prediction models, such as the one used in the current study, have several scientific advantages, at least in the initial stage of model development (Naselaris et al., 2011). The most salient advantage is that it can yield a functional characterization of specific brain regions that can be compared to regional characterizations in other studies.

Semantics-prediction models, on the other hand, have application advantages. They can serve as an interface for brain-reading and neuroprosthetic technologies (Aflalo et al., 2015; Collinger et al., 2013). They can also be used to assess knowledge of concepts in an educational setting (Mason and Just, 2016).

### 5.5. The possibility of neural-based machine translation

Most automated translation applications, including Google translate (Google Translate, 2016), use statistical machine translation (SMT) algorithms (Koehn, 2009). SMT relies on parallel bilingual corpora (i.e. sentence-to-sentence aligned texts between a pair of languages) as inputs. Since it is costly to obtain large parallel bilingual corpora for each pair of languages, the performance of SMT of many language pairs (e.g. Arabic and Filipino) are often hindered.

The findings of neural commonality in concept representation between speakers of different languages in this study may provide a neurally-based mediation for machine translation. Essentially, it can mediate translations between a pair of languages in terms of the commonality of the neural representation of concepts. With the advancement of mapping neural activation patterns evoked by various languages, as well as further development of NPSFs, neurally-based mediation might serve as a future alternative way of obtaining parallel corpora, or even be developed into a neurally-based machine translation technology.

### 5.6. Future directions

The findings of this study suggest that several lines of future work should be useful for developing a broader neurosemantic theory of cross language commonality and empirically applying it at a sentence or discourse level.

First, syntactic features as well as thematic role features need to be implemented in a way that aligns well with their respective neural encoding mechanisms. Second, such a model needs to be tested in other less similar languages, outside the Indo-European language family. Third, cross-modality (spoken versus written) prediction across languages needs to investigated.

## 6. Conclusions

The current study demonstrates the commonality of the neural representation of sentences across two languages. The model successfully predicted Portuguese sentences using brain locations and weights applied to Neurally Plausible Semantic Features from a mapping developed in English. The mapping between the neural activation patterns and Neurally Plausible Semantic Features can be obtained from any group of participants (Portuguese monolingual, English monolingual, or Portuguese-English bilingual) in either language (English or Portuguese) and yield successful prediction of the activation evoked by a new sentence composed of new words. The model also captured the gist of the described event or state rather than depending on any particular word class or any other idiosyncrasies. In sum, the

model demonstrated meta-language prediction capabilities across languages, people, and bilingual status. Future studies will have to determine the extensibility of this approach to other pairs of less similar languages and to other communication media, such as events depicted in videos.

## Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Air Force Research Laboratory (AFRL) contract number FA8650-13-C-7360. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, AFRL, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. During data collection Cyntia Bailer was supported by the Brazilian Ministry of Education, CAPES BEX 14636-13-1. We thank Nick Diana, Robert Vargas, and Zachary Anderson for their help in data collection, stimulus preparation and coding, and Leda Tomitch for guidance on bilingual issues.

## Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.neuroimage.2016.10.029.

## References

Aflalo, T., Kellis, S., Klaes, C., Lee, B., Shi, Y., Pejsa, K., Liu, C., 2015. Decoding motor imagery from the posterior parietal cortex of a tetraplegic human. Science 348 (6237), 906–910.

Bloch, C., Kaiser, A., Kuenzli, E., Zappatore, D., Haller, S., Franceschini, R., Luedi, G., Radue, E.-W., Nitsch, C., 2009. The age of second language acquisition determines the variability in activation elicited by narration in three languages in Broca's and Wernicke's area. Neuropsychologia 47 (3), 625–633.

Buchweitz, A., Prat, C., 2013. The bilingual brain: flexibility and control in the human cortex. Phys. Life Rev. 10 (4), 428–443.

Buchweitz, A., Shinkareva, S.V., Mason, R.A., Mitchell, T.M., Just, M.A., 2012. Identifying bilingual semantic neural representations across languages. Brain Lang. 120 (3), 282–289.

Collinger, J.L., Wodlinger, B., Downey, J.E., Wang, W., Tyler-Kabara, E.C., Weber, D.J., McMorland, A.J.C., Velliste, M., Boninger, M.L., Schwartz, A.B., 2013. High-performance neuroprosthetic control by an individual with tetraplegia. Lancet 381 (9866), 557–564.

Correia, J., Formisano, E., Valente, G., Hausfeld, L., Jansma, B., Bonte, M., 2014. Brain-based translation: fmri decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe. J. Neurosci. 34 (1), 332–338.

Costa, A., Sebastián-Gallés, N., 2014. How does the bilingual experience sculpt the brain? Nat. Rev. Neurosci. 15 (5), 336–345.

Dumais, S.T., 2004. Latent semantic analysis. Annu. Rev. Inf. Sci. Technol. 38 (1), 188–230.

Frankland, S.M., Greene, J.D., 2015. An architecture for encoding sentence meaning in left mid-superior temporal cortex. Proc. Natl. Acad. Sci. 112 (37), 11732–11737.

Fodor, J.A., Pylyshyn, Z.W., 1988. Connectionism and cognitive architecture: a critical analysis. Cognition 28 (1–2), 3–71.

Giuliani, N.R., Mann, T., Tomiyama, A.J., Berkman, E.T., 2014. Neural systems underlying the reappraisal of personally craved foods. J. Cogn. Neurosci. 26 (7), 1390–1402.

Glasgow, K., Roos, M., Haufler, A., Chevillet, M., Wolmetz, M., 2016. Evaluating semantic models with word-sentence relatedness. arXiv:1603.07253

Hastie, T., Tibshirani, R., Friedman, J., Franklin, J., 2005. The elements of statistical learning: data mining, inference and prediction. Math. Intell. 27, 83–85.

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

Huth, A.G., Nishimoto, S., Vu, A.T., Gallant, J.L., 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76 (6), 1210–1224.

Isel, F., Baumgaertner, A., Thrän, J., Meisel, J.M., Büchel, C., 2010. Neural circuitry of the bilingual mental lexicon: effect of age of second language acquisition. Brain Cogn. 72 (2), 169–180.

Johnson-Frey, S.H., 2004. The neural bases of complex tool use in humans. Trends Cogn. Sci. 8 (2), 71–78.

Jones, Ō.P., Green, D.W., Grogan, A., Pliatsikas, C., Filippopolitis, K., Ali, N., Seghier, M.L., 2011. Where, when and why brain activation differs for bilinguals and monolinguals during picture naming and reading aloud. Cereb. Cortex, [bhr161].

Just, M.A., Carpenter, P.A., 1980. A theory of reading: from eye fixations to comprehension. Psychol. Rev. 87 (4), 329.

Just, M.A., Cherkassky, V.L., Aryal, S., Mitchell, T.M., 2010. A neurosemantic theory of concrete noun representation based on the underlying brain codes. PLoS One 5 (1), e8622.

Just, M.A., Cherkassky, V.L., Buchweitz, A., Keller, T.A., Mitchell, T.M., 2014. Identifying autism from neural representations of social interactions: neurocognitive markers of autism. PLoS One 9 (12), e113879.

Kassam, K.S., Markey, A.R., Cherkassky, V.L., Loewenstein, G., Just, M.A., 2013. Identifying emotions on the basis of neural activation. PLoS One 8 (6), e66032.

Koehn, P., 2009. Statistical Machine Translation. Cambridge University Press, New York.

Kovelman, I., Baker, S.A., Petitto, L.A., 2008. Bilingual and monolingual brains compared: a functional magnetic resonance imaging investigation of syntactic processing and a possible "neural signature" of bilingualism. J. Cogn. Neurosci. 20 (1), 153–169.

Marcus, G., Marblestone, A., Dean, T., 2014. The atoms of neural computation. Science 346 (6209), 551–552.

Martin, A., Wiggs, C.L., Ungerleider, L.G., Haxby, J.V., 1996. Neural correlates of category-specific knowledge. Nature 379, 649–652.

Mason, R.A., Just, M.A., 2016. Neural representations of physics concepts. Psychol. Sci. 27 (6), 904–913.

Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.M., Malave, V.L., Mason, R.A., Just, M.A., 2008. Predicting human brain activity associated with the meanings of nouns. Science 320 (5880), 1191–1195.

Murphy, B., Talukdar, P., Mitchell, T., 2012. Selecting corpus-semantic models for neurolinguistic decoding. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. Association for Computational Linguistics, pp. 114–123.

Moseley, R.L., Pulvermüller, F., 2014. Nouns, verbs, objects, actions, and abstractions: local fMRI activity indexes semantics, not lexical categories. Brain Lang. 132, 28–42.

Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. NeuroImage 56 (2), 400–410.

Palomar-García, M.Á., Bueichekú, E., Ávila, C., Sanjuán, A., Strijkers, K., Ventura-Campos, N., Costa, A., 2015. Do bilinguals show neural differences with monolinguals when processing their native language? Brain Lang. 142, 36–44.

Parker Jones, O., Green, D.W., Grogan, a, Pliatsikas, C., Filippopolitis, K., Ali, N., Lee, H.L., Ramsden, S., Gazarian, K., Prejawa, S., Seghier, M.L., Price, C.J., 2012. Where, when and why brain activation differs for bilinguals and monolinguals during picture naming and reading aloud. Cereb. Cortex 22, 892–902.

Price, A.R., Peelle, J.E., Bonner, M.F., Grossman, M., Hamilton, R.H., 2016. Causal evidence for a mechanism of semantic integration in the angular gyrus as revealed by high-definition transcranial direct current stimulation. J. Neurosci. 36 (13), 3829–3838.

Proverbio, A.M., Čok, B., Zani, A., 2002. Electrophysiological measures of language processing in bilinguals. J. Cogn. Neurosci. 14 (7), 994–1017.

Rilling, J.K., Sanfey, A.G., Aronson, J.A., Nystrom, L.E., Cohen, J.D., 2004. The neural correlates of theory of mind within interpersonal interactions. NeuroImage 22 (4), 1694–1703.

Rustandi, I., Just, M.A., Mitchell, T., 2009. Integrating multiple-study multiple-subject fMRI datasets using canonical correlation analysis. In: Proceedings of the MICCAI 2009 Workshop: Statistical Modeling and Detection Issues in Intra-and Inter-subject Functional MRI Data Analysis.

Schilbach, L., 2015. The neural correlates of social cognition and social interaction. In: Brain Mapping: An Encyclopedic Reference, Elsevier, pp. 159–164.

Schilbach, L., Wohlschlaeger, A.M., Kraemer, N.C., Newen, A., Shah, N.J., Fink, G.R., Vogeley, K., 2006. Being with virtual others: neural correlates of social interaction. Neuropsychologia 44 (5), 718–730.

Schloss, B., Li, P., 2016. Disentangling narrow and coarse semantic networks in the brain: the role of computational models of word meaning. Behav. Res. Methods, 1–15.

Schuster, S., Hawelka, S., Hutzler, F., Kronbichler, M., Richlan, F., 2016. Words in context: the effects of length, frequency, and predictability on brain responses during natural reading. Cereb. Cortex.

Tranel, D., Kemmerer, D., Adolphs, R., Damasio, H., Damasio, A.R., 2003. Neural correlates of conceptual knowledge for actions. Cogn. Neuropsychol. 20 (3–6), 409–432.

Van der Cruyssen, L., Heleven, E., Ma, N., Vandekerckhove, M., Van Overwalle, F., 2015. Distinct neural correlates of social categories and personality traits. NeuroImage 104, 336–346.

Van der Laan, L.N., De Ridder, D.T., Viergever, M.A., Smeets, P.A., 2011. The first taste is always with the eyes: a meta-analysis on the neural correlates of processing visual food cues. NeuroImage 55 (1), 296–303.

Wang, J., Cherkassky, V.L., Just, M.A. Neural Structure of Complex Thoughts: Computational Modeling of Brain Representations of Sentences, submitted to Nature.

Zhu, Y., Zhang, L., Fan, J., Han, S., 2007. Neural basis of cultural influence on self-representation. NeuroImage 34, 1310–1316. http://dx.doi.org/10.1016/j.neuroimage.2006.08.047.

Zinszer, B.D., Anderson, A.J., Kang, O., Wheatley, T., Raizada, R.D., 2016. Semantic structural alignment of neural representational spaces enables translation between English and Chinese words. J. Cogn. Neurosci..