Semantic Structural Alignment of Neural Representational Spaces Enables Translation between English and Chinese Words

Benjamin D. Zinszer[1], Andrew J. Anderson[1], Olivia Kang[2]*, Thalia Wheatley[2], and Rajeev D. S. Raizada[1]

Abstract

■ Two sets of items can share the same underlying conceptual structure, while appearing unrelated at a surface level. Humans excel at recognizing and using alignments between such underlying structures in many domains of cognition, most notably in analogical reasoning. Here we show that structural alignment reveals how different people's neural representations of word meaning are preserved across different languages, such that patterns of brain activation can be used to translate words from one language to another. Groups of Chinese and English speakers underwent fMRI scanning while reading words in their respective native languages. Simply by aligning structures representing the two groups' neural semantic spaces, we successfully infer all seven Chinese–English word translations. Beyond language translation, conceptual structural alignment underlies many aspects of high-level cognition, and this work opens the door to deriving many such alignments directly from neural representational content. ■

INTRODUCTION

Two sets of items can differ in their surface appearance but share the same underlying conceptual structure. People have a powerful ability to discover such structures and infer new information by comparing overall the similarities of known structures with new structures, such as in analogical reasoning (Holyoak & Thagard, 1996), category learning (Gentner & Namy, 1999), and word learning (Gentner & Namy, 2004). This comparison across structures is commonly referred to as structural alignment because it is based not on commonalities of the items themselves but instead on matching the structural relations between those items. Perhaps the most important example is seen in language. Different languages' lexicons are composed of very different word forms, but these words describe the same world and thus shared conceptual structures. Here we test whether these shared concepts are similarly represented across brains, regardless of differences in their surface (linguistic) representation. Brain imaging offers the ability to probe people's neural activation and (more scientifically interesting) the representational content that this activation carries. In behavioral studies of structural alignment, correspondences are identified between two sets of items based on their common relational structure (Gentner & Smith, 2012). Here we investigate whether structural alignment can also be applied to neural representations. This offers a strong test of the degree to which neuroimaging techniques actually access the intended conceptual content by asking the following question: If we take two sets of neural activity and align their neural similarity structures, will that alignment reflect an accurate alignment of the conceptual content?

Previously, we have shown that by matching neural similarities across participants it was possible to perform neural decoding across participants exposed to several visual objects (Raizada & Connolly, 2012). However, in that study the different people were presented with stimuli that were the same not only conceptually but also perceptually. Here, we ask the new question of whether the process of matching neural similarity structures can yield alignment at the level of linguistic concepts, even when the surface appearances of the stimuli (word forms) are completely different. To address that question, we used the task of cross-language translation between English and Chinese. These two languages differ greatly in surface characteristics: A set of English words and their Chinese translation equivalents have essentially nothing in common phonetically or orthographically. We test whether the process of aligning the neural similarity structures of Chinese and English speakers is able to correctly pair English words with their correct Chinese translations.

Previous work has investigated correlations between neural similarity structures, namely, the important and increasingly popular fMRI approach of representational

[1]University of Rochester, [2]Dartmouth College
*O. Kang is now at the Department of Psychology, Harvard University.

Journal of Cognitive Neuroscience 28:11, pp. 1749–1759
doi:10.1162/jocn_a_01000

similarity analysis (RSA; Kriegeskorte, Mur, & Bandettini, 2008). However, although RSA assesses the overall degree of match between two sets of neural similarities, it does not in itself produce an alignment between them. In other words, RSA does not find pairwise correspondences between items in the structures that it compares. Outside the domain of neural decoding, algorithms have been developed to match sets of structural relations (Turney, 2008; Goldstone & Rogosky, 2002; Laakso & Cottrell, 2000; Falkenhainer, Forbus, & Gentner, 1989), and this approach can be used to achieve across-subject neural decoding (Raizada & Connolly, 2012). Going beyond RSA's correlation of overall neural similarity structure, structural alignment is used here to reveal the pairwise correspondences between individual items. Moreover, those items dissociate underlying meaning from surface appearance, as the shared linguistic concepts are represented in languages that are very different, both visually and phonetically.

In this study, we perform structural alignment between functional neural responses for speakers of different languages. Between these groups, the surface appearances for stimulus words differ dramatically, but we predict that underlying conceptual structures are similar enough to realign individual translation equivalent words. Using multivoxel pattern analysis and representational similarity structures, we compare distributed functional brain activity for separate groups of native speakers of Chinese and English. We ask whether the similarity structures among the functional brain responses to these word-elicited concepts in each language are close enough in structure across languages to perform neural decoding on group level data and translate words in one language into the other language.

METHODS

Participants

On the basis of the sample size of a previous, similar paradigm (Mitchell et al., 2008; $n = 9$), we aimed to recruit 10 participants per language at Dartmouth College. Additional participants were scheduled in case of failure to complete the task or meet eligibility requirements (see below), and thus 11 English-speaking participants and 12 Chinese-speaking participants were included in the study. Behavioral data analysis revealed that one Chinese-speaking participant did not provide complete behavioral responses, and this participant was excluded from the imaging analyses. Thus, a total of 11 native speakers of English (4 M/7 F) and 11 native speakers of Mandarin Chinese (3 M/8 F) were available for the imaging analyses. These analyses were performed after collection of the entire sample, with no subsequent changes in sample size due to exclusion or additional recruitment.

All participants were undergraduate students, graduate students, or postdoctoral researchers. Participants self-reported being native speakers of English or Mandarin Chinese, defined as being born in their native language environment and speaking that language as their earliest language. The English-speaking participants reported no knowledge of Chinese language in a verbal screening before the experiment. The Chinese-speaking participants were all Chinese–English bilinguals studying or working in the United States. A separate screening (described below) was used to evaluate the Chinese-speaking participants' knowledge of the English stimuli.

Materials

We selected seven translation equivalent words in English and Chinese before the study, meeting four criteria: (1) concrete nouns, (2) monosyllabic in both languages, (3) represented by a single Chinese character, and (4) unlikely for English translations to be known by the Chinese participants. To ensure that criterion (4) was met, Chinese participants completed a brief translation task in which they were asked to write the English translation for 20 Chinese words. Mean translation accuracy was 1.56 out of the seven critical stimuli. Although this result may be surprising at first, this low level of accuracy confirms that we were successful in our goal of choosing words that do not often arise in an academic context. The Chinese-speaking participants were largely graduate students or postdoctoral researchers from the computer science department and rarely needed to use these words in English.

The seven words that we used are listed in Table 1 with some common lexical parameters. Frequency data were obtained from film subtitle frequencies in American English and Chinese (Cai & Brysbaert, 2010; Brysbaert & New, 2009). Concreteness ratings were obtained from Brysbaert, Warriner, and Kuperman's (2014) norming study for English words, and all seven critical stimuli scored near ceiling (5). To our knowledge, an equivalent

Table 1. Seven Critical Word Stimuli in English and Chinese

English			*Chinese*	
Word	*Frequency*	*Concreteness*	*Word*	*Frequency*
axe	1.81	5.00	斧 (fǔ)	2.33
broom	2.13	4.89	帚 (zhǒu)	0.03
gown	3.14	4.61	袍 (páo)	4.98
hoof	0.72	4.89	蹄 (tí)	2.41
jaw	3.37	4.87	颚 (è)	0.63
mule	2.54	5.00	骡 (luó)	0.42
raft	1.56	5.00	筏 (fá)	0.21

Frequency data are per million words in the SUBTLEX film subtitles database. Concreteness ratings are on a scale of 1 (*most abstract*) to 5 (*most concrete*). See Brysbaert et al. (2014) for details.

database is not yet available in Chinese. The critical stimuli were presented in three different font faces (English: Helvetica, American Typewriter, and Times New Roman; Chinese: STFangSong, Kai, and STSong) to reduce the influence of visual similarity on neural representations of the stimuli. The functional activity elicited by these words forms the basis of all the analyses presented here.

Participants completed a semantic relatedness task involving catch trials and 42 filler words interspersed between the seven critical stimuli to encourage them to think about word meanings. Filler words were not used in any of the fMRI analysis and therefore did not conform to the criteria used to select critical stimuli. Of the 42 filler words (translation equivalents in both languages), half were semantically related to one of the critical stimuli (e.g., axe–log) and half were semantically unrelated (e.g., axe–moth) for a total of three related words and three unrelated words for each critical stimulus. Because semantic relatedness was only a distractor task and not a manipulation in the experiment, the related and unrelated conditions were based on the experimenters' judgment, and participants were not evaluated on their accuracy in reproducing the same judgments.

Stimuli for this task were presented as black or red text on a gray background via projector to a screen behind the MRI scanner. Participants viewed the projected words through a mirror attached to the scanner's head coil.

Procedure

Experimental procedures were approved by the Dartmouth Committee for the Protection of Human Subjects. Participants completed the semantic relatedness task while undergoing fMRI. Words were presented for 2000 msec, followed by a 4000–6000 msec jittered fixation cross. If a catchword was presented in red text with a "?" (e.g., "moth?", also presented for 2000 msec), participants responded by indicating whether the catchword was semantically related to the word immediately preceding. Catchwords were always filler words and occurred in approximately one third of trials to encourage participants to think about the meanings of each stimulus word. Each functional run was composed of 45–50 stimulus presentations, about 7 min in duration. Participants completed seven functional runs for a total of 35 presentations per critical stimulus word.

Scanning Parameters

The study was performed using a Philips Intera Achieva 3-T scanner (Philips Medical Systems, Bothell, WA) with a SENSE (SENSEitivity Encoding) 32-channel head coil. Anatomical images were collected using a high-resolution 3-D MPRAGE sequence (220 slices, 1 mm isotropic voxels, field of view = 240 mm, acquisition matrix = 256 × 256). Functional images were collected in seven runs using echo-planar functional images sensitive to BOLD contrast (repetition time = 2000 msec, echo time = 35 msec, flip angle = 90°, 3 mm in-plane resolution). During each of the functional runs, 175 sets of axial images (35 slices/volume) were collected in an interleaved fashion across the whole brain.

Preprocessing and Estimation

Preprocessing and model estimation were all completed within SPM8 (Penny, Friston, Ashburner, Kiebel, & Nichols, 2011). Functional images across seven runs were registered to the mean image, smoothed with a 5-mm FWHM kernel for realignment, realigned, and resliced. Data were then normalized to the ICBM space template and written as 3 mm cubic voxels. A general linear model was estimated with separate regressors for each of the seven critical stimuli and a regressor for the response type (catch trial or none). Separate parameters were estimated for each functional run and then averaged in contrasts defined for each of the critical stimuli.

Multivoxel Pattern Analysis and Neural Similarity

Individual participants' multivoxel patterns for the critical stimuli were computed separately in 96 anatomical ROIs (48 in each hemisphere), as defined by the Harvard-Oxford Atlas (www.fmrib.ox.ac.uk/fsl/).

Figure 1 illustrates the procedure for calculating neural similarity in a single participant. A general linear model is estimated in SPM for each participant with separate parameters for each of the seven critical stimulus words. The estimated multivoxel response pattern for each critical stimulus (a word) is defined by a contrast map (beta weights) estimated in the participant level (first-level) general linear model (Figure 1B). Thus, for each critical stimulus, we extract a $1 \times n$ vector of beta values for all

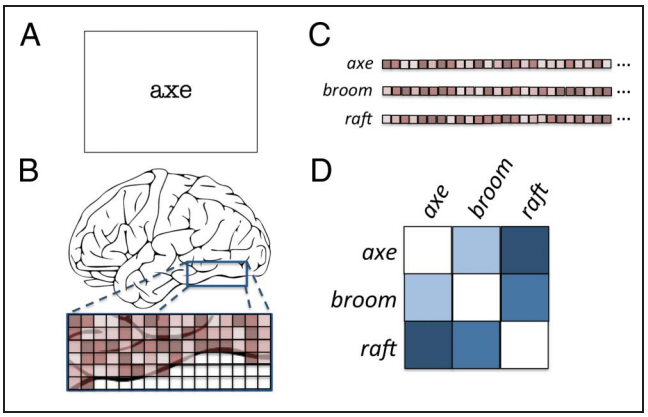

Figure 1. Procedure for computing a neural similarity matrix. (A) Stimulus is presented during functional imaging. (B) Individual voxel responses to stimulus are measured or estimated. (C) Responses for each stimulus are compared as n-dimensional vectors for n voxels. (D) Stimulus representations are correlated to generate the neural similarity matrix.

n voxels in the ROI (Figure 1C). An individual participant's multivoxel patterns for the critical stimuli are finally abstracted into similarity space. This step is critical because it eliminates the need for matching up patterns voxel-by-voxel across participants and instead allows comparison of the similarity structure to any set of multivariate patterns for the same stimuli. The response patterns to each of the seven critical stimuli are pairwise correlated (Pearson's *r*), resulting in a 7 × 7 neural similarity matrix in which each stimulus is described by the correlation of its functional response pattern to that of the other six stimuli (Figure 1D). The correlation values in each participant's 7 × 7 similarity matrix are transformed using Fisher's *r*-to-*z* for normalizing correlation coefficients, and the similarity matrices are averaged across participants in each group, resulting in a single 7 × 7 group level matrix for English and another for Chinese.

Permutation-based Decoding

To achieve neural translation, a reference matrix (e.g., the English group neural similarity structure) is compared with every possible permutation of stimuli in the test matrix (e.g., the Chinese group neural similarity structure). If the neural similarity structures are similar enough between two languages, the permutation of the test matrix most highly correlated with the reference matrix will be the correct set of translations. Statistical significance for the permutation test is determined by observing all possible permutations for an empirical probability distribution. The 95th percentile of this distribution represents a p value of .05. Note that it is important to distinguish between permutation testing for nonparametrically estimating statistical significance, a common procedure that has also been used to assess the significance of overall correlations between similarity matrices (e.g., Kriegeskorte, Mur, Ruff, et al., 2008), as opposed to our permutation-based decoding, which establishes structural alignments between specific pairs of items by picking as its alignment the permutation of labels that produces the highest correlation between the permuted and reference similarity matrices.

RESULTS

Participants were prompted to make semantic relatedness judgments in approximately 30% of stimulus presentations for seven target words and 42 additional filler words. Mean participant response rate was 83% (*SD* = 20%) with a mean RT of 1398 msec (*SD* = 167 msec). No measure of response accuracy was performed because the semantic relatedness judgments are subjective.

Figure 1 illustrates the procedure for calculating participants' neural similarity matrices based on their unique patterns of functional activity for word stimuli. For the seven word stimuli in each language (Table 1), similarity matrices were computed and transformed using Fisher's *r*-to-*z* (the inverse hyperbolic tangent) to normalize the *r* distribution, and a group similarity matrix was computed by averaging individual participants' matrices for each language. No voxel selection criteria were used in this or any other analyses in this article, and thus in whole-brain measures, the entire set of voxels from all parts of all anatomical regions were used to produce the structural alignments.

The English and Chinese group neural similarity structures were also estimated in each anatomical ROI of a standard and widely used brain atlas (the Harvard-Oxford atlas, www.fmrib.ox.ac.uk/fsl/) and used to attempt neural decoding of one language using the neural similarity structures estimated for the other language. For example, the group similarity matrices for English and Chinese in left postcentral gyrus are illustrated in Figure 2A. Between-group comparison of these structures provides a neurally mediated form of translation wherein Chinese words can be matched to English words based only on their respective functional brain response patterns, via the neural similarity structures for each language. These seven-by-seven matrices thus provide the sole source of information for alignment across languages and thus for the translations.

To provide an intuitive visualization of the structural alignments produced from these similarity matrices, we performed a classical multidimensional scaling of the similarity matrices into 2-D projections. The *z*-transformed correlations in the similarity matrices were transformed into distance matrices using 1/*z* (where *z* corresponds to the transformed correlation coefficient), and overall structures were plotted for each language. Alignment of these structures across languages provides some insight into the translation relationships that would be predicted based on the left postcentral gyrus. However, the MDS projections depict only the first two dimensions of these multidimensional data. Correlations between the distance matrices and their 2-D projections were strong, indicating that the MDS projections visualized for illustration in Figure 2 account for a large amount of the variance in the neural representations (Chinese: R^2 = .66; English: R^2 = .81). Furthermore, the distances in the 2-D projections were highly correlated between languages, giving an impression of how well the structures aligned (R^2 = .82) in the third panel of Figure 2B. Crucially, however, none of these diagnostics nor the MDS solutions as a whole were used to evaluate translation accuracy. They are presented strictly as a visualization of the end-results of the structural alignment analysis. However, that structural alignment analysis itself was performed by permuting the 7 × 7 similarity matrices, a process that occurred before and independently of the subsequent 2-D MDS visualization.

The threshold for statistical significance was computed by taking the 95th percentile of the full distribution of

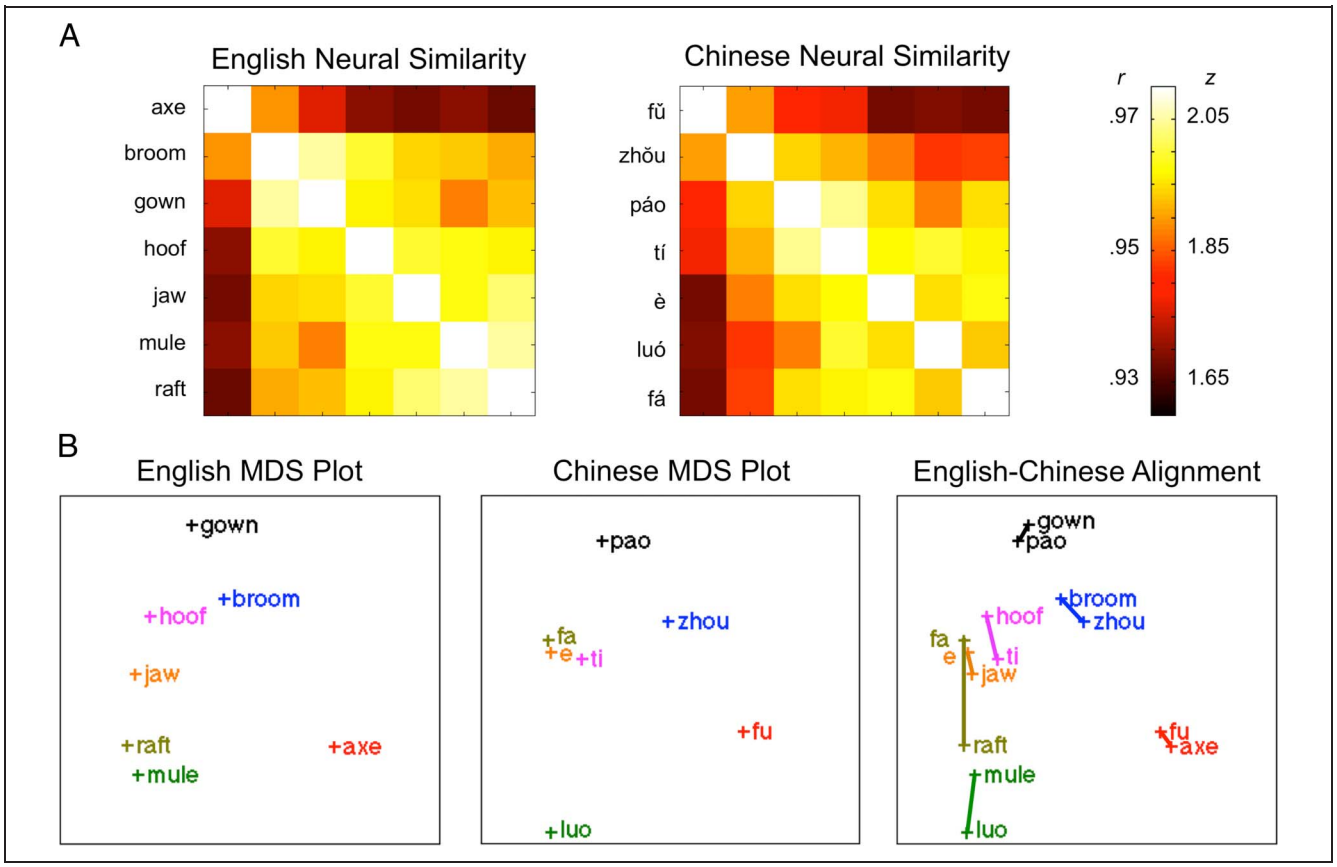

Figure 2. (A) Neural similarity matrices (*r*-to-*z* transformed) based on all voxels in left postcentral gyrus for each language group. (B) Conceptual structures represented as MDS projections of neural similarity structures, based on first two dimensions.

accuracy scores for all possible permutations of words in the test structure. The 95th percentile of the accuracy distribution for all permutations was 3 of 7 correct translations. Thus, scores above this threshold have less than a .05 probability of occurring by random selection (Raizada & Connolly, 2012). Bonferroni correction for multiple comparisons (96 ROIs) results in a significance threshold of 5 of 7 correct translations. Figure 2B illustrates a strong visual correspondence between the Chinese and English similarity structures in left postcentral gyrus.

All ROIs that achieved five or seven correct translations are outlined in Table 2. In this case, six correct translations are impossible because one translation error would entail swapping two words, resulting in only five correct translations. Switching the reference and test matrices yields identical results. Six regions produced 7 of 7 translations, and an additional 11 ROIs (listed in Table 2) correctly translated five of the seven word stimuli (the corrected threshold value for significance). Figure 3 (visualized using the xjView toolbox available at alivelearn.net/xjview) depicts a sample of the cortical regions tested and highlights all six ROIs with 7 of 7 translations.

Regions with successful neural decoding were largely located in temporal and parietal regions, although most regions were not bilaterally equal in decoding accuracy. Several frontal regions were also included in the set of successfully decoded ROIs. The majority of regions with 7/7 translations (4 of 6 regions) and slightly narrower majority for all regions with 5 of 7 or more translations (11 of 17 regions) were right hemisphere regions. The whole-brain similarity structures yielded 5 of 7 translations.

In one final analysis, we applied a jackknife resampling procedure to test the reliability of the decoding results. We excluded participants one at a time from the sample of all 22 participants and repeated the between-group decoding. In the whole-brain data, resampling did not significantly change decoding accuracy (mean accuracy = .73, $t(21) = 1.00$, $p = .38$). For each of the 17 significant ROIs, we compared the set of 22 resampled means to the original significance threshold of .71 (5 of 7 translations) using a single-sample t test. Only the temporooccipital part of right inferior temporal gyrus (Harv.-Oxf. #15) fell significantly below this threshold after correction for multiple comparisons. The mean and standard deviations for the resampled means are reported in Table 2.

DISCUSSION

We have demonstrated that representational similarity structures for word-elicited concepts are conserved across speakers of different languages. These similarity structures are effective for comparing functional brain

Table 2. Regions with Significant Decoding Accuracy

Harvard-Oxford ROI	*Anatomical ROI*	*Resampled Mean (SD)*
Regions with 7 of 7 Correct Translations (Accuracy = 1.0)		
14	R inferior temporal gyrus, posterior division	0.82 (0.22)
16	L postcentral gyrus	0.90 (0.14)
18	R supramarginal gyrus, anterior division	0.84 (0.19)
28	R cingulate gyrus, anterior division	0.75 (0.29)
32	R frontal orbital cortex	0.82 (0.25)
33	L parahippocampal gyrus, anterior division	0.96 (0.10)
Regions with 5 of 7 Correct Translations (Accuracy = 0.71)		
02	R superior frontal gyrus	0.68 (0.15)
04	L inferior frontal gyrus, pars triangularis	0.66 (0.21)
06	R precentral gyrus	0.66 (0.10)
09	R superior temporal gyrus, posterior division	0.75 (0.11)
11	R middle temporal gyrus, posterior division	0.69 (0.05)
11	L middle temporal gyrus, posterior division	0.73 (0.06)
15	R inferior temporal gyrus, temporooccipital part	0.62 (0.11)
17	R superior parietal lobule	0.69 (0.08)
18	L supramarginal gyrus, anterior division	0.69 (0.14)
29	R cingulate gyrus, posterior division	0.71 (0.14)
37	L temporal fusiform cortex, posterior division	0.77 (0.11)

ROIs are numbered according to the Harvard-Oxford anatomical atlas. Translation accuracy was also resampled using a jackknife (leave-one-participant-out) procedure. Mean and *SD* of resampled accuracy scores are reported.

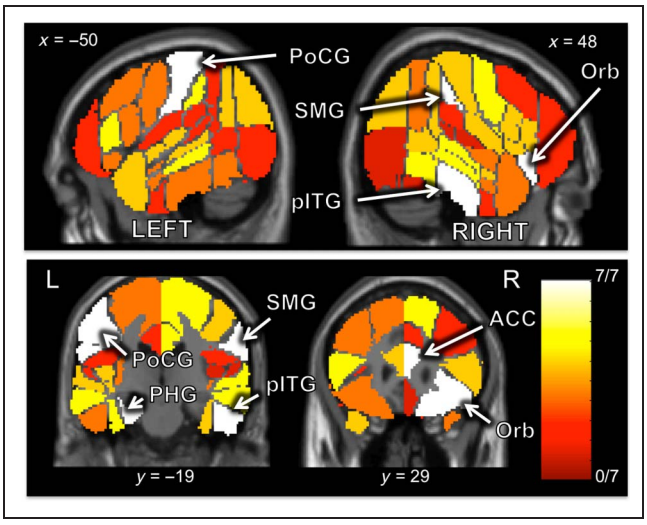

Figure 3. Decoding accuracy projected onto the cortical surface. See Table 2 for list of all ROIs with accuracy significantly greater than chance level. ROIs providing 7 of 7 correct translations are labeled as follows: ACC = anterior cingulate cortex; Orb = frontal orbital cortex; PHG = parahippocampal gyrus; pITG = posterior inferior temporal gyrus; PoCG = postcentral gyrus; SMG = supramarginal gyrus.

activity across speakers of different languages and enable cross-language decoding, that is, neural translation, via structural alignment. We thereby provide strong neurocognitive evidence for an intuitive (but heretofore untested) claim: Although the words we use differ in visual and auditory form across languages, we conceptualize the meanings of these words similarly, such that they can be reliably translated across brains.

This study is distinct from all previous work investigating the neural decoding of language because we translate semantic representations between independent groups for each language. Although previous studies have investigated neural decoding across languages (Correia, Jansma, Hausfeld, Kikkert, & Bonte, 2015; Correia et al., 2014; Buchweitz, Shinkareva, Mason, Mitchell, & Just, 2012), these studies have sidestepped one or more aspects of the translation challenge by using within-subject designs, which only allow comparisons within a single brain. In a similar vein, cross-modal (but within-language) comparison has found significant similarity for orthographic and auditory elicitations of word meanings (e.g., Akama, Murphy, Na, Shimizu, & Poesio, 2012) but also remains limited to within-subject decoding and

thus makes no generalization about regularities in representation across people. One notable exception, Honey, Thompson, Lerner, and Hasson's (2012) translation of narratives between different groups of Russian and English speaking participants examined only global semantic processing without addressing word meaning. The current work is distinct from this previous work because we translate meaning of individual word-elicited concepts between independent groups for each language.

Within-subject studies are interesting but reveal only associative matches between words in each language (like learning synonyms for a word) or idiosyncratic semantic relationships that are conserved only within an individual's brain. Here we use a between-subject design to demonstrate the generalizability of structural alignment across brains, allowing the identification of common conceptual spaces across speakers and languages. Our data expand upon dominant models of bilingual lexical representation and access, which maintain that bilinguals draw on a single, shared conceptual store for building lexical semantic connections (Kroll, van Hell, Tokowicz, & Green, 2010; Van Hell & De Groot, 1998; Kroll & Stewart, 1994; De Groot, 1992) by suggesting that these shared concepts are broadly preserved across entire groups of speakers of different languages.

The neuroanatomical regions that produce the highest decoding performance achieve successful translation because they encode the underlying conceptual structures evoked by word stimuli while all other aspects of the word stimuli (e.g., visual, orthographic, phonetic) differ between Chinese and English. Previous research in semantic representation and processing has produced a broad-reaching network of anatomical regions that integrate perceptual and functional information into concepts and decode auditory or visual word forms to activate those concepts in language comprehension. Many of the regions that produced significant cross-language decoding in this study correspond to areas widely recognized for the integration of multimodal semantic information, either in traditional semantic processing studies or in previous within-language neural decoding analyses. Here we review these regions and compare our findings with the existing semantic literature. We also found a handful of regions that are not typically reported for involvement in semantic processing, and we explore these findings in a later section.

Regions Classically Associated with Semantic Representation

The left fusiform gyrus has been repeatedly implicated in previous neural decoding studies of semantic representation (Anderson, Bruni, Lopopolo, Poesio, & Baroni, 2015; Chen, Garcea, & Mahon, 2015; Fernandino et al., 2015; Raizada & Connolly, 2012; Mitchell et al., 2008; Haxby et al., 2001) and translated five of seven words accurately in this study. Recent studies have also suggested that the parahippocampal gyri may be a multimodal hub for the conjunction of multiple sensory modalities (Fernandino et al., 2015), which is consistent with its role in functional representations instead of specific sensory modalities. The parahippocampal gyri have also been implicated in specific domains of concept representation: better for shelter- than tool-specific concepts (Just, Cherkassky, Aryal, & Mitchell, 2010) but also seem to represent functional information about tools (versus action/motor information; Chen et al., 2015).

The lateral and ventral temporal cortices are also widely recognized for integrating multimodal information in word representation (see reviews by Poeppel, Emmorey, Hickok, & Pylkkänen, 2012; Price, 2012; Binder, Desai, Graves, & Conant, 2009). A series of regions proceeding from the ventral temporooccipital cortex toward the anterior temporal lobe support visual wordform processing, as in this study's task. Price's (2012) account describes the most posterior regions (the temporooccipital cortex) as performing visual feature extraction whereas more anterior regions (the posterior inferior temporal cortex) access lexical semantics of the identified words. However, we have found that a large segment of posterior regions (both temporooccipital and posterior portions of the inferior temporal gyrus) in right hemisphere encode sufficient semantic information to allow cross-language decoding. This observation is inconsistent with a strict role of visual feature extraction for the temporooccipital region because the languages should provide no orthographic or visual cues to translation. Findings in the left and right middle temporal gyri are more consistent with conventional semantic processing models, wherein the middle temporal gyrus provides cross-modal semantic representation (Fairhall & Caramazza, 2013) and shows only weak left lateralization (Poeppel et al., 2012). Finally, the left posterior superior temporal gyrus has long been included in the anatomical definition of Wernicke's area and associated with word comprehension (Poeppel et al., 2012; although with some debate as to its involvement in auditory processing, auditory word recognition, and semantic access; Price, 2012; Binder et al., 2009). In this case, we find that translation is successful only in the right superior temporal gyrus (STG). Bilateral involvement of posterior STG is generally restricted to auditory processing functions in the aforementioned reviews, leaving open the possibility that the visual word forms are eliciting in STG some semantic representation adequate for achieving translation (perhaps mediated by auditory word form retrieval).

Several parietal regions identified in this study have been emerging in recent literature as playing important roles in integration of semantic representations across sensory modalities. Recent studies in neural decoding have converged upon the supramarginal gyrus as a multimodal (or transmodal) integrator of sensory information (Fernandino et al., 2015). Previous cross-language studies have found that this region produced stable responses

across languages in bilinguals (Buchweitz et al., 2012) and were correlated in Russian and English speakers' semantic processing (Honey et al., 2012), although these effects were left lateralized. In this study, the bilateral supramarginal gyrus reached or exceeded the significance criterion (5 of 7 translations).

The bilateral (left-dominant) activation of posterior cingulate gyrus was associated with semantic tasks in Binder et al.'s (2009) meta-analysis. On the basis of its involvement in visual, spatial, and emotional processes in other tasks, Binder and colleagues suggest that the posterior cingulate interacts with the hippocampus to process episodic memory and thus is highly correlated with semantic processing tasks that involve retrieval of episodic information. This claim is consistent with our finding that the right posterior cingulate produced accurate translations. Our results also identified the right anterior cingulate for high decoding accuracy. The anterior cingulate has more typically been described as providing conflict monitoring for cognitive control (Shenhav, Straccia, Cohen, & Botvinick, 2014; Botvinick, Cohen, & Carter, 2004), including in the case language conflict in bilingualism (Green & Abutalebi, 2013; Abutalebi & Green, 2007). The anterior cingulate cortices have not often been the focus on concept representation research, but we see some evidence of this modal specificity in that it is selectively responsive to information about object shape, while the posterior cingulate decodes across multiple modalities (Fernandino et al., 2015).

Recent and Novel Findings in Semantic Representation

Several other anatomical regions where we observed significant cross-language decoding are not typically included in neural accounts of semantic representation. According to the meta-analysis of Binder et al. (2009), pre- and postcentral gyri and the adjacent superior parietal lobule rarely appear in contrast-based neuroimaging analyses, perhaps suggesting a lack of overall change in activation levels across tasks. This absence of significant spatial contrasts does not rule out the possibility that multivoxel response patterns still encode important semantic information, and we find that these regions encode accurate translation information. According to previous decoding studies, the postcentral gyrus is, in fact, tuned for identifying tools (Just et al., 2010), suggesting that it may represent information related to manipulation of objects. Relatedly, a region spanning the left postcentral gyrus and superior parietal lobule are selectively responsive to shape information, and the right precentral gyrus to manipulation information (Fernandino et al., 2015). Both of these factors would be highly relevant to the classification of concrete objects, such as those described by the seven nouns in our study.

The left inferior frontal gyrus has been widely implicated in language processing tasks, but Binder et al. (2009) suggested that that only the pars orbitalis of the inferior frontal gyrus is directly involved in semantic processing and that any detected pars triangularis involvement is attributable to phonological or working memory aspects of the task. Our findings suggest that the pars triangularis may also contain semantic information. Previous research in picture naming has linked pars triangularis with lexical selection when objects have low name agreement (Kan & Thompson-Schill, 2004), and this selection may involve more semantically informed representations than revealed by contrast analyses, which focus on magnitude of activation.

Left dorsomedial and ventromedial pFC are frequent correlates of semantic processing, but their direct role in representation is not clear. According to Binder and colleagues' (2009) account, the dorsal and medial surfaces of the left superior frontal gyrus may have a role in retrieval processes, although these effects are primarily left-lateralized, while our findings were restricted to the region's right homologue. Price (2012) proposes that the left superior frontal gyrus is principally involved in various constraints on semantic meaning, either by syntax or by context. The frontal orbital cortex (in the wider ventromedial pFC) appears to be associated with affective processing and thus likely underlies affective information in semantic processing (Binder et al., 2009). Poeppel et al. (2012) more generally attribute ventromedial pFC as another semantic combinatory region, tasked with representing syntactically covert meaning. These proposals do not appear to be mutually exclusive and could be consistent with our current cross-language findings.

A right lateralization effect seems to be suggested by the relative prevalence of right hemisphere regions among significant ROIs, but this observation may be explained in light of the jackknife resampling results. The stability information provided by the resampling analysis supports a more balanced account: Although the left hemisphere regions from the 7 of 7 translation set were highly stable (mean accuracy ≥ .90 in left parahippocampal and post central gyri), right hemisphere regions with the same group level translation accuracy achieved markedly lower resampled accuracy and higher variability (larger standard deviations; see Table 2). Thus, the relative importance of these regions should be interpreted with the reliability of their performance across participants in mind.

Crucially, the limited number of words used in this study may simply result in too stark a contrast between regions with successful and unsuccessful translation performance. For example, a region that achieves five of seven accurate translations has a 50% chance of successfully translating the last pair of words (given that only two permutations of the remaining two words exist). This fact should limit the importance we assign to whether a region achieves five versus seven translations, and indeed the apparent right bias decreases slightly when considering all significant regions together, as well as the appearance of typical bilateral or left-lateralized regions for

semantic integration (middle temporal gyrus, supramarginal gyrus, left fusiform cortex) in the broader 5+/7 set of regions.

Previous neural decoding studies have not specifically investigated lateralization effects, but one cautious explanation for a right hemisphere advantage might be drawn from research on lateralization in semantic processing: Semantic information in the right hemisphere has long been hypothesized to represent coarser, message level semantic representations (Beeman, 1993) and more recently been associated with processing semantically distant or novel associations and semantic context (Vigneau et al., 2011; Jung-Beeman, 2005) such as in metaphor comprehension (Vigneau et al., 2011; Schmidt, DeBuse, & Seger, 2007). In this study, coarser representations may offer better cross-language symmetry than fine-grained language-specific or culturally specific information. Particularly, because our word stimuli were composed of only seven relatively distant concepts, the coarse representations for these concepts could be more consistent across languages than their left-lateralized, finer-grained representations (such as the exact shape and appearance of a prototypical *broom* or *raft*).

Future Directions

This study is the first to show that neural activation can be used to structurally align a set of words in speakers of one language with the translation-equivalent words in speakers of a very different language. Clearly this promising result in a new area is just the beginning of a much broader set of future investigations. In this study, we relied on group level averages of the neural similarity structures to perform the alignments. This approach reduces interparticipant noise in the similarity structures, but it also limits our ability to generalize about new participants or new words.

One aspect that could obviously be extended is the number of words being considered. Our study used only seven words. Now that cross-language structural alignment has been shown to be achievable, the constraints that led to that small set size could be loosened in future work. For example, each word could be presented a smaller number of times during the fMRI scanning (thereby allowing more words to be used without lengthening the overall scan), and our tight constraints on word choice (single character in Chinese, monosyllabic in English, etc.) could also potentially be relaxed. Including a greater number of words would provide more similarity data for the representation of each word, and these additional data could improve individual participant level translation rather than relying on group-averaged data to improve the signal-to-noise in the similarity matrices. Furthermore, we have evaluated the overall success of the structural alignment, but with a greater number of words, comparisons in translation accuracy could be drawn between words. Translation may be more or less successful for new sets of words based on their semantic similarity, translation equivalence, or other factors that we simply cannot explore with the current, small stimulus set. Clearly, more research is needed. This study demonstrates that across-subject neural translation is possible. Building upon that initial foundation, future work can explore how well such translation can succeed across larger numbers and varieties of words.

Moreover, although the permutation-based decoding worked well in this study, it has certain computational limits: It is computationally infeasible to perform the same exhaustive search for even slightly larger lexicons, as the number of permutations that must be compared expands factorially (e.g., 7 words have 5040 permutations, but 10 words have over 3.6 million permutations). However, search optimization strategies offer a number of opportunities to refine the existing algorithm, which would allow neural translation to scale up to much larger lexicons.

Our comparison of neural similarity structures in native speakers of Chinese and English yielded a successful translation between English and Chinese words based on the functional brain responses of separate groups of participants using each language. This ability to compare brain representations of words between speakers of different languages presents a new way of studying translation asymmetry, such as between abstract nouns for which experimental evidence indicates translation costs due to ambiguity (Van Hell & De Groot, 1998). Neurally informed translation permits comparison of multiple translation candidates for their relative fitness to brain responses elicited by the other language. Furthermore, language-specific and language-independent elements of brain representation can be contrasted by examining translation pairs for correlation to nonlinguistic measures (e.g., visual object information) and linguistic measures (e.g., word co-occurrence) in a similar fashion to Anderson et al. (2015).

Human cognition often draws on the ability to find alignments between structures that differ in their surface appearance (such as the relationships between words in Chinese and English) but share a deeper underlying structure (such as the relationships between concepts underlying words). Structural alignments of this sort support a broad range of cognitive functions, including analogical reasoning, scientific inference, and language learning (Gentner & Smith, 2012; Gentner & Namy, 1999; Holyoak & Thagard, 1996). In this study, we showed that such structural alignment of matching concepts could be performed directly on people's neural representations. This work may therefore open the door to the largely uncharted domain within cognitive neuroscience of the structural alignment of neural conceptual spaces.

Acknowledgments

This work was supported in part by NSF award 1228261 (P.I. Raizada).

Reprint requests should be sent to Benjamin D. Zinszer, Rochester Center for Brain Imaging, Room 2B201, 430 Elmwood Ave., Rochester, NY 14627, or via e-mail: bzinszer@gmail.com.

REFERENCES

Abutalebi, J., & Green, D. W. (2007). Bilingual language production: The neurocognition of language representation and control. *Journal of Neurolinguistics, 20,* 242–275.

Akama, H., Murphy, B., Na, L., Shimizu, Y., & Poesio, M. (2012). Decoding semantics across fMRI sessions with different stimulus modalities: A practical MVPA study. *Frontiers in Neuroinformatics, 6*.

Anderson, A. J., Bruni, E., Lopopolo, A., Poesio, M., & Baroni, M. (2015). Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induced by written text. *Neuroimage, 120,* 309–322.

Beeman, M. (1993). Semantic processing in the right hemisphere may contribute to drawing inferences from discourse. *Brain and Language, 44,* 80–120.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex, 19,* 2767–2796.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences, 8,* 539–546.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41,* 977–990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46,* 904–911.

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

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS One, 5,* e10729.

Chen, Q., Garcea, F. E., & Mahon, B. Z. (2015). The representation of object-directed action and function knowledge in the human brain. *Cerebral Cortex, 26,* 1609–1618.

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. *Journal of Neuroscience, 34,* 332–338.

Correia, J. M., Jansma, B., Hausfeld, L., Kikkert, S., & Bonte, M. (2015). EEG decoding of spoken words in bilingual listeners: From words to language invariant semantic-conceptual representations. *Frontiers in Psychology, 6*.

De Groot, A. M. B. (1992). Bilingual lexical representation: A closer look at conceptual representations. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 289–412). Amsterdam: Elsevier.

Fairhall, S. L., & Caramazza, A. (2013). Brain regions that represent amodal conceptual knowledge. *Journal of Neuroscience, 33,* 10552–10558.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence, 41,* 1–63.

Fernandino, L., Binder, J. R., Desai, R. H., Pendl, S. L., Humphries, C. J., Gross, W. L., et al. (2015). Concept representation reflects multimodal abstraction: A framework for embodied semantics. *Cerebral Cortex, 26,* 2018–2034.

Gentner, D., & Namy, L. (1999). Comparison in the development of categories. *Cognitive Development, 14,* 487–513.

Gentner, D., & Namy, L. L. (2004). The role of comparison in children's early word learning. In D. G. Hall & S. R. Waxman (Eds.), *Weaving a lexicon* (pp. 533–568). Cambridge, MA: MIT Press.

Gentner, D., & Smith, L. (2012). Analogical reasoning. In V. S. Ramachandran (Ed.), *Encyclopedia of Human Behavior* (2nd ed., pp. 130–136). Oxford, UK: Elsevier.

Goldstone, R. L., & Rogosky, B. J. (2002). Using relations within conceptual systems to translate across conceptual systems. *Cognition, 84,* 295–320.

Green, D. W., & Abutalebi, J. (2013). Language control in bilinguals: The adaptive control hypothesis. *Journal of Cognitive Psychology, 25,* 515–530.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293,* 2425–2430.

Holyoak, K. J., & Thagard, P. (1996). *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press.

Honey, C. J., Thompson, C. R., Lerner, Y., & Hasson, U. (2012). Not lost in translation: Neural responses shared across languages. *Journal of Neuroscience, 32,* 15277–15283.

Jung-Beeman, M. (2005). Bilateral brain processes for comprehending natural language. *Trends in Cognitive Sciences, 9,* 512–518.

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,* e8622.

Kan, I. P., & Thompson-Schill, S. L. (2004). Effect of name agreement on prefrontal activity during overt and covert picture naming. *Cognitive, Affective & Behavioral Neuroscience, 4,* 43–57.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., et al. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron, 60,* 1126–1141.

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language, 33,* 149–174.

Kroll, J. F., van Hell, J. G., Tokowicz, N., & Green, D. W. (2010). The revised hierarchical model: A critical review and assessment. *Bilingualism, 13,* 373–381.

Laakso, A., & Cottrell, G. (2000). Content and cluster analysis: Assessing representational similarity in neural systems. *Philosophical Psychology, 13,* 47–76.

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

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., & Nichols, T. E. (Eds.) (2011). *Statistical parametric mapping: The analysis of functional brain images*. Burlington, MA: Academic Press.

Poeppel, D., Emmorey, K., Hickok, G., & Pylkkänen, L. (2012). Towards a new neurobiology of language. *Journal of Neuroscience, 32,* 14125–14131.

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *Neuroimage, 62,* 816–847.
Raizada, R., & Connolly, A. (2012). What makes different people's representations alike: Neural similarity space solves the problem of across-subject fMRI decoding. *Journal of Cognitive Neuroscience, 24,* 868–877.
Schmidt, G. L., DeBuse, C. J., & Seger, C. A. (2007). Right hemisphere metaphor processing? Characterizing the lateralization of semantic processes. *Brain and Language, 100,* 127–141.
Shenhav, A., Straccia, M. A., Cohen, J. D., & Botvinick, M. M. (2014). Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value. *Nature Neuroscience, 17,* 1249–1254.
Turney, P. (2008). The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research, 33,* 615–655.
Van Hell, J. G., & De Groot, A. M. B. (1998). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Bilingualism: Language and Cognition, 1,* 193–211.
Vigneau, M., Beaucousin, V., Hervé, P. Y., Jobard, G., Petit, L., Crivello, F., et al. (2011). What is right-hemisphere contribution to phonological, lexico-semantic, and sentence processing? Insights from a meta-analysis. *Neuroimage, 54,* 577–593.