Neurocognitive Mechanisms Supporting the Generalization of Concepts Across Languages

Usman Ayub Sheikh[1], Manuel Carreiras[1,2,3], and David Soto[1,2]

[1] *Basque Center on Cognition, Brain and Language, San Sebastian, Spain*
[2] *Ikerbasque, Basque Foundation for Science, Bilbao, Spain*
[3] *University of the Basque Country, Bilbao, Spain.*
Correspondence to: David Soto (d.soto@bcbl.eu)

December 28, 2020

Acknowledgements

We thank Liv Hoversten for comments on a prior draft of the manuscript. We also thank Cesar Caballero for his advice and guidance during the review process. D.S. acknowledges support from the Basque Government through the BERC 2018-2021 program, from the Spanish Ministry of Economy and Competitiveness, through the 'Severo Ochoa' Programme for Centres/Units of Excellence in R&D (SEV-2015-490) and also from project grants PSI2016-76443-P from MINECO and PI-2017-25 from the Basque Government.

The authors declare no competing interests.

Abstract

The neurocognitive mechanisms that support the generalization of semantic representations across different languages remain to be determined. Current psycholinguistic models propose that semantic representations are likely to overlap across languages, although there is evidence also to the contrary. Neuroimaging studies observed that brain activity patterns associated with the meaning of words may be similar across languages. However, the factors that mediate cross-language generalization of semantic representations are not known. We here identify a key factor: the depth of processing. Human participants were asked to process visual words as they underwent functional MRI. We found that, during shallow processing, multivariate pattern classifiers could decode the word semantic category within each language in putative substrates of the semantic network, but there was no evidence of cross-language generalization in the shallow processing context. By contrast, when the depth of processing was higher, significant cross-language generalization was observed in several regions, including inferior parietal, ventromedial, lateral temporal, and inferior frontal cortex. These results are in keeping with distributed-only views of semantic processing and favour models based on multiple semantic hubs. The results also have ramifications for existing psycholinguistic models of word processing such as the BIA+, which by default assumes non-selective access to both native and second languages.

Keywords: *semantic representation, bilingualism, language, machine learning*

1 Introduction

A key unresolved question is whether different languages in bilingual people are integrated in the same system with shared/overlapping representations or rely on separate systems/representations for each language. Behavioral evidence from cross-language priming studies suggests that semantic representations are at least partially overlapping (Grainger, 1998; Perea, Dunabeitia, & Carreiras, 2008; Schoonbaert, Duyck, Brysbaert, & Hartsuiker, 2009). The evidence has led to the development of psycholinguistic models of bilingual language representation (Kroll & Stewart, 1994; Van Hell & De Groot, 1998). Although these models differ in their predictions about the mechanisms that underlie lexical processing and the links between lexical and semantic processing of the two languages, they agree that semantic representations are at least partially overlapping between languages. Yet, other studies have failed to support overlapping semantic systems (Grainger & Beauvillain, 1988; De Groot & Nas, 1991; Altarriba & Basnight-Brown, 2007). The mixed evidence between the cross-language priming studies is likely to originate from a lack of control of low-level properties of primes and targets (e.g. word length and frequency) (Balota & Chumbley, 1984), which can lead to cross-language priming effects not due to semantics.

Previous fMRI studies based on univariate activation-based approaches did not show reliable differences in task-related (i.e. word generation, picture naming) hemodynamic activity across languages (Abutalebi, Cappa, & Perani, 2001; Stowe & Sabourin, 2005; Indefrey, 2006). One limitation of these studies is that the experimental tasks and contrasts supposedly targeting semantic processing were often confounded by other untargeted orthographic/phonological processes (Binder, Desai, Graves, & Conant, 2009). Univariate fMRI-based priming studies (Chee, Soon, & Lee, 2003; Crinion et al., 2006) have found some evidence for both language-shared and language-specific brain responses, but the role of strategic factors such as expectancy lists of prime-target relations could not be determined (Basnight-Brown & Altarriba, 2007). Strategies linked to expectancy lists (i.e. involving participants constructing a list of expected targets), alongside the use of long SOAs, may also alter the depth of processing (Basnight-Brown & Altarriba, 2007). Moreover, mass-univariate approaches are not best suited to identify whether or not semantic processing is mediated by a similar system across the different languages. The observation that a cortical area is activated in both languages does not imply that the brain representations are also similar. Two recent studies used multivariate pattern analysis (MVPA) to assess whether the brain activity patterns elicited by words in one language can predict the patterns of equivalent words in the other language (Buchweitz, Shinkareva, Mason, Mitchell, & Just, 2012; Correia et al., 2014). They found language-shared representations in well-known semantic substrates including the left parietal lobe, inferior frontal gyrus, and posterior temporal lobe.

A key limitation of the studies reviewed so far is that the factors underlying the generalization of semantic representations across languages remain to be determined. Critically, none of the previous MVPA studies noted above (Buchweitz et al., 2012; Correia et al., 2014) considered the depth of processing during the task. Here we operationalize the depth of processing based on the contrast between covertly reading a visual word (henceforth shallow processing) and mentally simulating the properties associated with the word

concept (henceforth deep processing) based on the re-enactment of modality-specific representations. We note that while our manipulation of the depth of processing differs from the seminal experimental framework on 'levels of processing' (Craik & Lockhart, 1972) based on tasks targeting semantic vs. lower level phonemic/orthographic judgements, our experimental procedure is in keeping with different processing depths of processing; mental simulation is more likely to promote deeper semantic access, while the more shallow processing counterpart mainly taps onto phonological processing and rich semantic analyses is not mandatory.

Little research has examined the role of task-related factors on the brain representation of meaning. We here hypothesize that the depth of processing imposed by the task plays a critical role in the generalizability of semantic representations across languages. However, according to influential psycholinguistic models of word processing i.e. the Bilingual Integrated Activation model (Dijkstra & Van Heuven, 2002), the activation of language-shared representations may be independent of the depth of processing, and rather derived in parallel and non-selectively. Other theoretical accounts such as the perceptual symbols theory (Barsalou, 1999; Simmons & Barsalou, 2003) propose that semantic representations result from an implicit and automatic process of simulation in modality-specific sensory and action systems. This model therefore also predicts that semantic representations generalize across languages regardless of the depth of processing. Here we used fMRI-based MVPA to investigate how the depth of processing influences both within-language decoding and the generalization of semantic representations across languages in canonical substrates of the semantic network (Binder et al., 2009). The cross-language generalization of the decoder was taken as a proxy for language-shared representations (Dehghani et al., 2017). To pre-empt the results, we observed that while the decoding of the semantic category of words is significant within a given language regardless of the depth of processing, cross-language generalization of the brain representations of concepts was only found in the context of deeper levels of processing.

2 Materials and Methods

2.1 Participants

Thirty early and proficient Spanish-Basque bilinguals (mean age 24.2 ± 3.0 years; 19-34 years; 20 female) including twenty with Spanish as L1 were recruited through BCBL's own web portal specifically designed for this purpose: `https://www.bcbl.eu/participa`. They came from different educational backgrounds ranging from high school to postgraduate and professional training. All of them were healthy, had normal or corrected to normal vision, gave written informed consent prior to the experiment and were financially compensated with 20 euros for their time. The experiment was approved by the BCBL Ethics Review Board and conformed to the guidelines of the Helsinki Declaration.

All participants had acquired both languages before the age of 6. The age of acquisition of Spanish ($mean = 0.24 \pm 0.74$) was found to be statistically significantly lower ($t(29) = -2.60; p = 0.01$) than the age of acquisition of Basque (1.17 ± 1.61). Similarly, their reported performance in the two well known tests of language proficiency, i.e. LexTALE (Lemhöfer & Broersma, 2012) - available for only 27 out of 30 partic-

ipants - and BEST (De Bruin, Carreiras, & Duñabeitia, 2017) - available for only 29 out of 30 participants - was also found to be statistically significantly higher (LexTALE: $t(26) = 5.46; p < 0.05$, BEST: $t(28) = 5.40; p < 0.05$) in Spanish (LexTALE: 94.54 ± 4.93, BEST: 99.36 ± 1.27) as compared to Basque (LexTALE: 86.56 ± 9.13, BEST: 89.76 ± 9.20). This shows that participants were more proficient in Spanish than in Basque.

Basque and Spanish are two very different languages with different roots. While Spanish is a romance language, Basque has unknown linguistic roots. It is an isolated pre-indo-european language. In addition, Basque holds many prominent linguistic differences with Spanish in the canonical word order in sentences regarding the subject, verb and object, morphology (Basque: agglutinative), syntax (Basque: ergative), and lexicon (many different vocabulary and non-cognates).

2.2 MRI Acquisition

A SIEMENS's Magnetom Prisma-fit scanner, with 3 Tesla magnet and 64-channel head coil, was used to collect, for each participant, one high-resolution T1-weighted structural image and ten functional acquisition runs each lasting for about 7 minutes. The proposed MR sequence was set up and run using SIEMENS's software Numaris/4 (version: syngo MR E11). In each fMRI run, a multiband gradient-echo echo-planar imaging sequence with acceleration factor of 6, resolution of $2.4 \times 2.4 \times 2.4 mm^3$, TR of 850ms, TE of 35 ms, flip angle of 56 deg and bandwidth of 2582 Hz/Px was used to obtain 477 3D volumes of the whole brain (66 sagittal slices; FoV = 210mm). The high resolution T1-weighted structural image covering the whole brain (resolution of $1.0 \times 1.0 \times 1.0 mm^3$, TR of 2530ms, TE of 2.36 ms, flip angle of 7 deg) was collected after the fifth functional run using a fast 3D mprage sequence. The visual stimuli were projected on an MRI-compatible out-of-bore screen using a projector placed in the room adjacent to the MRI-room. A small mirror, mounted on the head coil, reflected the screen for presentation to the participants. The head coil was also equipped with a microphone that enabled the participants to communicate with the experimenters in between the runs.

2.3 Stimuli

A total of 16 words were used with 8 words per language. The Basque words were translational equivalents of Spanish words. Among 8 words, the 4 were living words including wolf, rooster, fox, and sheep, and the 4 were non-living words including candle, key, tube and mirror (for Spanish and Basque translations, see Figure 1). All the words were non-cognates and were balanced with respect to length and frequency (per million words; a standard measure independent of the corpus size) across categories (living and non-living) and languages ($t(7) = -1.16, p = 0.28$ for length and $t(7) = 0.28, p = 0.78$ for frequency per million; see Table 1 for details) based on the statistics provided by Espal (for Spanish; (Duchon, Perea, Sebastián-Gallés, Martí, & Carreiras, 2013) and E-Hitz databases (for Basque; (Perea et al., 2006)). The requirement of length and frequency balancing across categories and languages put some constraints on the number of words; nevertheless the number finally selected was in keeping with previous studies of semantic decoding (Shinkareva, Malave, Mason, Mitchell, & Just, 2011; Buchweitz et al., 2012; Correia et al., 2014). The semantic analysis of these words based on word embeddings

i.e. word2vec (see § S7) show the non-living things to be more similar between them (light blue shade) as compared to the living things and there is a room for increased separation in the semantic space.

	Spanish		Basque	
	Living	Non-living	Living	Non-living
Length	4.5±0.58	4.75±0.96	4.5±0.58	5.25±0.96
Frequency	28.73±19.90	19.90±6.12	23.53±17.90	24.55±8.01

Table 1: The table shows mean word length and frequency per million of stimuli with respect to both languages and semantic categories. These statistics were gathered using Espal for Spanish and E-Hitz for Basque. It can be seen that they are balanced across categories and languages.

2.4 Experimental Procedure

Each trial began with a fixation period of 250 ms followed by a blank screen of 500 ms (see Figure 2). The target word, randomly drawn from a pool of 4 living and 4 non-living words (see § 2.3), was presented for 1 s. Depending on a run's instructions (shallow or deep processing), the participants were supposed to either read and attend to the word, or to think about the characteristics of the living/non-living object it represented (e.g. its shape, its color etc.). Following a delay of 4 seconds, a red asterisk appeared at the center of the screen presented for a jittered time (see below) in which participants were instructed to do nothing. To ensure that the participants focused on the stimuli and the task, a maximum of two catch trials were set to appear at random points in each of the runs. These catch trials showed number words from among ZERO, ONE, and THREE in place of usual living/non-living words, and participants were supposed to respond by pressing any one of the four buttons on the fMRI response pad. The number TWO ("dos" in Spanish and "bi" in Basque) was not used due to different number of letters across languages. The total number of catch trials was kept equal across conditions.

To have as many trials as possible per each run, and at the same time maximize the separation between the brain activity corresponding to each of the trials, an event-related design was used and the time for which the asterisk stayed on the screen was jittered between 6 to 8s. This jitter was based on a pseudo-exponential distribution resulting in 50% of trials with the inter-trial interval of 6s, 25% with 6.5s, 12.5% with 7s and so on.

Both instructions and stimuli were presented at the center of the screen, in white against black background and in all uppercase Arial font. The experiment was programmed using Psychopy (Peirce, 2007) and is summarized in Figure 1. It comprised 10 runs (7 minutes each) and lasted for about 1.25 hours. In odd-numbered runs, participants were instructed to read and attend to the words (shallow processing), while in the even numbered ones, they were instructed to think about the characteristics of the living/non-living that the word represented (deep processing). Each fMRI run was subdivided into four language blocks with two Spanish (S) and two Basque (B) blocks, and the order of these blocks was counterbalanced across runs (SBSB, BSBS, and so on). In each

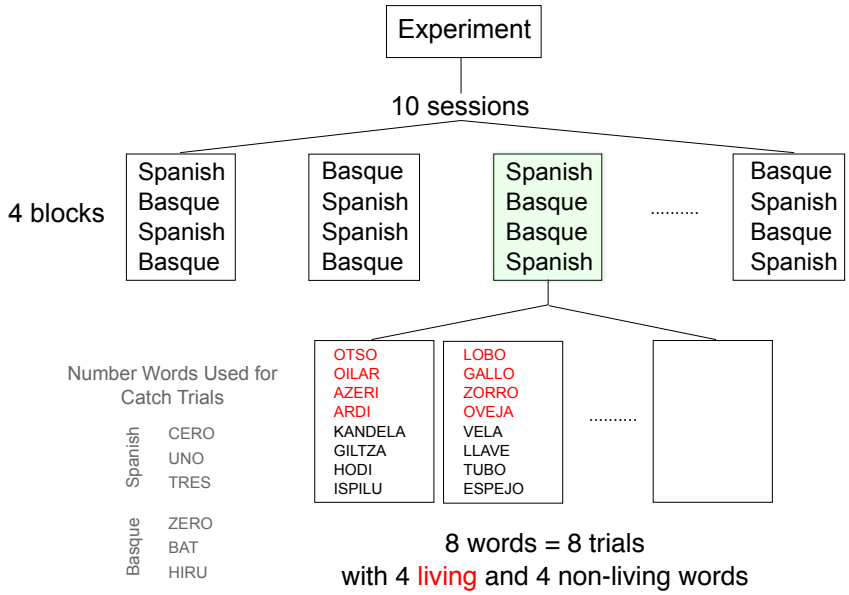

Figure 1: The figure summarizes the organization of runs, blocks and trials in the experiment. The experiment comprised of 10 runs with odd-numbered runs for shallow processing and even-numbered runs for deep processing. Each run was further subdivided into 4 language blocks (2 Spanish and 2 Basque). Each of these blocks was made up of 8 trials corresponding to single presentation of each of 4 living and 4 non-living words. The figure also shows the Spanish and Basque translations of both living/non-living words and number words.

of these blocks, eight words were presented (without repetition) in a random arrangement resulting in a total of thirty two trials per run.

2.5 MRI Data Preprocessing

The preprocessing of fMRI data was performed using FEAT (fMRI Expert Analysis Tool), a tool in FSL suite (FMRIB Software Library; v5.0). After converting all data from DICOM to NIfTI format using MRIConvert (`http://lcni.uoregon.edu/downloads/mriconvert`), the following steps were performed on each run's fMRI. To ensure steady state magnetisation, the first 9 volumes corresponding to the task instruction period were discarded; to remove non-brain tissue, brain extraction tool (BET) (Smith, 2002) was used; head-motion was accounted for using MCFLIRT (Jenkinson, Bannister, Brady, & Smith, 2002); minimal spatial smoothing was performed using a gaussian kernel with FWHM of 3mm. Next, ICA based automatic removal of motion artifacts (ICA-AROMA) was used to remove motion-induced signal variations (Pruim et al., 2015) and this was followed by a high-pass filter with a cutoff of 60s. All the runs were aligned to a reference volume of the first run. All further analyses were performed in native BOLD space.

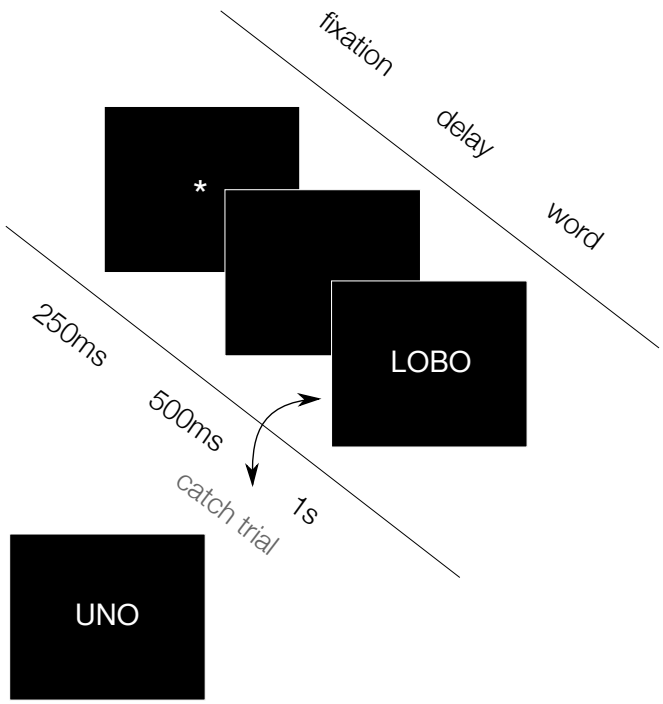

Figure 2: The figure illustrate the sequence of events on each trial. Following a fixation period, a word was presented for 1s. The participants were supposed to either read and attend to the word or think about the living/non-living thing it represented. Next, following a delay of 4s, a red asterisk appeared at the center of the screen and stayed there for a jittered interval of 6-8s. To ensure that the participants were engaged, catch trials were placed at random points in each of the runs. These catch trials showed number words from among ZERO, ONE, and THREE in place of living/non-living words, and participants were supposed to respond by pressing a button.

A set of 8 left-lateralized ROIs was pre-specified (see Figure S3) with 7 based on a meta-analysis of the semantic system by Binder et al. 2009 (Binder et al., 2009) and one anterior temporal lobe (ATL) due to its crucial role as a "semantic hub" (Damasio, Grabowski, Tranel, Hichwa, & Damasio, 1996; Patterson, Nestor, & Rogers, 2007; Correia et al., 2014). So, the ROIs included: inferior parietal lobe (IPL), lateral temporal lobe (LTL), ventromedial temporal lobe (VTL), dorsomedial prefrontal cortex (dmPFC), inferior frontal gyrus (IFG), ventromedial prefrontal cortex (vmPFC), posterior cingulate gyrus (PCG) and anterior temporal lobe (ATL). First, automatic segmentation of the high-resolution structural image was obtained using FreeSurfer's automated algorithm `recon-all`. Next, `mri_binarize` was used to extract individual gray matter masks from aparc+aseg volume using corresponding label indices in FreeSurferColorLUT text file (`https://surfer.nmr.mgh.harvard.edu/fswiki/FsTutorial/AnatomicalROI`). And finally, after visually inspecting these in FSLView, they were transformed to each run's func-

tional space using FLIRT (7 DoF global rescale transformation). (Jenkinson & Smith, 2001; Jenkinson et al., 2002) and were binarized (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT/FAQ).

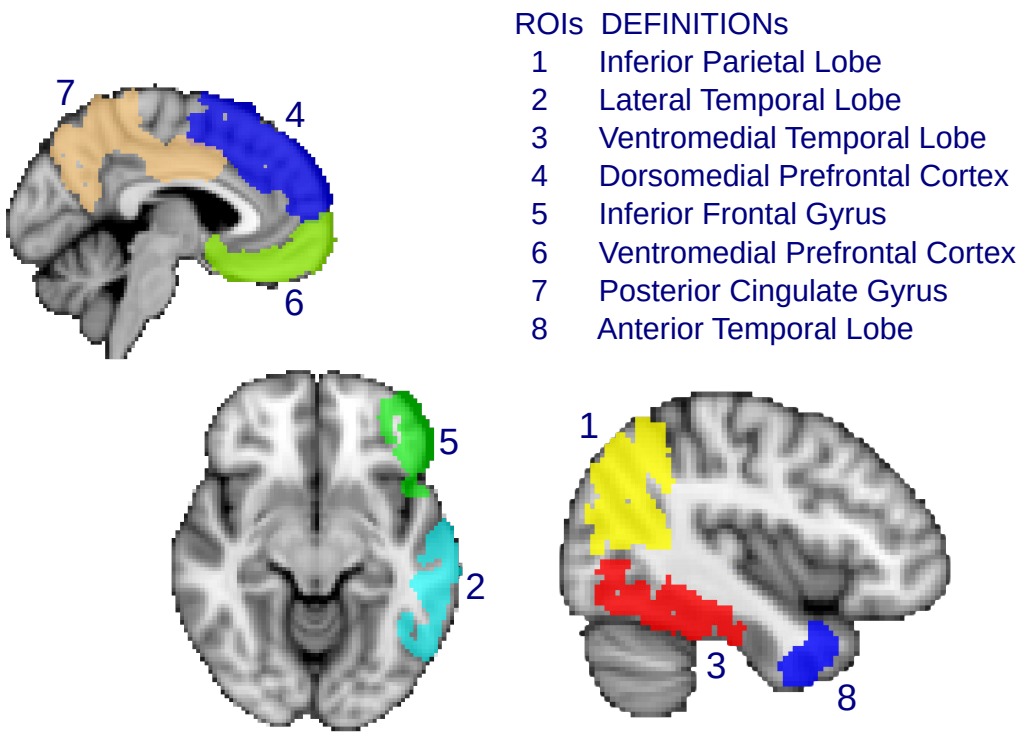

Figure 3: The figure shows the selected regions of interest projected on an MNI standard template image. These 8 left-lateralized areas were pre-specified and included inferior parietal lobe (IPL), lateral temporal lobe (LTL), ventromedial temporal lobe (VTL) including fusiform gyrus and parahippocampal gyrus, dorsomedial prefrontal cortex (dmPFC), inferior frontal gyrus (IFG), ventromedial prefrontal cortex (vmPFC), posterior cingulate gyrus (PCG) and anterior temporal lobe (ATL).

2.6 Multivariate Pattern Analysis

Multivariate pattern analysis was conducted using scikit-learn (Pedregosa et al., 2011) and PyMVPA (Hanke et al., 2009) libraries. Specifically, classification based on a supervised machine learning algorithm i.e. linear support vector machine (Fan, Chang, Hsieh, Wang, & Lin, 2008), was used to evaluate whether multi-voxel patterns in each of the eight ROIs carry information related to the semantic category (living, non-living) of the word in each of the conditions. Within-language (or language-specific) decoding involved restricting the analysis to trials of a specific language (either Spanish or Basque) while cross-language (or language-independent) decoding entailed training the classifier on trials from one language and testing it on trials from another language. Both of these analyses were done separately

for shallow and deep processing trials. Additional details related to the data preparation, feature selection, classification and statistics are presented in the following subsections.

2.6.1 Data Preparation

For each participant, the relevant time points or scans of the preprocessed fMRI data of each run were labeled with attributes such as word, category, language, and condition using Psychopy generated data files (CSVs). Invariant voxels (or features) were removed. These were the voxels/features whose value did not vary throughout the length of one run. If not removed, such features can cause numerical difficulties with procedures like z-scoring of features. Next, data from all ten runs were stacked and each voxel's time series was run-wise z-score normalized and linear detrended. Finally, following two recent cross-language generalization studies (Correia et al., 2014; Buchweitz et al., 2012), one example was created per trial by averaging the 4 volumes between the interval of 3.4s and 6.8s after the word onset, which corresponded to 1 second presentation of the word (see Figure 2). Importantly, this was the same in the shallow and deep processing conditions.

2.6.2 Pattern Classification

Linear support vector machine (SVM) classifier, with all parameters set to default values as provided by the scikit-learn package ($l2$ regularization, $C = 1.0$, $tolerance = 0.0001$), was used for both within- and cross-language decoding in both shallow and deep processing conditions. The following procedure was repeated for each ROI separately. To obtain an unbiased generalization estimate, following Varoquaux et al. 2016 (Varoquaux et al., 2016) the data was randomly shuffled and resampled multiple times to create 300 sets of balanced train-test (80%-20%) splits. Since each example was represented by a single feature vector with each feature a mean of voxel intensities across the sub-interval of 3.4s and 6.8s (see § 2.6.1), the length of a vector was equal to the number of voxels in the ROI. To further reduce the dimensionality of the data and thus reduce the chances of overfitting (Pereira, Mitchell, & Botvinick, 2009; Mitchell et al., 2004), Principal Component Analysis (PCA) with all parameters set to default values as provided by the scikit-learn was used. Since the `n_components` argument was set to `None`, the number of components was chosen to be the smaller from among the number of samples (m) and features (n). In our case, the n was always greater than m, hence, the first m components were selected. The size of the data matrix after PCA was therefore $m \times m$. These components were linear combinations of the preprocessed voxel data and since none of the components was excluded, it was an information loss-less change of the coordinate system to a subspace spanned by the examples (Mourão Miranda, Bokde, Born, Hampel, & Stetter, 2005). Features thus created were used to train the decoder, and its classification performance on the test set was recorded. This procedure was repeated separately for each of the 300 sets, and the mean of corresponding accuracies was collected for each of the participants. Note that PCA was performed on the training set; then the trained PCA was used to extract components in the test data and its classification performance was assessed. This procedure was repeated separately for each of the 300 sets, and the mean of corresponding accuracies was collected for each of the participants.

Our rationale to infer language-shared representations from the MVP classification analysis is based on the following logic: if a classier trained to discriminate stimulus classes

in context A (or language A) generalises to discriminate the stimulus classes of previously unseen items in context B, there are grounds to argue that the underlying representations are similar across the two contexts and the level of similarity is proportional to the level of generalization performance of the classifier.

2.6.3 Statistics

To determine whether the observed decoding accuracy in a given ROI is statistically significantly different from the chance-level of 0.5 (or 50%), a two-tailed t-test was performed with p-values corresponding to each of the ROIs corrected for multiple comparisons using a false discovery rate (FDR) method. To get the empirical estimate of chance-level, we ran the classification tests while randomly permuting over the category labels. The chance-level was computed across participants, ROIs, classification problems (within and cross-language) and conditions. For each case, 300 permutations were performed and the mean and standard deviation of the collected permutation scores was calculated across participants. For all ROIs, and classification problems, the chance-level was consistently found to be centered around 0.5. All effect sizes are reported as *mean effect size*±*standard error*, *t(degrees of freedom)=t-value, p-value* across all participants.

3 Results

3.1 Behavioral

To ensure that participants were attending to the items during the task, a few catch trials were randomly presented at different points in each run. These trials showed number words and required a response via button press. Further details related to the participants and procedure are provided in § 2. To ensure equal treatment of both conditions, the total number of catch trials ($mean = 6.8 \pm 1.6$) was kept equal in both shallow and deep processing runs. Catch trial data from two initial participants could not be obtained due to a technical issue. The proportion of correct responses on catch trials was 0.90 ± 0.13 in the shallow processing, and 0.93 ± 0.12 in deep processing conditions, which did not differ ($t(27) = 0.87, p = 0.39$), hence showing that participants were equally engaged with the task in both conditions.

3.2 FMRI-based MVPA Results

For each participant, we performed MVPA in 8 well-known left-lateralized semantic ROIs (see Figure S3). We asked whether shallow processing is sufficient for decoding the word semantic category within a given language and also to activate semantic representations that generalize across languages; or, whether higher depth of processing is needed for such cross-language generalization. Specifically, linear support vector machine (SVM) was used for classification of the semantic category in all ROIs in both shallow and deep processing conditions. Two different classification analyses were performed, namely within-language decoding and cross-language generalization. Both of these were performed separately for each of the conditions on each subject, and were restricted to eight pre-specified ROIs based on a prior meta-analysis (Binder et al., 2009). To determine whether the observed decoding accuracy in a specific ROI and condition is statistically significantly

above chance, a two-tailed t-test was performed. All t-tests reported below were corrected for multiple comparisons using FDR method.

3.2.1 Within-language Decoding

Within-language decoding was restricted to one language at a time whereby 80% of trials of that language were used to train the SVM-based classifier and the remaining 20% to test the learned model. Figures 4 and 5 therefore present the summary statistics of the ROIs for both shallow and deep processing conditions within Spanish and Basque respectively. It can be seen that in both the shallow and deep processing conditions, the decoding of the semantic category (living/non-living) was found to be statistically significantly above chance in almost all pre-specified ROIs (see Figures 4, 5 and Supplemental Results § S1 for statistics).

Deep processing also resulted in relatively higher decoding performance relative to the shallow processing condition in some of the ROIs. Specifically, deep processing was found to improve within-language decoding in IPL ($p = 0.004$), VTL ($p = 0.02$), and PCG ($p = 0.002$) for Spanish and IPL ($p = 0.001$), VTL ($p = 0.008$) and IFG ($p = 0.04$) for Basque. It can also be seen that an exception to this was ATL where decoding in the shallow condition was found to be higher than that in deep condition ($p = 0.002$ for Spanish, $p = 0.75$ for Basque).

We conducted further control analyses to address the following points. First, it may be argued that the decoding accuracy in the within-language classification could reflect low-level features of the items given that the same words (though different examples) were used in training and testing the classifier. We believe this is an unlikely explanation because we controlled for linguistic properties (i.e. length and frequency) of the items. Further, classification accuracy was quite distributed across the ROIs, including high-level semantic ROIs. Nevertheless, the within-language decoding analyses were re-run with the classifier trained on all words but one and tested on the left-out word. Similar results were observed, although the level of decoding accuracy was somehow weaker across ROIs and the within-language decoding was most evident in Spanish relative to Basque (see Supplemental Results § S3). It is possible that any seemingly stronger effect in Spanish may be due to the fact that most of our participants had Spanish as the first language. However, since this is not the focus of the study, this issue will not be discussed further. In summary, these results show that within-language decoding did not reflect low-level features of the words. Note that this issue does not apply in the case of cross-language generalization, which is the focus of the present study.

3.2.2 Cross-language Decoding

Cross-language generalization involved training the decoder on the examples of one language (training language) and testing it on the examples of the other language (test language). Figures 7 and 6 present summary statistics of the ROIs for Spanish to Basque and Basque to Spanish generalization respectively in both shallow and deep processing conditions. It can be seen that in the shallow processing condition, the cross-language generalization from both Spanish to Basque and Basque to Spanish was not different

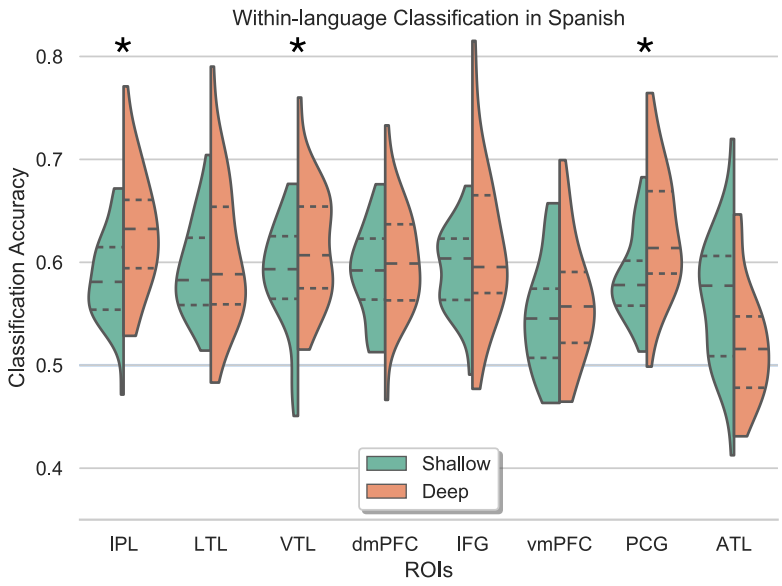

Figure 4: The figure shows summary statistics of the ROIs for within-language decoding in Spanish. It can be seen that the decoding was above chance in all ROIs in both conditions. The three dotted lines inside each violin are the quartiles. The black asterisks mark ROIs that showed statistically significant improvement in decoding accuracy in deep as compared to shallow processing condition. The p-values were corrected for multiple comparisons.

from chance-level in all pre-specified ROIs (see Supplemental Results § S2 for additional details).

In the deep processing condition on the other hand, the Spanish to Basque generalization was found to be statistically significantly above-chance and better than shallow condition (FDR corrected for multiple comparisons) in five out of eight ROIs including: IPL ($55.18 \pm 5.27, t(29) = 5.29, p = 2.99e-05$), LTL ($55.84 \pm 5.35, t(29) = 5.88, p = 1.78e-05$), VTL ($55.45 \pm 5.49, t(29) = 5.34, p = 2.99e-05$), dmPFC ($53.12 \pm 4.25, t(29) = 3.95, p = 0.0006$), IFG ($54.89 \pm 5.33, t(29) = 4.94, p = 6.00e-05$), vmPFC ($51.47 \pm 2.74, t(29) = 2.89, p = 0.008$), PCG ($53.57 \pm 4.62, t(29) = 4.15, p = 0.0004$), ATL ($50.01 \pm 4.69, t(29) = 0.02, p = 0.99$). Similarly, Basque to Spanish generalization was found to be statistically significantly above chance and better compared to shallow condition (FDR corrected for multiple comparisons) in four out of eight ROIs including: IPL ($54.50 \pm 5.12, t(29) = 4.74, p = 0.0002$), LTL ($54.47 \pm 5.72, t(29) = 4.21, p = 0.0005$), VTL ($55.34 \pm 6.36, t(29) = 4.52, p = 0.0003$), dmPFC ($53.05 \pm 4.38, t(29) = 3.75, p = 0.001$), IFG ($54.78 \pm 5.06, t(29) = 5.08, p = 0.0002$), vmPFC ($50.55 \pm 3.65, t(29) = 0.82, p = 0.48$), PCG ($53.03 \pm 5.95, t(29) = 2.74, p = 0.01$), ATL ($49.64 \pm 4.28, t(29) = -0.45, p = 0.65$)). Notably, above-chance cross-language generalization in the deep condition was not restricted to ROIs that showed superior within-language decoding as compared to the shallow condition (see Figures 4 and 5). We come back to this point in the Discussion.

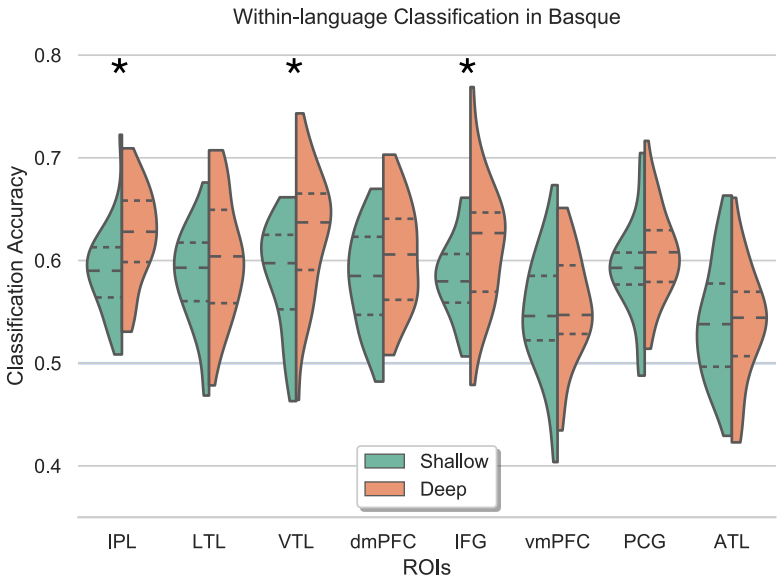

Figure 5: The figure shows summary statistics of the ROIs for within-language decoding in Basque. It can be seen that the decoding was above chance in all ROIs in both conditions. The three dotted lines inside each violin are the quartiles. The black asterisks mark ROIs that showed statistically significant improvement in deep as compared to shallow processing condition. The p-values were corrected for multiple comparisons.

The above results clearly show that cross-language generalization was stronger in the deep compared to the shallow processing condition. Because parametric statistical tests were used, additionally Shapiro-Wilk tests were run to check the normality assumption in the data. The results showed that normality assumption held in our dataset. Additionally, we also ran non-parametric statistical tests i.e. Wilcoxon signed-rank tests and found a similar pattern of results to those obtained using parametric t-tests. Furthermore, we also ran Bayesian analyses with all parameters set to default values in the JASP statistical package (Wagenmakers et al., 2018; Team et al., 2018)) to assess the extent of the evidence for the null hypothesis in the cross-language generalization in the shallow condition (see Supplemental Table S5). The results here showed that evidence for the null hypothesis in the shallow condition ranged from moderate to anecdotal in all of the ROIs. While this could be interpreted as non-conclusive evidence for the absence of generalization in the shallow processing case, the key observation is that generalization is far stronger in the deep relative to the shallow processing condition. The evidence for the alternative hypothesis in the deep processing context was found to be extreme in most of the ROIs (see Supplemental Table S6).

Given the evidence for ATL involvement as a semantic hub, we performed some additional analysis in the ATL. The results presented above showed that while significant

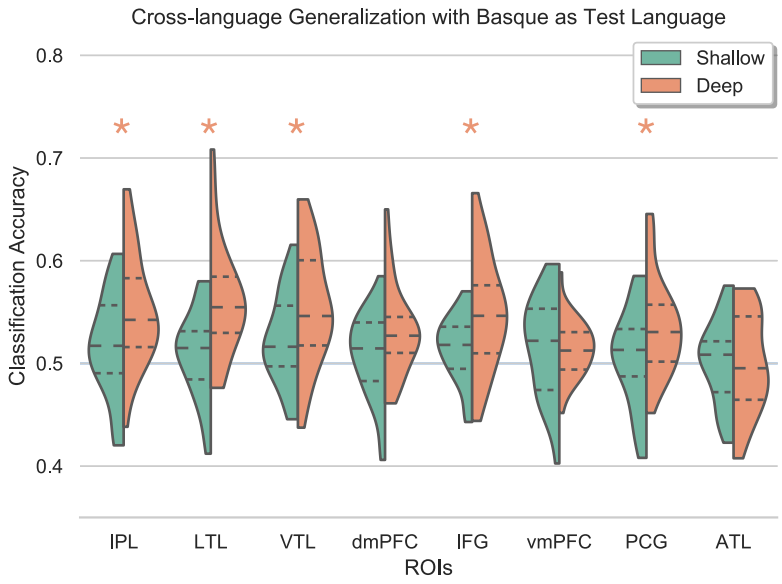

Figure 6: The figure shows summary statistics of the ROIs for cross-language generalization from Spanish to Basque in both shallow and deep processing conditions. It can be seen that whereas the generalization was not different from chance in all ROIs in the shallow condition, it was statistically significantly above-chance and better than shallow condition in deep condition in five out of eight ROIs including IPL, LTL, VTL, IFG and PCG. The three dotted lines inside each violin are the quartiles. The orange asterisks mark ROIs where cross-language generalization in deep was found to be statistically significantly above chance and better than shallow condition. The p-values were corrected for multiple comparisons.

decoding in the ATL was found in the shallow context, there was no evidence of cross-language generalization even during deep processing. This result was obtained with a mask of the ATL based on Freesurfer anatomical segmentation and is in keeping with the study of Damasio et al. 1996 (Damasio et al., 1996) and Correia et al. 2014 (Correia et al., 2014) in which cross-language generalization was found. However, there is a further, relatively more posterior ATL area that was also implicated as a multi-modal semantic hub (see Chen et al. 2017 (Chen, Ralph, & Rogers, 2017)). To re-run the decoding analysis on this area, we derived a 6 mm mask for each subject in native space based on registration from the corresponding MNI coordinates (-39, 18, -30), which lie between ROIs 3 and 8 in Figure S3. We found above-chance within-language decoding in both shallow (Spanish: $52.78 \pm 5.82, t(29) = 2.57, p = 0.02$; Basque: $53.48 \pm 6.00, t(29) = 3.12, p = 0.004$) and deep (Spanish: $54.60 \pm 4.48, t(29) = 5.53, p = 7.00e-06$; Basque: $54.14 \pm 5.08, t(29) = 4.40, p = 0.0001$) conditions with no significant differences between them ($p = 0.20$ for Spanish, and 0.63 for Basque), chance-level cross-language generalization was found in both conditions (shallow: $51.19 \pm 3.82, t(29) = 1.67, p = 0.19$ for Spanish to Basque and $50.66 \pm 3.52, t(29) = 1.01, p = 0.56$ for Basque to Spanish generalization;

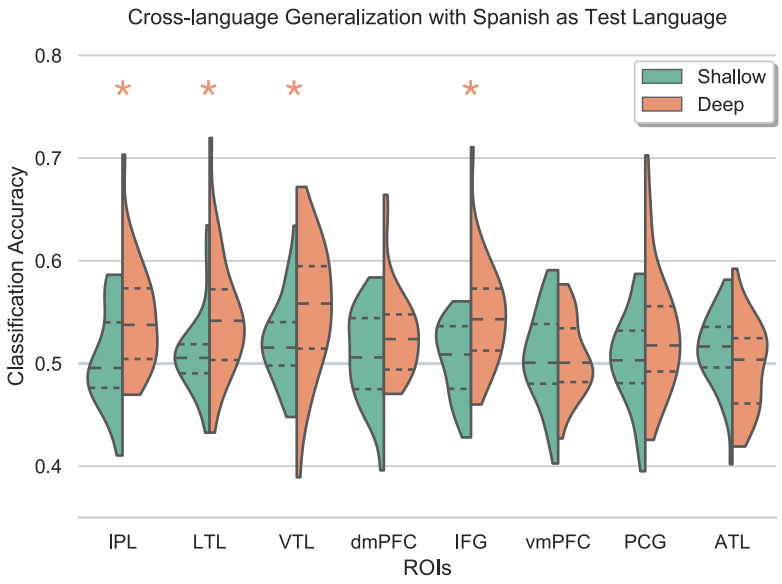

Figure 7: The figure shows summary statistics of the ROIs for cross-language generalization from Basque to Spanish in both shallow and deep processing conditions. It can be seen that while the generalization was not different from chance in all ROIs in the shallow condition, it was statistically significantly above-chance and better than shallow condition in deep condition in four out of eight ROIs including IPL, LTL, VTL and IFG. The three dotted lines inside each violin are the quartiles. The orange asterisks mark ROIs where cross-language generalization in deep was found to be statistically significantly above chance and better than shallow condition. The p-values were corrected for multiple comparisons.

deep: $50.16 \pm 4.15, t(29) = 0.20, p = 0.95$ and $50.03 \pm 4.21, t(29) = 0.03, p = 0.97$).

Next, we explore several factors that may account for the apparent absence of cross-language generalization in the shallow condition.

We wondered whether cross-language generalization in the shallow condition may be related to inter-individual differences in language proficiency scores in BEST and LeX-TALE tests. Hence, we assessed the correlation of language proficiency and cross-language decoding accuracy in the different ROIs. Specifically, we expected that balanced bilinguals, namely, participants with minimal difference in Spanish and Basque proficiency scores would display increased cross-language generalization accuracy (mean generalization scores across Spanish to Basque and vice versa). However, we did not find reliable evidence in support of this hypothesis (see Supplemental Results § S6).

It could be argued that the sub-interval of 3.4s-6.8s may not be the most optimal choice for creating examples (see § 2.6.1). As mentioned above, the choice of this time

window was based on previous cross-language generalization studies (Buchweitz et al., 2012; Correia et al., 2014) and standard guidelines in the field of fMRI-based multivariate pattern decoding (Pereira et al., 2009). However, we also re-ran the whole analysis taking the average of 2 volumes across the sub-interval of 4.25s and 5.95s. We found both within-language and cross-language generalization to be similar to those obtained using the sub-interval of 3.4s and 6.8s. Specifically, cross-language generalization was again found to be at chance-level in all ROIs in the shallow condition.

It could also be argued that information critical for cross-language generalization in the shallow condition is stored in spatially distributed, remote brain areas (Yamashita, Sato, Yoshioka, Tong, & Kamitani, 2008). Given that our ROI-based approach restricted the MVP analysis to one ROI at a time, it remains possible that significant cross-language generalization in the shallow condition is observed with a bigger ROI. To investigate this, we combined the data from all eight ROIs and repeated the analysis in the shallow condition. We found above-chance within-language decoding (Spanish: $59.12 \pm 4.47, t(29) = 10.98, p = 7.61e-12$; Basque: $58.24 \pm 4.88, t(29) = 9.10, p = 5.36e-10$), but cross-language decoding was not different from chance during Spanish to Basque generalization ($51.37 \pm 4.77, t(29) = 1.54, p = 0.13$) and Basque to Spanish generalization ($51.19 \pm 4.46, t(29) = 1.43, p = 0.16$).

Conversely, it could also be argued that the pre-specified ROIs were relatively large and the PCA merged features that were irrelevant for further classification analysis (Van Schooten, Harel, Ercan, & De Groot, 2014). This could be suggested as one possible reason for chance-level cross-language generalization in the shallow condition. In an attempt to address this point, the 8 ROIs were further subdivided into 15 more fine-grained ROIs based on individual anatomically segmented masks from Freesurfer (i.e. including inferior parietal lobe, inferior temporal lobe, medial temporal lobe, fusiform gyrus, parahippocampal gyrus, superior frontal gyrus, pars opercularis, pars orbitalis, pars triangularis, lateral orbitofrontal cortex, medial orbitofrontal cortex, posterior cingulate gyrus, precuneus, and anterior temporal lobe). Then, the same MVP analysis was repeated. However, cross-language generalization in the shallow condition was not different from chance in all ROIs for both Spanish to Basque and Basque to Spanish generalization, while crucially generalization was significantly above chance in the deep condition in a number of ROIs located in the anatomical spaces of the 8 ROIs described above (see Supplemental Results § S5).

4 Discussion

An important question in psychology and neuroscience is whether the acquisition of different languages is integrated within the same neurocognitive system and include shared/overlapping representations, or whether different languages are represented in separate brain systems. Previous investigations did not address the factors that underlie the generalization of semantic representations across languages. Hence it remained to be determined whether and how semantic representations generalise across languages. This fMRI study provides novel insights into this issue by uncovering how the depth of processing during semantic tasks influences within-language decoding of word category

and cross-language generalization based on multivoxel patterns of BOLD responses in putative substrates of the semantic network.

We found that the semantic category of words could be significantly decoded above chance levels when both Spanish and Basque languages were considered separately in all pre-specified semantic areas based on a prior meta-analysis (Binder et al., 2009). This happened even under shallow processing conditions when participants were merely asked to attend and read the words. However, the decoding performance was significantly better in deep compared to shallow processing in IPL, VTL, and PCG for Spanish and IPL, VTL, and IFG for Basque. The superior decoding performance in the deep relative to shallow processing condition aligns with other recent observations in our laboratory (Soto, Sheikh, Mei, & Santana, 2019) and indicates that the task requirement had an impact on the brain representation of meaning.

Cross-language generalization was not different from chance in all ROIs during shallow processing conditions (see also (Sheikh, Carreiras, & Soto, 2019)). Only in the context of deep information processing did brain activity patterns reliably generalize from Spanish to Basque in several brain regions (from Spanish to Basque: IPL, LTL, VTL, dmPFC, IFG, and PCG; from Basque to Spanish: IPL, LTL, VTL and IFG) known to be involved in semantic processing. For instance, the left IPL has been found to allow cross-language generalization in fMRI studies using visual (Buchweitz et al., 2012), auditory word comprehension with concrete nouns (Correia et al., 2014) and also during narrative comprehension task (Dehghani et al., 2017). PCG, and dmPFC have previously been found in cross-language generalization with visual stimuli (Buchweitz et al., 2012; Dehghani et al., 2017) but not in those using auditory stimuli (Correia et al., 2014). Similarly, LTL and VTL have been found to carry patterns that generalize across languages in studies using visual word comprehension (Buchweitz et al., 2012) as well as production tasks (Van de Putte, De Baene, Brass, & Duyck, 2017; Van de Putte, 2018).

It is worth noting that cross-language generalization in the deep condition was also found in ROIs which showed no difference in within-language decoding as function of the depth of processing. Specifically, multivoxel patterns in the lateral temporal lobe and dorsomedial prefrontal regions contained information that generalized across languages only in the context of a higher depth of processing but not during shallow processing, despite within-language decoding accuracy was the same in deep and shallow contexts. This pattern of results indicates that cross-language generalization in the deep processing case is not merely due to the increased signal to noise ratio of the multivoxel patterns corresponding to living and non-living items or merely based on modality-specific representations triggered by mental (e.g. visual) imagery processes occurring more strongly during deep relative to shallow processing (Soto et al., 2019). Our results also indicate that language-independent neural representations of semantic knowledge may not be easily generated during bottom-up information processing (i.e. automatically) but may require top-down strategic control processes (Stolz & Besner, 1999) such as those triggered during deep information processing and mental simulation.

The influential hub-and-spoke model suggests that sensory-motor representations of a concept are encoded in modality-specific brain regions (spokes), yet, unified and amodal

representations are formed within a single transmodal hub in anterior temporal lobes (ATL). On the other hand, the distributed-only model suggests that the higher-order generalizations from modality-specific (or language-specific) to amodal (or language-independent) semantic representations is not confined to a single semantic hub, rather distributed multiple brain regions are involved (Patterson et al., 2007; Ralph, Jefferies, Patterson, & Rogers, 2017). In our study, we found significant within-language decoding in ATL, yet cross-language generalization was not observed in this region in both shallow and deep conditions. These null results however must be taken with caution given that ATL is well-known to have susceptibility-induced signal dropout issues, and also considering the amount of evidence in the favour of the key role of ATL as a multi-modal semantic hub (Lambon Ralph, 2014). The critical finding however is that the cross-language generalization was found in multiple substrates of the semantic network. This is in keeping with previous neuroimaging studies (Buchweitz et al., 2012; Correia et al., 2014), though here we revealed the critical role of the depth of processing. We propose that the depth of information processing triggered the global sharing of information across a distributed set of brain areas implicated in semantic representation and this supported cross-language generalization. We suggest that the present results are in keeping with distributed-only views of semantic processing (Patterson et al., 2007).

We observed significant decoding of semantic category in inferior parietal, medial and inferior temporal and inferior frontal regions. These cortical association areas, also known as a transmodal cortex" (Luria, 1976), are thought to play a critical role in higher-order semantic processing (Lambon Ralph, 2014). Although the specific role of inferior frontal cortex still remains a topic of debate, previous studies indicate that it is not involved in the storage of semantic knowledge as such, but in semantic control (Thompson-Schill, D'Esposito, Aguirre, & Farah, 1997; Wagner, Paré-Blagoev, Clark, & Poldrack, 2001; Noonan, Jefferies, Corbett, & Lambon Ralph, 2010). In our study, word meaning could be decoded from patterns of activity in inferior frontal gyrus, namely, pars opercularis and pars triangularis, both within-, and also cross-languages. These results implicate this region in semantic representation (see also (Buchweitz et al., 2012; Shinkareva et al., 2011; Soto et al., 2019)). It is typically assumed that bilinguals are constantly switching between the two languages, selecting one and inhibiting the other based on task goals. However, it is hard to explain the within- and cross-language decoding of semantic categories based on this language switching account and semantic control view.

The present results have ramifications for psycholinguistic models of visual word recognition e.g. BIA+ (Dijkstra & Van Heuven, 2002). These models implement word processing in a purely bottom-up manner with parallel and non-selective (i.e. language independent) activation of linguistic codes not just at the level of semantics but orthography and phonology too. We propose that such models need to be revised to incorporate the influence of top-down factors related to the depth of processing. Our results indicate that non-selective access to word meaning across languages is not mandatory or intrinsic property of the semantic system. Instead, our results are in keeping with the view that depending on the depth of processing, the extent of parallel and non-selective access can be modulated. For instance, studies that did not encourage high depth of processing only found evidence for selective access (Rodriguez-Fornells, Rotte, Heinze, Nösselt, & Münte, 2002; Hoversten, Brothers, Swaab, & Traxler, 2015). More research is however needed

to elucidate the extent to which the depth of processing shapes how bilinguals access semantic representations, namely, the extent to which different language representations for a given concept are co-activated in parallel (Oppenheim, Wu, & Thierry, 2018) or whether, according to BIA+, bilinguals access to the lexical and semantic representation is delayed in the second language compared to the first language (Brysbaert, Van Wijnendaele, & Duyck, 2002). Furthermore, here we only used eight words per language in order to match them as much as possible in linguistic factors, however, the limited number of words imposes constraints on the scope of inferences that can be drawn about the neurocognitive architecture of the semantic system across different languages. Future studies using a larger corpus of words, time-resolved electrophysiology and computational models are needed to pinpoint the effect of the depth of processing and other task-related factors on the brain dynamics for accessing semantic representations in different languages. Ongoing work in the lab is being directed to test this view. An additional limitation of the present study may relate to the high sampling rate used (multiband acceleration factor of 6), which might have led to signal loss in some regions and geometric distortion that can affect the anatomical registration of the functional images. No field maps were obtained to correct for potential field inhomogeneities. However, inspection of our images did not reveal greater distortions compared to standard (i.e. no multiband) acquisitions. Additional research is needed to achieve a comprehensive evaluation of the relationship between acquisition parameters (MB factors, in-plane acceleration, voxel size, TR, flip angle) and MVPA decoding results, and benefits in event-related designs with short trial event have already been demonstrated through a comparison of multiband 2 and 3 (see (Demetriou et al., 2016), also (Chen et al., 2015)). Of note, however, the level of decoding performance in the present study was similar to previous MVPA decoding studies that used similar paradigms with standard MRI sequences (Shinkareva et al., 2011; Buchweitz et al., 2012; Correia et al., 2014).

References

Abutalebi, J., Cappa, S. F., & Perani, D. (2001). The bilingual brain as revealed by functional neuroimaging. *Bilingualism: Language and cognition*, *4*(2), 179–190.

Altarriba, J., & Basnight-Brown, D. M. (2007). Methodological considerations in performing semantic-and translation-priming experiments across languages. *Behavior Research Methods*, *39*(1), 1–18.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? the role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human perception and performance*, *10*(3), 340.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, *22*(4), 577–660.

Basnight-Brown, D. M., & Altarriba, J. (2007). Differences in semantic and translation priming across languages: The role of language direction and language dominance. *Memory & cognition*, *35*(5), 953–965.

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*(12), 2767–2796.

Brysbaert, M., Van Wijnendaele, I., & Duyck, W. (2002). On the temporal delay assumption and the impact of non-linguistic context effects. *Bilingualism: Language and Cognition*, *5*(3), 199–201.

Buchweitz, A., Shinkareva, S. V., Mason, R. A., Mitchell, T. M., & Just, M. A. (2012, Mar). Identifying bilingual semantic neural representations across languages. *Brain and language*, *120*(3), 282–9.

Chee, M. W., Soon, C. S., & Lee, H. L. (2003). Common and segregated neuronal networks for different languages revealed using functional magnetic resonance adaptation. *Journal of Cognitive Neuroscience*, *15*(1), 85–97.

Chen, L., Ralph, M. A. L., & Rogers, T. T. (2017). A unified model of human semantic knowledge and its disorders. *Nature human behaviour*, *1*(3), 0039.

Chen, L., Vu, A. T., Xu, J., Moeller, S., Ugurbil, K., Yacoub, E., & Feinberg, D. A. (2015). Evaluation of highly accelerated simultaneous multi-slice epi for fmri. *Neuroimage*, *104*, 452–459.

Correia, J. a., Formisano, E., Valente, G., Hausfeld, L., Jansma, B., & Bonte, M. (2014, Jan). Brain-based translation: fmri decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *34*(1), 332–8.

Craik, F. I., & Lockhart, R. S. (1972, December). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684. Retrieved from https://doi.org/10.1016/s0022-5371(72)80001-x doi: 10.1016/s0022-5371(72)80001-x

Crinion, J., Turner, R., Grogan, A., Hanakawa, T., Noppeney, U., Devlin, J. T., ... others (2006). Language control in the bilingual brain. *Science*, *312*(5779), 1537–1540.

Damasio, H., Grabowski, T. J., Tranel, D., Hichwa, R. D., & Damasio, A. R. (1996). A neural basis for lexical retrieval. *Nature*, *380*(6574), 499.

De Bruin, A., Carreiras, M., & Duñabeitia, J. A. (2017). The best dataset of language proficiency. *Frontiers in psychology*, *8*, 522.

De Groot, A. M., & Nas, G. L. (1991). Lexical representation of cognates and noncognates in compound bilinguals. *Journal of memory and language*, *30*(1), 90–123.

Dehghani, M., Boghrati, R., Man, K., Hoover, J., Gimbel, S. I., Vaswani, A., ... others (2017). Decoding the neural representation of story meanings across languages. *Human brain mapping*, *38*(12), 6096–6106.

Demetriou, L., Kowalczyk, O. S., Tyson, G., Bello, T., Newbould, R. D., & Wall, M. B. (2016). A comprehensive evaluation of multiband-accelerated sequences and their effects on statistical outcome measures in fmri. *BioRxiv*, 076307.

Dijkstra, T., & Van Heuven, W. J. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and cognition*, *5*(3), 175–197.

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013, Dec). Espal: one-stop shopping for spanish word properties. *Behavior research methods*, *45*(4), 1246–58.

Fan, R., Chang, K., Hsieh, C., Wang, X., & Lin, C. (2008). Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, *9*, 1871–1874.

Grainger, J. (1998). Masked priming by translation equivalents in proficient bilinguals. *Language and cognitive processes*, *13*(6), 601–623.

Grainger, J., & Beauvillain, C. (1988). Associative priming in bilinguals: Some limits of interlingual facilitation effects. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *42*(3), 261.

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009). Pymvpa: A python toolbox for multivariate pattern analysis of fmri data. *Neuroinformatics*, *7*(1), 37–53.

Hoversten, L. J., Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Language membership identification precedes semantic access: Suppression during bilingual word recognition. *Journal of Cognitive Neuroscience*, *27*(11), 2108–2116.

Indefrey, P. (2006). A meta-analysis of hemodynamic studies on first and second language processing: Which suggested differences can we trust and what do they mean? *Language learning*, *56*, 279–304.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. M. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, *17*(2), 825–841.

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical image analysis*, *5*(2), 143–156.

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*(2), 149–174.

Lambon Ralph, M. A. (2014). Neurocognitive insights on conceptual knowledge and its breakdown. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120392.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Lemhöfer, K., & Broersma, M. (2012). Introducing lextale: A quick and valid lexical test for advanced learners of english. *Behavior research methods*, *44*(2), 325–343.

Luria, A. R. (1976). *The working brain: An introduction to neuropsychology*. USA: Basic Books.

Mitchell, T., Hutchinson, R., Niculescu, R., Pereira, F., Wang, X., MA, J., & Newman, S. (2004, 01). Learning to decode cognitive states from brain images. , *13*, 667-668.

Mourão Miranda, J., Bokde, A. L. W., Born, C., Hampel, H., & Stetter, M. (2005, Dec). Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional mri data. *NeuroImage*, *28*(4), 980–95.

Noonan, K. A., Jefferies, E., Corbett, F., & Lambon Ralph, M. A. (2010). Elucidating the nature of deregulated semantic cognition in semantic aphasia: evidence for the roles of prefrontal and temporo-parietal cortices. *Journal of Cognitive Neuroscience*, *22*(7), 1597–1613.

Oppenheim, G., Wu, Y. J., & Thierry, G. (2018). Found in translation: Late bilinguals do automatically activate their native language when they are not using it. *Cognitive science*, *42*(5), 1700–1713.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), 976.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Peirce, J. W. (2007, May). Psychopy–psychophysics software in python. *Journal of neuroscience methods*, *162*(1-2), 8–13.

Perea, M., Dunabeitia, J. A., & Carreiras, M. (2008). Masked associative/semantic priming effects across languages with highly proficient bilinguals. *Journal of Memory and Language*, *58*(4), 916–930.

Perea, M., Urkia, M., Davis, C. J., Agirre, A., Laseka, E., & Carreiras, M. (2006, Nov). E-hitz: a word frequency list and a program for deriving psycholinguistic statistics in an agglutinative language (basque). *Behavior research methods*, *38*(4), 610–5.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fmri: a tutorial overview. *Neuroimage*, *45*(1), S199–S209.

Pruim, R. H., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J. K., & Beckmann, C. F. (2015). Ica-aroma: A robust ica-based strategy for removing motion artifacts from fmri data. *Neuroimage*, *112*, 267–277.

Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42.

Rodriguez-Fornells, A., Rotte, M., Heinze, H.-J., Nösselt, T., & Münte, T. F. (2002). Brain potential and functional mri evidence for how to handle two languages with one brain. *Nature*, *415*(6875), 1026.

Schoonbaert, S., Duyck, W., Brysbaert, M., & Hartsuiker, R. J. (2009). Semantic and translation priming from a first language to a second and back: Making sense of the findings. *Memory & cognition*, *37*(5), 569–586.

Sheikh, U. A., Carreiras, M., & Soto, D. (2019). Decoding the meaning of unconsciously processed words using fmri-based mvpa. *NeuroImage*, *191*, 430–440.

Shinkareva, S. V., Malave, V. L., Mason, R. A., Mitchell, T. M., & Just, M. A. (2011). Commonality of neural representations of words and pictures. *Neuroimage*, *54*(3), 2418–2425.

Simmons, W. K., & Barsalou, L. W. (2003). The similarity-in-topography principle: Reconciling theories of conceptual deficits. *Cognitive neuropsychology*, *20*(3-6), 451–486.

Smith, S. (2002). Fast robust automated brain extraction. *Human brain mapping*, *17*(3), 143–155.

Soto, D., Sheikh, U. A., Mei, N., & Santana, R. (2019). Decoding and encoding models reveal the role of the depth of processing in the brain representation of meaning. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/early/2019/11/21/830448 doi: 10.1101/830448

Stolz, J. A., & Besner, D. (1999, April). On the myth of automatic semantic activation in reading. *Current Directions in Psychological Science*, *8*(2), 61–65. Retrieved from https://doi.org/10.1111/1467-8721.00015 doi: 10.1111/1467-8721.00015

Stowe, L. A., & Sabourin, L. (2005). Imaging the processing of a second language: Effects of maturation and proficiency on the neural processes involved. *International Review of Applied Linguistics in Language Teaching*, *43*(4), 329–353.

Team, J., et al. (2018). *Jasp (version 0.8. 6)[computer software].*

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K., & Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. *Proceedings of the National Academy of Sciences*, *94*(26), 14792–14797.

Van de Putte, E. (2018). *The representation of language in bilinguals: Neural overlap*

as a function of modality, representational level, language proficiency and context (Unpublished doctoral dissertation). Ghent University.

Van de Putte, E., De Baene, W., Brass, M., & Duyck, W. (2017). Neural overlap of l1 and l2 semantic representations in speech: A decoding approach. *NeuroImage*, *162*, 106–116.

Van Hell, J. G., & De Groot, A. M. (1998). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Bilingualism: Language and cognition*, *1*(3), 193–211.

Van Schooten, S., Harel, R., Ercan, S., & De Groot, E. (2014). Applying feature selection methods on fmri data. *Student project report*.

Varoquaux, G., Raamana, P. R., A. Engemann, D., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2016, 06). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. , *145*.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ... others (2018). Bayesian inference for psychology. part ii: Example applications with jasp. *Psychonomic bulletin & review*, *25*(1), 58–76.

Wagner, A. D., Paré-Blagoev, E. J., Clark, J., & Poldrack, R. A. (2001). Recovering meaning: left prefrontal cortex guides controlled semantic retrieval. *Neuron*, *31*(2), 329–338.

Yamashita, O., Sato, M.-a., Yoshioka, T., Tong, F., & Kamitani, Y. (2008). Sparse estimation automatically selects voxels relevant for the decoding of fmri activity patterns. *NeuroImage*, *42*(4), 1414–1429.

Supplemental Results

S1 Within-Language Decoding

ROI	shallow		deep		deep - shallow
IPL	58.61±4.40	$p = 3.97\times10^{-11}$	62.98±5.85	$p = 8.19\times10^{-12}$	$p = 9.77\times10^{-3}$
LTL	59.45±4.54	$p = 1.27\times10^{-11}$	60.81±6.97	$p = 5.35\times10^{-9}$	$p = 0.45$
VTL	58.89±5.19	$p = 5.20\times10^{-10}$	61.62±5.57	$p = 1.21\times10^{-11}$	$p = 4.77\times10^{-2}$
dmPFC	58.80±4.89	$p = 5.40\times10^{-11}$	60.18±5.31	$p = 6.39\times10^{-11}$	$p = 0.44$
IFG	59.30±4.01	$p = 2.84\times10^{-12}$	61.21±7.38	$p = 6.69\times10^{-9}$	$p = 0.38$
vmPFC	55.01±5.64	$p = 4.59\times10^{-5}$	56.19±5.92	$p = 5.00\times10^{-6}$	$p = 0.44$
PCG	58.64±4.10	$p = 1.27\times10^{-11}$	62.84±6.07	$p = 1.21\times10^{-11}$	$p = 9.24\times10^{-3}$
ATL	56.57±6.29	$p = 5.05\times10^{-6}$	52.00±5.47	$p = 0.06$	$p = 9.24\times10^{-3}$

Table S1: The table presents within-language decoding results for Spanish in both shallow and deep processing conditions. The p-values were corrected for multiple comparisons.

ROI	shallow		deep		deep - shallow
IPL	58.95±4.18	$p = 1.30\times10^{-11}$	62.62±4.65	$p = 5.28\times10^{-14}$	$p = 0.01$
LTL	58.34±4.75	$p = 4.67\times10^{-10}$	60.44±5.86	$p = 2.29\times10^{-10}$	$p = 0.26$
VTL	58.50±5.27	$p = 1.95\times10^{-9}$	62.73±6.09	$p = 1.13\times10^{-11}$	$p = 0.03$
dmPFC	58.29±4.82	$p = 6.02\times10^{-10}$	60.56±5.12	$p = 1.16\times10^{-11}$	$p = 0.15$
IFG	58.45±4.00	$p = 1.30\times10^{-11}$	61.57±6.06	$p = 5.65\times10^{-11}$	$p = 0.11$
vmPFC	54.80±5.87	$p = 1.51\times10^{-4}$	55.54±5.12	$p = 2.88\times10^{-6}$	$p = 0.69$
PCG	59.12±4.93	$p = 1.94\times10^{-10}$	60.78±4.48	$p = 5.57\times10^{-13}$	$p = 0.30$
ATL	54.15±5.88	$p = 6.85\times10^{-4}$	53.71±5.63	$p = 0.001$	$p = 0.75$

Table S2: The table presents within-language decoding results for Basque in both shallow and deep processing conditions. The p-values were corrected for multiple comparisons.

S2 Cross-Language Generalization

ROI	shallow		deep		deep - shallow
IPL	50.62±4.74	p = 0.57	55.18±5.27	p = 2.99×10^{-5}	p = 0.03
LTL	51.04±4.08	p = 0.48	55.84±5.35	p = 1.78×10^{-5}	p = 0.02
VTL	52.22±4.21	p = 0.07	55.45±5.49	p = 2.99×10^{-5}	p = 4.94×10^{-2}
dmPFC	50.85±4.42	p = 0.57	53.12±4.25	p = 6.00×10^{-4}	p = 0.07
IFG	50.47±3.67	p = 0.57	54.89±5.33	p = 6.00×10^{-5}	p = 0.03
vmPFC	50.62±4.55	p = 0.57	51.47±2.74	p = 8.00×10^{-3}	p = 0.92
PCG	50.45±4.59	p = 0.60	53.57±4.62	p = 4.00×10^{-4}	p = 0.06
ATL	51.20±3.72	p = 0.37	50.01±4.69	p = 0.99	p = 0.92

Table S3: The table presents cross-language generalization results for Spanish to Basque generalization in both shallow and deep processing conditions. The p-values were corrected for multiple comparisons.

ROI	shallow		deep		deep - shallow
IPL	51.66±4.97	p = 0.16	54.50±5.12	p = 2.00×10^{-4}	p = 0.03
LTL	51.08±4.20	p = 0.28	54.47±5.72	p = 5.00×10^{-4}	p = 0.06
VTL	52.36±4.35	p = 0.05	55.34±6.36	p = 3.00×10^{-4}	p = 0.08
dmPFC	50.87±4.00	p = 0.34	53.05±4.38	p = 1.00×10^{-3}	p = 0.12
IFG	51.48±3.18	p = 0.07	54.78±5.06	p = 2.00×10^{-4}	p = 0.03
vmPFC	51.73±4.73	p = 0.16	50.55±3.65	p = 0.48	p = 0.96
PCG	50.67±4.73	p = 0.51	53.03±5.95	p = 0.01	p = 0.14
ATL	50.12±3.99	p = 0.88	49.64±4.28	p = 0.65	p = 0.13

Table S4: The table presents cross-language generalization results for Basque to Spanish generalization in both shallow and deep processing conditions. The p-values were corrected for multiple comparisons.

S3 Out of Sample Generalization

Figure S1 and S2 present the summary statistics of the ROIs for out-of-sample generalization in both shallow and deep processing conditions. It can be seen that in the shallow processing condition, the decoding of the semantic category (living/non-living) in Spanish was found to be above-chance (FDR corrected for multiple comparisons) in two out of eight ROIs including IPL (51.07±3.96; t(30) = 1.46; p = 0.27), LTL (52.46±4.35; t(30) = 3.05; p = 0.02), VTL (51.31±5.00; t(30) = 1.41; p = 0.27), dmPFC (51.50±4.61; t(30) = 1.76; p = 0.24), IFG (52.17±3.87; t(30) = 3.01; p = 0.02), vmPFC (50.29±4.64; t(30) = 0.34; p = 0.74), PCG (50.75±4.22; t(30) = 0.96; p = 0.39), ATL (51.31±5.47; t(30) = 1.29; p = 0.28). In Basque however, it was found to be at chance-level in all pre-specified ROIs including IPL (51.06±4.53; t(30) = 1.26; p = 0.82), LTL (50.57±4.85; t(30) = 0.63; p = 0.82), VTL (50.21±4.78; t(30) = 0.23; p = 0.82), dmPFC (50.23±4.29; t(30) = 0.29; p = 0.82), IFG (50.34±4.68; t(30) = 0.39; p = 0.82), vmPFC (49.66±5.29; t(30) = -0.34;

p = 0.82), PCG (51.48±4.50; t(30) = 1.77; p = 0.69), ATL (49.44±5.42; t(30) = -0.56; p = 0.82).

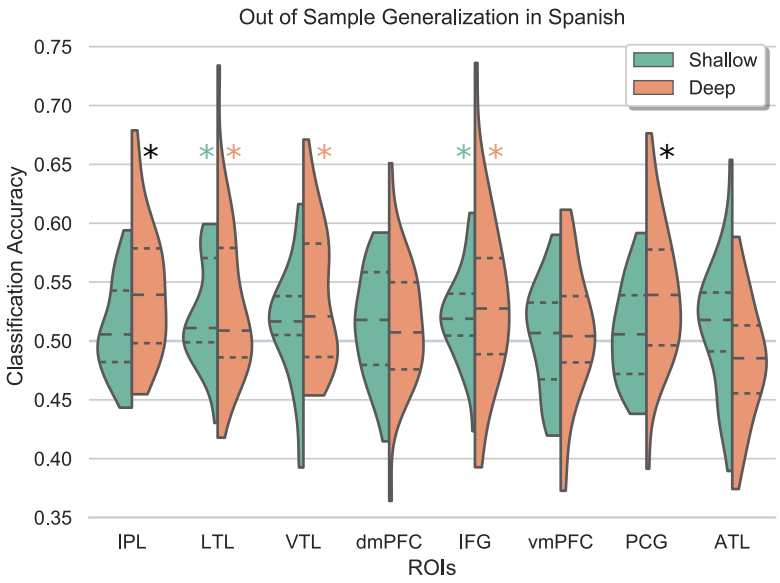

Figure S1: The figure shows summary statistics of the ROIs for out-of-sample generalization of the semantic category in Spanish. The three dotted lines inside each violin are the quartiles. The green and orange asterisks mark the ROIs that showed significantly above-chance performance in the shallow and deep conditions respectively and the black asterisks those with statistically significant improvement in deep as compared to shallow condition. The p-values were corrected for multiple comparisons.

On the other hand, in the deep processing condition, the decoding of the semantic category in Spanish was found to be above-chance and better than shallow condition (FDR corrected for multiple comparisons) in three out of eight ROIs including: IPL (54.39±5.57; t(30) = 4.24; p = 0.002), LTL (53.05±6.55; t(30) = 2.51; p = 0.029), VTL (53.77±5.80; t(30) = 3.49; p = 0.004), dmPFC (51.37±5.43; t(30) = 1.36; p = 0.21), IFG (53.37±7.22; t(30) = 2.51; p = 0.03), vmPFC (50.70±5.72; t(30) = 0.66; p = 0.51), PCG (53.95±5.98; t(30) = 3.56; p = 0.004), ATL (48.29±5.11; t(30) = -1.80; p = 0.11). In Basque however, it was found to be above-chance and better than shallow condition (FDR corrected for multiple comparisons) in one out of eight ROIs including: IPL (53.28±5.03; t(30) = 3.52; p = 0.006), LTL (51.87±5.73; t(30) = 1.76; p = 0.14), VTL (53.69±4.81; t(30) = 4.13; p = 0.002), dmPFC (51.45±5.33; t(30) = 1.47; p = 0.20), IFG (51.68±5.14; t(30) = 1.76; p = 0.14), vmPFC (50.11±4.84; t(30) = 0.12; p = 0.90), PCG (52.11±4.82; t(30) = 2.36; p = 0.07), ATL (49.53±4.23; t(30) = -0.59; p = 0.64).

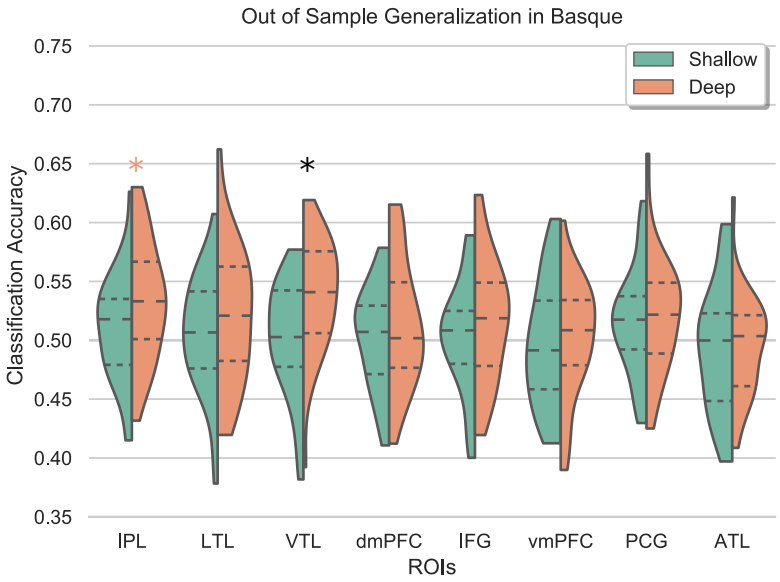

Figure S2: The figure shows summary statistics of the ROIs for out-of-sample generalization of the semantic category in Basque. The three dotted lines inside each violin are the quartiles. The orange asterisks mark those that showed above-chance performance in the deep condition and the black asterisks mark those with statistically significant improvement in deep as compared to shallow condition. The p-values were corrected for multiple comparisons.

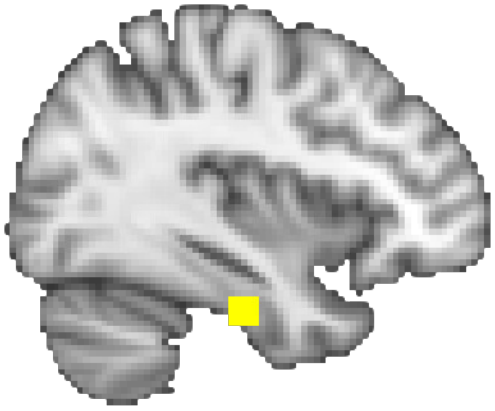

Figure S3: The figure shows another location of anterior temporal lobe projected on an MNI standard template image. It is a more posterior area previously implicated as a semantic hub by (Chen et al., 2017).

S4 Bayesian Analysis

S4.1 Cross-language Generalization

ROIs	Spanish	Basque
IPL	0.245; moderate	0.816; anecdotal
LTL	0.452; anecdotal	0.461; anecdotal
VTL	5.37; moderate support of H1	6.277; moderate support of H1
dmPFC	0.316; anecdotal	0.359; anecdotal
IFG	0.241; moderate	2.72; anecdotal support of H1
vmPFC	0.249; moderate	1.061; anecdotal
PCG	0.221; moderate	0.255; moderate
ATL	0.745; anecdotal	0.197; moderate

Table S5: Results of Bayesian analyses testing the evidence favor the null hypothesis in the cross-language generalization in the shallow condition, and the corresponding interpretation based on Lee and Wagenmakerś classification scheme. Regions in which the test moderately supported the alternative hypothesis (H1) are noted (Lee & Wagenmakers, 2014).

ROIs	Spanish	Basque
IPL	463.9; extreme	1902; extreme
LTL	124.1; extreme	8427; extreme
VTL	267.4; extreme	2165; extreme
dmPFC	40.91; very strong	66.76; very strong
IFG	1112; extreme	773.2; extreme
vmPFC	0.264; moderately support the null	5.961; moderate
PCG	4.389; moderate	108.8; extreme
ATL	0.214; moderately supports the null	0.194; moderately supports the null

Table S6: Results of Bayesian analyses testing the evidence favor the alternative hypothesis in the cross-language generalization in the deep condition, and the corresponding interpretation based on Lee and Wagenmakerś classification scheme (Lee & Wagenmakers, 2014).

S5 Cross-language Generalization with 15 ROIs

A set of 15 left-lateralized ROIs was pre-specified (see Figure S4) based on a meta-analysis of the semantic system by Binder et al. 2009 (Binder et al., 2009) and one anterior temporal lobe (ATL) due to its crucial role as a "semantic hub" (Damasio et al., 1996; Patterson et al., 2007; Correia et al., 2014). So, the ROIs included: inferior parietal lobe (IPL), inferior temporal lobe (ITL), middle temporal lobe (MTL), precuneus, fusiform gyrus (FFG), parahippocampal gyrus (PHG), superior frontal gyrus (SPG), posterior cingulate gyrus (PCG), pars opercularis (POP), pars triangularis (PTR), pars orbitalis (POR), frontal pole (FP), medial orbitofrontal cortex (MOFC), laterial orbitofrontal cortex (LOFC), and anterior temporal lobe (ATL).

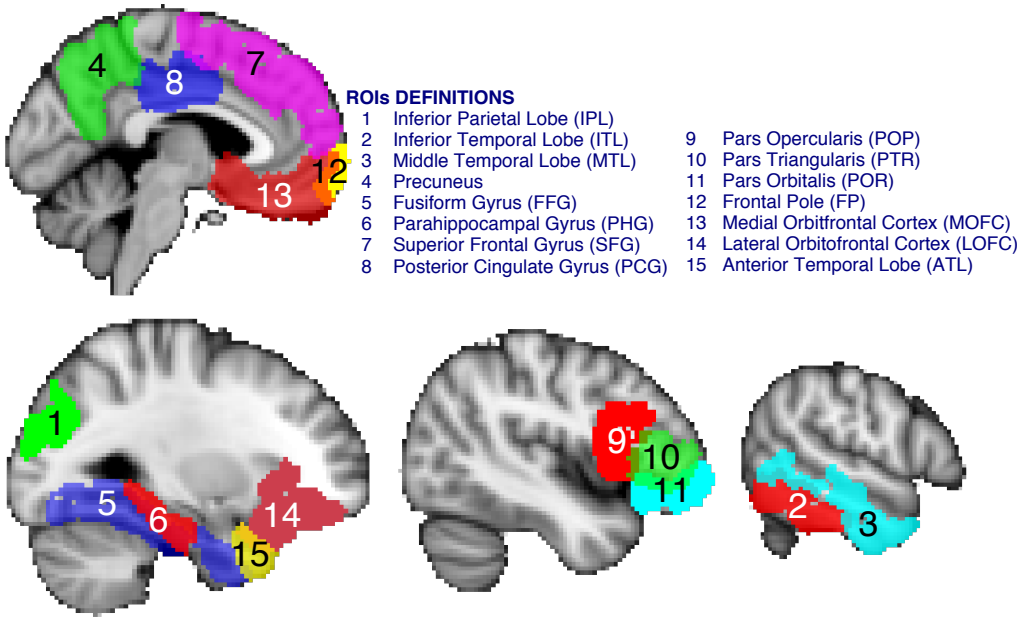

Figure S4: The figure shows the selected regions of interest projected on an MNI standard template image. The 15 left-lateralized areas were pre-specified and included regions: inferior parietal lobe, inferior temporal lobe, middle temporal lobe, precuneus, fusiform gyrus, parahippocampal gyrus, superior frontal gyrus, posterior cingulate gyrus, pars opercularis, pars triangularis, pars orbitalis, frontal pole, medial orbitofrontal cortex, laterial orbitofrontal cortex and anterior temporal lobe.

It can be seen that in the shallow processing condition, the cross-language generalization from Basque to Spanish (see Figure S6) was found to be not different from chance (FDR corrected for multiple comparisons) in all pre-specified ROIs including FP ($51.77 \pm 3.77; t(30) = 2.53; p = 0.26$), FFG ($51.71 \pm 4.41; t(30) = 2.09; p = 0.27$), IPL ($50.62 \pm 4.74; t(30) = 0.71 p = 0.74$), ITL ($50.76 \pm 4.14; t(30) = 0.99; p = 0.71$), LOFC ($50.39 \pm 3.81; t(30) = 0.55; p = 0.74$), MOFC ($50.36 \pm 4.70; t(30) = 0.42; p = 0.78$), MTL ($51.24 \pm 4.40; t(30) = 1.51; p = 0.42$), POP ($49.85 \pm 3.70; t(30) = -0.21; p = 0.86$), POR ($50.42 \pm 4.08; t(30) = 0.55; p = 0.74$), PTR ($50.47 \pm 3.88; t(30) = 0.65; p = 0.74$), PHG

$(51.65 \pm 4.42; t(30) = 2.01; p = 0.27)$, PCG $(49.84 \pm 4.76; t(30) = -0.18; p = 0.86)$, Precuneus $(50.74 \pm 4.89; t(30) = 0.81; p = 0.74)$, SFG $(50.86 \pm 4.43; t(30) = 1.05; p = 0.71)$, ATL $(51.20 \pm 3.72; t(30) = 1.74; p = 0.34)$. Similarly, the cross-language generalization from Spanish to Basque (see Figure S5) was also found to be not different from chance (FDR corrected for multiple comparisons) in all pre-specified ROIs including FP $(51.53 \pm 4.33; t(30) = 1.91; p = 0.35)$, FFG $(51.53 \pm 4.35; t(30) = 1.90; p = 0.35)$, IPL $(51.66 \pm 4.97; t(30) = 1.80; p = 0.35)$, ITL $(50.85 \pm 4.04; t(30) = 1.14; p = 0.38)$, LOFC $(51.00 \pm 3.99; t(30) = 1.35; p = 0.38)$, MOFC $(51.21 \pm 4.92; t(30) = 1.33; p = 0.38)$, MTL $(51.16 \pm 3.87; t(30) = 1.61; p = 0.35)$, POP $(51.31 \pm 4.173537473061428; t(30) = 1.69; p = 0.35)$, POR $(50.40 \pm 4.35; t(30) = 0.49; p = 0.72)$, PTR $(50.88 \pm 4.32; t(30) = 1.10; p = 0.38)$, PHG $(50.93 \pm 4.04; t(30) = 1.24; p = 0.38)$, PCG $(50.10 \pm 3.98; t(30) = 0.11; p = 0.91)$, Precuneus $(50.72 \pm 4.97; t(30) = 0.78; p = 0.55)$, SFG $(50.87 \pm 4.01; t(30) = 1.17; p = 0.38)$, ATL $(50.12 \pm 3.99; t(30) = 0.16; p = 0.91)$.

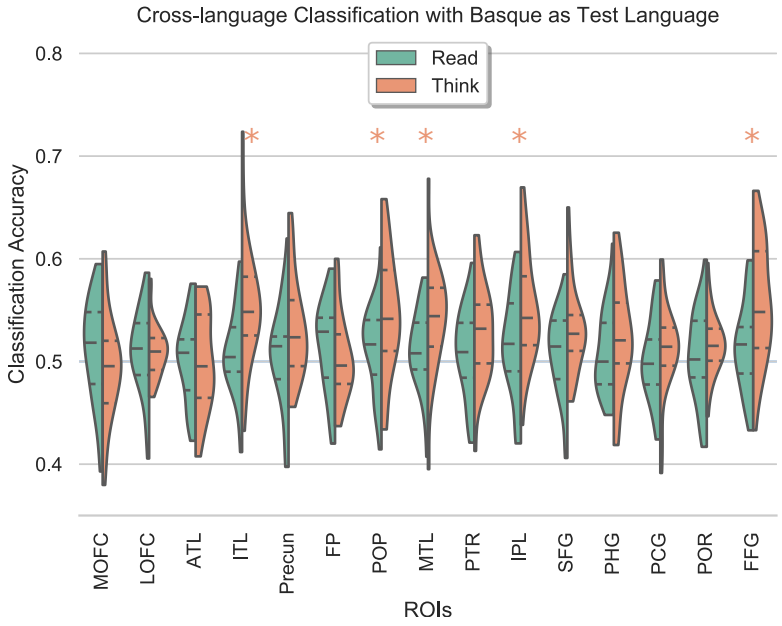

Figure S5: The figure shows summary statistics of the ROIs for cross-language generalization from Spanish to Basque in both shallow and deep processing conditions. It can be seen that while the generalization was at chance-level in all ROIs in the shallow condition, it was statistically significantly above-chance and better than shallow in deep condition in five out of fifteen ROIs including FFG, IPL, MTL, POP and ITL. The three dotted lines inside each violin are the quartiles. The orange asterisks mark ROIs where cross-language generalization in deep was found to be statistically significantly above chance and better than shallow condition. The p-values were corrected for multiple comparisons.

In the deep processing condition on the other hand, the Basque to Spanish generalization (see Figure S6) was found to be statistically significantly above-chance and better than shallow condition (FDR corrected for multiple comparisons) in 5 out of 15 ROIs

including FP ($50.30 \pm 4.21; t(30) = 0.38; p = 0.76$), FFG ($56.19 \pm 6.50; t(30) = 5.12; p = 0.0003$), IPL ($54.50 \pm 5.12; t(30) = 4.74; p = 0.0004$), ITL ($54.48 \pm 5.57; t(30) = 4.33; p = 0.0006$), LOFC ($50.24 \pm 2.97; t(30) = 0.43; p = 0.76$), MOFC ($50.13 \pm 4.46; t(30) = 0.15; p = 0.88$), MTL ($53.50 \pm 5.52; t(30) = 3.41; p = 0.004$), POP ($54.53 \pm 5.52; t(30) = 4.41; p = 0.0006$), POR ($51.91 \pm 4.12; t(30) = 2.49; p = 0.03$), PTR ($53.10 \pm 3.98; t(30) = 4.19; p = 0.0007$), PHG ($51.97 \pm 5.46; t(30) = 1.95; p = 0.09$), PCG ($50.46 \pm 4.55; t(30) = 0.55; p = 0.76$), Precuneus ($52.98 \pm 5.56; t(30) = 2.89; p = 0.01$), SFG ($53.04 \pm 4.39; t(30) = 3.73; p = 0.002$), ATL ($49.64 \pm 4.28; t(30) = -0.45; p = 0.76$). Similarly, Spanish to Basque generalization (see Figure S5) was found to be statistically significantly above chance and better compared to shallow condition (FDR corrected for multiple comparisons) in five out of fifteen ROIs including: FP ($50.36 \pm 4.07; t(30) = 0.48; p = 0.68$), FFG ($55.63 \pm 5.64; t(30) = 5.38; p = 8.41e-05$), IPL ($55.18 \pm 5.27; t(30) = 5.29; p = 8.41e-05$), ITL ($55.27 \pm 5.69; t(30) = 4.98; p = 0.0001$), LOFC ($51.05 \pm 2.56; t(30) = 2.21; p = 0.048$), MOFC ($49.45 \pm 5.04; t(30) = -0.59; p = 0.65$), MTL ($54.23 \pm 5.06; t(30) = 4.50; p = 0.0004$), POP ($54.43 \pm 5.86; t(30) = 4.08; p = 0.001$), POR ($51.72 \pm 2.97; t(30) = 3.11; p = 0.007$), PTR ($52.85 \pm 4.62; t(30) = 3.32; p = 0.005$), PHG ($52.43 \pm 5.28; t(30) = 2.48; p = 0.030$), PCG ($51.12 \pm 4.26; t(30) = 1.42; p = 0.21$), Precuneus ($53.26 \pm 4.79; t(30) = 3.66; p = 0.002$), SFG ($53.13 \pm 4.25; t(30) = 3.97; p = 0.001$), ATL ($50.01 \pm 4.69; t(30) = 0.02; p = 0.99$).

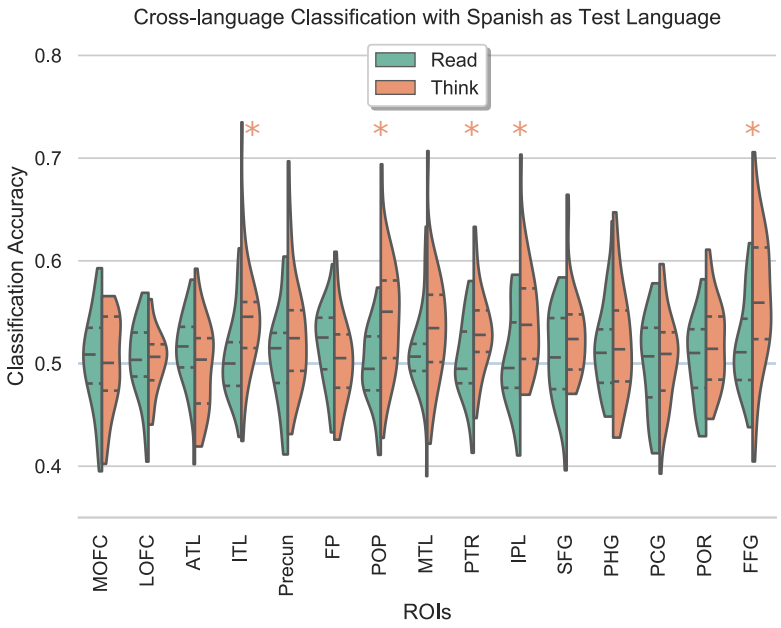

Figure S6: The figure shows summary statistics of the ROIs for cross-language generalization from Basque to Spanish in both shallow and deep processing conditions. It can be seen that while the generalization was at chance-level in all ROIs in the shallow condition, it was statistically significantly above-chance and better than shallow in deep condition in five out of fifteen ROIs including FFG, IPL, PTR, POP and ITL. The three dotted lines inside each violin are the quartiles. The orange asterisks mark ROIs where cross-language generalization in deep was found to be statistically significantly above chance and better than shallow condition. The p-values were corrected for multiple comparisons.

S6 Correlation between Cross-language Generalization and Language Proficiency

There were a few negative correlations between proficiency in Basque and Spanish indexed by the LeXTALE and cross-language generalization in LTL, IFG and dmPFC. However, these results should be taken with caution given that our study was not designed to explore inter-individual differences and that, while there were clear negative correlations, their statistical significance did not survive correction for multiple comparisons.

ROI	BEST scores	LeXTALE scores
IPL	$0.136; p = 0.497$	$-0.227; p = 0.255$
LTL	$-0.139; p = 0.489$	$-0.392; p = 0.043$
VTL	$-0.062; p = 0.759$	$-0.306; p = 0.120$
dmPFC	$-0.304; p = 0.123$	$-0.401; p = 0.038$
IFG	$-0.276; p = 0.164$	$-0.406; p = 0.036$
vmPFC	$-0.034; p = 0.866$	$-0.142; p = 0.479$
PCG	$-0.129; p = 0.522$	$-0.378; p = 0.052$
ATL	$0.026; p = 0.897$	$-0.282; p = 0.155$

Table S7: The table shows correlation between cross-language generalization score, and the difference between proficiency scores between Basque and Spanish in the shallow condition. The p-values are uncorrected.

S7 Semantic Analysis of the Stimuli

A matrix of word embeddings (word2vec) summarizing the semantic relationships between words within and across categories is presented in the Figure S7.

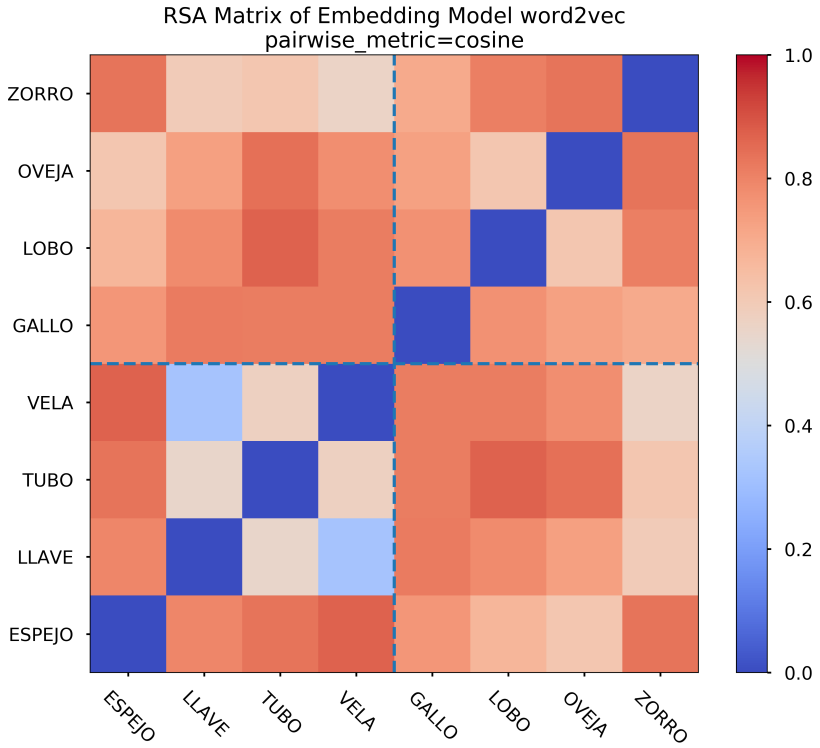

Figure S7: The figure shows a matrix of word embeddings i.e. word2vec summarizing the semantic relationships between stimuli within and across categories.