Decoding the Meaning of Unconsciously Processed Words Using fMRI-based MVPA

Usman Ayub Sheikh[a,*], Manuel Carreiras[a,b,c], David Soto[a,b,*]

[a] *Basque Center on cognition, Brain and Language (BCBL), 20009 Donostia, Spain.*
[b] *Ikerbasque, Basque Foundation for Science, Bilbao, Spain.*
[c] *University of the Basque Country, Bilbao, Spain.*

Abstract

Does the human brain elicit patterns of activity associated with the meaning of words in the absence of conscious awareness? Do such non-conscious semantic representations generalize across languages? This study aimed to address these questions using fMRI-based multivariate pattern analysis (MVPA) in a masked word paradigm. Animal and non-animal words were visually presented in two different languages (i.e. Spanish and Basque). Words were presented very briefly and were masked. On each trial, participants identified the semantic category and provided a visibility rating of the word. A support vector machine (SVM) was used to decode word category from multivoxel patterns of BOLD responses in seven canonical semantic regions of a left-lateralized network that were prespecified based on a previous meta-analysis. We show that the semantic category of non-conscious words (i.e. associated with null visual experience and chance-level discrimination performance) can be significantly decoded from BOLD response patterns. For Spanish, such discriminative patterns of BOLD responses were consistently found in inferior parietal lobe, dorsomedial prefrontal cortex, inferior frontal gyrus and posterior cingulate gyrus. While for Basque, these were found in ventromedial temporal lobe and posterior cingulate gyrus. All of the areas identified have previously been associated with semantic processing in studies involving animals-tools and animals-artifacts contrasts. In conscious trials, such patterns were found to be distributed over all seven regions of the semantic network in both Spanish and Basque. However, we found no evidence of across-language generalization. These results demonstrate that even in the absence of conscious awareness and lack of behavioural sensitivity to the words, putative semantic brain areas carry information related to the meanings of the words. The generalization of semantic representations across languages, however, may require deeper conscious semantic access.

Keywords: *unconscious, semantic processing, visual masking, machine learning, multivoxel pattern analysis, across-language*

*Correspondence:
Email addresses: u.sheikh@bcbl.eu (Usman Ayub Sheikh), d.soto@bcbl.eu (David Soto)

1. Introduction

Visual word processing entails activation of a number of different processes, these range from orthographic operations in low-level visual areas (posterior areas including left occipitotemporal region and superior temporal gyrus; Shaywitz et al. (2002), McCandliss et al. (2003), Dehaene and Cohen (2011), Booth et al. (2001), Lerma-Usabiaga et al. (2018)) to semantic processes in mainly high-level association areas (a left-lateralized network of 7 regions including dorsomedial prefrontal cortex, inferior frontal gyrus and ventromedial prefrontal cortex; see a meta-analysis Binder et al. (2009)) of the brain. An important question in the domain of non-conscious processing is: to what extent can the high-level cognitive processes unfold in the absence of conscious awareness [Dehaene and Changeux (2011), Sklar et al. (2012), van Gaal and Lamme (2012), Soto and Silvanto (2014)]. Whereas studies demonstrating non-conscious processing at relatively low levels of analysis (e.g. orthographic) are widely replicable and well-established now [Eriksen (1960), Johnston and Dark (1986), Greenwald (1992), Forster (1998)], most of the evidence implying non-conscious processing at higher levels (e.g. semantic) has been subject to many criticisms for reasons we discuss below (see also Kouider and Dupoux (2004), Kouider and Dehaene (2007)).

Ever since the seminal work by Marcel in 1980, non-conscious semantic processing has been investigated using visual masked priming paradigm. In a typical such experiment, participants engage in a lexical decision task, and non-conscious access to semantics is said to occur if they respond faster to targets preceded by a semantically-related unconscious prime (e.g. cat-dog) as compared to targets preceded by a semantically-unrelated unconscious prime (e.g. bag-dog). Initial studies using this paradigm were criticised on several grounds [Purcell et al. (1983), Holender (1986), Naccache et al. (2005), Dehaene (2014), Kouider et al. (2010)], the most prominent being the methodological shortcomings in how the threshold of prime awareness was established. In studies like Marcel (1983) and Devlin et al. (2004) for example, this threshold was established "offline", using separate blocks of detection trials prior to the semantic judgement trials, while in those including Abrams and Greenwald (2000), Abrams et al. (2002), Greenwald et al. (2003) and Nakamura et al. (2005), it was assessed after the semantic categorization task. These approaches do not assess sensitivity to the primes in an "online" manner at the time of prime-target presentation (see Dixon (1971); p. 18) and therefore are prone to either overestimation of awareness due to perceptual learning throughout the whole experiment [Schlaghecken et al. (2008)] or underestimation due to post-experiment fatigue and loss of motivation (Pratte and Rouder (2009); for a detailed review of such issues, see Lutz and Thompson (2003), Van den Bussche et al. (2013), and Haase and Fisk (2015)). Another important objection raised by Abrams and Greenwald (2000) and Damian (2001) regarding other semantic priming studies including Greenwald et al. (1996), Draine (1997), Dehaene et al. (1998b) and Draine and Greenwald (1998) is that the non-conscious semantic effects in such studies are explainable by a direct association mapping between the stimulus and the motor response (S-R mapping). Abrams and Greenwald argued that these experiments used the same set of words as primes and targets, and often with a strict response deadline, this enabled the brain to develop a shallow stimulus-response association that bypassed semantic analysis (but see Naccache and Dehaene (2001) for a study that circumvents this issue using a masked priming paradigm with number words). Another study by Gaillard

et al. (2006) presented masked emotional words with target-mask delay varied between the range of 33 milliseconds (ms) and 100ms. They presented emotionally negative (e.g. "pain") and neutral words (e.g. "color"), and collected response to the word naming task and a visibility rating (on a quasi-continuous visual scale) after each word presentation. They showed that emotional words enjoy a better access to consciousness as compared to neutral words which was intepreted as reflecting preferential non-conscious processing of emotional words. Although this study provided evidence for non-conscious semantic processing, emotional words were used which are known to be processed extraordinarily quickly and automatically [Ohman and Mineka (2001), Mineka and Öhman (2002), Whalen et al. (2004), Carretié et al. (2005)].

The goal of this study is to provide evidence of non-conscious semantic processing that circumvents the key issues noted above, most notably, the known difficulties in demonstrating the lack of awareness. First, we used a combination of moment-to-moment subjective reports of (un)awareness with signal detection measures and then analyzed the patterns of brain activity for words that observers rated as unaware and which critically were associated with null behavioural discrimination performance. This approach mitigates the concerns associated with the "offline" assessment of awareness which is standard in subliminal priming studies, even when objective measures of awareness based on signal detection theory are used. As noted above signal detection thresholds can vary across different testing sessions and thus any assessment of awareness must ideally occur concurrently with trials that will be used to demonstrate behavioral or neural evidence of unconscious information processing. In particular, here we sought to find out brain-based evidence that the semantic category of words was processed even though participants lack sensitivity to the relevant information. Accordingly, we used multivariate pattern analysis (MVPA) of functional MRI signals to decode the semantic category of the items. A similar approach has recently been taken by Axelrod et al. (2014), however, this study only involved the distinction between non-words and words embedded in sentences.

Thus, the first question we ask is whether the brain can encode the meaning of neutral words of animal and non-animal categories in the absence of conscious awareness. Additionally, we also aim to investigate the extent to which these non-conscious semantic representations of words are common or shared between different languages. Two recent fMRI studies Buchweitz et al. (2012) and Correia et al. (2014) showed that a decoder trained to classify the meaning of words in one language can predict with above-chance performance the meaning of words in the other language. Since they found shared patterns in well-known semantic areas of the brain, both studies claim to have pinned down language-independent semantic representations. This is an important line of research as it explores the existence of conceptual representations that are supposed to be more general and associated with language-free perceptual experience [Hauk et al. (2004)]. We argue that the key limitation of these studies is that the words were fully visible and participants were required to consciously think about the properties of the words. Therefore, it still remains to be seen whether such language-independent semantic representations can also emerge in the patterns of brain activity in the absence of conscious awareness and null behavioral sensitivity.

2. Materials and Methods

2.1. Participants

Twenty four early and proficient Spanish-Basque bilinguals (mean age 22.3±3.0 years; 17 female) including fourteen with Spanish as L1 were scanned using MRI. All of them had a normal or corrected to normal vision, gave written informed consent prior to the experiment and were financially compensated with 20 euros for their participation. The experiment lasted for about one and a half hour. Three of the participants were excluded before fMRI-based MVPA analysis. Two for failure to submit the category response in more than 50% of the trials, and one for failure to use the visibility ratings properly. The experiment was approved by the BCBL Ethics Review Board and complied with the guidelines of the Helsinki Declaration.

The online platform used for the recruitment (`www.bcbl.eu/participa`) also required participants to fill different questionnaires aimed at gathering information related to language proficiency of both languages. The collected data showed that all participants had acquired both languages before the age of 6. The mean age of acquisition was found to be 0.52 for Spanish and 1.05 for Basque with no statistically significant difference ($t(21) = -1.07, p = 0.30$). When considering their reported performance in the two well-known tests of language proficiency i.e. LexTALE (was available for only 20 out of 21 participants) [Lemhöfer and Broersma (2012)] and BEST [De Bruin et al. (2017)], statistically significant differences (LexTALE: $t(20) = 2.94; p < 0.05$, BEST: $t(21) = 5.15; p < 0.05$) were found between Spanish (LexTALE: 93.75 ± 4.62, BEST: 99.54 ± 1.13) and Basque (LexTALE: 87.22 ± 7.23, BEST: 86.46 ± 10.86). These scores thus show participants to be 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 subject, verb and object, morphology (Basque: agglutinative), syntax (Basque: ergative), and lexicon (many different vocabulary and non-cognates).

2.2. MRI Acquisition

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 eight functional images (corresponding to eight sessions). In each fMRI session, a multiband gradient-echo echo-planar imaging sequence with acceleration factor of 6, resolution of $2.4 \times 2.4 \times 2.4\text{mm}^3$, TR of 850ms, TE of 35ms and bandwidth of 2582Hz/Px was used to obtain 585 3D volumes of the whole brain (66 slices; FOV=210mm). The visual stimuli was 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 sessions.

2.3. Experimental Procedure

Each trial began with a fixation period of 500ms followed by a blank screen of another 500ms (see Figure 1). The target word, sandwiched between two 66ms circular white noise masks, was presented for 66ms and was followed by a response period of 3s. During this period, the participants were asked two questions, one after another, and were supposed to respond to each within the respective time window of 1.5s each. First, which semantic category does the word belong to, animals (A) or non-animals (nA)? To eliminate the effect of motor response difference on the choice of a semantic category, the mapping between choice and response button was randomly assigned on each trial. So, for some trials, A was on the right with nA on the left of the response screen, while for others, A was on the left with nA on the right. Participants were instructed to make their choice between left (i.e. button 1) and right (i.e. button 2) buttons based on the text displayed ("A nA" or "nA A") during the response period. Participants also provided an awareness rating of their visual experience of the word (1, 2 or 3); 1: I didn't see anything, 2: I think I saw a letter but not the word, or 3: I think I saw the word clearly or almost clearly. During training sessions, participants were given clear instructions that they were supposed to provide a forced-choice response to the category of the words in all the trials even for those in which they did not see the word at all (i.e. visibility rating of 1).

To ensure sufficient number of examples across the different states of visual awareness and to compensate for the changes in perceptual threshold across sessions [Gaillard et al. (2006)], the luminance of the words was varied based on an adaptive staircase procedure. Specifically, this procedure increased the value of luminance by 0.02 if the participant pressed 1, decreased it by 0.01 if he/she pressed 2, and decreased it by 0.02 if the he/she pressed 3 for the awareness rating in the previous trial. The starting point of the first session's staircase was based on a pre-experiment calibration session; for subsequent sessions, the final luminance from the previous session was used. The pre-experiment calibration involved running two staircases, first with a luminance step size of 0.1 and then with 0.02 (just like the experiment), and were used to determine a threshold of luminance that consistently coincided with the detection failure rate of 40% (or 40% of trials being labeled as 1-rating: "I did not see anything").

A total of 8 words were used in the whole experiment. In all of the blocks of all of the sessions of the experiment, the same 8 words were presented either in Spanish or in Basque. These were 4 animal words including wolf, rooster, fox, sheep, and 4 non-animal words including candle, key, tube and mirror (for Spanish and Basque translations, see Figure 2). All these 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 (animals and non-animals) and across languages (see Table 1 for details) based on the statistics provided by Espal (for Spanish; Duchon et al. (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 et al. (2011), Buchweitz et al. (2012), Correia et al. (2014)].

Both instructions and stimuli were presented at the center of the screen, in white against gray background and in all uppercase Arial font. The same stimuli was used for both the

calibration and the actual experiment.

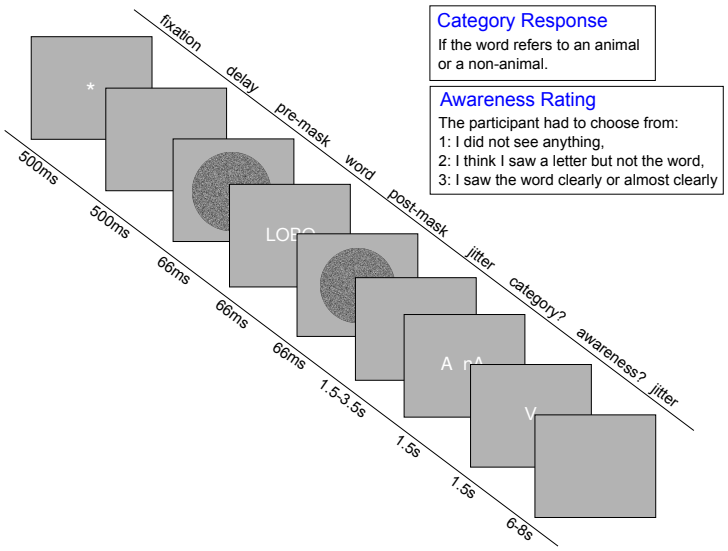


Figure 1: The figure summarizes the experimental design. A word was presented in the center of the screen for 66ms. This was both preceded and followed by circular white noise masks with each lasting for 66ms. Next, after a jittered interval of 1.5-3.5s, participants responded to two questions: 1. Which category from among animals and non-animals does the word belong to? and 2. Which awareness rating from among 1, 2, and 3 does best describe his/her perceptual awareness of the word? The inter-trial interval that followed these responses was a jittered interval of 6-8s.

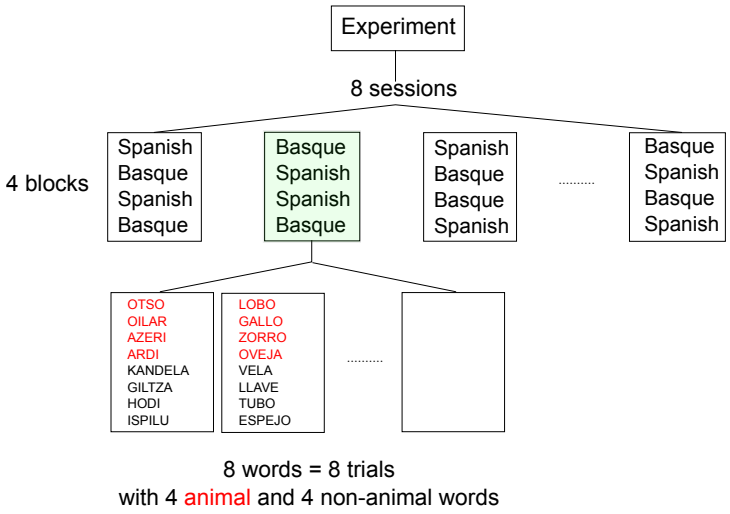


Figure 2: The figure summarizes the organization of sessions, blocks and trials in the experiment. Each experiment comprised of 8 sessions where each session was further subdivided into 4 language blocks (2 Spanish and 2 Basque). Each of these blocks was made up of 8 trials corresponding to single presentation of each of 4 animal and 4 non-animal Spanish/Basque words.

	Spanish		Basque	
	Animal	Non-animals	Animals	Non-animals
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 of stimuli i.e. 8 animal and non-animal words 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 were balanced across languages and categories.

The experiment was programmed and presented using Psychopy [Peirce (2007)] and is summarized in Figure 2. Each fMRI session was subdivided into four language blocks (see Figure 1) with two Spanish (S) and two Basque (B) blocks, the order of these blocks was counterbalanced across sessions (SBSB, BSBS, and so on). In each of these blocks, eight words were presented (without repetition) in a random arrangement resulting in a total of thirty two trials per session.

To maximize the separation between the brain activity corresponding to stimuli and that related to response, the interval between post-mask and response was jittered between 1.5s and 3.5s. Similarly, to further facilitate the estimation of HRF, the inter-trial interval (ITI) was also jittered between 6 and 8s. Both of these jitters were based on pseudo-exponential distributions resulting in 50% of trials with the ITI of 6s, 25% with 6.5s, 12.5% with 7s and so on.

2.4. 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 session's fMRI. To ensure steady state magnetisation, the first 9 volumes corresponding to the task instruction period were discarded; to remove non-brain tissue, FSL's brain extraction tool (BET) [Smith (2002)] was used; head-motion was accounted for using MCFLIRT [Jenkinson et al. (2002)]; minimal spatial smoothing was performed using a gaussian kernel with FWHM of 3mm and a high-pass filter with a cutoff of 90s (calculated by FEAT's "Estimate High Pass Filter Tool" based on the analysis of the frequency content of the design). The sessions were coaligned by aligning each session to a reference volume of the already preprocessed first session. Further analysis was performed in native BOLD space. However, to be able to transform the anatomical region of interest (ROI) masks generated using Freesurfer (see below for details), transformation matrices were obtained using linear registration of BOLD scans to the structural space (and vice versa) based on 7 DoF global rescale transformation.

A set of 7 left-lateralized ROIs were pre-specified (see Figure 3) based on a meta-analysis of the semantic system carried out by Binder et al. (2009). This meta-analysis is most relevant because it identifies the most critical semantic areas using only fMRI studies that used words as stimuli. These identified ROIs include: 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), and posterior cingulate gyrus

(PCG) along with precuneus. 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 session's functional space using FLIRT [Jenkinson and Smith (2001), Jenkinson et al. (2002)].

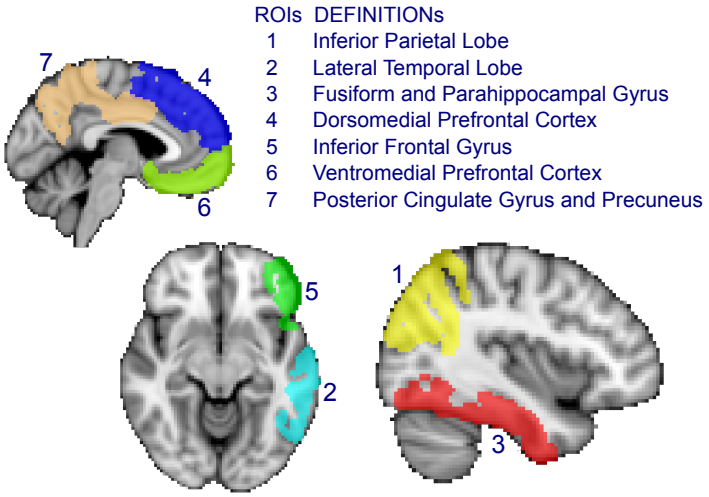


Figure 3: The figure shows the selected regions of interest projected on an MNI standard template image. These left-lateralized areas were pre-specified based on a meta-analysis by Binder et al. (2009) and included inferior parietal lobe (IPL), lateral temporal lobe (LTL), ventromedial temporal lobe (VTL) including fusiform gyrus and parahippocampal gyrus, dorsomedial prefrontal cortex (dmPFC), inferior frontal gyrus (IFG), ventromedial prefrontal cortex (vmPFC), and posterior cingulate gyrus (PCG) along with precuneus.

2.5. Multivariate Pattern Analysis

Multivariate pattern analysis were 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 et al. (2008)], was used to evaluate whether multi-voxel patterns in each of the seven ROIs carry information related to the semantic category (animal, non-animal) of the word in each state of awareness. Within-language (or language-specific) decoding involved restricting the analysis to trials of a specific language (either Spanish or Basque) while across-language decoding entailed training the classifier on trials from one language and testing it on trials from another language. Both of these analysis were done separately for each of the awareness conditions. Additional details related to the data preparation, feature selection, classification and statistics are presented in the following subsections.

2.5.1. Data Preparation

For each subject, the relevant time points or scans of the preprocessed fMRI data of each session were labeled with attributes such as category and language using a Python script

with corresponding Psychopy generated data files as input. The trial-by-trial awareness reports were used to separate the trials into 1-rating, 2-rating, and 3-rating trials. Invariant features were removed. These were the voxels/features whose value did not vary throughout the length of one session. If not removed, such features can cause numerical difficulties with procedures like z-scoring of features. Next, data from all eight sessions was stacked and each voxel's data points were session-wise z-score normalized and linear detrended. Finally, to account for the hemodynamic lag, one example was created per trial by averaging the 4 volumes between the interval of 3.4s and 6.8s after the word onset. Since the visibility rating of 1 represented the awareness report "I didn't see anything" and the mean behavioral performance in the corresponding 1-rating trials was also found to be at chance-level, these trials were considered as non-conscious trials. Similarly, trials with rating of 3 ("I think I saw the word clearly or almost clearly") were labeled as conscious trials. However, due to some participants having only a small number of 2-rating and others having a small number of 3-rating trials (see § 3.2), both 2-rating and 3-rating trials were collapsed and were considered to represent one condition. It is worth noting however that the rating of 2 ("I think I saw a letter but not the word") does not represent a conscious state. Hence, the resulting combination of both 2-rating and 3-rating conditions were labeled as partially conscious.

2.5.2. 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-language decoding and across-language decoding in both partially conscious and non-conscious. The following procedure was repeated for each ROI separately. To obtain an unbiased generalization estimate, following 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 § 2.5.1), 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 et al. (2009), Mitchell et al. (2004)], Principal Component Analysis (PCA) with all parameters set to default values as provided by the scikit-learn package (see `https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html`) was used. The number of components was equal to the number of examples thus resulting in all ROIs having equal number of components. 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 et al. (2005)]. Features thus created were used to train the decoder, and the 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 and averaged for each of the participants.

2.5.3. Statistics

To determine whether the observed decoding accuracy in an 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 across-language) and states of awareness. 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 Results

3.1.1. Awareness ratings were used properly

To establish whether the word in each trial was consciously perceived or not, participants were asked to submit both the objective categorization response (animal or non-animal) and the subjective visibility response (on the scale of 1 to 3; see § 2.3 for corresponding definitions) after each word presentation. Based on these responses, more than 40% of trials were found to be non-conscious (1-rating) in both Spanish ($41 \pm 4\%$) and Basque ($45 \pm 3\%$). Importantly, considering the objective performance in the animal vs. non-animal discrimination on these non-conscious trials, it was found to be at chance-level in both Spanish ($mean = 51 \pm 9\%; t(21) = 0.73; p = 0.47$) and Basque ($mean = 53 \pm 10\%; t(21) = 1.36; p = 0.18$; see Figure 4). Figure 4 also shows the objective performance for partially conscious trials. Specifically, it was found to be above chance for both 2-rating ($78 \pm 13\%; t(21) = 9.30; p < 0.05$ for Spanish and $74 \pm 14\%; t(21) = 7.67; p < 0.05$ for Basque), and 3-rating trials ($97 \pm 3\%; t(21) = 66.61; p < 0.05$ for Spanish and $97 \pm 4\%; t(21) = 49.99; p < 0.05$ for Basque). Taken together, this suggests that the participants used the awareness ratings correctly and the trials judged as not visible were genuinely invisible as per both subjective and objective behavioral measures.

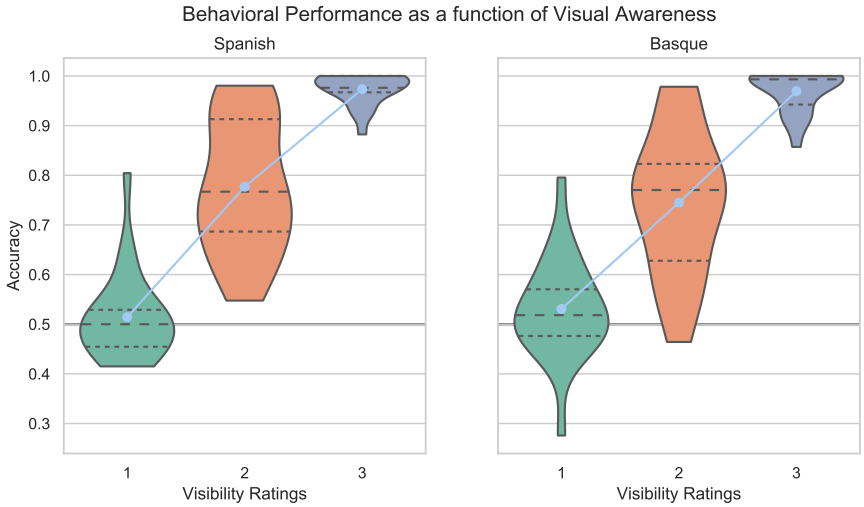


Figure 4: In trials with visibility rating of 1, the objective categorization performance was found to be at chance-level in both Spanish (left) and Basque (right). In those with visibility rating of 2, and 3, it was found to be clearly above chance in both the languages. The three dotted lines inside each violin are the quartiles. The black horizontal line in the background indicates the chance-level performance and the blue line shows the trend followed by the mean performance.

3.1.2. Stimulus strength was identical in all conditions

To compensate for variation in perceptual threshold across the experimental sessions, we decided to keep the adaptive luminance staircase (for details, see § 2.3) running throughout the whole experiment. The average luminance of the words however was found to be similar across the different visibility conditions (see Figure 5). One way ANOVA with three levels showed no statistically significant difference between conditions ($F(21) = 0.50; p = 0.61$ for Spanish, and $F(21) = 0.50; p = 0.66$ for Basque) and therefore can be used to conclude that the stimulus strength was similar in both partially conscious and non-conscious conditions.

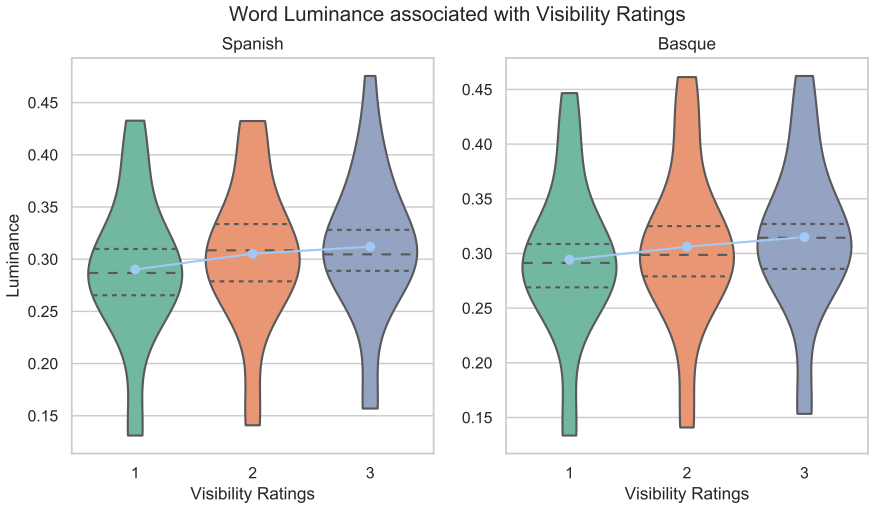


Figure 5: The figure shows the distribution of luminance values corresponding to each of the visibility ratings for both Spanish (left) and Basque (right). One way ANOVA with three levels showed no statistically significant difference between conditions. The three dotted lines inside each violin are the quartiles. The blue line passes through the means of the three distributions.

3.2. Brain Imaging Results

The primary goal of this study was to investigate whether the semantic category of non-conscious words can be predicted from BOLD response patterns, and which brain areas of the semantic network were involved. This classification problem was conducted separately for each language (henceforth within-language decoding). The second goal of the study was concerned with across-language generalization, namely, whether it is possible to decode the meaning of the non-conscious words in one language using a decoder trained to do the same in another language? [Buchweitz et al. (2012), Correia et al. (2014)].

Decoding was conducted separately for each of the awareness states. The average number of trials per subject was 82 for 3-rating (44 for Spanish, 38 for Basque), 65 for 2-rating (32 for Spanish, 33 for Basque), and 113 for 1-rating (54 for Spanish, 59 for Basque). Recall that since participants reported having no conscious awareness whatsoever in 1-rating trials and corresponding discrimination performance was also found to be at chance-level, these trials were considered as non-conscious trials. On the same lines, 3-rating trials (defined as "I saw a word clearly or almost clearly") reasonably qualified to be considered as conscious trials. However, since some participants had a small number of 3-rating trials ("I saw a word clearly or almost clearly"; $mean = 34\%; SD = 7\%$ in Spanish and $29\% \pm 7\%$ in Basque) and the others had a small number of 2-rating trials ("I think I saw a letter but not the word"; $mean = 25\%; SD = 10\%$ in Spanish and $26\% \pm 9\%$ in Basque), it was decided to combine both 2-rating and 3-rating trials and consider them as partially conscious trials (see § 2.5.1). The high variability in 2-rating and 3-rating trials was likely due to constraints imposed by our paradigm, namely the adaptive staircase procedure biased the luminance of the words to maximize the number of non-conscious trials. So, it is only after combining both 2-rating, and 3-rating trials that the mean and variability

became comparable to that of 1-rating trials and we obtained a more reasonable number for decoding.

3.2.1. Within-language Decoding

Within-language decoding involved restricting the SVM-based classification analysis to one language at a time, Figures 6 and 7 therefore present summary statistics of the ROIs for partially conscious and non-conscious conditions in both Spanish and Basque respectively. It can be seen that in non-conscious trials, considering Spanish results, the classification of the semantic category (animal/non-animal) was found to be statistically significantly above-chance in four out of seven ROIs including IPL ($mean = 54.7 \pm 7.0\%; t(21) = 2.97$; corrected $p = 0.02$; all p-values hereafter are FDR corrected), LTL ($53.0 \pm 7.4\%; t(21) = 1.82; p = 0.10$), VTL ($52.7 \pm 7.7\%; t(21) = 1.55; p = 0.14$), dmPFC ($56.1 \pm 6.4\%; t(21) = 4.25; p = 0.003$), IFG ($53.7 \pm 5.6\%; t(21) = 3.02; p = 0.02$), vmPFC ($53.9 \pm 7.6\%; t(21) = 2.27; p = 0.05$), and PCG ($54.0 \pm 7.5\%; t(21) = 2.42; p = 0.04$). In Basque, it was found to be above-chance in two of the seven ROIs including IPL ($mean = 52.8 \pm 6.5\%; t(21) = 1.94; p = 0.12$), LTL ($51.4 \pm 6.3\%; t(21) = 1.00; p = 0.38$), VTL ($54.4 \pm 6.5\%; t(21) = 3.04; p = 0.02$), dmPFC ($53.0 \pm 6.2\%; t(21) = 2.19; p = 0.09$), IFG ($51.8 \pm 6.0\%; t(21) = 1.34; p = 0.27$), vmPFC ($50.8 \pm 7.2\%; t(21) = 0.50; p = 0.07$), and PCG ($54.7 \pm 6.0\%; t(21) = 3.51; p = 0.02$).

In partially conscious trials, the classification of the semantic category was found to be statistically significantly above chance in all ROIs in both Spanish and Basque. Notably, while the decoding accuracies were similar in magnitude to that in non-conscious condition, above-chance accuracies were found to be distributed across all ROIs. For Spanish, these were: IPL ($mean = 54.1 \pm 4.5\%; t(21) = 4.08; p = 0.001$), LTL ($52.6 \pm 4.9\%; t(21) = 2.37; p = 0.03$), VTL ($52.9 \pm 4.4\%; t(21) = 2.93; p = 0.01$), dmPFC ($53.5 \pm 5.7\%; t(21) = 2.71; p = 0.02$), IFG ($55.9 \pm 4.8\%; t(21) = 5.51; p = 0.0002$), vmPFC ($52.3 \pm 4.7\%; t(21) = 2.24; p = 0.037$), and PCG ($55.7 \pm 5.7\%; t(21) = 4.51; p = 0.0008$). And for Basque, these were: IPL ($52.7 \pm 5.8\%; t(21) = 2.10; p = 0.049$), LTL ($54.8 \pm 4.7\%; t(21) = 4.59; p = 0.001$), VTL ($54.4 \pm 5.0\%; t(21) = 4.00; p = 0.002$), dmPFC ($52.8 \pm 5.4\%; t(21) = 2.28; p = 0.039$), IFG ($54.1 \pm 6.8\%; t(21) = 2.75; p = 0.02$), vmPFC ($52.9 \pm 4.7\%; t(21) = 2.66; p = 0.02$), and PCG ($53.9 \pm 6.3\%; t(21) = 2.80; p = 0.02$).

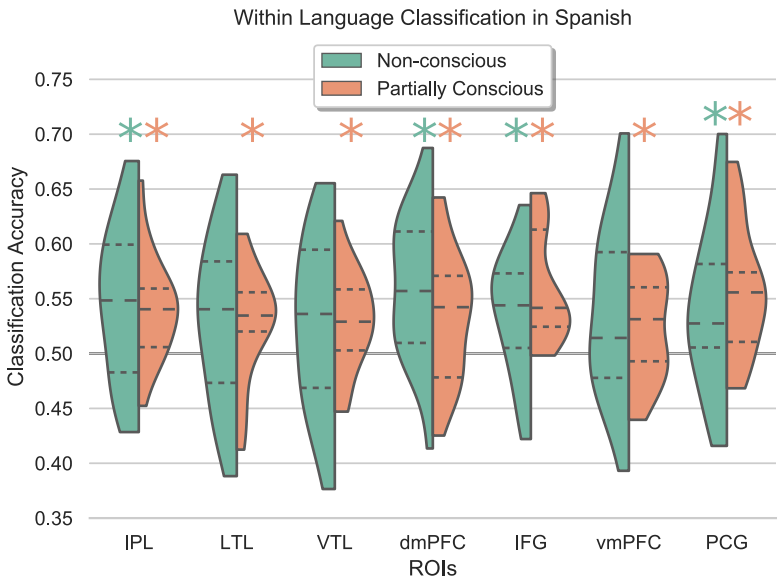


Figure 6: The figure shows the summary statistics of the ROIs for both partially conscious and non-conscious within-language decoding in Spanish. It can be seen that the decoding was above chance in four ROIs in non-conscious but all seven ROIs in partially conscious condition. The three dotted lines inside each violin are the quartiles. Orange and green asterisks signify statistically significantly above chance decoding in partially conscious and non-conscious conditions respectively. The black horizontal line in the background indicates the chance-level performance.

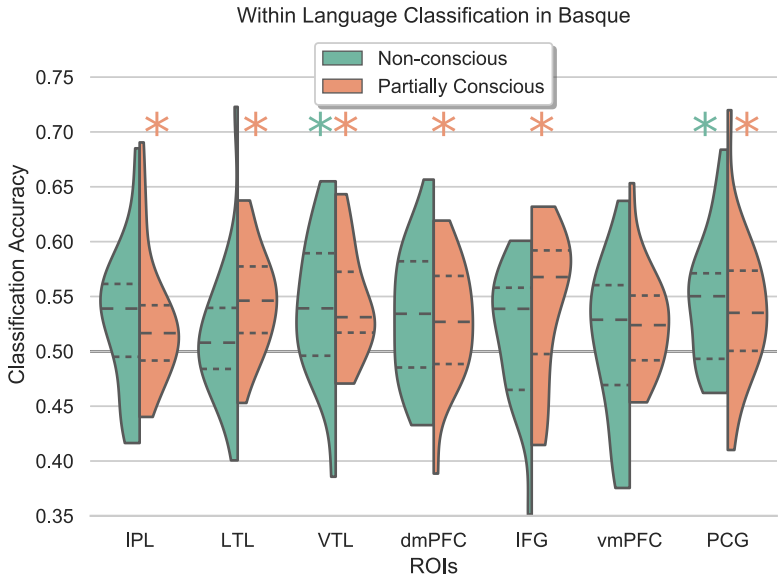
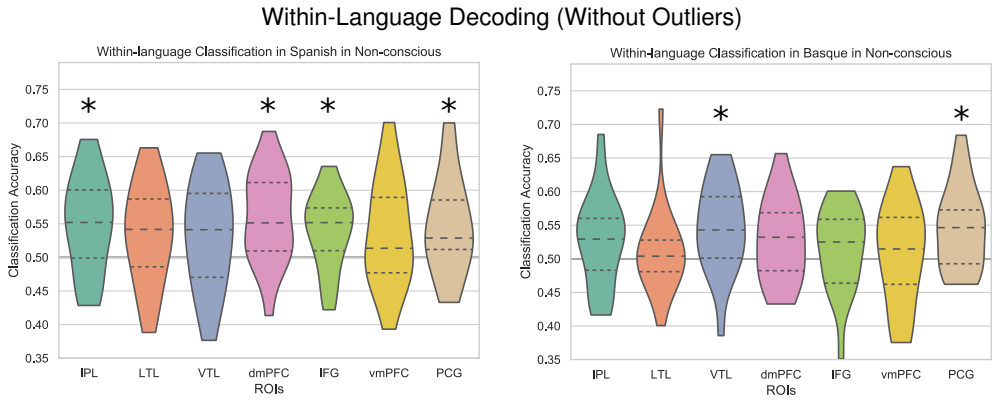


Figure 7: The figure shows the summary statistics of the ROIs for both partially conscious and non-conscious within-language decoding in Basque. It can be seen that the decoding was above chance in two ROIs in non-conscious but all seven ROIs in partially conscious condition. The three dotted lines inside each violin are the quartiles. Orange and green asterisks signify statistically significantly above chance decoding in partially conscious and non-conscious conditions respectively. The black horizontal line in the background indicates the chance-level performance.

It is noteworthy that there was one subject in Spanish and another in Basque whose behavioral discrimination performance was found to be around 80% in non-conscious condition. This is reminiscent of a "blindsight" effect or perception without awareness [Weiskrantz (1997), Sahraie et al. (1997)]. These represented outliers because their behavioral performance was 3 standard deviations higher than the mean. Although this was related to the behavioural performance, we wanted to ensure that these participants were not driving the above-chance decoding in non-conscious (see Figures 6 and 7). Therefore, within-language decoding procedure was re-run without including these outlier participants. Notably, it was found that that the pattern of results in the non-conscious remains intact in both Spanish (see Figure 8a) and Basque (Figure 8b). Specifically, in Spanish (see Figure 8a), the decoding of the semantic category (animal/non-animal) was again found to be statistically significantly above-chance in IPL ($mean = 55.0 \pm 7.0\%; t(21) = 3.08; p = 0.01$), dmPFC ($56.0 \pm 6.6\%; t(21) = 3.98; p = 0.006$), IFG ($54.0 \pm 5.6\%; t(21) = 3.08; p = 0.01$), and PCG ($54.7 \pm 7.1\%; t(21) = 2.86; p = 0.02$). Similarly in Basque (see Figure 8b), it was found to be above-chance in two of the seven ROIs i.e. VTL ($54.7 \pm 6.5\%; t(21) = 3.13; p = 0.02$) and PCG ($54.6 \pm 6.1\%; t(21) = 3.51; p = 0.02$).

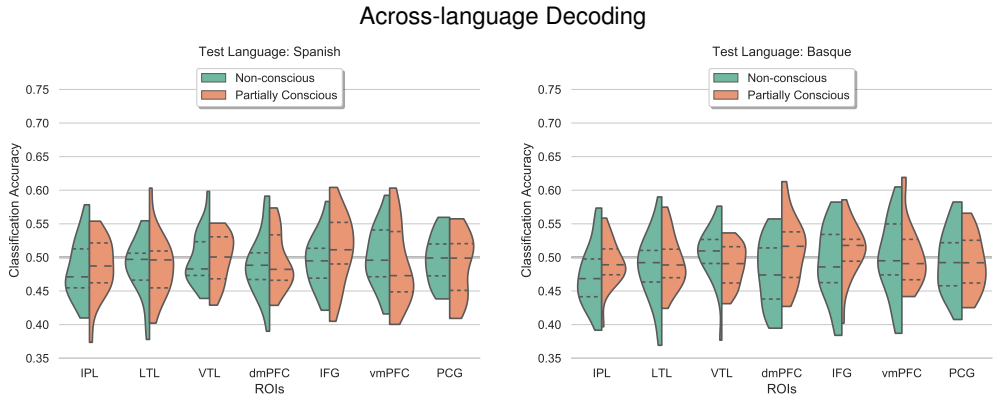


(a) (b)

Figure 8: There was one subject in Spanish and another in Basque whose behavioral discrimination performance was found to be around 80% in non-conscious condition. The figures show that the pattern of results in non-conscious decoding remained intact even after the removal of these outlier participants. Specifically, the same ROIs were found to be critical for the decoding of meaning in both Spanish (Figure 8a) and Basque (Figure 8b). The black asterisk signifies statistically significantly above chance decoding.

3.2.2. Across-language Decoding

Across-language decoding involved training the decoder on the examples of one language (train language) and testing it on the examples of the other language (test language). So, with Spanish as test language, Basque was the train language and vice versa. Figure 9 presents summary statistics of the ROIs for both partially conscious (for generalization from Spanish to Basque: IPL ($mean = 49.1 \pm 3.4\%; t(21) = -1.18; p = 0.49$), LTL ($49.4 \pm 3.8\%; t(21) = -0.69; p = 0.58$), VTL ($48.9 \pm 3.0\%; t(21) = -1.66; p = 0.49$), dmPFC ($51.0 \pm 4.8\%; t(21) = 0.96; p = 0.49$), IFG ($51.1 \pm 3.9\%; t(21) = 1.22; p = 0.49$), vmPFC ($50.3 \pm 4.6\%; t(21) = 0.30; p = 0.77$), and PCG ($49.1 \pm 4.1\%; t(21) = -1.04; p = 0.49$); for generalization from Basque to Spanish: IPL ($mean = 48.5 \pm 4.2\%; t(21) = -1.53; p = 0.51$), LTL ($48.6 \pm 4.5\%; t(21) = -1.43; p = 0.51$), VTL ($50.0 \pm 3.6\%; t(21) = -0.03; p = 0.98$), dmPFC ($49.5 \pm 4.1\%; t(21) = -0.57; p = 0.67$), IFG ($51.1 \pm 5.2\%; t(21) = 0.97; p = 0.51$), vmPFC ($48.8 \pm 5.6\%; t(21) = -0.93; p = 0.51$), and PCG ($48.8 \pm 4.6\%; t(21) = -1.13; p = 0.51$)) and non-conscious conditions (for generalization from Basque to Spanish: IPL ($mean = 48.3 \pm 4.4\%; t(21) = -1.76; p = 0.65$), LTL ($48.7 \pm 4.3\%; t(21) = -1.35; p = 0.67$), VTL ($49.8 \pm 3.8\%; t(21) = -0.22; p = 0.92$), dmPFC ($48.9 \pm 4.6\%; t(21) = -1.02; p = 0.75$), IFG ($49.4 \pm 4.0\%; t(21) = -0.70; p = 0.86$), vmPFC ($50.1 \pm 4.4\%; t(21) = 0.11; p = 0.92$), and PCG ($49.7 \pm 3.7\%; t(21) = -0.41; p = 0.92$); for generalization from Basque to Spanish: IPL ($mean = 47.4 \pm 4.6\%; t(21) = -2.53; p = 0.14$), LTL ($48.6 \pm 5.0\%; t(21) = -1.23; p = 0.54$), VTL ($50.4 \pm 4.1\%; t(21) = 0.40; p = 0.81$), dmPFC ($47.8 \pm 4.9\%; t(21) = -2.01; p = 0.20$), IFG ($49.2 \pm 5.4\%; t(21) = -0.70; p = 0.81$), vmPFC ($49.8 \pm 5.6\%; t(21) = -0.16; p = 0.88$), and PCG ($49.4 \pm 4.7\%; t(21) = -0.56; p = 0.81$)) and it can be seen that in both conditions, the across-language generalization was at chance-level in all ROIs.



(a) (b)

Figure 9: The figures show the summary statistics of the ROIs for both partially conscious and non-conscious across-language decoding. Specifically, the Figure 9a corresponds to when the decoder was trained on Basque and tested on Spanish, and the Figure 9b is for when the decoder was trained on Spanish and tested on Basque. From both of these figures, it is clear that the across-language generalization was at chance-level in both partially conscious and non-conscious conditions. The three dotted lines inside each violin are the quartiles. The black horizontal line in the background indicates the chance-level performance.

In a further test of across-language generalization, we combined the data from all ROIs in order to potentially increase the chances of decoding. However we found chance-level across-language generalization in both partially conscious ($mean = 49.7 \pm 3.5\%; p = 0.69$ for Spanish and $49.5 \pm 4.2\%; p = 0.57$ for Basque) and non-conscious conditions ($48.8 \pm 5.6\%; p = 0.36$ for Spanish and $mean = 49.0 \pm 4.5\%; p = 0.34$ for Basque). Within-language decoding was however statistically above chance-level in both partially conscious ($54.4 \pm 5.4\%; p = 0.001$ for Spanish and $mean = 55.2 \pm 5.5\%; p = 0.0004$ for Basque) and non-conscious conditions ($53.2 \pm 6.6\%; p = 0.04$ for Spanish and $mean = 55.0 \pm 6.9\%; p = 0.004$ for Basque) similar to the results found before.

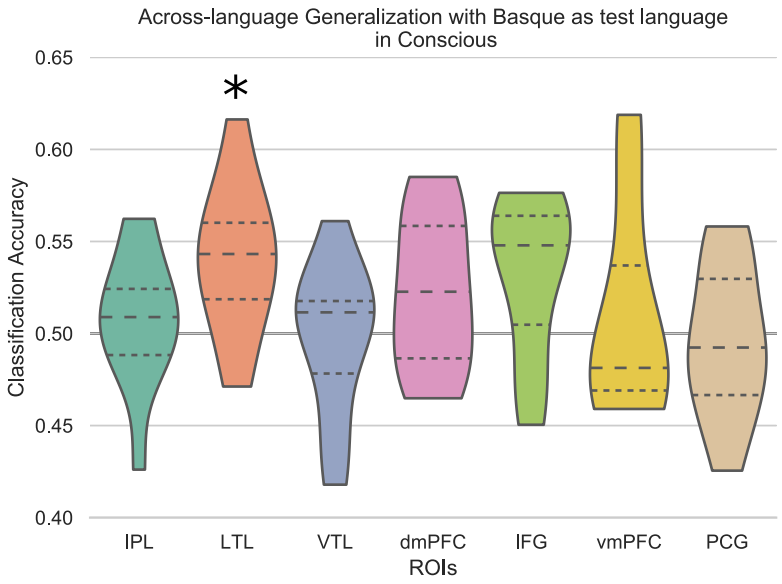


Figure 10: The figure shows the summary statistics of the ROIs for across-language generalization from Spanish to Basque. Only participants with relatively high within-language decoding performance were included. It can be seen that one ROI showed statistically significantly above-chance generalization from Spanish to Basque. The three dotted lines inside each violin are the quartiles. The black horizontal line in the background indicates the chance-level performance.

It could be argued that the absence of across-language decoding could be due to a floor effect, namely, given that classification accuracy was just above chance in the within-language decoding, it could only drop to chance level in the across-language generalization. To mitigate the presence of floor effects that could affect the ability to find across-language generalization, we ran an analysis including only those participants that had relatively high within-language classification performance (i.e. greater than or equal to 60% in the unconscious condition). There were 12 participants that satisfied this criteria for Spanish, and 7 that did it for Basque. Notably, in fully conscious condition, we found that there were two ROIs that showed statistically significantly above-chance across-language generalization from Spanish to Basque i.e. LTL ($54.0 \pm 3.9\%; p = 0.006; t(12) = 3.44$) and IFG ($53.1 \pm 4.1\%, p = 0.028, t(12) = 2.52$), with the LTL surviving the correction for multiple comparisons (see Figure 10). The generalization from Basque to Spanish ($N = 7$) was found to be at chance-level in all ROIs. We did not find any evidence of across-language generalization in partially conscious and non-conscious trials.

Finally, we also addressed decoding accuracy on fully conscious 3-rating trials. Because of the constraints in the number of 3-rating trials across participants (see § 3.2), we looked for those that had at least 20 animal and 20 non-animal examples with the visibility rating of 3. There were 11 participants that satisfied this criteria for Spanish and 5 that did it for Basque. In within-language, we did not find significant gain in performance as compared to corresponding partially conscious and non-conscious results. Across-language generalization was again found to be at chance-level in all ROIs.

4. Discussion

Our study investigated the brain basis of non-conscious semantic processing using masked word paradigm. Using multivariate pattern analysis of BOLD responses, we provide new insight into the brain substrates that support semantic representations across distinct states of visual awareness. Specifically, we showed that BOLD activity patterns associated with non-conscious words contain information that allows for decoding of the category of words both in Spanish and Basque (i.e. within language decoding). Notably, in the present study the words were non-conscious according to both subjective (i.e. rated as fully unaware on trial-by-trial basis [Overgaard et al. (2010)] as well as objective measures given that behavioural discrimination of the word category was found to be at chance level.

ROI analysis (see § 2.5.2) showed that such discriminative patterns for non-conscious items were found in canonical areas [Binder et al. (2009), Chen et al. (2017)] of the semantic network. Specifically, above-chance classification accuracies were found in a rather distributed set of brain regions including IPL, dmPFC, IFG and PCG for Spanish, and VTL and PCG for Basque. All of these areas have previously been associated with semantic processing of visible words in studies involving animals-tools and animals-artifacts contrasts [Cappa et al. (1998), Grossman et al. (2002), Kounios et al. (2003), Wheatley et al. (2005)]. We also showed that for partially conscious trials of both Spanish and Basque, such discriminative BOLD patterns were even more distributed, namely, significant decoding was found in all pre-specified left-lateralized seven ROIs of the semantic network. On the other hand, addressing the second question i.e. across-language generalization of semantic representations, we found little evidence for semantic generalization across languages, even on conscious trials.

All seven canonical areas of the semantic network were found to be implicated in the representation of word category under conditions where participants showed some awareness of the words in both Spanish and Basque. These results go in line with previous decoding studies of word meaning including Mitchell et al. (2008), Just et al. (2010) and Shinkareva et al. (2011). In the non-conscious condition, only one ROI was found to be common between Spanish and Basque i.e. PCG. Furthermore, while four ROIs (IPL, dmPFC, IFG, and PCG) were implicated in non-conscious semantic processing in Spanish, only two ROIs (VTL, PCG) were found for Basque. However, this pattern of results should not be taken to suggest that there are language-specific semantic representations. Different factors may have contributed to this pattern of results. For instance, our group of participants was mixed with most having Spanish as L1. Also, whereas no statistically significant difference was found between the age of acquisition of Spanish and Basque, the performance at both LexTALE and BEST tests of language proficiency was found to be statistically significantly superior in Spanish as compared to Basque (see § 2.1). Finally, we compared the decoding accuracy in those ROIs that did not overlap between Spanish and Basque but no difference between languages was found. Taken together, it is possible that inter-individual variability may have promoted the absence of a complete overlap between the informative ROIs between Spanish and Basque.

In the non-conscious trials of Spanish (see Figure 6), besides in IPL, and PCG, we also found significant decoding of meaning in dmPFC and IFG. What is interesting with the involvement of these frontal areas in non-conscious semantic representations is that it has implications for theoretical models of conscious and non-conscious processing i.e., Global Workspace theory [Dehaene et al. (1998a)]. According to this model, conscious representations result from widely distributed activity patterns involving both anterior (e.g. PFC) and posterior areas (e.g. object-selective brain areas), and information is broadcasted in these areas by means of top-down recurrent processing. Neuroimaging studies using masked priming paradigms indicate that non-conscious orthographic processing of words can occur in the left fusiform gyrus (i.e. the visual word form area; Dehaene et al. (2001), Kouider et al. (2006)). Priming experiments indicate that the non-conscious semantic priming implicates the left superior temporal areas [Devlin et al. (2004)]. Additional results from event related potentials indicate non-conscious semantic processing indexed by the N400 [Van Gaal et al. (2014), Eo et al. (2016), Heyman and Moors (2012)] but see Kang et al. (2011). Although these studies can be criticized based on the issues highlighted in the introduction (i.e. the absence of trial-by-trial measures of awareness, see § 1), the pattern of results suggests a relatively localized regional activity in non-conscious word processing that does not implicate higher-level prefrontal areas typically associated with conscious semantic processing (i.e. the left inferior frontal cortex) [Binder et al. (2009)]. The present results, on the other hand, indicate that the non-conscious semantic representations can be encoded in relatively distributed brain substrates involving the prefrontal cortex. A key difference between our paradigm and masked priming paradigm is that here the words were task-relevant and in masked priming, the primes are task-irrelevant. There is a limited data that supports the involvement of frontal areas during non-conscious word processing. One prior masked priming study Diaz and McCarthy (2007) showed regional BOLD response changes in a left-lateralized set of brain regions including the inferior frontal gyrus, inferior parietal and lateral temporal lobes during non-conscious processing of masked words. Another Axelrod et al. (2014) recently showed that the meaningful sentences rendered non-conscious by continuous flash suppression could be discriminated from non-words by using fMRI-based MVPA, specifically in left-lateralized brain areas including superior temporal sulcus and the middle frontal gyrus. However, our study goes beyond this finding, and shows that not only lower-level structural representations can be isolated [Axelrod et al. (2014)], but the semantic category of non-conscious words can also be classified. The present results also align with the prior research in visual working memory and executive control, which also indicates that dorsolateral prefrontal regions can be implicated in processing and brief maintenance of non-conscious visual stimuli ([Soto and Silvanto (2014), Bergström and Eriksson (2014, 2017), Dutta et al. (2014)]; though prefrontal activity in this later study occurred for subjectively unaware items unlike for items associated with null behavioral discrimination as demonstrated here). However, it is likely that non-conscious representations in prefrontal cortex are weak and hence unlikely to ignite sustained and strong feedback processing loops in distributed brain networks, which can be a requirement for information to become conscious [Van Vugt et al. (2018)]. Further research is needed to understand the limits and the functional scope of non-conscious semantic representations in the human brain, for instance, by testing its durability and the temporal dynamics of distributed semantic networks.

We now turn to the across-language generalization results. All of the ROIs showed chance-level decoding accuracy for semantic generalization from Spanish to Basque and vice versa. This happened not just for non-conscious words but also for partially conscious trials. We only found some evidence of across language generalization from Spanish to Basque in the conscious trials when we restricted our analysis to those participants with within-language decoding accuracies well exceeding chance level (i.e. 0.6), in order to avoid the presence of floor effects in across-language generalization.

This is the first time that MVPA-based across-language generalization has been used to investigate the scope of non-conscious semantic representations. However, the same approach has already been used with positive results in a number of different fMRI studies where words were available to conscious awareness [Buchweitz et al. (2012), Correia et al. (2014), Zinszer et al. (2015), Dehghani et al. (2017)]. The factors leading to across-language generalization are not well understood. There are a number of reasons that can explain why we did not find strong evidence for it. Firstly, the experiment was designed to maximize the number of non-conscious trials. The stimuli was briefly presented and masked, and luminance varied based on a staircase procedure that was biased towards decreasing luminance in response to ratings of partial or full awareness. Therefore, even though the participants reported partial and full visibility of the items, this does not mean that the stimuli strength was comparable to that of previous studies that reported across-language generalization [Buchweitz et al. (2012), Correia et al. (2014)], where stimuli were presented for much longer durations, were fully conscious and even observers were asked to think about the items to ensure that deep semantic analysis is taking place. Accordingly, our task may only have promoted shallow encoding of the words. Given the relatively small number of words used, it is also possible that the observers learned a mapping between the properties of the word stimuli and the semantic categorization response, which did not involve the level of processing required for across language generalization. We suggest that our task may have promoted a level of processing that is sufficient for within-language decoding but insufficient for across-language generalization.

It is also worth noting here that a significant amount of behavioural studies have addressed language-independent semantic representations by using translation and associative masked priming. Notably, while some of these studies have succeeded at showing cross-language semantic priming, most of them suffer from a number of methodological issues [Williams (1996), Basnight-Brown and Altarriba (2007)]. For instance, the reliance on post-hoc assessment of the visibility of prime words (and the absence of trial-by-trial measures of awareness) make it hard to establish that priming effects are not contaminated by some trials with prime awareness [Lutz and Thompson (2003), Van den Bussche et al. (2013), Haase and Fisk (2015)]. Further, the use of long SOAs (**s**timulus **o**nset **a**synchrony i.e. the time for which the prime gets displayed before it gets replaced by a target) do not rule out the operation of conscious strategic processes. Notably, non-replicable findings have been observed with most studies reporting absence of effect [Basnight-Brown and Altarriba (2007), Schoonbaert et al. (2009)] to a few reporting statistically significant cross-language facilitation [Perea et al. (2008)], yet trial-by-trial awareness assessment was not used in this study either. It is probably in the light of these issues that Basnight-Brown and Altarriba (2007) go so far as to conclude that all the cross-language priming effect seems to be the result of an improper control of additional conscious strategic fac-

tors that result in significant cross-language facilitation.

The current study demonstrated that the meaning of non-conscious words can be encoded in multi-voxel patterns of activity in putative semantic regions, including frontal areas. Whereas within-language classification of word meaning is possible in non-conscious contexts, across-language generalization (or evidence for language-independent semantic representations) seems harder to isolate; the latter may require not just conscious perception but a deeper semantic analysis too. Additional work is needed to make this determination.

References

Abrams, R. L. and Greenwald, A. G. (2000). Parts outweigh the whole (word) in unconscious analysis of meaning. *Psychological science*, 11(2):118–124.

Abrams, R. L., Klinger, M. R., and Greenwald, A. G. (2002). Subliminal words activate semantic categories (not automated motor responses). *Psychonomic Bulletin & Review*, 9(1):100–106.

Axelrod, V., Bar, M., Rees, G., and Yovel, G. (2014). Neural correlates of subliminal language processing. *Cerebral Cortex*, 25(8):2160–2169.

Basnight-Brown, D. M. and 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.

Bergström, F. and Eriksson, J. (2014). Maintenance of non-consciously presented information engages the prefrontal cortex. *Frontiers in human neuroscience*, 8:938.

Bergström, F. and Eriksson, J. (2017). Neural evidence for non-conscious working memory. *Cerebral Cortex*, pages 1–12.

Binder, J. R., Desai, R. H., Graves, W. W., and 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.

Booth, J. R., Burman, D. D., Santen, F. W. V., Harasaki, Y., Gitelman, D. R., Parrish, T. B., and Mesulam, M. M. (2001). The development of specialized brain systems in reading and oral-language. *Child Neuropsychology*, 7(3):119–141.

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

Cappa, S. F., Perani, D., Schnur, T., Tettamanti, M., and Fazio, F. (1998). The effects of semantic category and knowledge type on lexical-semantic access: a pet study. *Neuroimage*, 8(4):350–359.

Carretié, L., Hinojosa, J. A., Mercado, F., and Tapia, M. (2005). Cortical response to subjectively unconscious danger. *Neuroimage*, 24(3):615–623.

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

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

Damian, M. F. (2001). Congruity effects evoked by subliminally presented primes: Automaticity rather than semantic processing. *Journal of Experimental Psychology: Human Perception and Performance*, 27(1):154.

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

Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Penguin.

Dehaene, S. and Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2):200–227.

Dehaene, S. and Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in cognitive sciences*, 15(6):254–262.

Dehaene, S., Kerszberg, M., and Changeux, J.-P. (1998a). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences*, 95(24):14529–14534.

Dehaene, S., Naccache, L., Cohen, L., Le Bihan, D., Mangin, J.-F., Poline, J.-B., and Rivière, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature neuroscience*, 4(7):752.

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., van de Moortele, P.-F., and Le Bihan, D. (1998b). Imaging unconscious semantic priming. *Nature*, 395(6702):597.

Dehghani, M., Boghrati, R., Man, K., Hoover, J., Gimbel, S. I., Vaswani, A., Zevin, J. D., Immordino-Yang, M. H., Gordon, A. S., Damasio, A., et al. (2017). Decoding the neural representation of story meanings across languages. *Human brain mapping*, 38(12):6096–6106.

Devlin, J. T., Jamison, H. L., Matthews, P. M., and Gonnerman, L. M. (2004). Morphology and the internal structure of words. *Proceedings of the National Academy of Sciences*, 101(41):14984–14988.

Diaz, M. T. and McCarthy, G. (2007). Unconscious word processing engages a distributed network of brain regions. *Journal of Cognitive Neuroscience*, 19(11):1768–1775.

Dixon, N. F. (1971). Subliminal perception: The nature of a controversy.

Draine, S. C. (1997). *Analytic limitations of unconscious language processing*. PhD thesis.

Draine, S. C. and Greenwald, A. G. (1998). Replicable unconscious semantic priming. *Journal of Experimental Psychology: General*, 127(3):286.

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

Dutta, A., Shah, K., Silvanto, J., and Soto, D. (2014). Neural basis of non-conscious visual working memory. *Neuroimage*, 91:336–343.

Eo, K., Cha, O., Chong, S. C., and Kang, M.-S. (2016). Less is more: semantic information survives interocular suppression when attention is diverted. *Journal of Neuroscience*, 36(20):5489–5497.

Eriksen, C. W. (1960). Discrimination and learning without awareness: a methodological survey and evaluation. *Psychological review*, 67(5):279.

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

Forster, K. I. (1998). The pros and cons of masked priming. *Journal of psycholinguistic research*, 27(2):203–233.

Gaillard, R., Del Cul, A., Naccache, L., Vinckier, F., Cohen, L., and Dehaene, S. (2006). Nonconscious semantic processing of emotional words modulates conscious access. *Proceedings of the National Academy of Sciences*, 103(19):7524–7529.

Greenwald, A. G. (1992). New look 3: Unconscious cognition reclaimed. *American Psychologist*, 47(6):766.

Greenwald, A. G., Abrams, R. L., Naccache, L., and Dehaene, S. (2003). Long-term semantic memory versus contextual memory in unconscious number processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(2):235.

Greenwald, A. G., Draine, S. C., and Abrams, R. L. (1996). Three cognitive markers of unconscious semantic activation. *Science*, 273(5282):1699–1702.

Grossman, M., Koenig, P., DeVita, C., Glosser, G., Alsop, D., Detre, J., and Gee, J. (2002). The neural basis for category-specific knowledge: an fmri study. *Neuroimage*, 15(4):936–948.

Haase, S. J. and Fisk, G. D. (2015). Awareness of "invisible" arrows in a metacontrast masking paradigm. *The American journal of psychology*, 128(1):15–30.

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

Hauk, O., Johnsrude, I., and Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, 41(2):301–307.

Heyman, T. and Moors, P. (2012). Using interocular suppression and eeg to study semantic processing. *Journal of Neuroscience*, 32(5):1515–1516.

Holender, D. (1986). Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *Behavioral and brain Sciences*, 9(1):1–23.

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

Johnston, W. A. and Dark, V. J. (1986). Selective attention. *Annual review of psychology*, 37(1):43–75.

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

Kang, M.-S., Blake, R., and Woodman, G. F. (2011). Semantic analysis does not occur in the absence of awareness induced by interocular suppression. *Journal of Neuroscience*, 31(38):13535–13545.

Kouider, S., De Gardelle, V., Sackur, J., and Dupoux, E. (2010). How rich is consciousness? the partial awareness hypothesis. *Trends in cognitive sciences*, 14(7):301–307.

Kouider, S. and Dehaene, S. (2007). Levels of processing during non-conscious perception: a critical review of visual masking. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 362(1481):857–875.

Kouider, S., Dehaene, S., Jobert, A., and Le Bihan, D. (2006). Cerebral bases of subliminal and supraliminal priming during reading. *Cerebral Cortex*, 17(9):2019–2029.

Kouider, S. and Dupoux, E. (2004). Partial awareness creates the "illusion" of subliminal semantic priming. *Psychological science*, 15(2):75–81.

Kounios, J., Koenig, P., Glosser, G., DeVita, C., Dennis, K., Moore, P., and Grossman, M. (2003). Category-specific medial temporal lobe activation and the consolidation of semantic memory: evidence from fmri. *Cognitive Brain Research*, 17(2):484–494.

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

Lerma-Usabiaga, G., Carreiras, M., and Paz-Alonso, P. M. (2018). Converging evidence for functional and structural segregation within the left ventral occipitotemporal cortex in reading. *Proceedings of the National Academy of Sciences*, page 201803003.

Lutz, A. and Thompson, E. (2003). Neurophenomenology integrating subjective experience and brain dynamics in the neuroscience of consciousness. *Journal of consciousness studies*, 10(9-10):31–52.

Marcel, A. J. (1983). Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive psychology*, 15(2):197–237.

McCandliss, B., Cohen, L., and Dehaene, S. (2003). The visual word form area: expertise for reading in the fusiform gyrus. *Trends in cognitive sciences*, 7(7):293–299.

Mineka, S. and Öhman, A. (2002). Phobias and preparedness: The selective, automatic, and encapsulated nature of fear. *Biological Psychiatry*, 52(10):927–937.

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

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

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

Naccache, L. and Dehaene, S. (2001). Unconscious semantic priming extends to novel unseen stimuli. *Cognition*, 80(3):215–29.

Naccache, L., Gaillard, R., Adam, C., Hasboun, D., Clémenceau, S., Baulac, M., Dehaene, S., and Cohen, L. (2005). A direct intracranial record of emotions evoked by subliminal words. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7713–7.

Nakamura, K., Dehaene, S., Jobert, A., Le Bihan, D., and Kouider, S. (2005). Subliminal convergence of kanji and kana words: further evidence for functional parcellation of the posterior temporal cortex in visual word perception. *Journal of Cognitive Neuroscience*, 17(6):954–968.

Ohman, A. and Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review*, 108(3):483–522.

Overgaard, M., Timmermans, B., Sandberg, K., and Cleeremans, A. (2010). Optimizing subjective measures of consciousness. *Consciousness and cognition*, 19(2):682–684.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830.

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

Perea, M., Dunabeitia, J. A., and 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., and Carreiras, M. (2006). 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., and Botvinick, M. (2009). Machine learning classifiers and fmri: a tutorial overview. *Neuroimage*, 45(1):S199–S209.

Pratte, M. S. and Rouder, J. N. (2009). A task-difficulty artifact in subliminal priming. *Attention, perception & psychophysics*, 71(6):1276–83.

Purcell, D. G., Stewart, A. L., and Stanovich, K. E. (1983). Another look at semantic priming without awareness. *Perception & psychophysics*, 34(1):65–71.

Sahraie, A., Weiskrantz, L., Barbur, J., Simmons, A., Williams, S., and Brammer, M. (1997). Pattern of neuronal activity associated with conscious and unconscious processing of visual signals. *Proceedings of the National Academy of Sciences*, 94(17):9406–9411.

Schlaghecken, F., Blagrove, E., and Maylor, E. A. (2008). No difference between conscious and nonconscious visuomotor control: evidence from perceptual learning in the masked prime task. *Consciousness and cognition*, 17(1):84–93.

Schoonbaert, S., Duyck, W., Brysbaert, M., and 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.

Shaywitz, B. A., Shaywitz, S. E., Pugh, K. R., Mencl, W. E., Fulbright, R. K., Skudlarski, P., Constable, R. T., Marchione, K. E., Fletcher, J. M., Lyon, G. R., et al. (2002). Disruption of posterior brain systems for reading in children with developmental dyslexia. *Biological psychiatry*, 52(2):101–110.

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

Sklar, A. Y., Levy, N., Goldstein, A., Mandel, R., Maril, A., and Hassin, R. R. (2012). Reading and doing arithmetic nonconsciously. *Proceedings of the National Academy of Sciences*, 109(48):19614–19619.

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

Soto, D. and Silvanto, J. (2014). Reappraising the relationship between working memory and conscious awareness. *Trends in Cognitive Sciences*, 18(10):520–525.

Van den Bussche, E., Vermeiren, A., Desender, K., Gevers, W., Hughes, G., Verguts, T., and Reynvoet, B. (2013). Disentangling conscious and unconscious processing: a subjective trial-based assessment approach. *Frontiers in human neuroscience*, 7:769.

van Gaal, S. and Lamme, V. A. (2012). Unconscious high-level information processing implication for neurobiological theories of consciousness. *The neuroscientist*, 18(3):287–301.

Van Gaal, S., Naccache, L., Meuwese, J. D., Van Loon, A. M., Leighton, A. H., Cohen, L., and Dehaene, S. (2014). Can the meaning of multiple words be integrated unconsciously? *Phil. Trans. R. Soc. B*, 369(1641):20130212.

Van Vugt, B., Dagnino, B., Vartak, D., Safaai, H., Panzeri, S., Dehaene, S., and Roelfsema, P. R. (2018). The threshold for conscious report: Signal loss and response bias in visual and frontal cortex. *Science*, 360(6388):537–542.

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

Weiskrantz, L. (1997). *Consciousness lost and found: A neuropsychological exploration.* Oxford University Press.

Whalen, P. J., Kagan, J., Cook, R. G., Davis, F. C., Kim, H., Polis, S., McLaren, D. G., Somerville, L. H., McLean, A. A., Maxwell, J. S., and Johnstone, T. (2004). Human amygdala responsivity to masked fearful eye whites. *Science*, 306(5704):2061.

Wheatley, T., Weisberg, J., Beauchamp, M. S., and Martin, A. (2005). Automatic priming of semantically related words reduces activity in the fusiform gyrus. *Journal of cognitive neuroscience*, 17(12):1871–85.

Williams, J. N. (1996). Is automatic priming semantic? *European Journal of Cognitive Psychology*, 8(2):113–162.

Zinszer, B., Anderson, A. J., Kang, O., Wheatley, T., and Raizada, R. D. (2015). You say potato, i say tǔdòu: How speakers of different languages share the same concept. In *CogSci*.