Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Research Report

Neural decoding of speech with semantic-based classification

Yi Lin [a,*] and Po-Jang Hsieh [b,**]

[a] Taiwan International Graduate Program in Interdisciplinary Neuroscience, National Cheng Kung University and Academia Sinica, No. 128, Academia Road, Section 2, Nankan, 11529, Taipei, Taiwan
[b] Department of Psychology, National Taiwan University, No. 1, Roosevelt Road, Section 4, Da'an, 10617, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 1 March 2022
Reviewed 25 March 2022
Revised 17 April 2022
Accepted 9 May 2022
Action editor Alessandro Tavano
Published online 8 June 2022

Keywords:
Neural speech decoding
Language production
Language perception
Semantic representations

ABSTRACT

Speech is a complex cognitive process that begins with conceptualization, proceeds to word-level processing, and ends with articulation. Neural decoding of speech (i.e., using neural activity to decode the content of language production) has been mostly conducted by mapping neural activities in the later part of language production (i.e., phonological and motor processing). Here we show that neural decoding of speech can also be performed by mapping neural activities associated with semantic representations that occur in the early part of language production. Furthermore, we demonstrated that the classifier trained using the neural activity patterns of language perception was able to decode the content of language production, indicating a cross-modality similarity between language perception and language production in semantic representations.

1. Introduction

Speech is a complex cognitive process. The sequences of language production are generally assumed to include conceptualization, word-level processing (lemma), phonological-level processing, and articulation (Roelofs et al., 1998; Hickok, 2012; Levelt et al., 1999; Vigliocco et al., 1999; Levelt, 1992, 1993; Kempen & Huijbers, 1983). After an idea is formulated as a speech goal (conceptualization), the idea is mapped to a word with corresponding meaning (word-level or lemma-level processing). Following word-level processing, the phonological information is then prepared for articulation.

Studies on the neural decoding of speech (Anumanchipalli et al., 2019; Dash et al., 2020, 2021; Herff et al., 2015; Martin et al., 2014; Ramsey et al., 2018; Chakrabarti et al., 2015) have mostly focused on mapping the neural representations of the later part of the language production process. For instance, in

* *Corresponding author.*
** *Corresponding author.*
E-mail addresses: sa8081016@gs.ncku.edu.tw (Y. Lin), hsiehpj@ntu.edu.tw (P.-J. Hsieh).

https://doi.org/10.1016/j.cortex.2022.05.018

Herff et al. (2015), it is shown that the recorded electrocorticography (ECoG) signals produced while participants were speaking specific phrases can be successfully mapped to phonemes to predict language production. The success of this line of research suggests that human brains encode phonetic information during language production.

However, speech begins with conceptualization and word-level processing (Wiese, 1984). To meet a speech goal, the primary step for language production should be to retrieve the meaning of the corresponding word from the mental lexicon. On this view, semantic information is critical and directly encoded during language production. To directly test this hypothesis, we examined whether neural activities during language production are systematically associated with semantic representations.

This research strategy (i.e., mapping neural representations to semantic representations) has been widely used in decoding human perception (Anderson et al., 2017; Bonner & Epstein, 2021; Huth et al., 2016; Mitchell et al., 2008; Nishida & Nishimoto, 2018; Pereira et al., 2013, 2018; Schrimpf et al., 2021; Wang et al., 2020). For example, Pereira et al. (2018) first constructed a semantic representation database by using a natural language processing algorithm and successfully decoded the perceptual contents of brain activity. Bonner and Epstein (2021) used a similar research strategy to map low-dimensional statistical representations onto voxel-wise fMRI responses during object viewing and discovered that in the parahippocampal place area, cortical responses to specific objects could be predicted by their visual statistical contexts.

Based on these studies, we decided to adopt this well-established decoding method to test our hypothesis. Specifically, we attempted to 1) decode language *perception* by employing the same algorithm of Pereira et al. (2018) but with Mandarin stimuli on native Mandarin users; and 2) decode language *production* by using the same decoding algorithm.

We separately trained two classifiers (i.e., a perception classifier and a production classifier) with language perception data and language production data. We expected that both the perception classifier and the production classifier could successfully decode unlearned contents from brain activity. Furthermore, to examine whether neural representations were systemically associated with semantic representations during both language perception and language production, analyses were conducted to decode semantic content across classifiers (i.e., decoding the content of language perception with the production classifier and vice versa). If the neural activities during language perception and language production during word processing were similar, the content of language perception should be decodable by the production classifier and vice versa.

2. Methods

2.1. Subjects

Based on three related studies that reported 100% success rate in prediction (Bonner & Epstein, 2021; Mitchell et al., 2008; Pereira et al., 2018), we performed the proportion sign test (binominal test) with the null hypothesized proportion as 50% success rate and type one error rate as .05. One-tailed G*Power (Faul et al., 2009) showed that there is a 95% chance of correctly rejecting the null hypothesis that the sample proportion is equivalent to .5 (50%) with 5 participants.

A total of eight native Mandarin speakers were recruited for this study. All subjects (4 men and 4 women; mean age = 27.8; SD = 7.05) reported no mental or language-related disabilities. The study protocol was approved by the Ethics Committee of National Taiwan University, Taiwan. All the subjects signed informed consent forms and were asked to undergo five MRI sessions (1 h per session). Three additional prospective subjects were not included in the study because of excessive head motion and somnolence during the experiment.

2.2. Materials

2.2.1. Semantic vector and representative word selection

To model the semantic representations of Chinese words, we adopted the natural-language processing algorithm word2vec (Zhang et al., 2021; Mikolov et al., 2013a, 2013b). This algorithm was designed to represent words based on textual analysis, with the underlying assumption that words with similar meanings appear in similar contexts (Harris, 1954). The word2vec (skip-gram architecture) introduces a neural network to learn how a given target word predicts other contextual words accompanying the target word and gives each target word a vector (also called vector space or semantic vector) to represent this unique property/semantic relationship in the text. If two words have similar contexts (meaning that they are semantically similar according to the assumption), the distance between the vector spaces should be small. On the contrary, if two words are not related, the distance between the vector spaces should be wide. In short, by analyzing a big corpus of text, words can be represented by different vectors.

A text corpus from a Taiwan Wikipedia dump on April 05, 2021, was downloaded as the material. Since Chinese words cannot be directly segmented by space, a python package named jieba (https://github.com/fxsjy/jieba) was used to segment the text corpus into meaningful word chunks. The segmentation was based on the predefined dictionary to find the most possible word combinations. After segmentation, the Python package named gensim (Rehurek & Sojka, 2011; https://radimrehurek.com/gensim/index.html) was used to learn the representation of the words. The Skip-gram model was selected, and the dimension of the space vector was 300 in accordance with previous research (Bfaroni et al., 2014; Hollis, 2017; Landauer & Dumais, 1997; Mandera et al., 2017; Pereira et al., 2013). The window size was 5 (as default), and the words with a frequency lower than 10 were discarded (min_count = 10). A custom Mandarin Chinese word-embedding model was then constructed accordingly.

Within the custom word-embedding model was a total of 571,485 words. In addition, each word's frequency of occurrence in the corpus was calculated, and the 30,000 most frequently used words from the custom word-embedding model were selected from the word database. Based on the semantic vectors of these 30,000 words, a k-means analysis was performed to identify the 200 most representative categories on the basis of which to create a universal classifier (Pereira et al., 2018). Categories with unrecognizable words, foreign characters, or proper nouns were removed. Upon completion of the process, only 119 categories remained (Fig. 1).

A single word within each category needed to be selected to represent the corresponding category. Two steps were performed to determine the representative word of each category. 1) First, distances from the centroid of the specific category to each word within a category were calculated. After ranking the distance in ascending order, the top 20 words

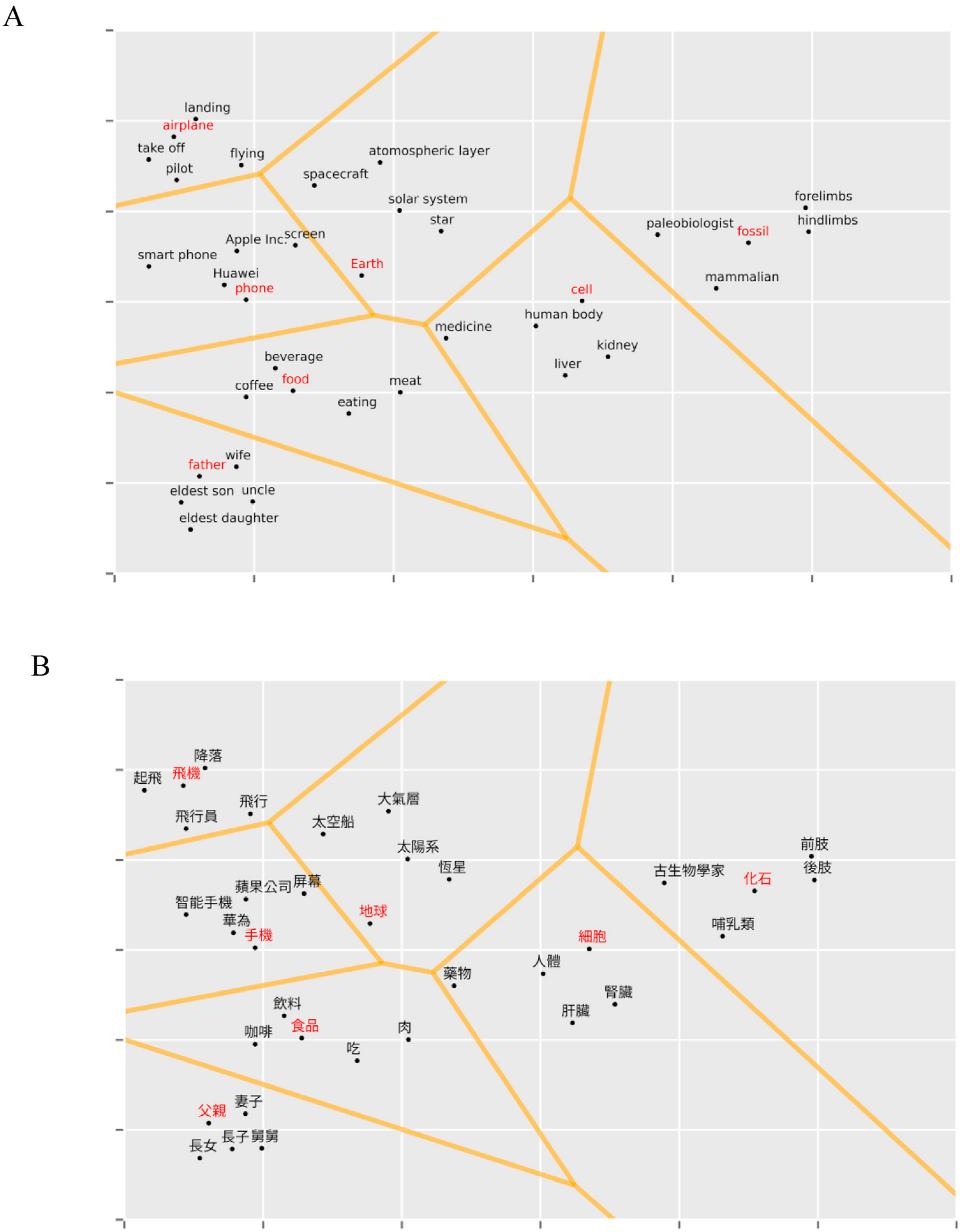


Fig. 1 – Visualization of semantic space. Seven selected categories of 300-dimensional semantic space were downgraded to two dimensions for visualization. The target words for each group were marked in red and the related words in black. (A) Stimuli translated in English (B) Stimuli in Mandarin.

nearest to the centroid of a category were selected. 2) Second, among these top 20 words, the order was resorted and reranked according to their frequency until the highest-ranking word with the highest frequency of occurrence was determined. The highest-ranking word with the highest frequency of occurrence was then used as the representative word. Moreover, all the highest-ranking words were manually checked by the investigators. If the highest-ranking word is not a superordinate word among other words to represent the category, the selection for the representative word (or target word) would move to the next word until the ideal (i.e., superior) word was found. For example, according to the frequency ranking, the highest-ranking word in category 27 is "red". And the fourth word is "color". Since color is superordinate than "red", we changed the representative word from the category to the "color" instead of "red". The representative words were used as the target words during the fMRI session in which subjects had to view or speak (fMRI paradigm).

2.2.2. *Experimental stimuli*

One representative word for each category was selected as the target word for the fMRI session. Each word was presented in two types with additional accompanying information; one was the sentence type, and the other was the picture-with-text type. The presentation methods were motivated from Pereira et al. (2018) and Bonner and Epstein (2021).

For the sentence type, a target word and one of the top 20 nearest-to-the-centroid words (excluding the target word and chosen arbitrarily) from the corresponding category (without the target words) were used to create a sentence. A total of five sentences were created for each semantic category. In addition, the target words were highlighted in bold and underlined (Fig. 2). For the picture-with-text type, five pictures related to the target word were selected from the Pexels webpage (https://www.pexels.com/zh-tw/). In addition, the top 4 words obtained from the 20 nearest-to-the-centroid words after frequency ranking (excluding the target words) were presented along with one picture and the target word (the target word was at the center in a larger font, with the four related words above it; Fig. 2). In brief, subjects were presented with five sentence-type stimuli and five picture-with-text stimuli to ensure that they could grasp the concepts. These 10 stimuli for a category (five sentences + five pictures-with-text) were shown only once during the experiment. However, the concept of each target word was repeated 10 times.

2.3. *fMRI paradigm*

Two types of cognition tasks (perception vs. production) were performed by subjects, and each task had two different stimulus presentation methods (sentence vs. picture-with-text), yielding four different fMRI conditions in total.

We mainly followed the experimental design of Pereira et al. (2018). In the perception–picture-with-text condition, a reminder frame (with the text 'reading') would first prompt the subject to perform a "perception" task in the current fMRI run. Subjects were asked to view the stimulus as it was displayed for 2 s and then think about the meaning of the target word with the accompanying information. The thinking frame lasted for 2 s and was followed by a 2-s fixation frame. The perception–sentence condition was identical to the perception–picture-with-text condition, except the stimuli were of the sentence type.

In the production–picture-with-text condition, a reminder frame (with the text 'speaking') cued subjects to perform a "production" task in the current fMRI run. Subjects were asked to view as well as think of the stimulus briefly (the stimuli were each displayed for 2 s) and then say aloud the target word. The speaking frame lasted for 2 s and was followed by a 2-s fixation frame. The production-sentence condition was identical to the production–picture-with-text condition (i.e., speaking out the target word only), except the stimuli were of the sentence type.

A total of 119 words were separated into two sets (Set1 and Set2). Set1 always contained 60 trials, whereas Set2 contained 59 trials. As a result, two runs were required to go through all the stimuli for each condition. Additionally, each run contained a 10-s fixation frame at the end and a 10-s break frame after 30 trials were completed. Therefore, one run required either 6 min and 40 s (for Set1 with 60 trials) or 6 min and 34 s (for Set2 with 59 trials).

Subjects were required to complete five MRI sessions. Within one MRI session, subjects would go through the four conditions, which were perception–picture-with-text, production–picture-with-text, perception–sentence, production–sentence. Two runs were required to complete all stimuli, and thus an MRI session included a total of eight MRI runs (Fig. 2). All conditions were repeated five times.

2.4. *fMRI data acquisition*

All MRI data were acquired using a 3 T scanner with a 32-channel head coil (MAGNETOM Skyra 3 T). Functional MRI runs were acquired using a gradient echo-planar imaging sequence (field of view = 220 × 220 mm; TR = 2 sec; TE = 24 ms; flip angle = 90°; slice thickness = 4 mm; acquisition matrix = 64 × 64). Anatomical image was acquired using an MPRAGE sequence (field needs of view = 256 × 256 mm; TR = 2 sec; TE = 2.98 ms; flip angle = 9°; slice thickness = 1 mm; acquisition matrix = 256 × 256).

2.5. *fMRI data analysis*

2.5.1. *Data preprocessing*

All MRI data were analyzed using SPM 12 (Penny et al., 2011). After slice-timing and realignment correction, the individual functional data were coregistered with their structural data, resampled into 2-mm isotropic voxels, and wrapped in the MNI template. The normalized data were smoothed with an 8-mm full-width half-maximum Gaussian kernel. A general linear model design for the four conditions was constructed independently for each subject to acquire the 119 target-word-related parameter estimations. These target word parameter estimations were then averaged across modalities (sentence + picture-with-text) and served as decoding features.

2.5.2. *Decoding methodology*

We followed the decoding methodology in the work of Pereira et al. (2018). In brief, classifiers were expected to generate correct or similar semantic vectors given specific brain image

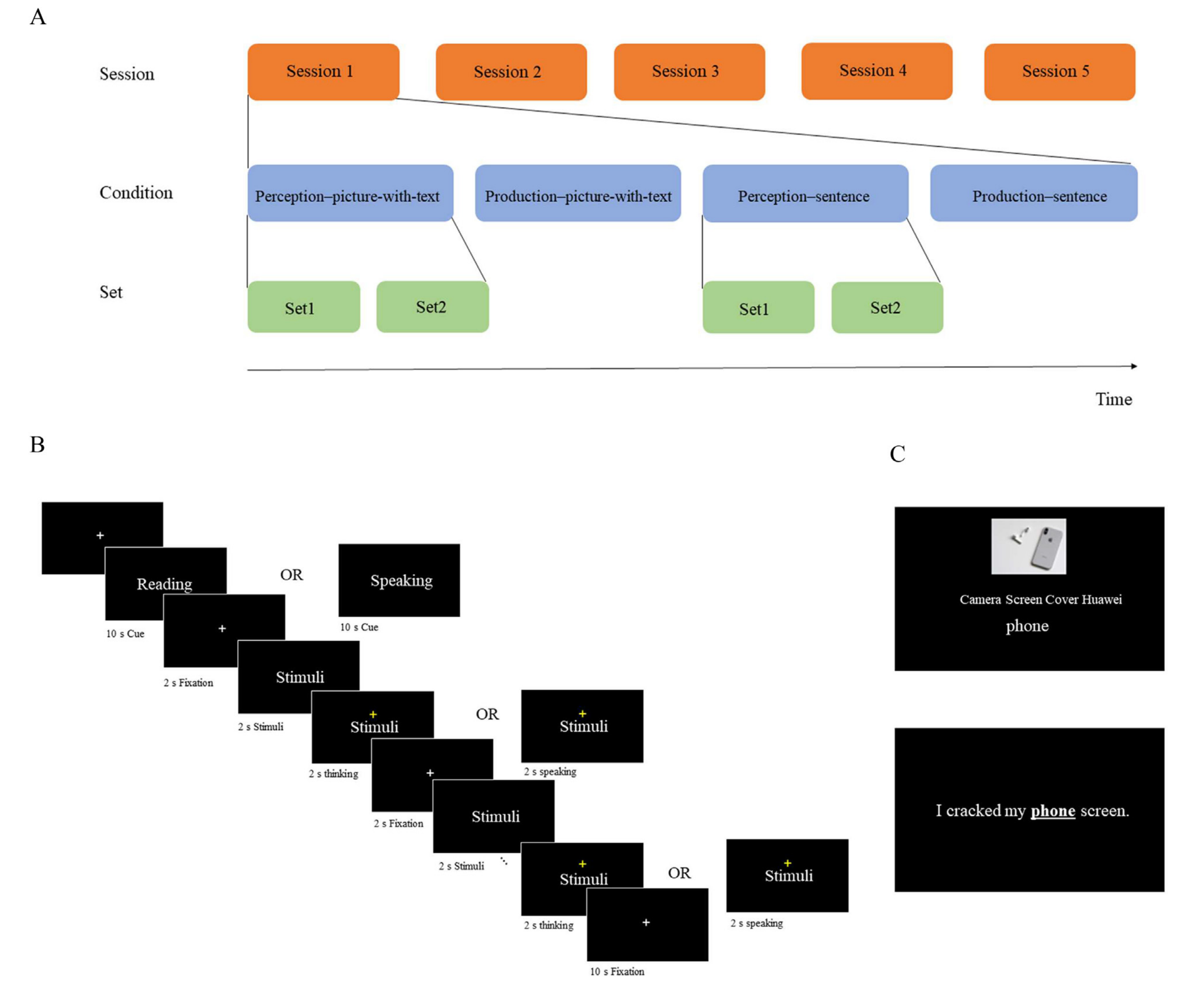


Fig. 2 – Presentation of the fMRI experimental design. (A) One MRI session consisted of four experimental conditions, namely perception–picture-with-text, production–picture-with-text, perception–sentence, and production–sentence. Each condition contained two sets. (B) Within a single MRI run, a reminder frame appeared first followed by the fixation frame, the stimuli, and then the task frame (each frame lasted for 2 s). (C) Display of picture-with-text type (upper image) stimuli and sentence type (lower image) stimuli.

data. Classifiers were thus expected to learn an association between brain activity and semantic vectors. Ridge regression was used to fulfill the goal.

During a model training session, given a series of brain activity matrix X (training set) and the corresponding semantic vectors Z, we expected to acquire a string of regression coefficients b and constant b_0 that minimized $\| Xb + b_0 - z \|_2^2 + \lambda \| b \|_2^2$ for each column of the semantic vectors Z. The term λ was the regularization parameter for each dimension using generalized cross-validation within the training set. Each voxel as well as each semantic vector dimension was normalized across training stimuli.

The training of these language classifiers (i.e., the production classifiers and the perception classifiers) was based on leave-10%-stimuli-out cross-validation. Hence, in each fold, brain images of 11 or 12 words were left out for later classifier validation. The rest of the brain images (107 or 108) served as a training data set for learning the regression parameters.

Moreover, in each fold, a voxel selection procedure was performed before classifier training. This procedure aimed to reduce the size of the training data to 5000 voxels per image. Voxels containing information concerning text-derived semantic vectors for training set images were selected. The vowel selection procedure mainly followed the method from Pereira et al. (2018). In short, leave-10%-stimuli-out cross-validation was performed within the training sets for voxel selection. Ridge regression models were used to learn and later predict each semantic dimension from the image data of each voxel. Because the image data were smoothed beforehand, in this step, we chose not to include the information from the adjacent voxels during training—this approach slightly differed from that in Pereira et al. (2018). Moreover, to identify the 5000 best informative voxels, cosine similarity was considered in calculating the similarity between the predicted semantic vector and the real semantic vector (Mitchell et al., 2008). Cosine similarity entails computing the dot product of the two given vectors divided by the length of the vectors. Higher cosine similarity indicates higher informativeness. Thus, the top 5000 voxels sorted according to high-to-low cosine similarity were selected and saved as a map for language classifier training.

Furthermore, in this study, two decoding directions were used. The first was the basic direction, which stated that the classifier was trained and validated on the image data from the same MRI condition (for example, the decoding of perception data by perception classifiers). The second direction was cross decoding, in which the image data for training and validation came from different data streams (i.e., decoding perception data by production classifier and vice versa).

2.5.3. *Statistical testing for decoding performance*

To evaluate the performance of our classifiers, the rank accuracy score (Pereira et al., 2018) was calculated. The similarity between the predicted semantic vectors and a set of candidate semantic vectors ($n = 119$) were compared. The similarity values were sorted from highest to lowest to determine the rank score of the correct stimuli. This rank score was then normalized ($1 - \frac{rank - 1}{119 - 1}$) to the range [0,1] as the rank accuracy score. Higher rank accuracy scores indicated more favorable decoding performance. The average rank accuracy score for all decoded vectors represented the performance of the classifier (Fig. 3).

To examine the statistical significance of the performance of the classifiers, we subsequently conducted permutation tests (Bonner & Epstein, 2021). The labels of the correct stimuli were randomly permutated within each cross-validation fold, and then the permutated rank accuracy was determined. We repeated the permutation 2000 times and reported the p value for the performance of the classifiers based on the permutated rank accuracy scores. Because all the language classifiers were independent entities, we reported the language classifier as successful if the p value was smaller than .05. Finally, we performed a binomial test to calculate the significance of successful decoding times. The probability of a successful outcome was assumed to be .5.

3. Results

3.1. *Performance of perception decoding with perception classifier*

All the perception classifiers were successfully trained and were able to decode the unlearned perception content on the basis of brain maps across the eight subjects (Fig. 4A). The rank accuracy scores and corresponding p values for subject 01 to subject 08 were .5574 ($p = .0005$), .5032 ($p = .0005$), .4738 ($p = .0215$), .4763 ($p = .0085$), .5232 ($p = .0005$), .6208 ($p = .0005$), .521($p = .0005$), and .5199 ($p = .0005$), respectively (Table 1). The result of the binomial test for successful classification ($p = .004$) again corroborated that the development of the perception classifier was successful.

3.2. *Performance of production decoding with production classifier*

Seven out of eight production classifiers were successfully trained (Fig. 4B). The rank accuracy scores and corresponding p values for subject 01 to subject 08 were: .5678 ($p = .0005$), .4662 ($p = .016$), .4189 ($p = .6407$), .4999 ($p = .0005$), .487 ($p = .0015$), .5051 ($p = .0005$), .4949 ($p = .001$), and .4609 ($p = .005$), respectively (Table 1). Although one production classifier was unable to decode the untrained language production content, the result of the binomial test ($p = .004$) still corroborated the successful development of the production classifier.

3.3. *Performance of production decoding with perception classifier*

To investigate the similarity of neural representations between language perception and language production, we tested the trained perception classifiers for their generalizability to brain maps acquired under production conditions. We found that all unlearned production content could be decoded by the

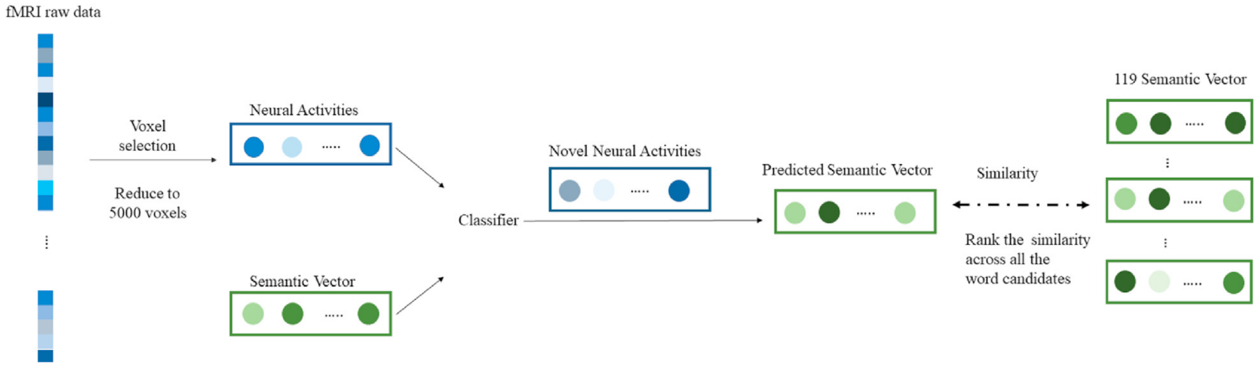
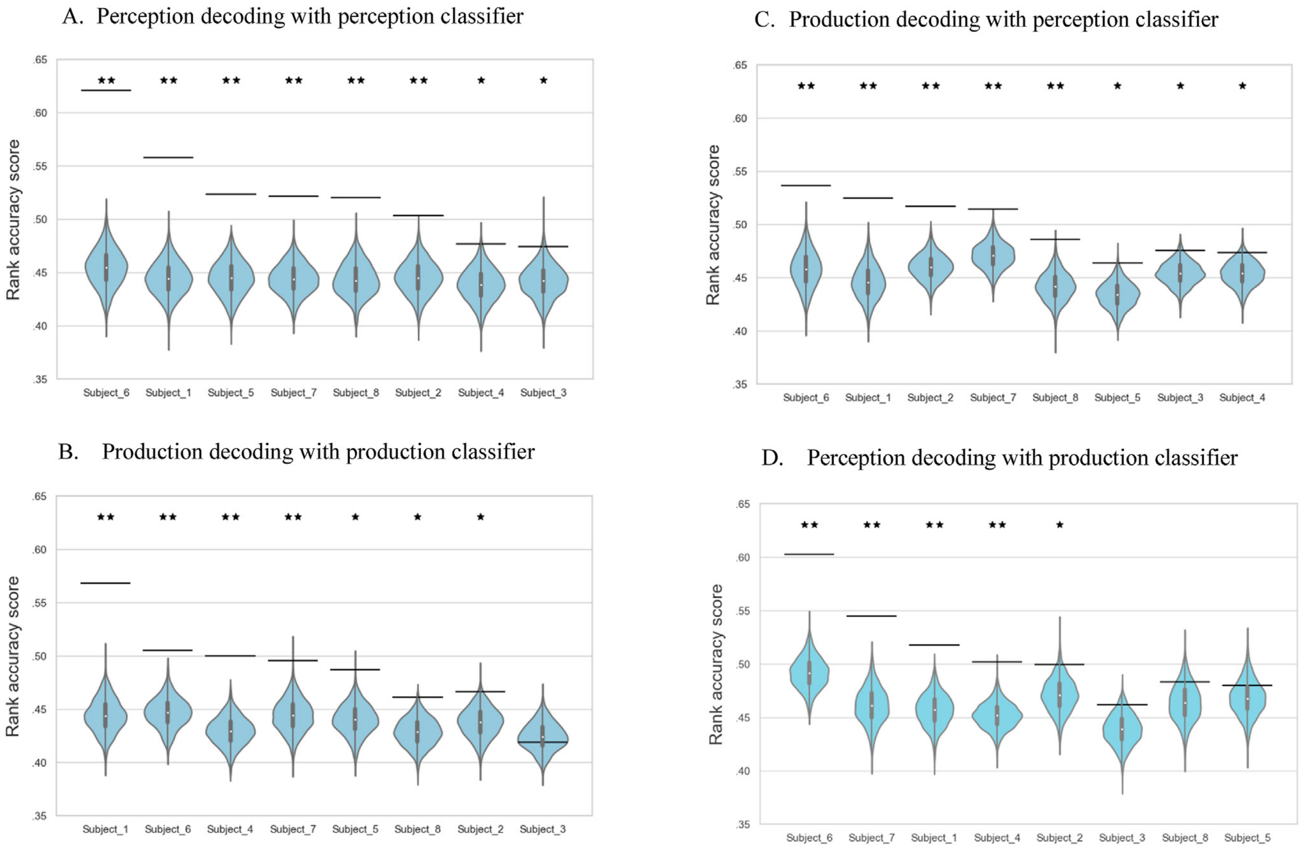


Fig. 3 – Decoding schematic. During the training session, 90% of brain maps and corresponding semantic vectors were used to train the language classifier. To reduce the data dimensions, voxel selection was firstly performed before classifier training based on the training data. The remaining 10% of brain maps were used to acquire the predicted semantic vectors for testing, which were compared to real semantic vectors for a range of stimuli.



A. Perception decoding with perception classifier

C. Production decoding with perception classifier

B. Production decoding with production classifier

D. Perception decoding with production classifier

Fig. 4 – Decoding performance. The violin plots show the distribution of the permutation results. The black lines above the violin plots indicate the real rank accuracy scores of the classifiers. (A) Performance of perception decoding with perception classifier. (B) Performance of production decoding with the production classifier. (C) Performance of production decoding with the perception classifier. (D) Performance of perception decoding with the production classifier. The results were ranked and displayed from the highest to lowest performance. *p < .05, **p < .001 uncorrected, one-sided permutation test.

Table 1 – The performance of the perception classifiers and production classifiers across subjects.

Subject ID	Performance of perception decoding with perception classifier		Performance of production decoding with production classifier		Performance of production decoding with perception classifier		Performance of perception decoding with production classifier	
	Rank accuracy scores	*p*-values	Rank accuracy scores	*p*-values	Rank accuracy scores	*p*-values	Rank accuracy scores	*p*-values
Subject 01	.5574	.0005*	.5678	.0005*	.5245	.0005*	.517	.0005*
Subject 02	.5032	.0005*	.4662	.016*	.517	.0005*	.4994	.032*
Subject 03	.4738	.0215*	.4189	.6407	.4754	.0235*	.4617	.0535
Subject 04	.4763	.0085*	.4999	.0005*	.4734	.036*	.502	.01*
Subject 05	.5232	.0005*	.487	.0015*	.4632	.0065*	.4798	.2199
Subject 06	.6208	.0005*	.5051	.0005*	.5362	.0005*	.6027	.0005*
Subject 07	.521	.0005*	.4949	.001*	.5139	.0005*	.5446	.0005*
Subject 08	.5199	.0005*	.4609	.005*	.4857	.001*	.4829	.1339

*p < .05.

perception classifiers (Fig. 4C). The rank accuracy scores and corresponding p values for subject 01 to subject 08 were .5245 (p = .0005), .517 (p = .0005), .4754 (p = .0235), .4734 (p = .036), .4632 (p = .0065), .5362 (p = .0005), .5139 (p = .0005), and .4857 (p = .001), respectively (Table 1). The binomial test for successful classification further confirmed the robustness of the generalizability of perception classifiers (p = .004).

3.4. *Performance of perception decoding with production classifier*

We also examined the generalizability of trained production classifiers on unlearned brain maps acquired under perception conditions. Five out of eight classifiers could be generalized to perception (Fig. 4D). The rank accuracy scores and corresponding p values for subject 01 to subject 08 were .517 ($p = .0005$), .4994 ($p = .032$), .4617 ($p = .0535$), .502 ($p = .01$), .4798 ($p = .2199$), .6027 ($p = .0005$), .5446 ($p = .0005$), and .4829 ($p = .1339$), respectively (Table 1). The binomial test for success classification was not statistically significant ($p = .219$).

Lastly, the 5000 most informative voxels for training classifiers were combined across the eight subjects (Fig. 5A and B). The value inside a voxel represented the selection probability for classifier development.

4. Discussion

To develop a universal speech decoder and investigate the neural basis of linguistic concepts during language perception and production, we mapped neural activities with semantic representations. Our results demonstrated that language production was decodable on the basis of corresponding neural data, suggesting that neural activities during language production were systematically associated with semantic representations. Moreover, we succeeded in decoding the untrained language production content with the perception classifiers.

The success of developing a language production classifier by mapping between semantic representations and neural activity are strong evidence for the speech model that proposes mental representing/processing of meaning before articulation (Roelofs et al., 1998; Hickok, 2012; Levelt et al., 1999; Vigliocco et al., 1999; Levelt, 1992, 1993; Kempen & Huijbers, 1983). Compared to motion-based and phonetic based classifiers that utilize neural activity from frontal and temporal lobes (i.e., motor and auditory cortex) while subjects spoke with corresponding phone representation (phonetic-based) or articulatory kinematic features (motion-based), our result extended the methodology to incorporate semantic representations in neural decoding of speech (Anumanchipalli et al., 2019; Herff et al., 2015; Martin et al., 2014; Ramsey et al., 2018; Chakrabarti et al., 2015; Sharon et al., 2020). Moreover, such a semantic-based classifier is better than motion-based and phonetic based classifiers under certain conditions. For example, the phonological level and motion level can fail during language production and cause the tip-of-the-tongue phenomenon (Vigliocco et al., 1998), apraxia of speech (Ogar et al., 2005) as well as dysarthria (Duffy, 2013). Only a semantic-based classifier targeting pre-articulation/utterance information can overcome the aforementioned obstacles.

Cross-decoding analyses were performed to investigate the similarity between the cognitive processes underlying language perception and language production. The perception classifiers in each of our subjects could successfully decode the subject's own language production content. This result suggests that language perception and production share a similar neural representation of semantics. Notably, this finding leads to a new question: whether humans share similar neural representations of semantics even across languages. Neural representations for higher semantic levels (i.e., narrative comprehension) overlap across languages. Dehghani et al. (2017) trained their classifier with neural activities and the high-semantic-level vector (i.e., the story vector) from other languages and reported above-chance-level decoding accuracy. For the lower semantic level (i.e., the word level), Van de Putte et al. (2017) and Correia et al. (2014) reported neural overlap in the semantic representations of bilingual individuals. However, due to the specificity of the bilingual participants (Riehl, 2010) and the experimental designs, these results can only support weak conclusions about cross-language neural representations. Future research is required to address this question.

By contrast, the generalizability of the trained production classifiers to unlearned brain maps acquired under language perception conditions was less successful. This result might be because language production entails not only semantic representations but also representations of articulation (Anumanchipalli et al., 2019) and phonetic representations (Herff et al., 2015).

Apart from the generalizability across modalities, the other intriguing issue is whether the semantic representations can be generalized across participants. By computing the selection possibility for the 5000 most informative voxels across subjects, our result showed that to some extent, semantic representations were overlapped across subjects. Yet, not

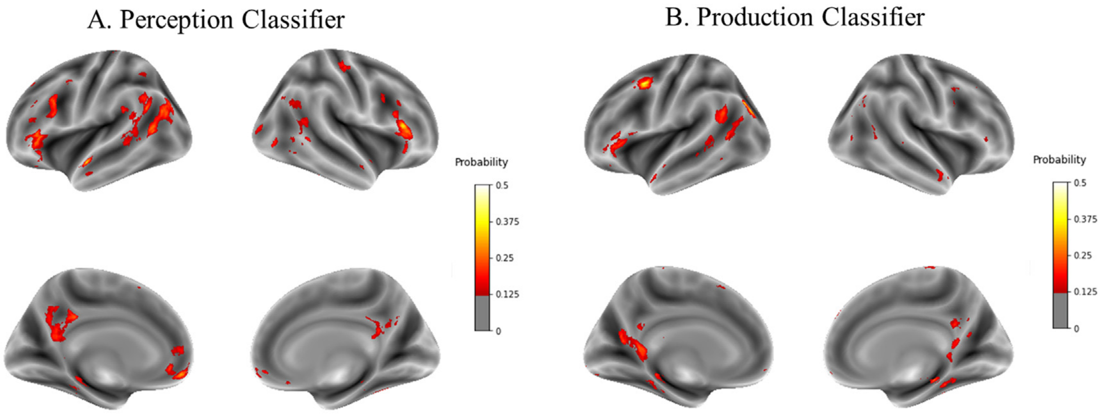


Fig. 5 – Distribution of the 5000 most informative voxels across subjects. (A) Most informative voxels for training the perception classifier. (B) Most informative voxels for training the production classifier.

surprisingly, there are individual differences that limit generalizability (Pereira et al., 2018; Wang et al., 2020). Future work could address this point to develop a more universal language classifier.

The successful decoding of the unlearned Mandarin words with the perception classifier confirmed the robustness of the decoding procedure proposed by Pereira et al. (2018). This result not only indicated the cross-language generalizability of the decoding method but also provided evidence for the plausible neural basis of the mental representation of linguistic concepts. Overall, our work demonstrates the potential for applying fMRI as a universal speech decoder for patients with speech impairment (Ho et al., 1998; Bruno et al., 2011; Ylvisaker, 1993; Baldo, Klostermann, & Dronkers, 2008). This decoder only requires brain data during perception and does not require motion experience or healthy phonological processing and motion planning mechanisms. The results of the present study may contribute to a better understanding of language storage and further advance the development of speech decoders.

Data and code availability

The scripts and data can be found on Mendeley Data, V3, https://data.mendeley.com/datasets/bt683rhwtf/3 (Lin and Hsieh, 2022). We adopted the scripts leanDecoder.m and trainVoxelwiseTargetPredictionModels.m from Pereira et al. ‘s work (2018) that is available at https://osf.io/crwz7/wiki/home/[project page], https://www.dropbox.com/sh/5z1ikn8osaao57w/AABg9LIlJfEOrQgZN7Sj7WHRa?dl=0&preview=learnDecoder.m [leanDecoder.m], and https://www.dropbox.com/sh/5z1ikn8osaao57w/AABg9LIlJfEOrQgZN7Sj7WHRa?dl=0&preview=trainVoxelwiseTargetPredictionModels.m [trainVoxelwiseTargetPredictionModels.m].

Author contributions

Yi Lin: Conceptualization; Investigation; Formal analysis; Software; Writing - original draft.

Po-Jang Hsieh: Conceptualization; Funding acquisition; Writing - review & editing.

Funding

This research was funded by the Ministry of Education to National Taiwan University (110L9A00701), Yushan Young Scholar Program (NTU-110V0202) and Ministry of Science and Technology (109-2410-H-002-004-MY3).

Open practices

We report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study. No part of the study procedures or analyses was preregistered prior to the research being conducted.

The study in this article earned Open Data and Open Materials badges for transparent practices. Materials and data for the study are available at https://data.mendeley.com/datasets/bt683rhwtf/3

Declaration of competing interest

The authors declare no competing interests.

Acknowledgment

We thank the Taiwan Mind and Brain Imaging Center at National Chengchi University and the Imaging Center for Integrative Body, Mind, and Culture Research at National Taiwan University.

REFERENCES

Anderson, A. J., Kiela, D., Clark, S., & Poesio, M. (2017). Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics, 5*, 17–30.

Anumanchipalli, G. K., Chartier, J., & Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature, 568*(7753), 493–498.

Baldo, J. V., Klostermann, E. C., & Dronkers, N. F. (2008). It’s either a cook or a baker: Patients with conduction aphasia get the gist but lose the trace. *Brain and Language, 105*(2), 134–140.

Bfaroni, M., Dinu, G., & Kruszewski, G. (2014, June). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 238–247). Long Papers.

Bonner, M. F., & Epstein, R. A. (2021). Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature Communications, 12*(1), 1–16.

Bruno, M. A., Bernheim, J. L., Ledoux, D., Pellas, F., Demertzi, A., & Laureys, S. (2011). A survey on self-assessed well-being in a cohort of chronic locked-in syndrome patients: Happy majority, miserable minority. *BMC Ophthalmology, 1*(1).

Chakrabarti, S., Sandberg, H. M., Brumberg, J. S., & Krusienski, D. J. (2015). Progress in speech decoding from the electrocorticogram. *Biomedical Engineering Letters, 5*(1), 10–21.

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

Dash, D., Ferrari, P., & Wang, J. (2020). Decoding imagined and spoken phrases from non-invasive neural (MEG) signals. *The Florida Nurse, 14*, 290.

Dash, D., Ferrari, P., & Wang, J. (2021, January). Role of brainwaves in neural speech decoding. In *2020 28th European Signal Processing Conference* (EUSIPCO) (pp. 1357–1361). IEEE.

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

Duffy, J. R. (2013). *Motor speech disorders: Substrates, differential diagnosis, and management*. St. Louis, MO: Elsevier.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160.

Harris, Z. S. (1954). Distributional structure. *Word, 10*(2–3), 146–162.

Herff, C., Heger, D., De Pesters, A., Telaar, D., Brunner, P., Schalk, G., & Schultz, T. (2015). Brain-to-text: Decoding spoken phrases from phone representations in the brain. *The Florida Nurse, 9*, 217.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience, 13*(2), 135–145.

Ho, A. K., Iansek, R., Marigliani, C., Bradshaw, J. L., & Gates, S. (1998). Speech impairment in a large sample of patients with Parkinson's disease. *Behavioural Neurology, 11*(3), 131–137.

Hollis, G. (2017). Estimating the average need of semantic knowledge from distributional semantic models. *Memory & Cognition, 45*(8), 1350–1370.

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

Kempen, G., & Huijbers, P. (1983). The lexicalization process in sentence production and naming: Indirect election of words. *Cognition, 14*, 185–209.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211.

Levelt, W. J. (1992). Accessing words in speech production: Stages, processes and representations. *Cognition, 42*(1–3), 1–22.

Levelt, W. J. (1993). *Speaking: From intention to articulation* (Vol. 1). MIT press.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *The Behavioral and Brain Sciences, 22*(1), 1–38.

Lin, Y., & Hsieh, P.-J. (2022). *Neural decoding of speech with semantic-based classification* (Vol. 3). Mendeley Data. https://data.mendeley.com/datasets/bt683rhwtf/3.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Medicine and Life, 92*, 57–78.

Martin, S., Brunner, P., Holdgraf, C., Heinze, H. J., Crone, N. E., Rieger, J., … Pasley, B. N. (2014). Decoding spectrotemporal features of overt and covert speech from the human cortex. *The Florida Nurse, 7*, 14.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). *Efficient estimation of word representations in vector space*. arXiv. Preprint arXiv:1301.3781.

Mikolov, T., Le, Q. V., & Sutskever, I. (2013b). *Exploiting similarities among languages for machine translation*. arXiv. Preprint arXiv:1309.4168.

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

Nishida, S., & Nishimoto, S. (2018). Decoding naturalistic experiences from human brain activity via distributed representations of words. *Neuroimage, 180*, 232–242.

Ogar, J., Slama, H., Dronkers, N., Amici, S., & Luisa Gorno-Tempini, M. (2005). Apraxia of speech: An overview. *Neurocase, 11*(6), 427–432.

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., & Nichols, T. E. (Eds.). (2011). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Elsevier.

Pereira, F., Botvinick, M., & Detre, G. (2013). Using Wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. *Artificial Intelligence, 194*, 240–252.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., … Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications, 9*(1), 1–13.

Ramsey, N. F., Salari, E., Aarnoutse, E. J., Vansteensel, M. J., Bleichner, M. G., & Freudenburg, Z. V. (2018). Decoding spoken phonemes from sensorimotor cortex with high-density ECoG grids. *Neuroimage, 180*, 301–311.

Rehurek, R., & Sojka, P. (2011). Gensim—python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3*(2).

Riehl, C. M. (2010). The mental representation of bilingualism. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(5), 750–758.

Roelofs, A., Meyer, A. S., & Levelt, W. J. (1998). A case for the lemma/lexeme distinction in models of speaking: Comment on Caramazza and Miozzo (1997). *Cognition, 69*(2), 219–230.

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., … Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences, 118*(45), Article e2105646118.

Sharon, R. A., Narayanan, S. S., Sur, M., & Murthy, A. H. (2020). Neural speech decoding during audition, imagination and production. *IEEE Access, 8*, 149714–149729.

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.

Vigliocco, G., Antonini, T., & Garrett, M. F. (1998). Grammaticalgender is on the tip of Italian tongues. *Psychological Science, 8*, 314–317.

Vigliocco, G., Vinson, D. P., Martin, R. C., & Garrett, M. F. (1999). Is "count" and "mass" information available when the noun is not? An investigation of tip of the tongue states and anomia. *Journal of Medicine and Life, 40*(4), 534–558.

Wang, S., Zhang, J., Wang, H., Lin, N., & Zong, C. (2020). Fine-grained neural decoding with distributed word representations. *Information Sciences, 507*, 256–272.

Wiese, R. (1984). *Language production in foreign and native languages: Same or different* (pp. 11–25). Second language productions.

Ylvisaker, M. (1993). Communication outcome in children and adolescents with traumatic brain injury. *Neuropsychological Rehabilitation, 3*(4), 367–387.

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2021). *Dive into deep learning*. arXiv. Preprint arXiv:2106.11342.