Transcutaneous Auricular Vagus Nerve Stimulation Strengthens Semantic Representations of Foreign Language Tone Words during Initial Stages of Learning

Ian Phillips[1,2,3], Regina C. Calloway[1], Valerie P. Karuzis[1], Nick B. Pandža[1,4], Polly O'Rourke[1], and Stefanie E. Kuchinsky[1,2]

Abstract

■ Difficulty perceiving phonological contrasts in a second language (L2) can impede initial L2 lexical learning. Such is the case for English speakers learning tonal languages, like Mandarin Chinese. Given the hypothesized role of reduced neuroplasticity in adulthood limiting L2 phonological perception, the current study examined whether transcutaneous auricular vagus nerve stimulation (taVNS), a relatively new neuromodulatory technique, can facilitate L2 lexical learning for English speakers learning Mandarin Chinese over 2 days. Using a double-blind design, one group of participants received 10 min of continuous priming taVNS before lexical training and testing each day, a second group received 500 msec of peristimulus (peristim) taVNS preceding each to-be-learned item in the same tasks, and a third group received passive sham stimulation. Results of the lexical recognition test administered at the end of each day revealed evidence of learning for all groups, but a higher likelihood of accuracy across days for the peristim group and a greater improvement in response time between days for the priming group. Analyses of N400 ERP components elicited during the same tasks indicate behavioral advantages for both taVNS groups coincided with stronger lexico-semantic encoding for target words. Comparison of these findings to pupillometry results for the same study reported in Pandža, N. B., Phillips, I., Karuzis, V. P., O'Rourke, P., and Kuchinsky, S. E. (Neurostimulation and pupillometry: New directions for learning and research in applied linguistics. *Annual Review of Applied Linguistics*, *40*, 56–77, 2020) suggest that positive effects of priming taVNS (but not peristim taVNS) on lexico-semantic encoding are related to sustained attentional effort. ■

INTRODUCTION

Few individuals who begin learning a foreign or second language (L2) as adults ultimately attain native-like perception of target language phonological categories (e.g., Díaz, Mitterer, Broersma, & Sebastián-Gallés, 2012; Abrahamsson & Hyltenstam, 2009; Long, 1990). This finding is often attributed to early closure of a sensitive period for phonological development, after which it becomes increasingly difficult to perceive and produce new phonological contrasts (e.g., Kuhl, 2010; Mattock, Molnar, Polka, & Burnham, 2008). Enhanced neuroplasticity—the ability of neural circuits in the brain to change in response to experience—is thought to underlie the rapid learning that characterizes sensitive periods (for a review, see White, Hutka, Williams, & Moreno, 2013). As first language (L1) phonemic categories become entrenched, a diminished capacity for neural reorganization affects the degree to which new categories are learned (Werker & Hensch, 2015; Kuhl, 2004), which has consequences across adult L2 learning domains, including the lexicon. The ability to learn new L2 words is generally not thought to be maturationally constrained (e.g., Hellman, 2011; but see Granena & Long, 2013, for other age-related lexical constraints); however, difficulty perceiving nonnative phonological contrasts impedes initial L2 lexical learning (Poltrock, Chen, Kwok, Cheung, & Nazzi, 2018; Cooper & Wang, 2012; Wong & Perrachione, 2007) and can have persistent negative effects on L2 lexical processing (Pelzl, Lau, Guo, & DeKeyser, 2021; Ling & Grüter, 2020; Pelzl, 2019; Sebastián-Gallés, Echeverría, & Bosch, 2005). To test whether enhancing neuroplasticity facilitates initial learning of L2 words featuring nonnative phonological contrasts, we administered transcutaneous auricular vagus nerve stimulation (taVNS)—a relatively new approach to inducing neuroplasticity by electrically stimulating peripheral afferent fibers of the vagus nerve—to native speakers of English while they learned a set of Mandarin pseudowords distinguished in part by lexical tone—a phonological contrast notoriously difficult for speakers of nontonal languages to master. In a previous article, we reported novel findings that taVNS modulates attentional effort and improves lexical learning in this

[1]University of Maryland Applied Research Laboratory for Intelligence & Security, College Park, MD, [2]Walter Reed National Military Medical Center, Audiology and Speech Pathology Center, Bethesda, MD, [3]The Geneva Foundation, Bethesda, MD, [4]University of Maryland Program in Second Language Acquisition, College Park, MD

paradigm (Pandža, Phillips, Karuzis, O'Rourke, & Kuchinsky, 2020). In this study, we build on these findings and report results for concurrently recorded electrophysiological data that further illuminate the mechanisms by which taVNS benefits L2 lexical learning.

Over half the world's languages are tonal (Yip, 2002), using differences in pitch to distinguish word meaning, as illustrated in the following common example from Mandarin Chinese, which has four contrastive tones plus a fifth, neutral tone (Wang, Spence, Jongman, & Sereno, 1999): /ma/ means "mother" when produced with a high flat tone (Tone 1), "hemp" when produced with a tone rising from mid to high (Tone 2), "to scold" when produced with a tone falling from high to low (Tone 4), and "horse" when produced with a tone that falls and then rises (Tone 3). Studies of short-term vocabulary training, focused on associating meanings with small sets of highly controlled tone words, show naive learners can achieve high levels of accuracy identifying tone words heard in isolation over several training sessions (Antoniou & Wong, 2016; Perrachione, Lee, Ha, & Wong, 2011; Chandrasekaran, Sampath, & Wong, 2010). Illustrating the important role of L2 phonetic perceptual ability and lexical learning, short-term lexical tone perception training has been shown to facilitate tone word learning (Cooper & Wang, 2013; Ingvalson, Barr, & Wong, 2013) similarly to benefits of musical experience and ability in discriminating and identifying both linguistic and nonlinguistic pitch categories (e.g., Poltrock et al., 2018; Bowles, Chang, & Karuzis, 2016; Dittinger et al., 2016; Wong & Perrachione, 2007).

During initial stages of L2 lexical learning, evidence of lexico-semantic development following fairly limited word exposure has been obtained using implicit measures, such as the N400 ERP component (e.g., Dittinger et al., 2016; Pu, Holcomb, & Midgley, 2016). The N400 is a negative-going deflection usually largest over centro-parietal scalp regions that peaks around 400 msec following word onset and is sensitive to a word's expectancy or plausibility (Nieuwland et al., 2020; Kutas & Federmeier, 2011; Kutas & Hillyard, 1980). The difference in N400 amplitude between unexpected and expected words given a preceding context, dubbed the N400 effect, reflects ease of lexico-semantic processing. Items that match an expectation set by the preceding context have reduced N400 amplitude and are easier to access and/or integrate than unexpected items (Lau, Phillips, & Poeppel, 2008). During word learning, increases in N400 amplitude (Borovsky, Elman, & Kutas, 2012; McLaughlin, Osterhout, & Kim, 2004) and shifts from frontal to centro-parietal topography with repeated word exposure are also thought to reflect the development of semantic representations (Dittinger et al., 2016; Rodríguez-Fornells, Cunillera, Mestres-Missé, & de Diego-Balaguer, 2009). Frontal effects early during learning may reflect increased working memory or cognitive control required to access a new item's semantic representation (Elgort, Perfetti, Rickles, & Stafura, 2015; Mestres-Missé, Rodriguez-Fornells, & Münte, 2007; Rodríguez-Fornells, De Diego Balaguer, & Münte, 2006) with demands lessening as items become lexicalized.

In a study of the impact of musical training on initial L2 word learning, Dittinger et al. (2016) used the N400 to measure lexico-semantic development over a single training session in which French-speaking professional musicians and nonmusicians learned L2 Thai words distinguished by several nonnative phonological contrasts, including lexical tone. Although musicians outperformed nonmusicians in only two behavioral tests of word learning, N400 results indicated faster and better learning for musicians across all tasks: Only musicians showed an increase in N400 amplitude at centro-parietal electrodes during the passive word learning task (indicating faster lexico-semantic encoding) and a centro-parietal N400 effect in the short-term matching task (larger amplitude for unexpected vs. expected words, indicating stronger semantic encoding) and semantic relatedness task (larger amplitude for unrelated vs. related words, indicating stronger integration of target words in existing semantic networks). These findings and those of other studies (e.g., Borovsky et al., 2012; Perfetti, Wlotko, & Hart, 2005; McLaughlin et al., 2004) illustrate the ability of the N400 to track initial development of lexical knowledge, even before behavioral changes occur, and support using the N400 to evaluate effects of taVNS on initial lexico-semantic encoding.

Electrical stimulation of the vagus, the tenth cranial nerve within the autonomic nervous system, has been explored over the past 30 years for treating neurological and neuropsychiatric disorders, including refractory epilepsy and depression. Findings from this research show that direct vagus nerve stimulation administered via implanted devices (iVNS) can alter autonomic nervous system activity (e.g., Desbeaumes Jodoin, Lespérance, Nguyen, Fournier-Gosselin, & Richer, 2015) and improve cognitive function, including learning and memory (e.g., Sun et al., 2017; Clark, Naritoku, Smith, Browning, & Jensen, 1999; see Vonck et al., 2014, for a summary). More recently, applying low voltage electrical stimulation via taVNS to the skin of the external auditory canal, inner tragus, or cymba conchae, which are innervated to varying degrees by the auricular branch of the vagus nerve (Butt, Albusoda, Farmer, & Aziz, 2020), has been shown to increase activity in vagal brainstem projections in adult humans, including the nucleus of the solitary tract and the locus coeruleus (LC; Yakunina, Kim, & Nam, 2017; Frangos, Ellrich, & Komisaruk, 2015; Kraus et al., 2013) and to confer similar cognitive benefits for healthy adults as compared to iVNS (e.g., Jacobs, Riphagen, Razat, Wiese, & Sack, 2015).

The exact mechanism by which VNS influences cognition is not fully understood, but evidence suggests increased LC activity plays an important role. The LC is the primary source of the neurotransmitter norepinephrine (NE) in cortex, modulating various cortical and

subcortical circuits involved in arousal, attention, sensory processing, and memory formation (Berridge & Waterhouse, 2003). The LC-NE system is thought to influence memory formation by regulating NE in the hippocampus and prefrontal cortex that supports long-term potentiation (LTP; Vonck et al., 2014), and is thought to optimize task performance by shifting between two modes of neural activity that modulate responsivity of task-relevant cortical circuits: a slow baseline, tonic pattern of firing that shifts arousal state, and a more rapid, task-evoked phasic burst of activity that facilitates task-relevant responses and is maximal at moderate levels of tonic LC activity (Aston-Jones & Cohen, 2005). The LC-NE system also modulates basal forebrain cholinergic activity, which is thought to play a critical role in regulating attentional effort necessary to maintain performance in difficult tasks by providing top–down support via the activation of attentional systems and related executive functions (Klinkenberg, Sambeth, & Blokland, 2011; Sarter, Gehring, & Kozak, 2006).

Enhancing specific cognitive functions and learning via VNS likely depends on a complex interplay of these neuromodulatory circuits (Hulsey, Shedd, Sarker, Kilgard, & Hays, 2019; Hulsey et al., 2016) and the degree to which each system contributes to plasticity and learning may depend on VNS timing. It has been found that taVNS improves associative memory when delivered continuously during encoding and consolidation phases of an association memory task (Jacobs et al., 2015) and L2 letter–sound mapping when paired with learning feedback (Thakkar, Engelhart, Khodaparast, Abadzi, & Centanni, 2020). In a study of Mandarin lexical tone perception, perceptual categorization was improved for tones that were paired with taVNS during training but not for tones that occurred in the same training task but were not paired with taVNS (Llanos et al., 2020), which parallels findings in animal models of iVNS affecting auditory processing only when temporally coupled to stimuli (Engineer, Engineer, Riley, Seale, & Kilgard, 2015).

The potential for taVNS to support L2 lexical learning both by inducing more global, slow-changing effects in tonic LC-NE activity and more rapid, transient effects in LC-NE phasic activity motivated the comparison of two taVNS timings in this study: delivering taVNS continuously for 10 min before naive L2 Mandarin learners (L1 English) completed tone categorization and lexical learning tasks (*priming taVNS*) or for 500 msec before each to-be-learned item within the same tasks (*peristimulus [peristim] taVNS*). We previously reported behavioral and pupillometry results for this study in Pandža et al. (2020) that showed better learning for priming and peristim taVNS over a passive sham taVNS control group and smaller task-evoked pupil responses for peristim taVNS, reflecting reductions in the allocation of cognitive effort as participants engaged in successful learning. These findings provide some of the first evidence linking taVNS-related learning improvements to changes in cognitive effort; however, it has yet to be shown how these differences in effort allocation relate to the development of L2 lexical knowledge. This study takes a first step in addressing this question by analyzing N400 ERP components elicited during the lexical learning tasks reported in Pandža et al. (2020) and interpreting them in the context of the previous behavioral and pupillometry results.

With this analysis, we sought to answer two research questions (RQs): RQ1. Does taVNS support initial development of lexico-semantic representations for L2 words as evidenced by N400 amplitude and topography? RQ2. Do priming and peristim taVNS have differential effects on initial L2 lexico-semantic development? For RQ1, we hypothesized that both taVNS timings would lead to more robust lexico-semantic encoding for novel L2 words earlier in training, reflected in stronger centro-parietal N400s earlier during learning and larger N400 effects in later recognition testing. For RQ2, we hypothesized that priming and peristim taVNS might have differential effects on N400 amplitude and topography given potential timing-related differences in underlying neuromodulatory mechanisms, but we could not make specific predictions because of the lack of literature characterizing VNS timing effects on the N400 component.

METHODS

This study was approved by the University of Maryland's Institutional Review Board and the U.S. Department of Navy Human Research Protection Program. Priming and peristim taVNS were tested in separate experiments that used identical materials and procedures except for aspects of taVNS delivery, described below in *taVNS Parameters*. Participants completed either the priming or the peristim experiment with active taVNS (priming or peristim) compared to a separate passive taVNS (sham) control group in each experiment. To directly compare the effects of priming to peristim taVNS in this study, all data were combined into a single analysis. The method and results are reported for the overall study with differences between the two experiments noted when necessary.

Participants

Participants gave informed consent before enrolling in this study and were paid for their time. Eighty-two participants completed the study. All reported being right-handed, native speakers of English, with normal or corrected-to-normal vision, normal hearing, no previous exposure to any tone languages, and no history of psychological or neurological disorders (see Pandža et al., 2020, for additional inclusion criteria). To balance taVNS groups on nonlinguistic pitch contour identification ability and musicianship, participants were assigned to a taVNS group (priming/peristim vs. sham) based on their overall accuracy on a shortened version (126 trials)

Table 1. taVNS Group Descriptive Statistics

				OMSI Response Count			
taVNS Group	*N (female)*	*Age (SD)*	*PCID Score (SD)*	*NM*	*ML*	*AM*	*SA*
Priming	12 (8)	22.7 (4.19)	0.67 (0.12)	2	5	3	2
Peristim	13 (9)	21.7 (2.87)	0.68 (0.10)	4	5	3	1
Sham	20 (12)	22.1 (4.01)	0.65 (0.11)	4	10	3	3

NM = nonmusician; ML = music-loving nonmusician; AM = amateur musician; SA = serious amateur musician.

of a pitch contour identification task (PCID; Bowles et al., 2016; Bent, Bradlow, & Wright, 2006) and their response to one item from the Ollen Musical Sophistication Index (OMSI; Zhang & Schubert, 2019; Ollen, 2006): *What title best describes you? 1 = Nonmusician, 2 = music-loving nonmusician, 3 = amateur musician, 4 = serious amateur musician, 5 = semiprofessional musician, 6 = professional musician*.

To link behavioral and electrophysiological indices of learning, the present analysis includes only participants with complete data sets for each measure and experimental task (see Analysis Approach section). Forty-five participants (n = 12 priming, n = 13 peristim, n = 20 sham) are included in the present analysis. These participants (29 female) were 18–34 years old (M = 22.13, SD = 3.70), and taVNS groups did not differ significantly on mean PCID score (ANOVA: $F(2, 42) = 0.34$, $p = .71$), PCID score variance (Levene's test: $F(2, 42) = 0.54$, $p = .59$), or OMSI score (Kruskal–Wallis rank sum test: $\chi^2(2) = 0.77$, $p = .68$; see Table 1). Whereas all taVNS groups included fewer participants than in Pandža et al. (2020; n = 17 priming, n = 17 peristim, n = 35 sham), the smaller sample sizes for the priming and peristim taVNS conditions in the present analysis are similar to that of Llanos et al. (2020; 12 participants per group), which found peristim taVNS effects on Mandarin tone categorization. Thus, the final sample size in the present analysis was determined a priori to be powerful enough to detect taVNS-related differences of lexical learning.

Materials

The target items in this study were nine phonologically plausible monosyllabic Mandarin pseudowords, consisting of the syllables /ba/, /bi/, and /pi/, each produced with Mandarin Tones 1 (flat), 2 (rising), and 4 (falling) by two native speakers of Mandarin (one female, one male). Tone 3 was not included in this study because of a creaky quality often observed in Tone 3, which could lead to easier identification from the other tone stimuli and to avoid the added difficulty of learning to discriminate between Tones 2 and 3, which are acoustically similar and more easily confused by L2 and L1 speakers of Mandarin (Hao, 2012; Shen & Lin, 1991). These items were recorded along with stimuli created for a previous study of Mandarin tone word learning (Bowles et al., 2016) using the same recording procedure. Each auditory recording was root-mean-squared normalized and presented at 70 dBA sound pressure level (SPL). The same recordings of the Mandarin pseudowords were used across tasks without any manipulations to stimulus duration or other acoustic features. Each pseudoword was paired with an English noun (*COIN, GOWN, LENS, MENU, OVEN, RAFT, SOFA, TRAY, VASE*), which appeared only in written form across tasks. English pairings were selected to control frequency (logSUBTLEX: 2.29–2.71; Brysbaert & New, 2009), concreteness (4.61–5.00; Brysbaert, Warriner, & Kuperman, 2014), word length (four letters, one to two syllables), and animacy. Three lists of Mandarin–English word pairings were created such that across lists, each English word occurred once with each tone (1, 2, 4) and once with each syllable (/ba/, /bi/, /pi/). Each participant encountered only one list across all tasks and sessions.

Procedures

Participants completed all tasks in a sound-attenuated, dimly lit room, seated 65 cm from a 24-in. LCD computer monitor. Tasks were delivered by E-Prime 2.0 software (Psychology Software Tools, Inc., 2012) with auditory Mandarin pseudowords presented via Neuvana (Neuvana [previously Nervana], LLC) first generation earbuds connected by a 3.5-mm audio plug to a Chronos device, which was used to collect behavioral responses, and by a 2.5-mm electrode plug to a Digitimer DS8R Biphasic Constant Current Stimulator (Digitimer North America, LLC), which generated the taVNS. The left earbud had an embedded electrode with two contact areas that made contact with the superior and inferior walls of the outer ear canal when inserted into the ear. The silicone tip fitting over this earbud was modified by replacing the silicone located over the electrode contact areas with 0.5–1 cm^2 pieces of Axelgaard AG735 and/or AG2550 hydrogel (Axelgaard Manufacturing Co., Ltd) as a transmission medium to maintain consistent contact between the electrode and outer ear canal wall. Once the left earbud was inserted, the taVNS conducting electrode made contact with the most lateral 1 cm of the superior and inferior walls of the outer ear canal via the hydrogel. Each participant was fit for earbud tip size, and a new pair of earbuds was prepared each session. EEG was acquired with an Electrical Geodesics Inc. 64-channel

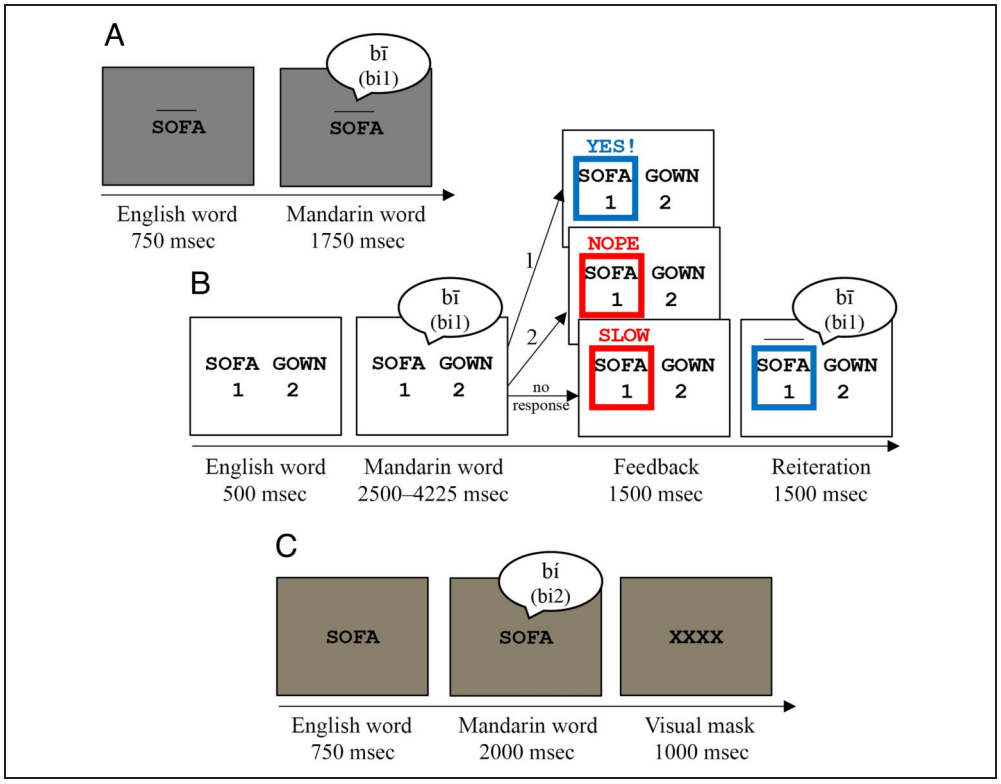

Figure 1. Sequence and durations of events in each trial for (A) the passive word learning task, (B) active word learning task, and (C) lexical recognition test. For each task, "Mandarin word" indicates the time period when the Mandarin pseudoword is auditorily presented. For active word learning (B) and lexical recognition (C), participant responses were recorded during the "Mandarin word" period of each trial.

HydroCel Geodesic Sensor Net and an NA300 high-impedance amplifier using Net Station software, Version 5.4.1.1 (r26882, Electrical Geodesics Inc.). Data were sampled at 1000 Hz, referenced online to the vertex (Cz), using a 400-Hz analog low-pass filter. Impedances were kept below 50 KΩ where possible and otherwise under 100 KΩ, as is customary with this system. EEG was preprocessed off-line in MATLAB 2017b using EEGLab Version 14.1.2 (Delorme & Makeig, 2004) and ERPLab Toolbox Version 7.0.0 (Lopez-Calderon & Luck, 2014).

During a pretraining session, at least 1 day before training began, participants completed several tasks to establish study eligibility and collect individual difference measures (including PCID and OMSI) and were assigned a taVNS group number by a member of the research team not involved in data collection or analysis. Training tasks were programmed to deliver the appropriate taVNS (priming/peristim/sham) based on participant number, and this design ensured that both proctors and participants were blind to the taVNS condition (Pandža et al., 2020). The two training sessions were completed on consecutive days or with 1 day in between. In each training session, participants first completed a task to familiarize them with the Mandarin tones appearing in the stimuli. In this task, participants read descriptions of each tone contour on the monitor and were shown corresponding visual representations (flat, rising, or falling lines) and listened to recorded examples of each tone produced with the vowel /a/ by a male native speaker of Mandarin. No data were collected during this task. After familiarization, participants completed phonological categorization and discrimination pretests (Day 1 only), a phonological training task, the passive and active word learning tasks followed by the lexical recognition test described below, and phonological categorization and discrimination posttests (Day 2 only). The data presented here are from a larger taVNS study, but only the tasks relevant to word learning are detailed here because of the scope of the present RQs (see Pandža et al., 2020, for the full task sequence and details of the PCID, OMSI, and tone familiarization tasks). Each session lasted approximately 3 hr. Illustrations of each task are shown in Figure 1.

Passive Word Learning

The passive word learning task comprised 90 trials and lasted 7–8 min including instructions. Each trial began with a 750-msec period in which one of the English words appeared in the center of the monitor (40 pt. black Courier New, gray background) with an image of its tonal contour above the word, followed by a 1750-msec period that began with the auditory presentation of the corresponding Mandarin pseudoword while the English translation equivalent remained on screen. Visual depictions of the tone contours were included based on results of pilot testing, which indicated the task may have been too difficult for participants to learn the tonal contrasts when relying solely on auditory input. Each Mandarin pseudoword was presented 10 times (5 times per speaker), and trial order was pseudorandomized so that no more than four consecutive trials had the same

syllable, tone, or speaker, and no consecutive trials had the same Mandarin pseudoword. Participants were instructed to try to memorize the meaning of each Mandarin pseudoword and to focus on the center of the screen and limit movements during the task. There was no practice for this task. EEG was recorded continuously during this task, and participants did not make behavioral responses.

Active Word Learning

The active word learning task comprised 36 trials, lasting 3–4 min with instructions. Each trial began with a 500-msec period in which two of the English words appeared side by side in the center of the monitor above the numbers 1 and 2 (35 pt. black Courier New, white background); followed by a 2500- to 4225-msec period that began with the auditory presentation of one Mandarin pseudoword while the English words remained on screen and ended once participants pressed Button 1 or 2 to indicate the correct translation; then a 1500-msec period in which a box appeared around the correct response and a word ("YES!", "NOPE," or "SLOW") appeared above the English words to indicate performance (blue if correct, red if incorrect); and a 1500-msec period in which the Mandarin pseudoword was presented again with an image of its tone contour above the correct English word. Each button was correct in half of the trials. In each trial, the Mandarin translation equivalents for the distractor and correct item differed in one of three ways: same tone, different syllable; different tone, same syllable; and different tone, different syllable. Each Mandarin pseudoword occurred with each distractor type at least once, and there were three to five occurrences of each tonal confusion pair (e.g., correct word is Tone 1, distractor is Tone 2) and syllable confusion pair (e.g., correct word is /ba/, distractor is /bi/). Each Mandarin pseudoword was presented 4 times (2 times per speaker), and trial order was pseudorandomized so no more than four consecutive trials contained the same syllable, tone, speaker, or distractor type, and no consecutive trials contained the same English words. Participants were instructed to select the correct English equivalent for each Mandarin pseudoword, and there was no practice for this task. Accuracy and RT were recorded for each trial.

Lexical Recognition Test

The lexical recognition test included 216 trials, lasting about 20 min with instructions. Each trial began with a 750-msec baseline period in which one of the English words appeared in the center of the monitor (30 pt. black Courier New, tan background); followed by a 2,000-msec period in which one Mandarin pseudoword was auditorily presented while the English word remained on screen and participants pressed a button to indicate whether stimuli were translation equivalents (match, Button 1) or not (mismatch, Button 2); and a 1000-msec period in which a visual mask ("XXXX") replaced the English word on screen. Each Mandarin pseudoword was presented 24 times, split over two testing blocks, and occurred in an equal number of match and mismatch trials. Trial order was pseudorandomized so that no more than four consecutive trials contained the same syllable, tone, speaker, or trial condition, and no consecutive trials contained the same Mandarin pseudoword or English word. Because of these requirements, the number of times each auditory pseudoword was produced by each speaker could not be perfectly balanced. Whereas some pseudowords were spoken 12 times by each speaker, others were spoken 16 times by one speaker and 8 times by the other speaker. This varied across presentation lists. The order of trial conditions was consistent within each block, but the specific stimulus pair appearing in each trial was randomized for each session. Participants were instructed to indicate as quickly and accurately as possible whether the English word and the Mandarin pseudoword in each trial were translation equivalents and to limit blinking to the mask portion of the trial. There was no practice or feedback. Accuracy, RT, and continuous EEG were recorded.

taVNS Parameters

The taVNS consisted of a biphasic square wave (50-μs pulse width, 350-μs interphase dwell, 100% recovery phase ratio) triggered at 300 Hz by a TTL pulse over a BNC jack from a custom-programmed Arduino UNO board controlled by E-Prime. All participants completed a calibration and ramping procedure before each task where taVNS was possible to determine their perceptual threshold (see Pandža et al., 2020, for details). In the subsequent tasks, priming and peristim participants received taVNS at 0.2 mA below their perceptual threshold to prevent them from feeling any sensation because of taVNS that might unblind them to their taVNS condition and affect their task performance. For the priming experiment, taVNS was delivered to participants in the priming group 3 times each session while they watched a 10-min silent animated video used for resting-state fMRI scans (Inscapes; Vanderwal, Kelly, Eilbott, Mayes, & Castellanos, 2015). This task was administered about 9 min before passive word learning, 18 min before active word learning, and immediately before the lexical recognition test, but only participants in the priming group received continuous taVNS during the video. In the peristim experiment, the video task was not administered. Instead, participants in the peristim group received 500-msec bursts of taVNS immediately preceding each trial in both word learning tasks and the lexical recognition test. Participants in the sham taVNS group for each experiment completed the same tasks but did not receive taVNS outside of calibration and ramping.

During piloting, it was discovered that taVNS at higher intensity levels occasionally produced an audible noise artifact. To mask this noise, a 60 dB SPL pink noise mask was played whenever taVNS was possible in the priming experiment, and the pink noise mask overlaid with a recording of the taVNS sound artifact was played in the peristim experiment. More specifically, the priming group and the sham group matched to the priming group heard the same pink noise mask during the priming video, whereas the peristim group and the sham group matched to the peristim group heard the same pink noise mask overlaid with the taVNS sound artifact for 500 msec prior to the onset of the Mandarin pseudowords in the word learning tasks and lexical recognition test. All participants also heard their respective sound masks during each taVNS calibration and ramping.

Analysis Approach

Preliminary analyses of active word learning and lexical recognition accuracy and RT, following the procedure outlined here, did not reveal major behavioral differences between sham groups—the only significant difference was a slowdown in RT to mismatch trials on Day 2 for the peristim sham group but not the priming sham group. Thus, data from the priming and peristim experiments were combined into a single, three-taVNS-group analysis, with one sham group containing all participants who received sham taVNS across the priming and peristim experiments. Accuracy was analyzed using logistic mixed-effects models, and log RT and ERP mean amplitude were analyzed using linear mixed-effects models, in R Version 3.6.3 (R Core Team, 2020). The buildmer package (v. 1.5; Voeten, 2020) was used to automatically determine the best-fitting model for each measure by first determining the maximal random- and fixed-effects structures that allowed the model to converge, ordering effects by the magnitude of their contribution to model fit, and then removing terms in a backward stepwise procedure until the model contained only factors that significantly improved model fit. This method provides a more objective and replicable way to fit exploratory mixed-effects models. Improvements in fit between nested models were assessed with likelihood ratio tests, and p values reported for linear model fixed effects were calculated using Satterthwaite's approximation for degrees of freedom. For significant fixed effects, model-predicted values shown in the below tables were obtained using the effects package (v. 4.1–4; Fox & Weisberg, 2018), which weights factor levels by sample size, absorbs lower-order terms for interactions, and averages over noninteracting terms. In all models, factors were treatment coded and PCID (centered, z-score transformed) and OMSI (centered on the mean) scores were included to model individual differences known to impact tone word learning. To interpret model fixed effects involving three-level factors, variables were releveled as necessary (and indicated in the text) to obtain model estimates for all factor level comparisons.

Prior to the analysis of accuracy and RT in the lexical recognition test, the time until voicing in each sound file was subtracted from the RT for corresponding trials because these portions of the Mandarin pseudowords did not carry tone information, which was required to determine whether the items were a match or mismatch. This was not done for the active word learning task because the syllable-initial phoneme could distinguish the correct item for some trials. For accuracy analyses in both tasks, trials were excluded if the response was recorded within 60 msec of the adjusted onset of the Mandarin pseudoword (0.03% of trials for active word learning; 0.05% for lexical recognition) and for RT analyses, trials that received incorrect responses were also excluded (28.31% of trials for active word learning; 27.34% for lexical recognition).

Prior to analysis, EEG was resampled at 250 Hz, high-pass filtered (0.1 Hz, Butterworth, second order) and notch filtered at 60 Hz to remove line noise. Loose channels were interpolated from surrounding sites, epochs were extracted from −200 to 1000 msec relative to the onset of the Mandarin pseudoword in each trial, and ocular artifacts were corrected based on an independent components analysis (Luck, 2014). Individual epochs were then low-pass filtered (30 Hz, Butterworth, second order), channels were rereferenced to the average of all sites, and epochs were baseline-corrected against the prestimulus period. Epochs with deflections exceeding ±100 μV on any channel were excluded from analysis (mean 9% for passive word learning; mean 6% for lexical recognition). EEG data for participants who had greater than 25% of epochs rejected due to artifacts in either session were excluded from analyses. The 300–500 msec following the onset of voicing in each trial was selected as the analysis window based on the N400 literature. For the passive word learning task, the dependent variable was mean amplitude over the analysis window at each of nine electrodes over lateral (left hemisphere: F3, C3, P3; right hemisphere: F4, C4, P4) and midline sites (Fz, Cz, Pz) covering frontal, central, and parietal regions (based on the 10–10 system; Luu & Ferree, 2005) calculated by participant and session. For the lexical recognition test, the dependent variable was the difference in mean amplitude between mismatch and match trials (mismatch minus match) at each of the nine electrode sites calculated by participant and session.

The maximal fixed-effects structures for accuracy and RT analyses included the three-way interaction between taVNS group (Group: Sham/Priming/Peristim), training session (Sess: Day 1/Day 2), and trial condition (Cond: Tone/Segment/Tone + Segment for active word learning; Cond: Match/Mismatch for lexical recognition) and all lower-order terms, plus noninteracting terms for musicianship (OMSI) and nonlinguistic pitch contour identification task (PCID) ability. The maximal random-effects

structures included by-participant random intercepts and slopes for the Sess × Cond interaction and lower-order terms plus by-item random intercepts and slopes for the three-way Group × Sess × Cond interaction and all lower-order terms. The maximal fixed-effects structure for ERP analyses included the four-way interaction between Group, Sess, anterior–posterior electrode region (Region: Fnt/Cnt/Par [frontal/central/parietal]), and electrode laterality (Lat: Left/Midline/Right) and all lower-order terms, plus three-way and lower-order interactions between the two electrode location factors and OMSI and PCID; the maximal random-effects structure included by-participant random intercepts and slopes for the three-way Sess × Region × Lat interaction and all lower-order interactions. Treatment coding electrode region and laterality factors provided direct comparisons between all three factor levels (e.g., right vs. midline vs. left) in the same model, and estimates for all comparisons were obtained by releveling.

Following the main analyses, post hoc models were run to more directly explore the relationship between behavioral outcomes in the lexical recognition test and differences in semantic encoding reflected in passive word learning N400 amplitude and topography. Lexical recognition mean accuracy (Acc) and RT by participant and session (both centered, z score transformed across sessions) were tested in two separate linear mixed-effects models. In each model, the behavioral predictor was added to the best-fitting passive word learning N400 model as a simple fixed effect and as interactions with all other fixed effects. To determine whether the inclusion of the behavioral predictor improved N400 model fit, the maximal model underwent the same model fitting procedure described above, with only model terms involving behavioral predictors subject to removal.

RESULTS

Passive Word Learning Task

N400

By-participant N400 mean amplitude (300–500 msec) was analyzed for the passive word learning task to determine whether there were differences in N400 magnitude or topography (by region and laterality) between taVNS groups and sessions that would indicate differences in the amount of exposure needed to encode novel tone words in semantic memory. Topographic plots of mean amplitude in Figure 2 reveal a prominent negativity centered over central or centro-parietal midline sites across taVNS groups and sessions, with the exception of the priming group on Day 1 where the negativity spans central, frontal, and frontopolar sites.

The best-fitting model for these data (Table 2) retained a significant Region × Sess × Group interaction, indicating differences in N400 anterior–posterior topography shifts from Day 1 to 2 between taVNS groups. For the priming group on Day 1, frontal sites were negative ($b = -1.06$, $SE = 0.36$, $p = .004$) and not different from central ($b = -0.16$, $SE = 0.38$, $p = .67$) and parietal sites ($b = 0.63$, $SE = 0.57$, $p = .27$), although parietal sites were less negative than central sites (with Priming, Cnt, Day 1 reference levels: Region(Par): $b = 0.79$, $SE = 0.38$, $p = .039$). For priming on Day 2, frontal sites did not differ from zero and central and parietal sites were more negative (with Priming, Fnt, Day 2 reference levels: Intercept: $b = 0.03$,

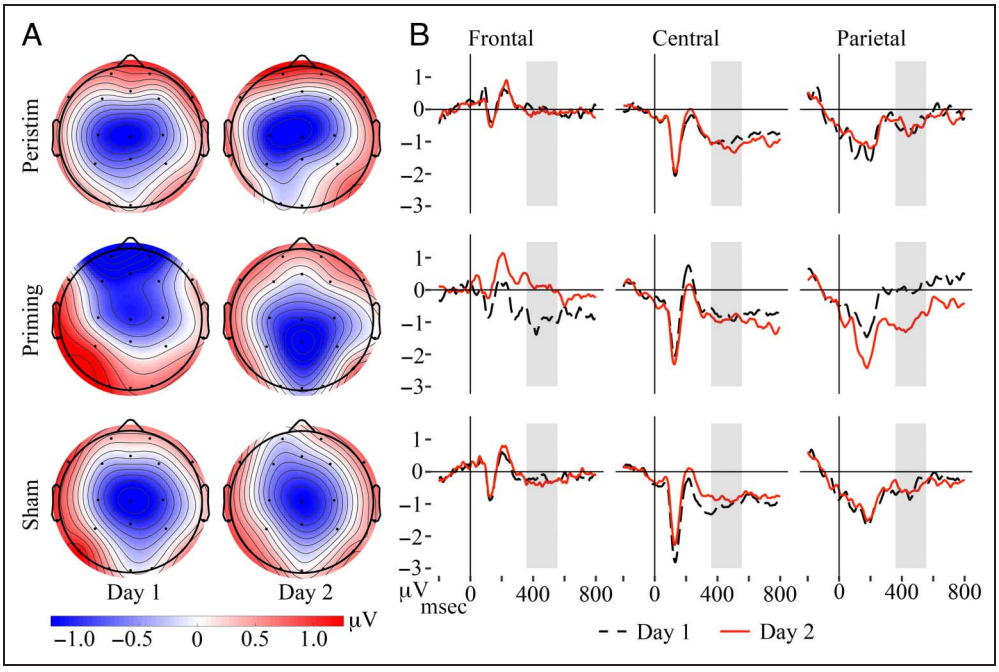

Figure 2. (A) Topographic plots of by-participant N400 mean amplitude during the analysis window in the passive word learning task. (B) By-participant ERPs averaged over all sites in each region comparing taVNS group and session for the passive word learning task. The onset of the Mandarin pseudoword is at 0 msec; gray boxes indicate the mean onset and offset for the adjusted analysis window across trials.

Table 2. Best-Fitting Model for Passive Word Learning N400 Mean Amplitude

Fixed Effect	*Est.*	*SE*	*df*	*t value*	*p value*	
(Intercept)	−1.06	0.36	58.79	−2.99	.004	**
Region(Cnt)	−0.16	0.38	67.19	−0.42	.67	
Region(Par)	0.63	0.57	50.88	1.11	.27	
Lat(Left)	0.16	0.17	656.00	0.91	.36	
Lat(Right)	0.09	0.17	656.00	0.51	.61	
Sess(Day 2)	1.10	0.20	656.00	5.55	< .001	***
Group(Sham)	0.82	0.44	55.65	1.85	.070	
Group(Peristim)	1.21	0.49	55.65	2.49	.016	*
Region(Cnt) × Lat(Left)	0.61	0.18	656.00	3.42	< .001	***
Region(Par) × Lat(Left)	0.81	0.18	656.00	4.58	< .001	***
Region(Cnt) × Lat(Right)	0.53	0.18	656.00	2.98	.003	**
Region(Par) × Lat(Right)	0.45	0.18	656.00	2.52	.012	*
Region(Cnt) × Sess(Day 2)	−1.20	0.28	656.00	−4.28	< .001	***
Region(Par) × Sess(Day 2)	−2.14	0.28	656.00	−7.66	< .001	***
Lat(Left) × Group(Sham)	−0.10	0.18	656.00	−0.56	.58	
Lat(Right) × Group(Sham)	−0.08	0.18	656.00	−0.46	.65	
Lat(Left) × Group(Peristim)	−0.58	0.19	656.00	−3.00	.003	**
Lat(Right) × Group(Peristim)	−0.12	0.19	656.00	−0.61	.54	
Region(Cnt) × Group(Sham)	−1.04	0.46	57.83	−2.27	.027	*
Region(Par) × Group(Sham)	−1.32	0.71	47.70	−1.86	.069	
Region(Cnt) × Group(Peristim)	−1.23	0.50	57.83	−2.45	.017	*
Region(Par) × Group(Peristim)	−1.54	0.78	47.70	−1.97	.055	
Sess(Day 2) × Group(Sham)	−1.22	0.25	656.00	−4.87	< .001	***
Sess(Day 2) × Group(Peristim)	−1.26	0.27	656.00	−4.60	< .001	***
Region(Cnt) × Sess(Day 2) × Group(Sham)	1.52	0.35	656.00	4.29	< .001	***
Region(Par) × Sess(Day 2) × Group(Sham)	2.25	0.35	656.00	6.37	< .001	***
Region(Cnt) × Sess(Day 2) × Group(Peristim)	1.24	0.39	656.00	3.21	.001	**
Region(Par) × Sess(Day 2) × Group(Peristim)	2.35	0.39	656.00	6.07	< .001	***

Random Effect	*Variance*	*SD*	*Correlation*	
Participant(Intercept)	1.17	1.08		
Region(Cnt)	1.11	1.05	−.76	
Region(Par)	3.32	1.82	−.93	.87
Residual	0.70	0.84		

Reference levels: Fnt, Mid, Priming, Day 1. Number of obs.: 810, participants: 45.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

Table 3. Model-Predicted Values for Passive Word Learning N400 Amplitude (μV) by taVNS Group, Region, and Session

taVNS Group		*Day 1*	*Day 2*
Peristim	Frontal	0.00	−0.17
	Central	−1.02	−1.14
	Parietal	−0.49	−0.44
Priming	Frontal	−0.98	0.11
	Central	−0.76	−0.86
	Parietal	0.07	−0.97
Sham	Frontal	−0.22	−0.34
	Central	−1.04	−0.84
	Parietal	−0.49	−0.50

Table 4. Model-Predicted Values for Passive Word Learning N400 Amplitude (μV) by Region and Laterality

	Left	*Midline*	*Right*
Frontal	−0.31	−0.25	−0.24
Central	−0.76	−1.31	−0.77
Parietal	−0.13	−0.88	−0.42

$SE = 0.36$, $p = .93$; Region(Cnt): $b = -1.35$, $SE = 0.38$, $p < .001$; Region(Par): $b = -1.51$, $SE = 0.57$, $p = .011$) with no difference between central and parietal (with Priming, Cnt, Day 2 reference levels: Region(Par): $b = -0.15$, $SE = 0.38$, $p = .69$). Going from Day 1 to Day 2 for the priming group, the frontal negativity disappeared ($b = 1.10$, $SE = 0.20$, $p < .001$), the negativity at central sites did not change (with Priming, Cnt, Day 1 reference levels: Sess(Day 2): $b = -0.10$, $SE = 0.20$, $p = .62$), and parietal sites became more negative (with Priming, Par, Day 1 reference levels: Sess(Day 2): $b = -1.04$, $SE = 0.20$, $p < .001$).

For the peristim and sham groups, N400 amplitude was largest over central and parietal sites, with frontal sites less negative and no difference between central and parietal sites on both days (with Peristim, Cnt, Day 1 reference levels: Intercept: $b = -1.25$, $SE = 0.26$, $p < .001$; Region(Fnt): $b = 1.39$, $SE = 0.36$, $p < .001$; Region (Par): $b = 0.49$, $SE = 0.36$, $p = .18$; with Peristim, Cnt, Day 2 reference levels: Intercept: $b = -1.36$, $SE = 0.26$, $p < .001$; Region(Fnt): $b = 1.35$, $SE = 0.36$, $p < .001$; Region(Par): $b = 0.65$, $SE = 0.36$, $p = .076$; with Sham, Cnt, Day 1 reference levels: Intercept: $b = -1.44$, $SE = 0.21$, $p < .001$; Region(Fnt): $b = 1.20$, $SE = 0.30$, $p < .001$; Region(Par): $b = 0.51$, $SE = 0.30$, $p = .094$; with Sham, Cnt, Day 2 reference levels: Intercept: $b = -1.24$, $SE = 0.21$, $p < .001$; Region(Fnt): $b = 0.88$, $SE = 0.30$, $p = .004$; Region(Par): $b = 0.30$, $SE = 0.30$, $p = .33$). Going from Day 1 to Day 2, there were no changes at frontal, central, or parietal regions for peristim and sham groups (with Peristim, Fnt, Day 1 reference levels: Sess (Day 2): $b = -0.16$, $SE = 0.20$, $p = .39$; with Peristim, Cnt, Day 1 reference levels: Sess(Day 2): $b = -0.12$, $SE = 0.19$, $p = .54$; with Peristim, Par, Day 1 reference levels: Sess(Day 2): $b = 0.05$, $SE = 0.19$, $p = 0.80$; with Sham, Fnt, Day 1 reference levels: Sess(Day 2): $b = -0.12$, $SE = 0.15$, $p = .43$; with Sham, Cnt, Day 1 reference levels: Sess(Day 2): $b = 0.2$, $SE = 0.15$, $p = .19$; with Sham, Par, Day 1 reference levels: Sess(Day 2): $b = -0.01$, $SE = 0.15$, $p = .95$). Together, these effects reflect stable centro-parietal N400 topography for sham and peristim groups across days, whereas the priming group's N400 shifted from fronto-central on Day 1 to centro-parietal on Day 2. Model-predicted values are shown in Table 3.

Considering the remaining significant fixed effects, the Region × Lat interactions reflect a reduced amplitude negativity for lateral versus midline central and parietal sites, which together indicate a midline focus of the centro-parietal N400 across taVNS groups and sessions (Region(Cnt) × Lat(Left): $b = 0.61$, $SE = 0.18$, $p < .001$; Region(Par) × Lat(Left): $b = 0.81$, $SE = 0.18$, $p < .001$; Region(Cnt) × Lat(Right): $b = 0.53$, $SE = 0.18$, $p = .003$; Region(Par) × Lat(Right): $b = 0.45$, $SE = 0.18$, $p = .012$). Corresponding model predictions are shown in Table 4. The nonsignificant Group × Lat interactions indicate no difference in N400 laterality between priming and sham groups (Lat(Left) × Group(Sham): $b = -0.10$, $SE = 0.18$, $p = .58$; Lat(Right) × Group(Sham): $b = -0.08$, $SE = 0.18$, $p = .65$). However, the significant Lat(Left) × Group(Peristim) interaction reveals that the relative difference in negativity between midline and left sites observed for the priming and sham groups was attenuated for the peristim group ($b = -0.58$, $SE = 0.19$, $p = .003$). This reflects a more left-lateralized N400 distribution for the peristim group compared to priming and sham (with Sham, Midline reference levels: Lat(Left) × Group(Peristim): $b = -0.48$, $SE = 0.17$, $p = .005$; Lat(Right) × Group(Peristim): $b = -0.04$, $SE = 0.17$, $p = .83$). Corresponding model-predicted values are shown in Table 5.

Table 5. Model-Predicted Values for Passive Word Learning N400 Amplitude (μV) by taVNS Group and Laterality

	Left	*Midline*	*Right*
Peristim	−0.61	−0.66	−0.36
Priming	−0.28	−0.91	−0.50
Sham	−0.33	−0.86	−0.53

Table 6. Best-Fitting Model for Active Word Learning Accuracy

Fixed Effect	*Est.*	*SE*	*z Value*	*p Value*	
(Intercept)	1.20	0.15	7.96	< .001	***
OMSI	0.32	0.12	2.76	.006	**
PCID	0.26	0.11	2.41	.016	*
Sess(Day 2)	0.52	0.08	6.22	< .001	***
Cond(Tone)	−0.65	0.11	−5.89	< .001	***
Cond(Syllable)	−0.48	0.13	−3.74	< .001	***

Random Effect	*Variance*	*SD*
Participant(Intercept)	0.34	0.58
Item(Intercept)	0.12	0.34

Reference levels: Day 1, Tone + Syllable. Number of obs.: 3240, participants: 45, items: 18.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

Active Word Learning Task

Accuracy

Active word learning task accuracy and RT were analyzed to determine whether taVNS influenced behavior during word learning. Across groups, mean accuracy improved between training sessions: by 8% for the sham taVNS group (Day 1: *M* (*SD*) = 0.66 (0.15); Day 2: *M* (*SD*) = 0.74 (0.13)), 9% for priming (Day 1: *M* (*SD*) = 0.71 (0.12); Day 2: *M* (*SD*) = 0.80 (0.12)), and 11% for peristim (Day 1: *M* (*SD*) = 0.65 (0.17); Day 2: *M* (*SD*) = 0.76 (0.15)). The best-fitting model for these data (Table 6) reveals that the likelihood of selecting the correct response was higher on Day 2 than Day 1 across taVNS groups (predicted = .79 vs. .69: $b = 0.52$, $SE = 0.08$, $p < .001$) and for individuals with higher PCID and OMSI scores (PCID: $b = 0.26$, $SE = 0.11$, $p = .016$; OMSI: $b = 0.32$, $SE = 0.12$, $p = .006$). Compared to the likelihood of accuracy for trials where response items' translation equivalents differed in both syllable and tone (pred. = .81), accurate responses were less likely for trials where response items differed only in tone (pred. = .69; $b = -0.65$, $SE = 0.11$, $p < .001$) or syllable (pred. = .73; $b = -0.48$, $SE = 0.13$, $p < .001$) with no difference between these conditions (with Tone reference level: Cond(Syllable): $b = 0.17$, $SE = 0.13$, $p = .18$).

RT

Across groups, mean RT decreased from Day 1 to Day 2: by 67 msec for the sham taVNS group (Day 1: *M* (*SD*) = 1185 (352); Day 2: *M* (*SD*) = 1118 (342)), 153 msec for priming (Day 1: *M* (*SD*) = 1178 (387); Day 2: *M* (*SD*) = 1025 (324)), and 159 msec for peristim (Day 1: *M* (*SD*) = 1182 (339); Day 2: *M* (*SD*) = 1023 (335)). The best-fitting model for RT (Table 7) reveals that, across taVNS groups, accurate responses were given faster on Day 2 than Day 1 (pred. = 1126 msec vs. 1017 msec: $b = -0.10$, $SE = 0.03$, $p < .001$) and, compared to trials were response item translation equivalents differed only in tone (pred. = 1099 msec), accurate responses were given faster for trials with response items that differed in syllable (pred. = 1042 msec: $b = -0.05$, $SE = 0.02$, $p < .001$) or syllable

Table 7. Best-Fitting Model for Active Word Learning Log RT

Fixed Effect	*Estimate*	*SE*	*df*	*t Value*	*p Value*	
(Intercept)	7.06	0.03	61.17	232.43	< .001	***
Sess(Day 2)	−0.10	0.03	43.92	−3.67	< .001	***
Cond(Tone + Syllable)	−0.04	0.01	1782.35	−2.58	.010	**
Cond(Syllable)	−0.05	0.02	613.64	−3.35	< .001	***

Random Effect	*Variance*	*SD*	*Correlation*
Participant(Intercept)	0.03	0.18	
Sess(Day 2)	0.03	0.17	−.43
Item(Intercept)	0.00	0.04	
Residual	0.07	0.26	

Reference levels: Day 1, Tone. Number of obs.: 2315, participants: 45, items: 18.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

Table 8. Best-Fitting Model for Lexical Recognition Accuracy

Fixed Effect	*Est.*	*SE*	*z Value*	*p Value*	
(Intercept)	1.03	0.25	4.13	< .001	***
OMSI	0.40	0.11	3.61	< .001	***
Sess(Day 2)	1.03	0.13	7.76	< .001	***
Cond(Match)	0.34	0.22	1.55	.12	
Group(Priming)	−0.93	0.34	−2.74	.006	*
Group(Sham)	−0.92	0.29	−3.21	.001	**
Group(Priming) × Cond(Match)	0.80	0.28	2.87	.004	**
Group(Sham) × Cond(Match)	0.57	0.24	2.40	.017	*
Sess(Day 2) × Cond(Match)	−0.04	0.16	−0.23	.82	

Random Effect	*Variance*	*SD*	*Correlation*		
Participant(Intercept)	0.86	0.93			
Cond(Match)	0.45	0.67	−.68		
Sess(Day 2)	0.37	0.61	.24	−.02	
Cond(Match) × Sess(Day 2)	0.34	0.59	.38	−.37	−.64
Item(Intercept)	0.08	0.29			
Cond(Match)	0.17	0.41	−.28		
Sess(Day 2)	0.10	0.31	−.23	.21	
Cond(Match) × Sess(Day 2)	0.17	0.41	.49	−.18	−.66

Reference levels: Day 1, Peristim, Mismatch. Number of obs.: 19429, participants: 45, items: 18.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

and tone (pred. = 1060 msec: $b = -0.04$, $SE = 0.01$, $p = .010$) with no difference between these conditions (with Tone + Syllable reference level: Cond(Syllable): $b = -0.02$, $SE = 0.02$, $p = .28$).

Lexical Recognition Test

Accuracy

Accuracy and RT during the lexical recognition test were analyzed to determine whether taVNS affected the level of attainment for the Mandarin pseudowords at the end of each session. Mean accuracy improved on Day 2 for all groups: by 16% for the sham taVNS group (Day 1: *M* (*SD*) = 0.61 (0.14); Day 2: *M* (*SD*) = 0.77 (0.14)), 16% for priming (Day 1: *M* (*SD*) = 0.67 (0.16); Day 2: *M* (*SD*) = 0.83 (0.15)), and 10% for peristim (Day 1: *M* (*SD*) = 0.69 (0.16); Day 2: *M* (*SD*) = 0.79 (0.14)). The best-fitting model for these data (Table 8) reveals that the likelihood of being accurate was higher on Day 2 than Day 1 for all groups (pred. = .86 vs. .69: $b = 1.03$, $SE = 0.13$, $p < .001$) and for participants who reported higher musicianship ($b = 0.40$, $SE = 0.11$, $p < .001$). The peristim group was equally likely to give correct responses for match (pred. = .87) and mismatch trials (pred. = .82; $b = 0.34$, $SE = 0.22$, $p = .12$), but both priming and sham were less likely to be accurate on mismatch trials (pred. priming = .65, sham = .65) than match trials (pred. priming = .85, sham = .82; with Priming, Match reference levels: Cond(Mismatch): $b = -1.14$, $SE = 0.23$, $p < .001$; with Sham, Match reference levels: Cond(Mismatch): $b = -0.91$, $SE = 0.18$, $p < .001$). The likelihood of accuracy did not differ between taVNS groups for match trials (with Peristim, Match reference levels: Group(Priming): $b = -0.13$, $SE = 0.29$, $p = 0.66$; Group(Sham): $b = -0.35$, $SE = 0.25$, $p = 0.15$; with Sham, Match reference levels: Group(Priming): $b = 0.27$, $SE = 0.28$, $p = .33$). However, the priming and sham groups were less likely than the peristim group to give correct responses to mismatch trials (Group(Priming): $b = -0.93$, $SE = 0.34$, $p = .006$; Group(Sham): $b = -0.92$, $SE = 0.29$, $p = .001$), with no difference between these groups (with

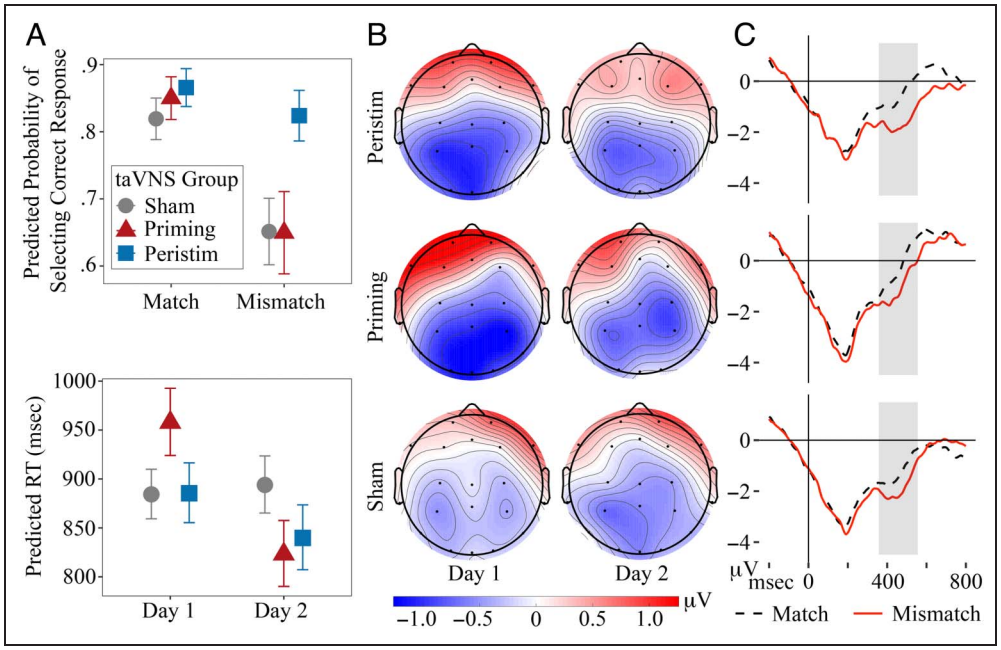

Figure 3. (A) Model-predicted values for accuracy likelihood and RT in the lexical recognition test, plotted with standard error bars. (B) Topographic plots of by-participant N400 effect (mismatch–match) mean amplitude during the adjusted analysis window, comparing taVNS group and session. (C) By-participant ERPs averaged over parietal sites (P3, Pz, P4) and sessions comparing taVNS group and trial condition for the lexical recognition test. The onset of the Mandarin pseudoword is at 0 msec; gray boxes indicate the mean onset and offset for the adjusted analysis window across trials.

Sham, Mismatch reference levels: Group(Priming): $b = -0.01$, $SE = 0.30$, $p = .98$). Model-predicted values are plotted in Figure 3.

RT

Across groups, mean RT decreased between Day 1 and Day 2: by 6 msec for the sham taVNS group (Day 1: M (SD) = 931 (313); Day 2: M (SD) = 925 (312)), 121 msec for priming (Day 1: M (SD) = 990 (323); Day 2: M (SD) = 869 (293)), and 66 msec for peristim (Day 1: M (SD) = 951 (315); Day 2: M (SD) = 885 (288)). The best-fitting model for these data (Table 9) reveals priming group RTs sped up from Day 1 to Day 2 ($b = -0.15$, $SE = 0.04$, $p < .001$) more so than RTs for peristim ($b = 0.10$, $SE = 0.05$, $p = .050$) and sham ($b = 0.16$, $SE = 0.05$, $p < .001$). RTs did not change between days for sham (pred. Day 1 = 884 msec, Day 2 = 894 msec; with Sham, Day 1 reference levels: Sess(Day 2): $b = 0.02$, $SE = 0.03$, $p = .62$) or peristim (pred. Day 1 = 885 msec, Day 2 = 840 msec), and RT changes did not differ between these groups (with Peristim, Day 1 reference levels: Sess(Day 2): $b = -0.05$, $SE = 0.04$, $p = .17$; Group(Sham) × Sess(Day 2): $b = 0.06$, $SE = 0.04$, $p = .15$). There were no RT differences between groups on Day 1 (Group(Sham): $b = -0.08$, $SE = 0.04$, $p = .073$; Group(Peristim): $b = -0.08$, $SE = 0.05$, $p = .11$; with Peristim, Day 1 reference levels: Group(Sham): $b = -0.001$, $SE = 0.04$, $p = .98$) or Day 2 (with Priming, Day 2 reference levels: Group(Peristim): $b = 0.02$, $SE = 0.05$, $p = .71$; Group (Sham): $b = 0.08$, $SE = 0.05$, $p = .10$; with Peristim, Day 2 reference levels: Group(Sham): $b = 0.06$, $SE = 0.05$, $p = .20$). RTs were faster for match (pred. = 835 msec) than mismatch trials (pred. = 934 msec; Cond(Match): $b = -0.11$, $SE = 0.02$, $p < .001$), but this did not differ between days ($b = -0.01$, $SE = 0.02$, $p = .71$) or groups (indicated by removal of the Group × Cond term during model selection because it did not improve model fit). Model-predicted values are plotted in Figure 3.

N400

Mean amplitude of by-participant N400 effects (mismatch–match difference) was analyzed for the lexical recognition test to determine whether there were differences between taVNS groups in the strength of semantic representations for the Mandarin pseudowords at the end of each session. Topographic plots of by-participant N400 effects averaged by taVNS condition and session are shown in Figure 3 with time-series plots for the same data averaged over parietal sites. These plots show broad centro-parietal negativity on both days for all taVNS groups, with comparatively larger amplitude for priming and peristim.

The best-fitting model for these data (Table 10) reveals a larger N400 effect amplitude at parietal sites for peristim and priming compared to sham (with Sham, Par reference levels: Group(Priming): $b = -0.40$, $SE = 0.20$, $p = .050$; Group(Peristim): $b = -0.43$, $SE = 0.20$, $p = .030$) but no difference between peristim and priming (with Peristim, Par reference levels: Group(Priming): $b = 0.03$, $SE = 0.22$, $p = .88$). N400 effect amplitude was not different between groups at frontal sites (Group(Priming): $b = 0.25$, $SE = 0.20$, $p = .22$; Group(Peristim): $b = 0.32$, $SE = 0.20$, $p = .11$; with Peristim, Ant reference levels: Group(Priming): $b = -0.07$, $SE = 0.22$, $p = .76$) or central sites (with Sham, Cnt reference levels: Group(Priming): $b = -0.24$, $SE = 0.20$, $p = .25$; Group(Peristim): $b = -0.11$, $SE = 0.20$, $p = .58$; with Peristim, Cnt reference levels: Group(Priming): $b = -0.13$, $SE = 0.22$, $p = .58$). Central and parietal sites were more negative than frontal sites for

Table 9. Best-Fitting Model for Lexical Recognition Log RT

Fixed Effect	*Estimate*	*SE*	*df*	*t Value*	*p Value*	
(Intercept)	6.92	0.04	47.72	192.08	< .001	***
Cond(Match)	−0.11	0.02	31.44	−5.36	< .001	***
Sess(Day 2)	−0.15	0.04	46.64	−4.02	< .001	***
Group(Peristim)	−0.08	0.05	40.30	−1.66	.11	
Group(Sham)	−0.08	0.04	40.83	−1.84	.073	
Group(Peristim) × Sess(Day 2)	0.10	0.05	41.28	2.02	.050	*
Group(Sham) × Sess(Day 2)	0.16	0.05	42.03	3.63	< .001	***
Cond(Match) × Sess(Day 2)	−0.01	0.02	28.10	−0.37	.71	

Random Effect	*Variance*	*SD*	*Correlation*		
Participant(Intercept)	0.01	0.12			
Sess(Day 2)	0.02	0.13	−.32		
Cond(Match)	0.01	0.08	−.08	−.25	
Sess(Day 2) × Cond(Match)	0.01	0.07	.18	−.17	−.48
Item(Intercept)	0.00	0.04			
Sess(Day 2)	0.00	0.02	−.08		
Cond(Match)	0.00	0.06	−.47	−.44	
Sess(Day 2) × Cond(Match)	0.00	0.05	−.28	−.12	.64
Residual	0.09	0.30			

Reference levels: Day 1, Priming, Mismatch. Number of obs.: 14008, participants: 45, items: 18.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

all groups (with Sham, Fnt reference levels: Region(Cnt): $b = -0.39$, $SE = 0.15$, $p = .012$; Region(Par): $b = -0.62$, $SE = 0.15$, $p < .001$; with Priming, Fnt reference levels: Region(Cnt): $b = -0.87$, $SE = 0.20$, $p < .001$; Region(Par): $b = -1.28$, $SE = 0.20$, $p < .001$; with Peristim, Fnt reference levels: Region(Cnt): $b = -0.82$, $SE = 0.19$, $p < .001$; Region(Par): $b = -1.38$, $SE = 0.19$, $p < .001$). The difference between frontal and central sites did not differ between groups (Region(Cnt) × Group(Peristim): $b = -0.43$, $SE = 0.25$, $p = .080$; Region(Cnt) × Group(Priming): $b = -0.48$, $SE = 0.25$, $p = .054$; with Priming, Fnt reference levels: Region(Cnt) × Group (Peristim): $b = 0.06$, $SE = 0.28$, $p = .84$), but peristim and priming showed larger differences between frontal and parietal sites compared to sham (Region(Par) × Group(Peristim): $b = -0.75$, $SE = 0.25$, $p = .002$; Region (Par) × Group(Priming): $b = -0.65$, $SE = 0.25$, $p = .010$) with no difference between these groups (with Priming, Fnt reference levels: Region(Par) × Group(Peristim): $b = -0.10$, $SE = 0.28$, $p = .71$). The Sess term was removed from the model because it did not improve model fit, indicating there was no difference in N400 effect topography between days for any group. Model-predicted values are given in Table 11.

Post Hoc Analyses Linking Behavioral Training Outcomes to ERP Indices of Learning

Post hoc analyses including lexical recognition accuracy and RT as predictors of N400 amplitude and topography during passive word learning were conducted to determine whether there is a direct link between electrophysiological measures of semantic encoding during training and behavioral training outcomes. The best-fitting model testing inclusion of RT as a predictor for the passive word learning N400 did not retain any fixed-effects involving RT, indicating RT did not improve the fit of the passive word learning N400 model without RT, reported above. In contrast, the best-fitting model testing inclusion of accuracy as a predictor for the passive word learning N400 retained three-way Region × Group × Acc and Region × Sess × Acc interactions and all lower-order terms

Table 10. Best-Fitting Model for Lexical Recognition N400 Effect (Mismatch–Match) Mean Amplitude

Fixed Effect	*Est.*	*SE*	*df*	*t Value*	*p Value*	
(Intercept)	0.08	0.12	165.44	0.64	.52	
Region(Cnt)	−0.39	0.15	759.00	−2.51	.012	*
Region(Par)	−0.62	0.15	759.00	−4.06	< .001	***
Group(Peristim)	0.32	0.20	165.44	1.62	.10	
Group(Priming)	0.25	0.20	165.44	1.23	.22	
Region(Cnt) × Group(Peristim)	−0.43	0.25	759.00	−1.75	.080	
Region(Par) × Group(Peristim)	−0.75	0.25	759.00	−3.08	.002	**
Region(Cnt) × Group(Priming)	−0.48	0.25	759.00	−1.93	.054	
Region(Par) × Group(Priming)	−0.65	0.25	759.00	−2.59	.010	**

Random Effect	*Variance*	*SD*
Participant(Intercept)	0.07	0.27
Residual	1.42	1.19

Reference levels: Fnt, Sham. Number of obs.: 810, participants: 45.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

(Table 12, predicted values plotted in Figure 4). Removal of the four-way Region × Group × Sess × Acc term during model fitting indicates that group differences in Region × Acc interactions were consistent between days, but, because of the retention of Sess in the model, estimates for the Region × Group × Acc interactions are given separately for each day below.

For the peristim group, on both days, higher lexical recognition accuracy predicted greater negativity over parietal sites and reduced negativity over frontal sites with the effect of Acc differing between frontal and central/parietal sites but not between central and parietal sites (with Peristim, Fnt, Day 1 reference levels: Acc: $b = 0.51$, $SE = 0.24$, $p = .034$; Region(Cnt) × Acc: $b = -0.73$, $SE = 0.27$, $p = .008$; Region(Par) × Acc: $b = -0.94$, $SE = 0.38$, $p = .015$; with Peristim, Fnt, Day 2 reference levels: Acc: $b = 0.67$, $SE = 0.24$, $p = .005$; Region(Cnt) × Acc: $b = -0.78$, $SE = 0.27$, $p = .005$; Region(Par) × Acc: $b = -1.29$, $SE = 0.38$, $p < .001$; with Peristim, Par, Day 1 reference levels: Acc: $b = -0.44$, $SE = 0.22$, $p = .046$; Region(Cnt) × Acc: $b = 0.21$, $SE = 0.27$, $p = .43$; with Peristim, Par, Day 2 reference levels: Acc: $b = -0.62$, $SE = 0.22$, $p = .005$; Region(Cnt) × Acc: $b = 0.51$, $SE = 0.27$, $p = .056$).

For the sham group, higher lexical recognition accuracy predicted greater negativity over parietal sites on both days and reduced negativity over frontal sites on Day 2 but not Day 1. On Day 1, the effect of Acc differed between frontal and parietal sites but not between central and frontal or parietal sites whereas, on Day 2, the effect of Acc differed between frontal and central/parietal sites but not between central and parietal sites (with Sham, Fnt, Day 1 reference levels: Acc: $b = 0.28$, $SE = 0.22$, $p = .20$; Region(Cnt) × Acc: $b = -0.56$, $SE = 0.25$, $p = .025$; Region(Par) × Acc: $b = -0.70$, $SE = 0.36$, $p = .055$; with Sham, Fnt, Day 2 reference levels: Acc: $b = 0.45$, $SE = 0.22$, $p = .043$; Region(Cnt) × Acc: $b = -0.61$, $SE = 0.25$, $p = .015$; Region(Par) × Acc: $b = -1.05$, $SE = 0.36$, $p = .004$; with Sham, Par, Day 1 reference levels: Acc: $b = -0.41$, $SE = 0.20$, $p = .039$; Region(Cnt) × Acc: $b = 0.14$, $SE = 0.24$, $p = .56$; with Sham, Par, Day 2 reference levels: Acc: $b = -0.60$, $SE = 0.20$, $p = .003$; Region(Cnt) × Acc: $b = 0.44$, $SE = 0.24$, $p = .070$).

For the priming group, higher lexical recognition accuracy predicted reduced negativity over parietal sites and greater negativity over frontal sites with the effect of Acc differing between frontal, central, and parietal sites on Day 1 but only differing between frontal and central/parietal but not between central and parietal sites on Day 2

Table 11. Model-Predicted Values for Lexical Recognition N400 Effect Amplitude (μV) by Region and taVNS Group

	Peristim	*Priming*	*Sham*
Frontal	0.40	0.33	0.08
Central	−0.42	−0.54	−0.31
Parietal	−0.98	−0.94	−0.54

Table 12. Best-Fitting Post Hoc Model of Passive Word Learning N400 Mean Amplitude Including Lexical Recognition Accuracy as a Fixed-Effect Predictor

Fixed Effect	*Est.*	*SE*	*df*	*t Value*	*p Value*	
(Intercept)	−1.06	0.41	42.65	−2.62	.012	*
Region(Cnt)	−0.16	0.42	45.13	−0.38	.71	
Region(Par)	0.63	0.68	37.93	0.93	.36	
Lat(Left)	0.16	0.17	640.12	0.94	.35	
Lat(Right)	0.09	0.17	640.12	0.53	.60	
Sess(Day 2)	1.10	0.19	640.12	5.74	< .001	***
Group(Sham)	0.82	0.51	40.98	1.62	.11	
Group(Peristim)	1.21	0.56	40.98	2.17	.036	*
Acc	−0.88	0.25	230.32	−3.55	< .001	***
Region(Cnt) × Lat(Left)	0.60	0.17	640.12	3.54	< .001	***
Region(Par) × Lat(Left)	0.81	0.17	640.12	4.74	< .001	***
Region(Cnt) × Lat(Right)	0.53	0.17	640.12	3.08	.002	**
Region(Par) × Lat(Right)	0.44	0.17	640.12	2.60	.009	**
Region(Cnt) × Sess(Day 2)	−1.20	0.27	640.12	−4.42	< .001	***
Region(Par) × Sess(Day 2)	−2.14	0.27	640.12	−7.91	< .001	***
Lat(Left) × Group(Sham)	−0.10	0.17	640.12	−0.58	.56	
Lat(Right) × Group(Sham)	−0.08	0.17	640.12	−0.47	.64	
Lat(Left) × Group(Peristim)	−0.58	0.19	640.12	−3.10	.002	**
Lat(Right) × Group(Peristim)	−0.12	0.19	640.12	−0.63	.53	
Region(Cnt) × Group(Sham)	−1.04	0.52	40.27	−2.01	.051	
Region(Par) × Group(Sham)	−1.32	0.85	36.32	−1.56	.13	
Region(Cnt) × Group(Peristim)	−1.23	0.57	40.27	−2.18	.035	*
Region(Par) × Group(Peristim)	−1.53	0.93	36.32	−1.66	.11	
Sess(Day 2) × Group(Sham)	−1.22	0.24	640.12	−5.04	< .001	***
Sess(Day 2) × Group(Peristim)	−1.26	0.27	640.12	−4.76	< .001	***
Group(Peristim) × Acc	1.39	0.33	215.17	4.15	< .001	***
Group(Sham) × Acc	1.16	0.32	180.79	3.61	< .001	***
Region(Par) × Acc	1.81	0.40	307.74	4.49	< .001	***
Region(Cnt) × Acc	0.98	0.28	135.03	3.46	< .001	***
Sess(Day 2) × Acc	0.17	0.11	656.45	1.55	.12	
Region(Cnt) × Sess(Day 2) × Group(Sham)	1.52	0.34	640.12	4.44	< .001	***
Region(Par) × Sess(Day 2) × Group(Sham)	2.25	0.34	640.12	6.58	< .001	***
Region(Cnt) × Sess(Day 2) × Group(Peristim)	1.24	0.37	640.12	3.31	< .001	***
Region(Par) × Sess(Day 2) × Group(Peristim)	2.35	0.37	640.12	6.27	< .001	***
Region(Cnt) × Group(Sham) × Acc	−1.54	0.36	106.92	−4.27	< .001	***
Region(Par) × Group(Sham) × Acc	−2.51	0.53	234.63	−4.72	< .001	***
Region(Cnt) × Group(Peristim) × Acc	−1.71	0.38	119.40	−4.52	< .001	***

Table 12. (*continued*)

Fixed Effect	*Est.*	*SE*	*df*	*t Value*	*p Value*	
Region(Par) × Group(Peristim) × Acc	−2.75	0.55	294.87	−5.04	< .001	***
Region(Cnt) × Sess(Day 2) × Acc	−0.05	0.15	666.79	−0.35	.73	
Region(Par) × Sess(Day 2) × Acc	−0.35	0.15	645.57	−2.32	.021	*

Random Effect	*Variance*	*SD*	*Correlation*		
Participant(Intercept)	1.65	1.28			
Region(Cnt)	1.56	1.25	−.83		
Region(Par)	4.93	2.22	−.95	.90	
Residual	0.66	0.81			

Reference levels: Fnt, Mid, Priming, Day 1. Number of obs.: 810, participants: 45.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

(with Priming, Fnt, Day 1 reference levels: Acc: $b = -0.88$, $SE = 0.25$, $p < .001$; Region(Cnt) × Acc: $b = 0.98$, $SE = 0.28$, $p < .001$; Region(Par) × Acc: $b = 1.81$, $SE = 0.40$, $p < .001$; with Priming, Fnt, Day 2 reference levels: Acc: $b = -0.71$, $SE = 0.25$, $p = .004$; Region(Cnt) × Acc: $b = 0.93$, $SE = 0.28$, $p = .001$; Region(Par) × Acc: $b = 1.46$, $SE = 0.40$, $p < .001$; with Priming, Par, Day 1 reference levels: Acc: $b = 0.93$, $SE = 0.23$, $p < .001$; Region(Cnt) × Acc: $b = -0.82$, $SE = 0.28$, $p = .003$; with Priming, Par, Day 2 reference levels: Acc: $b = 0.74$, $SE = 0.23$, $p = .001$; Region(Cnt) × Acc: $b = -0.52$, $SE = 0.28$, $p = .061$).

The Acc × Region interaction did not differ between peristim and sham nor did the simple effect of Acc at any region (with Peristim, Fnt reference levels: Group(Sham) × Acc: $b = -0.22$, $SE = 0.31$, $p = .48$; Region(Cnt) × Group(Sham) × Acc: $b = 0.17$, $SE = 0.35$, $p = .63$; Region(Par) × Group(Sham) × Acc: $b = 0.24$, $SE = 0.52$, $p = .64$; with Peristim, Par reference levels: Group(Sham) × Acc: $b = 0.02$, $SE = 0.28$, $p = .94$; Region(Cnt) × Group (Sham) × Acc: $b = -0.07$, $SE = 0.34$, $p = .84$; with Peristim, Cnt reference levels: Group(Sham) × Acc: $b = -0.05$, $SE = 0.25$, $p = .84$). In contrast, the simple effect of Acc differed at frontal and parietal sites but not central sites between priming and both peristim and sham (with Priming, Fnt reference levels: Group(Peristim) × Acc: $b = 1.39$, $SE = 0.33$, $p < .001$; Group(Sham) × Acc: $b = 1.16$, $SE = 0.32$, $p < .001$; with Priming, Par reference levels: Group(Peristim) × Acc: $b = -1.36$, $SE = 0.30$, $p < .001$; Group(Sham) × Acc: $b = -1.34$, $SE = 0.29$, $p < .001$; with Priming, Cnt reference levels: Group(Peristim) × Acc: $b = -0.33$, $SE = 0.27$, $p = .23$; Group(Sham) × Acc: $b = -0.38$,

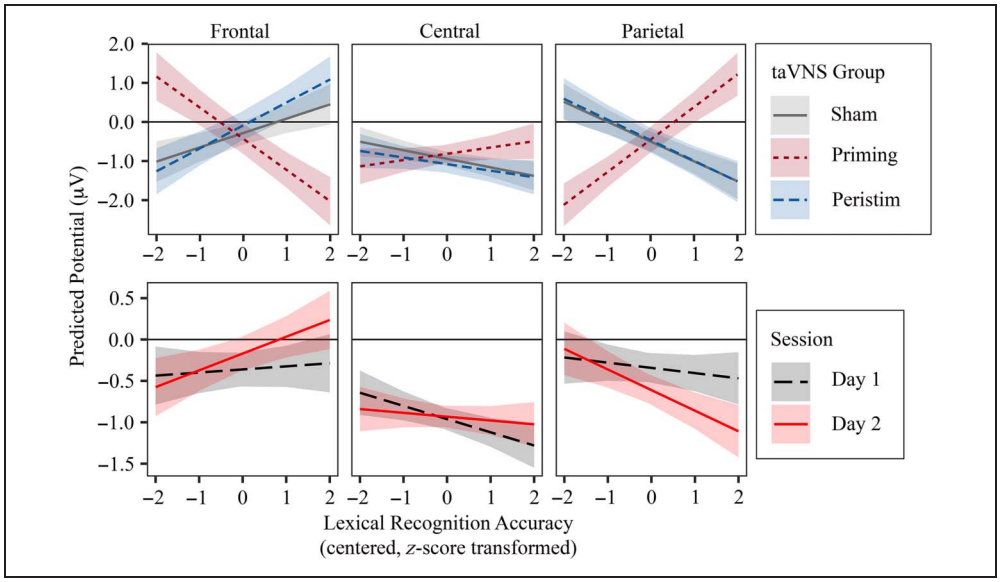

Figure 4. Model-predicted N400 mean amplitudes for the effect of taVNS group (top row) and session (bottom row) on the interaction of lexical recognition accuracy and electrode region during passive word learning, plotted with standard error shading.

$SE = 0.26, p = .14$) driving the significant Region × Group × Acc interactions (Region(Cnt) × Group(Peristim) × Acc: $b = -1.71$, $SE = 0.38$, $p < .001$; Region(Par) × Group(Peristim) × Acc: $b = -2.75$, $SE = 0.55$, $p < .001$; Region(Cnt) × Group(Sham) × Acc: $b = -1.54$, $SE = 0.36$, $p < .001$; Region(Par) × Group(Sham) × Acc: $b = -2.51$, $SE = 0.53$, $p < .001$).

Across taVNS groups, the effect of Acc did not differ between days at any region (with Fnt, Day 1 reference levels: Sess(Day 2) × Acc: $b = 0.17$, $SE = 0.11$, $p = .12$; with Cnt, Day 1 reference levels: Sess(Day 2) × Acc: $b = 0.11$, $SE = 0.11$, $p = .29$; with Par, Day 1 reference levels: Sess(Day 2) × Acc: $b = -0.19$, $SE = 0.11$, $p = .081$). However, the Region × Sess × Acc interaction reveals that between-days trends in the effect of Acc differed between parietal and frontal/central sites but not between frontal and central sites (Region(Par) × Sess(Day 2) × Acc: $b = -0.35$, $SE = 0.15$, $p = .021$; Region(Cnt) × Sess(Day 2) × Acc: $b = -0.05$, $SE = 0.15$, $p = .73$; with Par, Day 1 reference levels: Region(Cnt) × Sess(Day 2) × Acc: $b = 0.30$, $SE = 0.15$, $p = .046$). Over parietal sites, the trend was toward higher lexical recognition accuracy predicting comparatively greater negativity on Day 2 than Day 1 (more negative slope on Day 2 compared to Day 1 in Figure 4) but this trend was reversed over frontal and central sites.

DISCUSSION

The results of this study demonstrate that priming and peristim taVNS enhance behavioral outcomes of initial L2 lexical training for words partly distinguished by Mandarin tone contrasts and that these improvements are linked to stronger lexico-semantic encoding. All groups showed evidence of learning, reflected in higher likelihood of accuracy and improved processing speed reflected in faster RTs on Day 2 across tasks, but pairing taVNS with training led to greater behavioral gains in the lexical recognition test than training alone in line with findings reported for the larger data set (Pandža et al., 2020). Differences in N400 effect amplitude between taVNS groups in the lexical recognition test support the RQ1 hypothesis that pairing taVNS with lexical training results in more robust lexico-semantic representations. The N400 results for the passive word learning task do not clearly support the hypothesis that taVNS leads to faster semantic encoding during training, but they do provide some support for the RQ2 hypothesis that priming and peristim taVNS benefit lexico-semantic development in different ways. Behavioral and ERP results relevant to each RQ are considered in turn below.

taVNS Improves Recognition of Newly Learned L2 Lexical Items

The priming group made accurate responses in the lexical recognition test faster on Day 2 than Day 1, reflecting improved recognition speed or perhaps also participants' confidence in their judgments, whereas accurate response speed did not change for the other two taVNS groups. The peristim group was highly and equally likely to accurately identify matches and mismatches between English words and Mandarin pseudowords, reflecting improved recognition memory for word pairs, whereas the priming and sham groups were less likely to correctly identify mismatches than matches. Higher performance on match trials in the priming and sham groups, on par with the peristim group, likely reflects bias toward responding "match" that was observed for priming and sham (and to a lesser extent, peristim) rather than learning per se, because each Mandarin pseudoword appeared in the same number of match and mismatch trials.

Contrary to expectations, taVNS did not influence accuracy or RT for the active word learning task. One potential explanation is that taVNS may have had a comparatively weak effect on modulating the formation and strengthening of lexico-semantic representation during the passive compared to active word learning task. In this case, priming and peristim would perform similarly to sham at the start of active word learning (at least on Day 1), but improve more rapidly during the task. Within-task differences in learning trajectory may not be captured in the analysis of mean performance over the task. Another potential explanation is that the active word learning task (unlike the lexical recognition test) provided accuracy feedback each trial, which could inform following responses. This may have allowed participants to perform well without relying on memories formed in prior learning tasks, thereby attenuating any effects of taVNS on prior learning. This attenuation combined with a low trial number ($n = 36$) may have reduced the taVNS effect size below a detectable level in these data. However, it should be noted that identical accuracy and RT model structures were selected, with very similar estimates, when the same model fitting procedure was applied to the full data set ($n = 82$ participants).

Aside from the null taVNS effect in active word learning, participants were slower to pick the correct response when Mandarin translation equivalents for response items differed only in lexical tone versus when they had different syllables. Cue status (nonnative tone vs. native syllable phonemes) may have contributed to this effect, but differences in cue temporal features likely also played a role. In two thirds (66.8%) of the trials with different syllables, the difference occurred on the first phoneme. The first phoneme also carried tone information in two thirds of the stimuli (/ba/, /bi/), but tone direction continued to unfold over the second phoneme. Given that speakers of nontonal languages are less sensitive to tone direction (e.g., Chandrasekaran et al., 2010), participants likely needed to hear more than the first phoneme to differentiate items differing only in tone, while they more often than not could distinguish items with different syllables after hearing only the first phoneme. Accuracy

results also suggest that cue status played a comparatively minor role in performance in this task. The likelihood of responding accurately was higher when response item translation equivalents differed in tone and syllable than when they differed in tone or syllable alone, with no difference between these conditions. This pattern suggests the number of cues (two vs. one) distinguishing response items was more important than whether the cue was contrastive in participants' L1. If, instead, performance was driven by cue status, accuracy differences would be expected between trials where response item translation equivalents differed only in tone or only in syllable.

Lastly, this study replicated some previous findings in the literature regarding individual differences in lexical tone–word learning (e.g., Bowles et al., 2016; Dittinger et al., 2016). Higher likelihood of accuracy on the active word learning task was predicted by higher self-rated musicianship and ability to identify nonlinguistic pitch, whereas higher accuracy on the lexical recognition test was predicted by higher self-rated musicianship. Notably, these factors did not affect RT or ERP results. That these factors did not predict N400 amplitude or topography in the passive word learning task or N400 effect amplitude in the lexical recognition test is unexpected because very similar tasks in Dittinger et al. (2016) showed increased centro-parietal N400 amplitude in passive word learning and strong centro-parietal N400 effects in lexical recognition for professional musicians. One possible explanation for the difference is that Dittinger et al. (2016) compared professional musicians to nonmusicians, whereas there was more variation in musicianship and no professional musicians in this study.

taVNS Strengthens Lexico-Semantic Encoding Beyond Training Alone

A strong centro-parietal N400 effect (mismatch–match) was observed in the lexical recognition test on both days for priming and peristim. The N400 effect was not statistically different between priming and peristim, but was larger over parietal sites for both groups on both days compared to sham. Given N400 amplitude correlates negatively with a word's expectancy or predictability from prior context (Lau, Holcomb, & Kuperberg, 2013; Van Petten & Luka, 2012; Kutas & Federmeier, 2011; Lau et al., 2008; Kutas & Hillyard, 1980), a larger N400 effect at parietal sites indicates the English translation equivalents primed Mandarin pseudowords to a greater degree on both days for priming and peristim versus sham. This difference in N400 effect strength largely parallels that in Dittinger et al. (2016), where musicians showed a centro-parietal N400 effect (mismatch–match) in both blocks of a lexical recognition task very similar to that used in this study, but nonmusicians showed the opposite effect over fronto-central regions (larger N400 for match than mismatch). The sham group in this study showed a centro-parietal N400 effect (possibly indicating stimuli in this study were easier to learn than those in Dittinger et al., 2016), although the larger N400 effect at parietal sites for priming and peristim suggests pairing either method of taVNS delivery with lexical training provided an advantage.

The N400 results in the passive word learning task suggest taVNS did not reduce the number of exposures needed to encode meaning of the Mandarin pseudowords. While the priming group showed a fronto-central to centro-parietal N400 shift between days, perstim and sham showed centro-parietal N400s on both days. The shift observed for priming—consistent with previous findings of frontal N400s during initial stages of word learning and central-parietal N400s during later stages, when learners are presumed to have developed stronger semantic representations for learned words (Dittinger et al., 2016; Rodríguez-Fornells et al., 2009)—may suggest weaker semantic encoding of Mandarin pseudowords for the priming group on Day 1 of passive word learning compared to peristim and sham but comparable encoding strength by Day 2. There is some evidence that iVNS delivered continuously at higher intensity levels during learning and testing can negatively affect recognition memory for figures, although verbal memory was unaffected and RT was improved in the same study (Helmstaedter, Hoppe, & Elger, 2001). Thus, it is possible that priming taVNS could have impaired semantic encoding during passive word learning on Day 1. However, any negative effects of priming taVNS in this study would result from differences in total amount of current received rather than current amplitude because taVNS amplitude was calibrated to the same perceptual threshold for peristim and priming. It is also not clear why similar levels of priming taVNS would be detrimental to encoding 1 day but not the next. Thus, an alternative explanation for the priming group's fronto-central N400 on Day 1 of passive word learning that considers potential taVNS-related changes in cognitive functions thought to contribute to frontal N400 topography during word learning is given below.

The centro-parietal N400 on Day 1 for peristim and sham suggests these individuals developed semantic representations for target words with fewer word exposures than were needed for participants in Dittinger et al. (2016), where N400 amplitude was largest fronto-centrally during Block 1 of a very similar passive word learning task, only increasing at centro-parietal regions during Block 2 for musicians. This again suggests stimuli in this study were easier to learn, perhaps because of items differing along only one nonnative phonological contrast. Although we did not observe typical N400 topographic shifts between days that would indicate faster lexico-semantic encoding for taVNS, the N400 for the peristim group showed a left-lateralized distribution on both days whereas N400s for priming and sham were focused over midline sites. Similarly, Dittinger et al. (2016)

showed an increased left-lateralized N400 amplitude for learned words during Block 2 of passive word learning for musicians, but not nonmusicians. Other studies investigating individual differences in word learning have also shown left-lateralized N400 effects (e.g., Perfetti et al., 2005). Within a single session of learning rare words, Perfetti et al. (2005) found that, although more- and less-skilled readers showed similar right-lateralized N400 effects for related versus unrelated words during a semantic judgment task, only more-skilled readers also showed a left-lateralized N400 effect. The left-lateralized differences in this study, perhaps distinguishable by some measure of skill, may indicate a greater integration of new lexical items into an existing semantic network that was facilitated by peristim taVNS. However, it should be noted that post hoc analyses did not show a direct relationship between these laterality differences and behavioral performance during lexical recognition.

Recognition Accuracy (But Not RT) Predicts N400 Changes during Passive Word Learning

Post hoc modeling uncovered direct relationships between lexical recognition accuracy and N400 indices of lexico-semantic encoding during passive word learning. Across taVNS groups, for individuals with higher recognition accuracy compared to those with lower accuracy, parietal sites were numerically more negative whereas frontal and central sites were numerically less negative on Day 2 than Day 1. This reflects a parietal shift in the passive word learning N400 between days that was greater for individuals with higher recognition, which aligns with previous observations of N400s elicited during early word learning and between novel and known words (e.g., Dittinger et al., 2016; Rodríguez-Fornells et al., 2009). This finding suggests behavioral learning improvements in this study are tied to changes in strength of lexico-semantic representations of the Mandarin pseudowords, with individuals who showed better recognition of learned words also showing increases in strength of lexico-semantic representations during learning (via shifts in N400 topography) whereas those with lower recognition showing less change in N400 topography between days.

Of greater interest to this study, post hoc modeling also revealed differences between taVNS groups in the link between lexical recognition accuracy and passive word learning N400 topography. The same pattern was observed for peristim and sham, with higher recognition accuracy predicting increased negativity over parietal sites and reduced negativity over frontal sites during both days of passive word learning. This pattern aligns with the literature linking stronger lexico-semantic representations to centro-parietal N400s and weaker representations to fronto-central N400s and suggests that behavioral learning improvements for the peristim and sham groups are tied directly to the strength of the Mandarin pseudoword lexico-semantic representations. For peristim but not sham, there were other parallels between ERP and behavioral measures of lexical learning across days (higher accuracy for mismatch trials and strong centro-parietal N400 effects in lexical recognition, left-lateralized N400s in passive word learning) that may suggest other potential links between stronger lexico-semantic encoding and improved lexical recognition that result from pairing peristim taVNS with lexical training. If present, however, these links appear to be indirect as the interaction between passive word learning N400 laterality and lexical recognition accuracy was not retained in the post hoc model, indicating that recognition accuracy does not predict N400 laterality differences (or the relationship is nonlinear) in the present data.

Compared to peristim and sham, the priming group showed a qualitatively different relationship between lexical recognition accuracy and passive word learning N400 topography, with higher accuracy predicting greater negativity over frontal sites and reduced negativity over parietal sites. This pattern does not align with an interpretation of fronto-central N400 topography reflecting weaker lexico-semantic representations, but it is perhaps consistent with the idea that frontal negativity during initial lexical learning may result from increased engagement of certain cognitive functions. In this view, the fronto-central focus of the passive word learning N400s elicited for priming on Day 1 may more accurately reflect other aspects of performance, potentially more effortful processing related to cognitive control (Elgort et al., 2015; Mestres-Missé et al., 2007; Rodriguez-Fornells et al., 2006). Increased effort may be associated with executive functions necessary to promote controlled semantic retrieval (Mestres-Missé et al., 2007), including interference resolution, in which features of the nontarget language are suppressed to promote processing of the target language (Rodriguez-Fornells et al., 2006). This controlled processing becomes more automatic as learned words become more integrated into the lexicon (Bakker, Takashima, van Hell, Janzen, & McQueen, 2015). In our study, participants who received priming taVNS may have had greater capacity to sustain attentional effort that resulted in more controlled lexical access/retrieval processes on Day 1, eliciting a frontal negativity. As the Mandarin items became more integrated with more training, processing would be expected to become more automatic and the frontal distribution of the N400 would be expected to lessen, consistent with the observed results on Day 2.

For the priming group, the improvement in lexical recognition RT between days would also be consistent with the above interpretation, although post hoc modeling indicates any potential link between passive word learning N400 topography and recognition RT is indirect. More controlled processing could be expected to result in slower RTs on Day 1 whereas more automatic processing would result in faster RTs on Day 2. However, all terms

involving recognition RT were removed from the post hoc passive word learning N400 model, indicating that lexical recognition RT does not predict passive word learning N400 topography (or the relationship is nonlinear) in these data. Attributing the priming group's RT improvement to a shift from more controlled to more automatic processing (rather than purely reflecting semantic representation strength) also makes sense in light of the apparent inconsistency between the lexical recognition RT improvement from Day 1 to Day 2 and the strong centro-parietal N400 effect that was observed on both days in the same task. The strong N400 effect on Day 1 suggests priming taVNS strengthened semantic encoding for Mandarin pseudowords without parallel effects seen in explicit behavioral responses. This mirrors effects seen in other studies of L2 early word learning (e.g., McLaughlin et al., 2004) and aligns with other studies of lexico-semantic processing that find ERP but not behavioral differences (e.g., Barber, Otten, Kousta, & Vigliocco, 2013; Balass, Nelson, & Perfetti, 2010).

Priming and Peristim taVNS May Differentially Influence Initial L2 Lexical Learning

The behavioral learning advantages observed for priming and peristim over sham and the striking similarities between N400 effects elicited for peristim and priming in the lexical recognition test suggest both taVNS timings enhanced lexico-semantic encoding to a similar degree. However, the finding that priming taVNS improved RTs but not accuracy, and peristim taVNS improved accuracy but not RT, paired with distinct N400 patterns during passive word learning may suggest priming and peristim taVNS support L2 lexical learning through different underlying mechanisms. One way taVNS could potentially lead to stronger semantic encoding is by enhancing LTP during word learning tasks, thus impacting memory formation for the Mandarin pseudowords. Although this explanation has been suggested for memory improvements following VNS administered during memory encoding and consolidation, *after* stimulus presentation (Jacobs et al., 2015), priming and peristim taVNS could potentially affect LTP because increased NE levels in the basolateral amygdala have been observed to last more than an hour after administering 30 sec of VNS in animal models (Hassert, Miyashita, & Williams, 2004). Stimuli in our training tasks were presented within a few seconds of peristim taVNS and within 30 min of priming taVNS. Another possibility is that, by modulating LC-NE activity, taVNS helped participants to maintain an optimal state of focused attention during the word learning tasks, thus indirectly leading to better encoding of target items.

The data collected in this study do not permit direct examination of LC-NE activity or LTP. However, the results of the post hoc analyses contextualized with respect to the passive word learning task pupillometry results presented in Pandža et al. (2020) suggest that beneficial effects of priming taVNS, but not peristim taVNS, may be in part related to sustained attentional effort during learning. The potential links drawn below between the present results and maintenance of attentional effort are necessarily speculative as the multidimensional nature of pupillary changes that characterize differences in effort deployment in Pandža et al. (2020) does not easily permit linking the present results to a unidimensional indicator of sustained effort, but it is worth some discussion for its potential value in guiding future research. Although N400 amplitude has been linked to concurrently recorded task-evoked pupillary response amplitude previously (Kuipers & Thierry, 2011), the design of this study does not permit the same statistical comparisons.

During the passive word learning task, participants were instructed to memorize the meaning of the Mandarin pseudowords. Thus, stimulus-evoked changes in pupil size were interpreted in Pandža et al. (2020) to reflect the level of attention or effort that participants allocated to memorizing Mandarin–English word pairs. With this interpretation, the pupillometry results for passive word learning indicated the priming group maintained similar levels of sustained effort during trials on Day 1 and 2, whereas peristim and sham groups showed less sustained effort during trials on Day 2. Peristim and sham had similar between-days changes in pupillary responses compared to priming, showing an earlier deployment of cognitive effort that was less sustained on Day 2 compared to Day 1. This between-days reduction in sustained effort was significantly greater for peristim than sham and priming, and greater for sham compared to priming. This pattern of results suggests the lexical learning advantage for the priming group in this study is related to sustaining attentional effort during learning whereas the advantage for the peristim group is not. The post hoc analyses in this study lend further support to this conclusion by revealing a direct link between recognition accuracy and N400 amplitude during learning, with a qualitatively distinct pattern observed for priming where accuracy predicted a frontal negativity thought to reflect potentially effortful cognitive control processes. Explaining stronger lexico-semantic encoding and faster lexical processing for the priming group as a result of better sustained attentional effort because of taVNS is straightforward given the role of attention in learning: "…learning, particularly after a sensitive period, appears to be a gated system, through which attention (via acetylcholine) can facilitate or restrict plasticity" (White et al., 2013, p. 12).

The explanation for stronger lexico-semantic encoding and better learning for the peristim group seems unlikely to involve sustained attentional effort because these participants had the biggest drop off in effort between days yet they were the most likely to respond correctly in the lexical recognition test. These results would be expected, however, if peristim taVNS enhanced memory formation

directly via its effects on LTP. The N400s elicited for peristim and sham on Day 1 of passive word learning suggest individuals in these groups had developed some level of lexico-semantic representation for the Mandarin pseudowords during this task and the fact that there was no significant change between days in any ERP or behavioral measure for either group indicates that the semantic representations that peristim and sham developed early during training on Day 1 remained stable through Day 2. Because there is no indication from the passive word learning task pupillometry findings that the peristim group exerted more effort than the sham group during learning, the stronger lexico-semantic representations developed by the peristim group evidenced in the lexical recognition test results are presumably because of peristim taVNS increasing learning efficiency, which is consistent with the idea of taVNS benefiting memory encoding more directly via its effects on LTP.

Limitations

As an initial exploration into potential taVNS effects, taVNS was administered prior to (priming) and during (peristim) the word learning tasks and the lexical recognition test in this study. Thus, we cannot rule out the possibility that taVNS effects on task engagement or memory retrieval during the lexical recognition test contributed to better performance without necessarily influencing learning. Although lexical recognition test ERP results indicate participants who received taVNS developed stronger lexico-semantic representations, suggesting taVNS affected learning itself, future work should tease apart the effects of taVNS administered during training and testing phases to further support this conclusion. As pointed out by one reviewer, the inclusion of visual representations of the lexical tones in this study may have potentially diminished taVNS effects. Future work may gain further insight into the efficacy of taVNS by studying its effects in more ecologically valid language learning contexts. The design of this study also does not provide evidence on whether taVNS facilitates integration of lexico-semantic representations for newly learned L2 words into existing semantic networks or the longevity of taVNS-related learning improvements. Dittinger et al. (2016) found evidence from a semantic priming task that musical experience facilitated the integration of target item lexico-semantic representations into existing semantic networks; however, this study did not include a similar semantic priming task and therefore does not provide data necessary to support a similar claim for taVNS. Similarly, Dittinger et al. found that the subset of professional musicians who repeated the matching and semantic tasks 5 months after training showed higher accuracy on the match task compared to the subset of nonmusicians who completed the same posttest. (The advantage for musicians in the semantic task, however, disappeared at 5 months.) Although this study provides useful initial evidence of the impact of taVNS on language learning, future studies should assess semantic integration and the longevity of taVNS effects. Lastly, it is possible that the effects of taVNS on Mandarin pseudoword learning are partly because of taVNS enhancing low-level auditory processing of the novel phonological features distinguishing Mandarin pseudowords. Recent studies have shown that pairing direct VNS with pure tones and speech sounds can strengthen responses of primary auditory cortex neurons to these stimuli in humans and animal models (Engineer et al., 2011, 2015; De Ridder, Vanneste, Engineer, & Kilgard, 2014). Although these changes have been observed following several weeks of iVNS, they were not found within a single day of taVNS-paired training of Mandarin tones (Llanos et al., 2020), and it is unclear whether changes could be expected following just 2 days of training in this study.

Conclusion

Individuals lose sensitivity at an early age to phonological features that do not distinguish word meaning in their native language, which can negatively impact L2 lexical learning and use. In this study, we found that combining two timings of taVNS with lexical training markedly improved native English speakers' recognition ability for a relatively small set of L2 pseudowords distinguished in part by a particularly challenging nonnative phonological contrast: Mandarin lexical tone. Our novel ERP findings show that taVNS strengthened lexico-semantic encoding of the target items, which likely contributed to behavioral improvements although the links appear to be indirect. Both priming and peristim taVNS strengthened lexico-semantic representations to a similar degree, although they likely accomplish this via different underlying mechanisms. These findings contribute to the limited literature examining the effects of neuromodulatory interventions on L2 lexical learning and suggest that taVNS holds promise as a safe and effective technique for supporting adult L2 acquisition. Although these data are not able to elucidate the neurophysiological mechanisms driving taVNS-enhanced learning, interpreting these novel findings in the context of pupillometry results from the same study (Pandža et al., 2020) suggests that stronger semantic encoding is in part facilitated by modulation of attentional effort and thus the positive effects of taVNS likely have more general applicability beyond the specific case of L2 lexical learning.

Acknowledgments

We thank Eric Pelzl for his assistance with stimuli selection and design; Matthew Turner and Sara McConnell for their assistance with data collection; Meredith Hughes, Jason Struck, and Alison Tseng for their assistance with blinding the data and assigning groups; Jarrett Lee for assistance programming taVNS; and Henk Haarmann and Greg Colflesh for contributions to earlier portions of the overall project. This material is based upon work

supported by the Naval Information Warfare Center and Defense Advanced Research Projects Agency under Cooperative Agreement No. N66001–17-2-4009. The identification of specific products or scientific instrumentation is considered an integral part of the scientific endeavor and does not constitute endorsement or implied endorsement on the part of the author, DoD, or any component agency. The views expressed in this article are those of the author and do not reflect the official policy of the Department of Army/Navy/Air Force, Department of Defense, or U.S. Government.

Reprint requests should be sent to Ian Phillips, Audiology and Speech Pathology Center, Walter Reed National Military Medical Center, Room 5400, 4954 North Palmer Road, America Building, Bethesda, MD 20889, United States, or via e-mail: iphillip@umd.edu.

Author Contributions

Ian Phillips: Data curation; Formal analysis; Investigation; Methodology; Project administration; Visualization; Writing—Original draft; Writing—Review & editing. Regina C. Calloway: Data curation; Formal analysis; Writing—Original draft; Writing—Review & editing. Valerie P. Karuzis: Data curation; Formal analysis; Investigation; Methodology; Project administration; Software; Visualization; Writing—Original draft; Writing—Review & editing. Nick B. Pandža: Data curation; Formal analysis; Investigation; Methodology; Writing—Review & editing. Polly O'Rourke: Conceptualization; Funding acquisition; Methodology; Project administration; Supervision; Writing—Review & editing. Stefanie E. Kuchinsky: Conceptualization; Funding acquisition; Methodology; Project administration; Supervision; Writing—Review & editing.

Funding Information

Polly O'Rourke and Stephanie E. Kuchinsky, Defense Advanced Research Projects Agency (https://dx.doi.org/10.13039/100000185), grant number: N66001-17-2-4009.

Diversity in Citation Practices

A retrospective analysis of the citations in every article published in this journal from 2010 to 2020 has revealed a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .408, W(oman)/M = .335, M/W = .108, and W/W = .149, the comparable proportions for the articles that these authorship teams cited were M/M = .579, W/M = .243, M/W = .102, and W/W = .076 (Fulvio et al., *JoCN*, 33:1, pp. 3–7). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance. The authors of this article report its proportions of citations by gender category to be as follows: M/M = .462; W/M = .262; M/W = .108; W/W = .169.

REFERENCES

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning*, *59*, 249–306. https://doi.org/10.1111/j.1467-9922.2009.00507.x

Antoniou, M., & Wong, P. C. (2016). Varying irrelevant phonetic features hinders learning of the feature being trained. *Journal of the Acoustical Society of America*, *139*, 271–278. https://doi.org/10.1121/1.4939736, PubMed: 26827023

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*, 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709, PubMed: 16022602

Bakker, I., Takashima, A., van Hell, J. G., Janzen, G., & McQueen, J. M. (2015). Tracking lexical consolidation with ERPs: Lexical and semantic-priming effects on N400 and LPC responses to newly-learned words. *Neuropsychologia*, *79*, 33–41. https://doi.org/10.1016/j.neuropsychologia.2015.10.020, PubMed: 26476370

Balass, M., Nelson, J. R., & Perfetti, C. A. (2010). Word learning: An ERP investigation of word experience effects on recognition and word processing. *Contemporary Educational Psychology*, *35*, 126–140. https://doi.org/10.1016/j.cedpsych.2010.04.001, PubMed: 22399833

Barber, H. A., Otten, L. J., Kousta, S. T., & Vigliocco, G. (2013). Concreteness in word processing: ERP and behavioral effects in a lexical decision task. *Brain and Language*, *125*, 47–53. https://doi.org/10.1016/j.bandl.2013.01.005, PubMed: 23454073

Bent, T., Bradlow, A. R., & Wright, B. A. (2006). The influence of linguistic experience on the cognitive processing of pitch in speech and nonspeech sounds. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 97–103. https://doi.org/10.1037/0096-1523.32.1.97, PubMed: 16478329

Berridge, C. W., & Waterhouse, B. D. (2003). The locus coeruleus–noradrenergic system: Modulation of behavioral state and state-dependent cognitive processes. *Brain Research Reviews*, *42*, 33–84. https://doi.org/10.1016/S0165-0173(03)00143-7, PubMed: 12668290

Borovsky, A., Elman, J. L., & Kutas, M. (2012). Once is enough: N400 indexes semantic integration of novel word meanings from a single exposure in context. *Language Learning and Development*, *8*, 278–302. https://doi.org/10.1080/15475441.2011.614893, PubMed: 23125559

Bowles, A. R., Chang, C. B., & Karuzis, V. P. (2016). Pitch ability as an aptitude for tone learning. *Language Learning*, *66*, 774–808. https://doi.org/10.1111/lang.12159

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–990. https://doi.org/10.3758/BRM.41.4.977, PubMed: 19897807

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904–911. https://doi.org/10.3758/s13428-013-0403-5, PubMed: 24142837

Butt, M. F., Albusoda, A., Farmer, A. D., & Aziz, Q. (2020). The anatomical basis for transcutaneous auricular vagus nerve stimulation. *Journal of Anatomy*, *236*, 588–611. https://doi.org/10.1111/joa.13122, PubMed: 31742681

Chandrasekaran, B., Sampath, P. D., & Wong, P. C. M. (2010). Individual variability in cue-weighting and lexical tone learning. *Journal of the Acoustical Society of America*, *128*, 456–465. https://doi.org/10.1121/1.3445785, PubMed: 20649239

Clark, K. B., Naritoku, D. K., Smith, D. C., Browning, R. A., & Jensen, R. A. (1999). Enhanced recognition memory following vagus nerve stimulation in human subjects. *Nature Neuroscience*, *2*, 94–98. https://doi.org/10.1038/4600, PubMed: 10195186

Cooper, A., & Wang, Y. (2012). The influence of linguistic and musical experience on Cantonese word learning. *Journal of the Acoustical Society of America*, *131*, 4756–4769. https://doi.org/10.1121/1.4714355, PubMed: 22712948

Cooper, A., & Wang, Y. (2013). Effects of tone training on Cantonese tone–word learning. *Journal of the Acoustical Society of America*, *134*, EL133–EL139. https://doi.org/10.1121/1.4812435, PubMed: 23927215

De Ridder, D., Vanneste, S., Engineer, N. D., & Kilgard, M. P. (2014). Safety and efficacy of vagus nerve stimulation paired with tones for the treatment of tinnitus: A case series. *Neuromodulation: Technology at the Neural Interface*, *17*, 170–179. https://doi.org/10.1111/ner.12127, PubMed: 24255953

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*, 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009, PubMed: 15102499

Desbeaumes Jodoin, V., Lespérance, P., Nguyen, D. K., Fournier-Gosselin, M.-P., & Richer, F. (2015). Effects of vagus nerve stimulation on pupillary function. *International Journal of Psychophysiology*, *98*, 455–459. https://doi.org/10.1016/j.ijpsycho.2015.10.001, PubMed: 26437126

Díaz, B., Mitterer, H., Broersma, M., & Sebastián-Gallés, N. (2012). Individual differences in late bilinguals' L2 phonological processes: From acoustic-phonetic analysis to lexical access. *Learning and Individual Differences*, *22*, 680–689. https://doi.org/10.1016/j.lindif.2012.05.005

Dittinger, E., Barbaroux, M., D'Imperio, M., Jäncke, L., Elmer, S., & Besson, M. (2016). Professional music training and novel word learning: From faster semantic encoding to longer-lasting word representations. *Journal of Cognitive Neuroscience*, *28*, 1584–1602. https://doi.org/10.1162/jocn_a_00997, PubMed: 27315272

Elgort, I., Perfetti, C. A., Rickles, B., & Stafura, J. Z. (2015). Contextual learning of L2 word meanings: Second language proficiency modulates behavioural and event-related brain potential (ERP) indicators of learning. *Language, Cognition and Neuroscience*, *30*, 506–528. https://doi.org/10.1080/23273798.2014.942673, PubMed: 25984550

Engineer, C. T., Engineer, N. D., Riley, J. R., Seale, J. D., & Kilgard, M. P. (2015). Pairing speech sounds with vagus nerve stimulation drives stimulus-specific cortical plasticity. *Brain Stimulation*, *8*, 637–644. https://doi.org/10.1016/j.brs.2015.01.408, PubMed: 25732785

Engineer, N. D., Riley, J. R., Seale, J. D., Vrana, W. A., Shetake, J. A., Sudanagunta, S. P., et al. (2011). Reversing pathological neural activity using targeted plasticity. *Nature*, *470*, 101–104. https://doi.org/10.1038/nature09656, PubMed: 21228773

Fox, J., & Weisberg, S. (2018). Visualizing fit and lack of fit in complex regression models with predictor effect plots and partial residuals. *Journal of Statistical Software*, *87*, 1–27. https://doi.org/10.18637/jss.v087.i09

Frangos, E., Ellrich, J., & Komisaruk, B. R. (2015). Non-invasive access to the vagus nerve central projections via electrical stimulation of the external ear: fMRI evidence in humans. *Brain Stimulation: Basic, Translational, and Clinical Research in Neuromodulation*, *8*, 624–636. https://doi.org/10.1016/j.brs.2014.11.018, PubMed: 25573069

Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research*, *29*, 311–343. https://doi.org/10.1177/0267658312461497

Hao, Y.-C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, *40*, 269–279. https://doi.org/10.1016/j.wocn.2011.11.001

Hassert, D. L., Miyashita, T., & Williams, C. L. (2004). The effects of peripheral vagal nerve stimulation at a memory-modulating intensity on norepinephrine output in the basolateral amygdala. *Behavioral Neuroscience*, *118*, 79–88. https://doi.org/10.1037/0735-7044.118.1.79, PubMed: 14979784

Hellman, A. B. (2011). Vocabulary size and depth of word knowledge in adult-onset second language acquisition. *International Journal of Applied Linguistics*, *21*, 162–182. https://doi.org/10.1111/j.1473-4192.2010.00265.x

Helmstaedter, C., Hoppe, C., & Elger, C. E. (2001). Memory alterations during acute high-intensity vagus nerve stimulation. *Epilepsy Research*, *47*, 37–42. https://doi.org/10.1016/S0920-1211(01)00291-1, PubMed: 11673019

Hulsey, D. R., Hays, S. A., Khodaparast, N., Ruiz, A., Das, P., Rennaker, R. L., et al. (2016). Reorganization of motor cortex by vagus nerve stimulation requires cholinergic innervation. *Brain Stimulation*, *9*, 174–181. https://doi.org/10.1016/j.brs.2015.12.007, PubMed: 26822960

Hulsey, D. R., Shedd, C. M., Sarker, S. F., Kilgard, M. P., & Hays, S. A. (2019). Norepinephrine and serotonin are required for vagus nerve stimulation directed cortical plasticity. *Experimental Neurology*, *320*, 112975. https://doi.org/10.1016/j.expneurol.2019.112975, PubMed: 31181199

Ingvalson, E. M., Barr, A. M., & Wong, P. C. (2013). Poorer phonetic perceivers show greater benefit in phonetic-phonological speech learning. *Journal of Speech, Language, and Hearing Research*, *56*, 1045–1050. https://doi.org/10.1044/1092-4388(2012/12-0024), PubMed: 23275405

Jacobs, H. I., Riphagen, J. M., Razat, C. M., Wiese, S., & Sack, A. T. (2015). Transcutaneous vagus nerve stimulation boosts associative memory in older individuals. *Neurobiology of Aging*, *36*, 1860–1867. https://doi.org/10.1016/j.neurobiolaging.2015.02.023, PubMed: 25805212

Klinkenberg, I., Sambeth, A., & Blokland, A. (2011). Acetylcholine and attention. *Behavioural Brain Research*, *221*, 430–442. https://doi.org/10.1016/j.bbr.2010.11.033, PubMed: 21108972

Kraus, T., Kiess, O., Hösl, K., Terekhin, P., Kornhuber, J., & Forster, C. (2013). CNS BOLD fMRI effects of sham-controlled transcutaneous electrical nerve stimulation in the left outer auditory canal—A pilot study. *Brain Stimulation*, *6*, 798–804. https://doi.org/10.1016/j.brs.2013.01.011, PubMed: 23453934

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*, 831–843. https://doi.org/10.1038/nrn1533, PubMed: 15496861

Kuhl, P. K. (2010). Brain mechanisms in early language acquisition. *Neuron*, *67*, 713–727. https://doi.org/10.1016/j.neuron.2010.08.038, PubMed: 20826304

Kuipers, J. R., & Thierry, G. (2011). N400 amplitude reduction correlates with an increase in pupil size. *Frontiers in Human Neuroscience*, *5*, 61. https://doi.org/10.3389/fnhum.2011.00061, PubMed: 21747766

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of*

Psychology, *62*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123, PubMed: 20809790
Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*, 203–205. https://doi.org/10.1126/science.7350657, PubMed: 7350657
Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single word contexts. *Journal of Cognitive Neuroscience*, *25*, 484–502. https://doi.org/10.1162/jocn_a_00328, PubMed: 23163410
Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De)constructing the N400. *Nature Reviews Neuroscience*, *9*, 920–933. https://doi.org/10.1038/nrn2532, PubMed: 19020511
Ling, W., & Grüter, T. (2020). From sounds to words: The relation between phonological and lexical processing of tone in L2 Mandarin. *Second Language Research*. https://doi.org/10.1177/0267658320941546
Llanos, F., McHaney, J. R., Schuerman, W. L., Yi, H. G., Leonard, M. K., & Chandrasekaran, B. (2020). Non-invasive peripheral nerve stimulation selectively enhances speech category learning in adults. *NPJ Science of Learning*, *5*, 1–11. https://doi.org/10.1038/s41539-020-0070-0, PubMed: 32802406
Long, M. H. (1990). Maturational constraints on language development. *Studies in Second Language Acquisition*, *12*, 251–285. https://doi.org/10.1017/S0272263100009165
Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*, 213. https://doi.org/10.3389/fnhum.2014.00213, PubMed: 24782741
Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd ed.). Cambridge, MA: MIT Press.
Luu, P., & Ferree, T. (2005). *Determination of the HydroCel Geodesic Sensor Nets' average electrode positions and their 10–10 International equivalents* (p. 11) [Technical Note]. Electrical Geodesics, Inc. https://www.egi.com/images/HydroCelGSN_10-10.pdf.
Mattock, K., Molnar, M., Polka, L., & Burnham, D. (2008). The developmental course of lexical tone perception in the first year of life. *Cognition*, *106*, 1367–1381. https://doi.org/10.1016/j.cognition.2007.07.002, PubMed: 17707789
McLaughlin, J., Osterhout, L., & Kim, A. (2004). Neural correlates of second-language word learning: Minimal instruction produces rapid change. *Nature Neuroscience*, *7*, 703–704. https://doi.org/10.1038/nn1264, PubMed: 15195094
Mestres-Missé, A., Rodriguez-Fornells, A., & Münte, T. F. (2007). Watching the brain during meaning acquisition. *Cerebral Cortex*, *17*, 1858–1866. https://doi.org/10.1093/cercor/bhl094, PubMed: 17056648
Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., et al. (2020). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, *375*, 20180522. https://doi.org/10.1098/rstb.2018.0522, PubMed: 31840593
Ollen, J. E. (2006). *A criterion-related validity test of selected indicators of musical sophistication using expert ratings* (Doctoral dissertation). Retrieved from https://www.ohiolink.edu/etd/view.cgi?osu1161705351.
Pandža, N. B., Phillips, I., Karuzis, V. P., O'Rourke, P., & Kuchinsky, S. E. (2020). Neurostimulation and pupillometry: New directions for learning and research in applied linguistics. *Annual Review of Applied Linguistics*, *40*, 56–77. https://doi.org/10.1017/S0267190520000069
Pelzl, E. (2019). What makes second language perception of Mandarin tones hard? Chinese as a second language. *Journal of the Chinese Language Teachers Association*, *54*, 51–78. https://doi.org/10.1075/csl.18009.pel
Pelzl, E., Lau, E. F., Guo, T., & DeKeyser, R. (2021). Even in the best-case scenario L2 learners have persistent difficulty perceiving and utilizing tones in Mandarin: Findings from behavioral and event-related potentials experiments. *Studies in Second Language Acquisition*, *43*, 268–296. https://doi.org/10.1017/S027226312000039X
Perfetti, C. A., Wlotko, E. W., & Hart, L. A. (2005). Word learning and individual differences in word learning reflected in event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1281–1292. https://doi.org/10.1037/0278-7393.31.6.1281, PubMed: 16393047
Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *Journal of the Acoustical Society of America*, *130*, 461–472. https://doi.org/10.1121/1.3593366, PubMed: 21786912
Poltrock, S., Chen, H., Kwok, C., Cheung, H., & Nazzi, T. (2018). Adult learning of novel words in a non-native language: Consonants, vowels, and tones. *Frontiers in Psychology*, *9*, 1211. https://doi.org/10.3389/fpsyg.2018.01211, PubMed: 30087631
Psychology Software Tools, Inc. (2012). *E-Prime* (Version 2.0). [Computer software]. https://www.pstnet.com
Pu, H., Holcomb, P. J., & Midgley, K. J. (2016). Neural changes underlying early stages of L2 vocabulary acquisition. *Journal of Neurolinguistics*, *40*, 55–65. https://doi.org/10.1016/j.jneuroling.2016.05.002, PubMed: 28983152
R Core Team. (2020). *R: A language and environment for statistical computing. R Foundation for Statistical Computing* (Version 3.6.3). R Foundation for Statistical Computing. https://www.R-project.org/
Rodríguez-Fornells, A., Cunillera, T., Mestres-Missé, A., & de Diego-Balaguer, R. (2009). Neurophysiological mechanisms involved in language learning in adults. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *364*. https://doi.org/10.1098/rstb.2009.0130, PubMed: 19933142
Rodríguez-Fornells, A., De Diego Balaguer, R., & Münte, T. F. (2006). Executive control in bilingual language processing. *Language Learning*, *56*, 133–190. https://doi.org/10.1111/j.1467-9922.2006.00359.x
Sarter, M., Gehring, W. J., & Kozak, R. (2006). More attention must be paid: The neurobiology of attentional effort. *Brain Research Reviews*, *51*, 145–160. https://doi.org/10.1016/j.brainresrev.2005.11.002, PubMed: 16530842
Sebastián-Gallés, N., Echeverría, S., & Bosch, L. (2005). The influence of initial exposure on lexical representation: Comparing early and simultaneous bilinguals. *Journal of Memory and Language*, *52*, 240–255. https://doi.org/10.1016/j.jml.2004.11.001
Shen, X. S., & Lin, M. (1991). A perceptual study of Mandarin tones 2 and 3. *Language & Speech*, *34*, 145–156. https://doi.org/10.1177/002383099103400202
Sun, L., Peräkylä, J., Holm, K., Haapasalo, J., Lehtimäki, K., Ogawa, K. H., et al. (2017). Vagus nerve stimulation improves working memory performance. *Journal of Clinical and Experimental Neuropsychology*, *39*, 954–964. https://doi.org/10.1080/13803395.2017.1285869, PubMed: 28492363
Thakkar, V. J., Engelhart, A. S., Khodaparast, N., Abadzi, H., & Centanni, T. M. (2020). Transcutaneous auricular vagus nerve stimulation enhances learning of novel letter–sound relationships in adults. *Brain Stimulation*, *13*, 1813–1820. https://doi.org/10.1016/j.brs.2020.10.012, PubMed: 33127581

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*, 176–190. https://doi.org/10.1016/j.ijpsycho.2011.09.015, PubMed: 22019481

Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L. C., & Castellanos, F. X. (2015). Inscapes: A movie paradigm to improve compliance in functional magnetic resonance imaging. *Neuroimage*, *122*, 222–232. https://doi.org/10.1016/j.neuroimage.2015.07.069, PubMed: 26241683

Voeten, C. C. (2020). *buildmer: Stepwise elimination and term reordering for mixed-effects regression* (Version 1.5). https://CRAN.R-project.org/package=buildmer

Vonck, K., Raedt, R., Naulaerts, J., De Vogelaere, F., Thiery, E., Van Roost, D., et al. (2014). Vagus nerve stimulation…25 years later! What do we know about the effects on cognition? *Neuroscience & Biobehavioral Reviews*, *45*, 63–71. https://doi.org/10.1016/j.neubiorev.2014.05.005, PubMed: 24858008

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *Journal of the Acoustical Society of America*, *106*, 3649–3658. https://doi.org/10.1121/1.428217, PubMed: 10615703

Werker, J. F., & Hensch, T. K. (2015). Critical periods in speech perception: New directions. *Annual Review of Psychology*, *66*, 173–196. https://doi.org/10.1146/annurev-psych-010814-015104, PubMed: 25251488

White, E. J., Hutka, S. A., Williams, L. J., & Moreno, S. (2013). Learning, neural plasticity and sensitive periods: Implications for language acquisition, music training and transfer across the lifespan. *Frontiers in Systems Neuroscience*, *7*, 90. https://doi.org/10.3389/fnsys.2013.00090, PubMed: 24312022

Wong, P. C., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied PsychoLinguistics*, *28*, 565–585. https://doi.org/10.1017/S0142716407070312

Yakunina, N., Kim, S. S., & Nam, E.-C. (2017). Optimization of transcutaneous vagus nerve stimulation using functional MRI. *Neuromodulation: Technology at the Neural Interface*, *20*, 290–300. https://doi.org/10.1111/ner.12541, PubMed: 27898202

Yip, M. (2002). *Tone*. Cambridge University Press. https://doi.org/10.1017/CBO9781139164559

Zhang, J. D., & Schubert, E. (2019). A single item measure for identifying musician and nonmusician categories based on measures of musical sophistication. *Music Perception*, *36*, 457–467. https://doi.org/10.1525/mp.2019.36.5.457