



Structural variation in the temporal lobe predicts learning and retention of non-native speech sounds

Pamela Fuhrmeister & Emily B. Myers


To cite this article: Pamela Fuhrmeister & Emily B. Myers (2021): Structural variation in the temporal lobe predicts learning and retention of non-native speech sounds, *Language, Cognition and Neuroscience*, DOI: [10.1080/23273798.2021.1944658](https://doi.org/10.1080/23273798.2021.1944658)

To link to this article: <https://doi.org/10.1080/23273798.2021.1944658>

 [View supplementary material](#) 

 Published online: 22 Jun 2021.

 [Submit your article to this journal](#) 

 [View related articles](#) 

 [View Crossmark data](#) 



Structural variation in the temporal lobe predicts learning and retention of non-native speech sounds

Pamela Fuhrmeister and Emily B. Myers

Department of Speech, Language, and Hearing Sciences, University of Connecticut, Storrs, CT, USA

ABSTRACT

Studies of non-native speech sound learning report considerable individual variability in learning new sounds and retaining them in memory. The current study tested whether individual variation in brain structure (measured using MRI) accounts for differences in learning or retention of non-native speech sounds. Fifty-seven participants were tested on identification and discrimination of difficult non-native speech sounds in the evening before training, after training, and tested again the next morning. Surface area and volume of the left superior temporal gyrus positively predicted discrimination learning, whereas surface area of the left transverse temporal gyrus negatively predicted overnight improvement of identification. Hippocampal volume as well as gyrification of bilateral transverse temporal gyri positively predicted overnight improvement of discrimination. Findings suggest that individual differences in non-native speech sound learning can be traced to differences in brain structure supporting perception, while differences in retention are linked to the structure of hippocampal regions important for memory consolidation.

ARTICLE HISTORY

Received 8 January 2021
Accepted 4 June 2021

KEYWORDS

Non-native speech sound learning; memory consolidation; brain structure; individual differences; structural MRI

Learning a second language in adulthood can be a challenging process, and individuals vary substantially in their ultimate success (e.g. Bradlow et al., 1999; Flege et al., 1995; Flege et al., 1999). Much of the variability in second language learning can be attributed to factors that are fairly easily observed, such as age of acquisition, time spent in a country or relative time spent using the second language, or motivation to improve proficiency (see Flege et al., 1995, 1999; Piske et al., 2001). After some time spent learning a second language, many adult learners achieve at least moderate success in some domains of language acquisition, such as syntax or word learning (e.g. Flege et al., 1999; Granena & Long, 2013). However, a common but perplexing finding persists in this literature: Adult learners exhibit a wide range of variability in their ability to learn to perceive and produce the speech sounds of a second language (e.g. Bradlow et al., 1999; Lim & Holt, 2011; Myers & Swan, 2012; Yi et al., 2016). In fact, substantial individual variability can be found in almost all published studies on non-native speech sound learning, but as a field, we still have an incomplete picture of where these differences come from.

Individual differences in non-native speech sound learning

Many studies have found that both naive listeners and experienced second-language learners vary in how well they can perceive or produce speech sounds in a second language. For example, early studies that trained experienced second-language learners to improve their perception and production found a wide range of scores in both learning (e.g. Bradlow et al., 1997) and long-term retention (Bradlow et al., 1999). Studies of naive listeners have similarly found a great deal of individual variability in learning a non-native speech sound contrast using a variety of training paradigms. For instance, individual variability has been observed in studies using high-variability training, in which participants hear speech sounds spoken by several talkers during training (Perrachione et al., 2011; Sadakata & McQueen, 2014), in that pre-training aptitude predicts whether learners benefit from this type of training. Variability among participants has also been observed from study designs using implicit or incidental learning paradigms (Lim & Holt, 2011; Luthra et al., 2019), perceptual fading (Myers & Swan, 2012),

or explicit identification training (Earle et al., 2017; Earle & Arthur, 2017; Fuhrmeister & Myers, 2020). Thus, findings of substantial individual differences in non-native speech sound learning are robust and are found in naive listeners as well as experienced learners, and in multiple training paradigms.

Several studies have sought to elucidate the sources of individual variability in non-native speech sound learning, especially by relating individual differences in brain structure to behaviour. Individual differences in behaviour must originate somewhere, and individual variability in brain structure is a logical place to look for factors that could explain these differences. Finding *which* brain structures predict non-native category learning may give us clues as to what mechanism underlies individual differences in behaviour (in this case, non-native speech sound learning abilities).

Some recent work, however, suggests that learners do not only vary in their ability to *learn* non-native speech sounds, but they also vary in how well they *retain* what they learn after a delay or period of offline consolidation (e.g. Earle et al., 2017; Fuhrmeister & Myers, 2020). For instance, even two learners who show similar performance on a speech perception task before training (i.e. a pretest/baseline measure) may show very different patterns of learning and retention of the sounds over time (Earle et al., 2017; Fuhrmeister & Myers, 2020). One recent study suggests that individual differences in sleep duration may be one variable that predicts overnight improvement on non-native speech sound learning tasks. Specifically, Earle et al. (2017) found that total sleep duration predicted overnight gains on a non-native discrimination task and duration of slow wave sleep predicted overnight gains in identification performance. Therefore, the individual variability seen in overnight improvement in non-native speech sound learning studies may at least partially be explained by sleep duration. An open question, however, is whether individual variation in brain structure also predicts retention of newly learned non-native speech sounds.

Functional and anatomical correlates of non-native speech sound learning

Converging evidence from several structural and functional MRI studies suggests that individual variability in brain structure or amount of activation in response to non-native speech sounds predicts individual differences in non-native speech sound *learning*. Less is known about whether these relationships hold for *retention*. To make predictions for retention, we draw on the structural and functional MRI literature from non-native

speech sound learning, as well as the memory consolidation literature.

Several studies of brain structure have found that variation in regions specialised for both speech and general auditory processing, especially temporal and parietal regions, is related to non-native speech sound learning. For example, Golestani et al. (2006) found that white matter density in the transverse temporal gyrus was greater for faster vs. slower learners of the Hindi dental/retroflex contrast. Wong et al. (2008) found that grey matter density was greater in the transverse temporal gyrus in more successful learners of a non-native tonal contrast. Early auditory/sensory areas such as the transverse temporal gyrus typically process fine-grained acoustic details of a stimulus (e.g. Binder et al., 2000; Okada et al., 2010), and these findings suggest that individual variability in non-native speech sound learning may in part be explained by an individual's auditory acuity or the ability to attend to subtle acoustic details in the speech stream. An earlier study by Golestani et al. (2002) found greater white matter density in a region anterior to the parieto-occipital sulcus in faster learners than in slower learners of the Hindi dental/retroflex contrast. Complementing structural studies, Golestani and Zatorre (2004) found that functional activation in the bilateral angular gyri predicted individual differences in non-native speech sound learning. Thus, structure or function of temporal and parietal areas seem to predict individual differences in *learning* of non-native speech sounds, and we expect they may be related to *retention* of new sounds as well.

Other studies have found anatomical variation in frontal regions that predicts non-native speech perception or learning. Sebastián-Gallés et al. (2012) found differences between good and poor perceivers of native and non-native vowels in a region encompassing the right insula and frontal operculum in bilingual participants. Specifically, poor perceivers had more white matter density in this region, which the authors interpreted as possibly resulting from the use of compensatory strategies in speech perception (i.e. greater reliance on frontal regions as opposed to sensory areas in temporal regions). Rodriguez et al. (2018) found somewhat similar results, namely that cortical thickness of the left insula predicted non-native speech sound learning in bilinguals, though this pattern did not hold for monolinguals. Golestani et al. (2011) found that the volume of the pars opercularis region of the inferior frontal gyrus predicted years of phonetic training in a group of expert phoneticians. Though these findings do not specifically relate to non-native speech sound learning, they suggest that the pars opercularis subregion of the inferior frontal gyrus is related to phonetic expertise,

and it is logical to suppose that phonetic expertise may aid in non-native speech sound learning. In an fMRI study of non-native phonetic learning, Golestani and Zatorre (2004) found that activation in frontal regions including the left inferior frontal gyrus was related to learning of a non-native speech sound contrast. Based on these findings, we predict that anatomical variability in frontal regions such as the pars opercularis may predict non-native speech sound learning or retention.

Memory consolidation of non-native speech sounds

In many studies of non-native speech sound learning, participants complete several training sessions (e.g. Golestani et al., 2002; Golestani & Zatorre, 2004; Wong et al., 2008). Therefore, any study that involves outcomes after multiple days/sessions of training likely collapses across learning and retention. Because of this, we predict that the relationships between structural variation in frontal and temporal regions and non-native speech sound *learning* described in the previous section will also predict a participant's *retention* of the sounds. However, in the current study, we collected behavioural measures of non-native speech sound learning before and after a period of offline consolidation (an overnight interval containing sleep). This design allows us to test whether structural variation in these regions predicts learning or retention separately. It also allows us to test a new question, namely, whether hippocampal volume predicts improvement on the non-native speech tasks after a period of sleep at the individual level (see Earle & Myers, 2014, for a review on the contributions of sleep to non-native speech sound learning). The hippocampus has been found to play a role in sleep-related consolidation of newly formed memory traces (Davis et al., 2009; Davis & Gaskell, 2009; McClelland et al., 1995), and an open question is whether hippocampal volume predicts individual differences in improvement on a non-native speech sound learning task after a period of offline consolidation. Such a finding would highlight the importance of domain-general memory consolidation processes for explaining individual differences in speech sound learning.

Morphological variability of the transverse temporal gyrus

The size and gyrification patterns (the folding patterns of the cerebral cortex) of the transverse temporal gyrus (or Heschl's gyrus) vary among individuals. Common morphological variations include split, duplicate, and sometimes even multiple transverse temporal gyri (Marie et al., 2016). A large-scale study with over 200 right-

handed participants found that approximately 64% of the sample had split or duplicate Heschl's gyri in either the right or left hemisphere (Marie et al., 2016). The transverse temporal gyrus is of interest to speech research because it contains primary auditory cortex. Indeed, several studies have found that variation in gyrification patterns in the transverse temporal gyri predict non-native speech sound learning (Golestani et al., 2006), musical ability (Turker et al., 2017), and phonetic expertise (Golestani et al., 2011). Golestani et al. (2006) found that faster learners of a challenging non-native phonetic contrast were more likely to have multiple or split transverse temporal gyri in the left hemisphere. Turker et al. (2017) obtained similar findings: Participants who scored higher on a Hindi speech sound imitation task were more likely to have multiple or split transverse temporal gyri, but in the right hemisphere. Paradoxically, split or duplicate transverse temporal gyri have been linked to both phonological deficits (Leonard et al., 2001) and phonetic expertise (Golestani et al., 2011). Leonard et al. (2001) found that individuals with dyslexia who exhibited a phonological deficit were more likely to have multiple or split Heschl's gyri compared to a group of typical readers. However, Golestani et al. (2011) found that a group of expert phoneticians had more occurrences of multiple or split Heschl's gyri compared to a group of controls with a comparable education background. Findings from these two studies are difficult to reconcile, but neither study had a large sample size. For instance, the study by Leonard et al. (2001) had only 11 participants in the group with phonological dyslexia, and the study by Golestani et al. (2011) only had 17 in their group of expert phoneticians. Low-powered studies due to small sample sizes can result in overestimates of effects or even effects that have the wrong sign (Gelman & Carlin, 2014). Therefore, it is not surprising that we see conflicting evidence in the MRI literature, where many studies suffer from small sample sizes (see Button et al., 2013, for review). In addition, the criteria for phonological dyslexia in the study by Leonard et al. (2001) was determined by a pseudoword decoding task. It is possible that this skill is different from skills that the group of phoneticians in the Golestani study had acquired from their phonetic training. Another possibility is that people with phonological dyslexia are actually adept at distinguishing subtle differences in speech sounds, and this ability to detect subtle differences in sound is related to gyrification patterns in primary auditory areas. This would be consistent with Serniclaes' theory of allophonic perception in dyslexia (Serniclaes et al., 2004), which posits that people with dyslexia often perceive allophonic variants of speech categories as separate sounds, which makes it difficult to map the sounds to a common grapheme.

Overall, we lack the evidence to conclude whether multiple or split transverse temporal gyri are predictive of disordered phonological processing or phonological expertise, or whether these relationships are reflective of different skills. On the whole, it seems that individuals who have more occurrences of multiple or split transverse temporal gyri are better at detecting subtle differences in sounds, whether those sounds are non-native speech sounds or native-language speech sounds. Because gyrification patterns are established very early in development (Rakic, 2000; White et al., 2010), this may suggest that some individuals have a predisposition for attending to subtle differences in sound. Based on these studies, we predict that listeners who have more gyrified transverse temporal gyri will show more learning of a non-native speech sound contrast. We also extend this prediction to overnight improvement, as Fuhrmeister and Myers (2020) found that more accurate pre-training discrimination of non-native speech sounds positively predicted overnight improvement on a discrimination task.

Current study

In the current study, we aim to extend previous findings on the structural neural correlates of non-native speech sound learning to predict individual differences in both learning and retention (or consolidation) of non-native speech sounds. To test this, we trained participants to learn the voiced dental and retroflex stop consonants found in Hindi and measured their baseline discrimination ability (pretest), immediate learning, and retention after an overnight delay of approximately 12 h. Sleep duration was measured, and brain structure was measured using MRI. Based on the previous studies reviewed above, we predicted that individual variation in brain structure in typical speech regions (both frontal and temporo-parietal regions) would predict non-native speech sound learning. In our own work, we have found that a measure of baseline discrimination of non-native speech sounds predicts learning and retention of the sounds (i.e. naive perception, learning, and retention seem to be related, Fuhrmeister & Myers, 2020); therefore, we predict that measurements of brain structure from typical language regions will similarly predict retention of the sounds after a delay. Borrowing from the memory consolidation literature, we predict that hippocampal volume will predict retention after a period of offline consolidation. A finding that structural variation in typical language regions predicts retention would suggest that individual differences in speech perception or language ability contribute to both learning and retention of non-native speech sounds. In contrast, a

relationship between hippocampal volume and retention may imply that individual differences in memory consolidation processes drive differences in retention of speech sounds. If we find that both structural variation in language regions and hippocampal volume predict retention, this might suggest that individual differences in perceptual and memory processes work in tandem to enhance (or perhaps contribute to different aspects of) developing memory representations of speech sounds.

Method

Participants

Fifty-eight monolingual, native speakers of English (43 female, 15 male; ages 18–40, mean age = 23.32, $SD = 4.67$) were recruited from the University of Connecticut community by means of flyers posted in university buildings and advertisements in the daily email announcements for students, faculty, and staff. Data from one participant was excluded from all analyses due to an equipment error. One participant's data was excluded from the MRI analyses because the participant did not complete that session of the experiment. Data from the remaining participants are reported below (behavioural analyses, $N = 57$; MRI analyses, $N = 56$). Participants reported having no history of speech or language disorders and typical hearing. We additionally administered a pure tone audiometric hearing screening at 25 dB HL for frequencies of 500, 1000, 2000, and 4000 Hz in a quiet room with an Earscan 3 Audiometer (Micro Audio-metrics Corp, Murphy, North Carolina). All participants passed the hearing screening at all tested frequency levels, with the exception of one participant whose data was lost due to experimenter error. This participant did not show any abnormal patterns of speech sound learning or retention and is included in the analyses. All participants indicated that they were right-handed via self-report, and they had either at least a bachelor's degree or were undergraduate students at the time of the study. No participants reported substantial experience with a second language; only nine participants reported exposure to either French or Spanish in school before the age of thirteen. We obtained informed consent from participants, following the guidelines of the University of Connecticut Institutional Review Board. Participants were compensated \$10 per hour for behavioural tasks and \$30 per hour for the MRI.

Stimuli and materials

To assess non-native speech sound learning, participants were trained on the voiced dental (/d/) and retroflex (/ɖ/)

stop consonants found in Hindi (a difficult phonetic contrast for native English speakers to learn, e.g. Best et al., 2001). These stimuli were recorded in a sound-attenuated booth by a female, native speaker of Hindi at the University of Connecticut in the Brain Imaging Research Center. Five recordings of each minimal pair nonword (/ɖʊg/ and /ɖʊg/) were obtained. Stimuli were scaled to a mean amplitude of 65 dB SPL using Praat (Boersma & Weenink, 2013). All auditory stimuli were presented using over-ear headphones (SONY MDR-7506, New York) at a comfortable listening level that participants could adjust themselves. Visual stimuli consisted of “Fribbles”, (novel objects that participants should have no familiarity with, stimulus images courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University, <http://www.tarrlab.org/>). The experiment was presented using OpenSesame experimental software (Mathôt et al., 2012) on a desktop computer.

Procedure

Participants made a total of three visits to the lab: two behavioural sessions and one MRI session. The behavioural sessions were completed on two consecutive days. The first session took place between the hours of 5 and 9 PM, and the second session took place between 8 and 10 AM (see Figure 1). Two recent studies on the effects of sleep-mediated consolidation on non-native speech sound learning have found more improvement on behavioural tasks following an interval of sleep when the initial learning session took place in the evening, rather than the morning hours (Earle & Myers, 2015; Qin & Zhang, 2019, 2019). We therefore chose to train all participants in the evening hours to give participants a better opportunity to show consolidation-based improvements. In the first session, participants gave informed consent, then completed a pretest to assess their discrimination ability for the sounds, a phonetic training task to learn the Hindi sounds, and two posttests to assess their identification and discrimination of the sounds (see Figure 1). Sleep duration was also measured between the two behavioural sessions via an Actigraph GT3XP-BTLE wristwatch device (ActiGraph LLC, Pensacola, FL, USA). In the second session, participants self-reported their total sleep duration to check against the ActiGraph data and were reassessed on their identification and discrimination of the Hindi sounds to measure retention. This data set is part of a larger study, for which we also collected data to measure perception of native-language speech sounds, working memory, and phonological skills; these data are not reported here. The MRI session could take

place at any time (before or after the behavioural sessions), as brain structure does not change rapidly as a result of phonetic training (see Golestani, 2014, for review). Structural MRI images were acquired from a 3-T Siemens Prisma with a 64-channel head coil. T1-weighted images were acquired sagittally using an MPRAGE sequence (TR = 2300 ms, TE = 2.98 ms, FOV = 256 mm, flip angle = 9 degrees, voxel size = 1 × 1 × 1 mm³). Diffusion weighted images and magnetic resonance spectroscopy data were also collected but are not reported here.

Non-native speech sound learning tasks

AX discrimination

In order to assess pre- and post-training perceptual sensitivity to the Hindi sounds, participants completed an AX discrimination task. In this task, two of the minimal pair nonwords were presented auditorily (e.g. /ɖʊg/ ... /ɖʊg/), and participants indicated whether they thought the words sounded the same or different. Participants completed 64 trials total with no feedback. For half of the trials, the initial speech sounds of each nonword came from the same speech category but were acoustically distinct recordings to discourage participants from using low-level details of the acoustic signal to differentiate the sounds. Among the same trials, half of those consisted of two exemplars of the dental category, and half of the retroflex category. For the different trials, the onset speech sounds were from two different categories, and on half of these trials, the dental token was presented first, and for the other half, the retroflex was presented first.

Identification training and test

Immediately following the baseline AX discrimination measure, participants were familiarised with the nonwords that correspond to each novel visual stimulus. After that, participants completed 400 training trials with a two-minute break after the first 200 trials. On each training trial, participants saw two novel visual images on the screen and heard one word beginning with either the dental or retroflex sound. Minimal feedback was provided visually (e.g. “Correct!” or “Incorrect!”). Identification tests consisted of 50 trials identical to training, except feedback was not provided.

Analysis approach

Non-native speech sound learning tasks

For data from the discrimination tasks, *d* prime (*d'*) scores were calculated [$z(\text{hits}) - z(\text{false alarms})$] to account for response bias (Macmillan & Creelman,

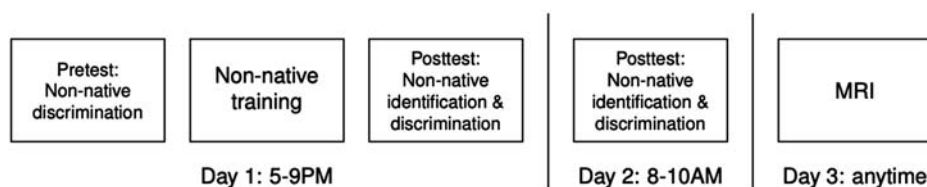


Figure 1. Experiment procedure and schedule of tasks.

2004). For identification assessments without feedback, we occasionally see that participants confuse the category labels even though they can differentiate the sounds, and this results in accuracy scores less than what would be expected by chance performance. We calculated below-chance performance using a binomial test, which for our data resulted in a threshold of less than 38% accuracy ($p < .05$). This threshold was applied to all subjects, such that trial-level data for participants whose total accuracy scores at either time point were less than 38% were recoded to reflect the label switching (i.e. 0 was recoded as 1 and 1 was recoded as 0). This affected four participants' data on the next-day posttest (no participants switched the labels at the immediate posttest). To test whether participants improved over time on the non-native learning tasks, we used mixed effects linear regression models¹ to predict discrimination performance (d' scores) and mixed effects logistic regression models to predict identification accuracy (0 or 1) using the lme4 package (Bates et al., 2015) in R (R Core Team, 2020). Since we averaged over trials in the discrimination data (d' scores), we did not have enough data to estimate random slopes; therefore, we only included random intercepts for participant. In linear mixed effects models, p -values were estimated with the Satterthwaite method using the afex package (Singmann et al., 2019). To determine the random effects structure of the model of identification data (where we had trial-level data), we used a backwards stepping procedure as in Matuschek et al. (2017). For any model convergence issues, we used the glmerControl optimiser "bobyqa" to increase iterations to 200,000. All raw data and analysis scripts for behavioural analyses are publicly available at <https://osf.io/vbtxd/>.

MRI data

Structural MRI data were preprocessed with FreeSurfer's automated preprocessing pipeline (Dale et al., 1999; Fischl, 2012). FreeSurfer enables surfaced-based analyses of brain structure by reconstructing MRI images into a two-dimensional surface map consisting of a triangle mesh containing information at each vertex. The pial surface (boundary between grey matter and cerebral spinal fluid) and white matter surface (boundary

between white matter and grey matter) are estimated, which enable precise calculations of surface area, cortical thickness (distance between the pial and white matter surfaces), and volume (surface area multiplied by cortical thickness). Surface-based analyses have advantages over voxel-based morphometry because they allow analysis of several different structural metrics, such as surface area, cortical thickness, volume (which is the product of surface area and cortical thickness), curvature, and gyrification. There is evidence that cortical thickness and surface area result from different genetic processes (Wierenga et al., 2014; Winkler et al., 2010), so considering these metrics separately will give us more information about possible genetic differences that underlie non-native speech sound learning. In the present study, we are interested in measures of surface area and cortical thickness in order to expand on previous findings, specifically, to test whether volumetric relationships with non-native speech sound learning stem from thickness or surface area. To compare our results with previous studies that have used voxel-based approaches, we also included measures of volume in our analyses.

Whole-brain exploratory analyses

We first did a whole-brain analysis in addition to planned region of interest analyses to explore whether non-native speech sound learning is predicted by clusters of surface area, cortical thickness, or volume that do not fall within our a priori defined regions of interest. A whole-brain analysis identifies clusters of vertices of structural metrics that differ as a function of group or are correlated with a continuous measure, as was done in the present study. To that end, we fit a series of generalised linear models using the `mri_glmfit` command in FreeSurfer. Separate analyses were carried out to test relationships between behavioural measures of non-native learning (difference score of immediate posttest pretest) and retention (difference score of next-day posttest – immediate posttest) and measures of surface area, cortical thickness, and volume, and for each hemisphere. Surfaces were smoothed with a Gaussian kernel with a full-width/half-max of 10mm. We used `mri_glmfit-sim` to implement a vertex-wise cluster

forming threshold of .001 (Greve & Fischl, 2018) and a cluster-wise p threshold of .05 using non-directional tests. Bonferroni correction was applied to correct for tests over two hemispheres.

Region of interest analyses

For region of interest analyses, each vertex of the cortical surface is probabilistically assigned to a region according to an atlas. Regions of interest for the current study were selected from the Destrieux atlas in Freesurfer (Destrieux et al., 2010). Based on previous literature, we identified the following bilateral regions of interest for our analyses of non-native measures: the pars opercularis region of the inferior frontal gyrus (Golestani & Zatorre, 2004; Lee et al., 2012; Myers, 2007; Myers et al., 2009), the supramarginal gyrus (Golestani et al., 2002), the angular gyrus (Golestani & Zatorre, 2004), the superior temporal gyrus (Golestani & Zatorre, 2004; Myers, 2007), the transverse temporal gyrus (Golestani et al., 2006, 2011; Turker et al., 2017; Wong et al., 2008). Regions of interest can be found in Figure 2 and the FreeSurfer labels for these regions can be found in Table 1.²

To test whether structural measures of these regions were related to non-native speech sound learning, mixed effects models were fit to predict learning and retention of non-native discrimination and identification from structural metrics (surface area, cortical thickness, and volume) of each region of interest. Unless specified otherwise below, brain measures were mean-centred and scaled using the scale function in R (scaled and centred values were derived by subtracting the mean of the vector and dividing by the standard deviation). For each dependent variable (non-native

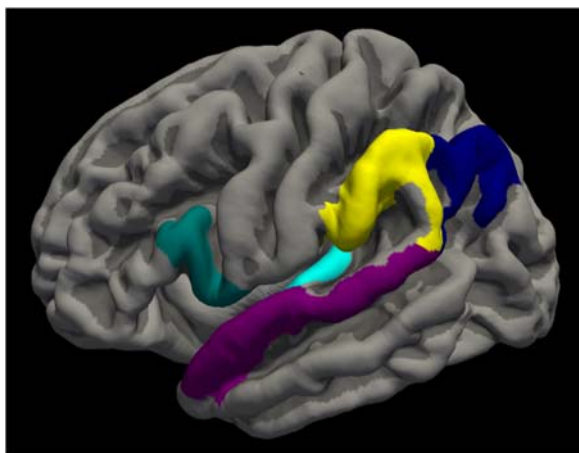


Figure 2. Regions of interest included in analyses: the pars opercularis region of the inferior frontal gyrus, supramarginal gyrus, angular gyrus, superior temporal gyrus, and transverse temporal gyrus. All regions were tested bilaterally.

identification and discrimination), we fit three separate models (ROI predictors of surface area, cortical thickness, and volume measurements); Bonferroni correction was applied to p -values to correct for these three comparisons. To account for differences in head size, total intracranial volume was added as a predictor to the models using surface area and volume as predictors. Cortical thickness is not as related to head size, so models were fit with measures of thickness as predictors without including total intracranial volume as a predictor. Details of each model can be found in the results section below.

We also tested whether hippocampal volume predicted overnight change in non-native discrimination or identification, as sleep-mediated memory consolidation may be beneficial for non-native speech sound learning (see Earle & Myers, 2014, for review). Subcortical structures require a volumetric segmentation procedure (described in detail in Fischl et al., 2002), so in contrast to cortical regions, we can only obtain volume measurements for subcortical structures. Note that the hippocampus is not listed in Table 1 because the volume measurements were derived from a different segmentation process, and these data were analysed separately because we had specific a priori hypotheses about the contributions of hippocampal volume to *overnight change* in non-native speech sound learning tasks. Total intracranial volume was included as a fixed effect in this model to account for differences in head size.

Gyrification

We used Freesurfer to compute the local gyrification index using the `-localGI` flag in `recon -all`. The local gyrification index is the ratio of the pial surface to a smoothed outer surface, and it is calculated at each vertex of the two-dimensional cortical surface (Schaer et al., 2012). The local gyrification index for a region of interest is the mean of the local gyrification indices at each vertex in each region of the cortical parcellation. The local gyrification index provides us a continuous measure of gyrification, which we chose to use to increase statistical power, as opposed to manually identifying the number of gyri (e.g. Golestani et al., 2006, 2011; Turker et al., 2017; Wong et al., 2008). Based on

Table 1. Regions of interest and Freesurfer Destrieux atlas labels. All regions were tested bilaterally.

Region of interest	Destrieux atlas label
Inferior frontal gyrus (pars opercularis region)	G_front_inf-Opercular
Supramarginal gyrus	G_pariet_inf-Supramar
Angular gyrus	G_pariet_inf-Angular
Transverse temporal gyrus	G_temp_sup-G_T_transv
Superior temporal gyrus	G_temp_sup-Lateral

previous work, we were interested in testing whether gyrification of the bilateral transverse temporal gyri predicts learning or retention of non-native speech sounds (Golestani et al., 2006, 2011; Turker et al., 2017; Wong et al., 2008). All data and analysis scripts for region of interest and gyrification analyses are publicly available at <https://osf.io/vbtxd/>.

Results

Non-native behavioural measures

Learning and retention of non-native speech sounds

First, we examined non-native speech sound learning and retention. Data from all 57 participants contributed to these analyses.

Identification

We first tested whether participants improved after a delay on the identification task (the trained task). The model predicted accuracy (0 or 1 for each trial), and time was included as a fixed effect (immediate posttest or next-day posttest), which was deviation coded³ (immediate posttest = $-.5$, next-day posttest = $.5$). The final model included random intercepts for participant. The intercept of the model was significantly greater than zero, $\beta = 2.16$ (95% CI [1.81, 2.54]), $SE = .18$, $z = 11.85$, $p < .001$, indicating that participants identified the sounds at above chance levels. There was no significant difference in performance in the two time points, $\beta = .05$ (95% CI [$-.10$, $.21$]), $SE = .08$, $z = .71$, $p = .48$, suggesting that participants maintained training-induced gains after the overnight interval, but did not further improve overnight (see Figure 3A).

Discrimination

We next tested for changes in discrimination performance over time. Time (pretest, immediate posttest, next-day posttest) was included as a fixed effect in the model and was backwards difference coded using the `contr.sdif()` function from the MASS package (Venables & Ripley, 2002) to test the following contrasts: immediate posttest – pretest (improvement after training) and next-day posttest – immediate posttest (overnight improvement). The intercept of the model (the grand mean) was significantly greater than zero, $\beta = 1.26$ (95% CI [1.01, 1.52]), $SE = .13$, $t = 9.79$, $p < .001$, indicating that participants discriminated the sounds at above chance levels. There was a significant difference between pretest and the posttest immediately following training, $\beta = .72$ (95% CI [.51, .92]), $SE = .11$, $t = 6.79$, $p < .001$, suggesting participants improved their

discrimination as a result of training. However, the difference between the immediate posttest and the next-day posttest did not reach significance, $\beta = .18$ (95% CI [$-.03$, $.38$]), $SE = .11$, $t = 1.69$, $p = .09$, despite the numerical increase between these two time points (see Figure 3B). This indicates that participants maintained training-induced gains, but did not improve further overnight.

Sleep duration and overnight improvement

Although we did not see that participants as a group improved on the non-native tasks after an interval of sleep, we saw a large amount of individual variability in overnight change (range in discrimination overnight change in d' scores: [-1.18 , 1.74]). Therefore, we may see that an individual's sleep duration predicts overnight change. Sleep duration was measured in the current study with an Actigraph wristwatch device and we also asked participants for a self-report of their sleep duration to check the accuracy of the Actigraph data. Actigraph devices measure total sleep duration only. In the study by Earle et al. (2017), however, they measured sleep duration with a different device that measured individual sleep stages. One participant's sleep data was not recorded due to experimenter error, so the remaining 56 participants were included in these analyses. Sleep duration did not predict overnight change in either discrimination or identification. Full details on the analyses of sleep data can be found in supplementary materials.

MRI analyses

Whole brain analyses

To explore whether structural measurements (surface area, cortical thickness, or volume) were related to learning or retention of non-native speech sounds, we conducted a whole-brain analysis as described above that tested relationships between brain structure and non-native phonetic learning (immediate posttest – pretest) and retention (next-day posttest – immediate posttest). Anatomical regions for each cluster were determined by the cortical parcellations from the Desikan-Killiany atlas (Desikan et al., 2006). No clusters of surface area, cortical thickness, or volume survived correction.

Region of interest analyses.

Identification

This model tested whether surface area, cortical thickness, or volume of the regions of interest predicted retention on the non-native identification task. Time (immediate posttest, next-day posttest) was included as a fixed factor (deviation coded as before). Structural

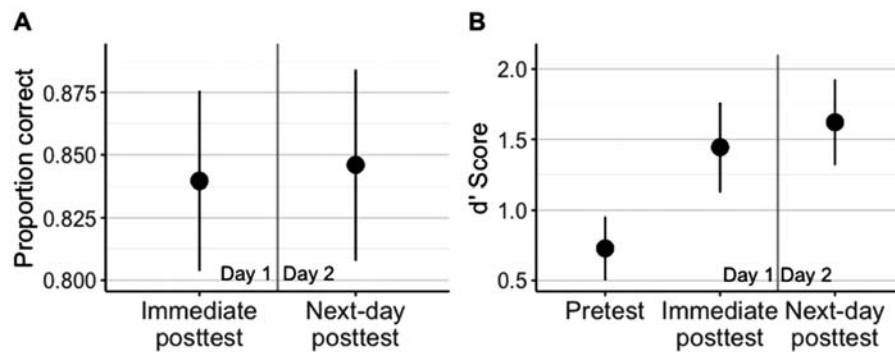


Figure 3. A. Non-native identification performance at each time point. Participants maintained learning but did not significantly improve after a period of offline consolidation. B. Non-native discrimination performance at each time point. Participants showed a significant increase in discrimination performance after training, but they did not significantly improve after a period of offline consolidation. Error bars represent 95% confidence intervals.

metrics of each region of interest were also included as predictors, as well as the interaction of time and the structural measurements (lmer syntax: accuracy ~ time*(LIFG + RIFG + LAG + RAG + LSM + RSM + LSTG + RSTG + LTTG + RTTG) + total intracranial volume).

The model predicting identification accuracy from surface area of the regions of interest revealed an interaction between surface area of the left transverse temporal gyrus and time, $\beta = -.28$ (95% CI [-.50, -.07]), $SE = .12$, $z = -2.62$, $p = .03$ (see Figure 4A), suggesting surface area of this region was inversely related to overnight change on the identification task. No cortical thickness or volume measurements of the regions of interest predicted retention of the non-native speech sounds, as measured by the identification task.

Discrimination

This analysis tested whether surface area, cortical thickness, or volume of the pre-selected regions of interest

predicted learning or retention on the non-native discrimination task. Time (pretest, immediate posttest, and next-day posttest) was included as a fixed effect (backwards difference coded as before). Structural metrics (surface area, cortical thickness, or volume) of each region of interest were included as predictors, as well as the interaction of time and the structural measurements (lmer syntax: $d'prime \sim time*(LIFG + RIFG + LAG + RAG + LSM + RSM + LSTG + RSTG + LTTG + RTTG) + total\ intracranial\ volume$).

Results revealed that surface area of the left superior temporal gyrus interacted with the time points immediate posttest – pretest (i.e. predicted learning of the contrast), $\beta = .36$ (95% CI [.11, .61]), $SE = .14$, $t = 2.58$, $p = .03$, as did volume of the left superior temporal gyrus, $\beta = .43$ (95% CI [.18, .68]), $SE = .14$, $t = 3.05$, $p = .01$, (see Figure 5). This suggests that greater surface area and volume of the left superior temporal gyrus predicted learning of the contrast, as measured by the discrimination task.

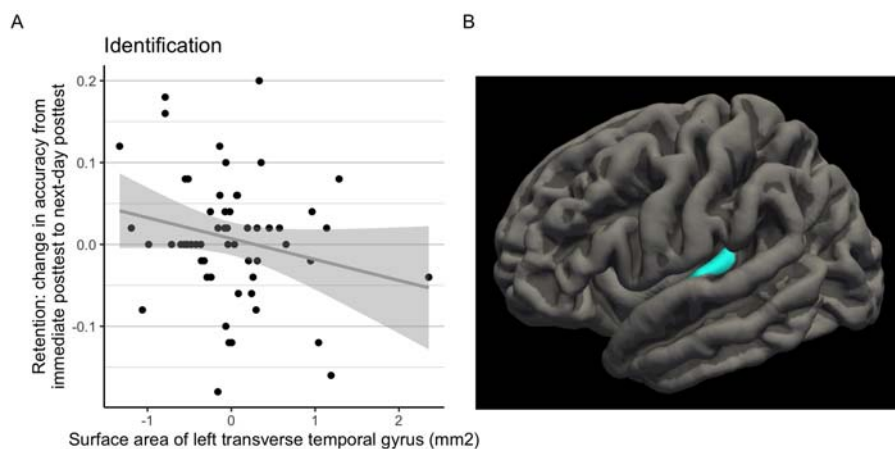


Figure 4. A. Surface area of the left transverse temporal gyrus negatively predicted retention/improvement after a delay on the identification task. Note that learning and retention scores are plotted as difference scores for ease of visualisation. B. Left transverse temporal gyrus region of interest.

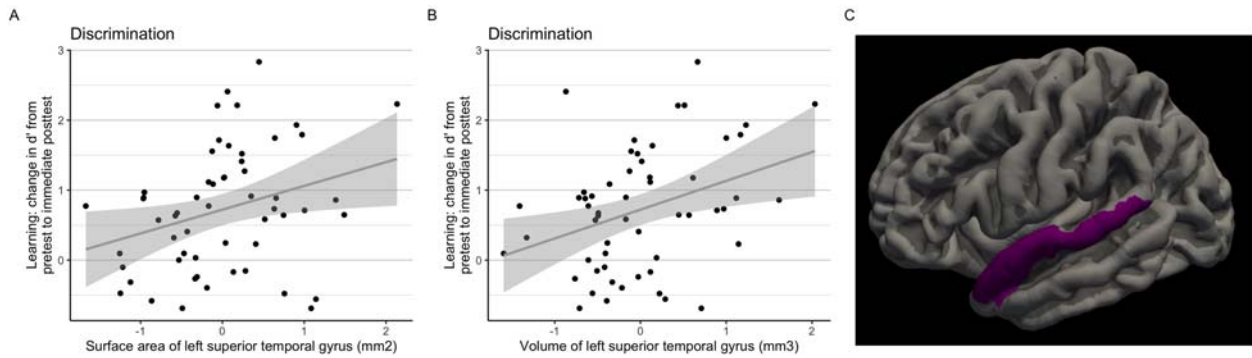


Figure 5. A. Surface area and B. volume of the left superior temporal gyrus positively predicted learning on the discrimination task. C. Left superior temporal gyrus region of interest.

No other significant effects or interactions were found, and cortical thickness did not predict learning or overnight change.

Hippocampal volume

Identification

This analysis tested whether hippocampal volume predicted overnight change on the identification task. Fixed effects included time (deviation coded as before), the interaction of hippocampal volume and hemisphere (hemisphere was deviation coded: left = $-.5$, right = $.5$), and their interactions, and total intracranial volume was included as a fixed factor to account for head size (glmer syntax: $\text{accuracy} \sim \text{time} * (\text{hemisphere} : \text{hippocampal volume}) + \text{total intracranial volume}$). This allowed us to estimate a main effect of time, and simple effects of hippocampal volume on identification scores at each time point in each hemisphere independently without estimating a main effect for hemisphere and removing the correlations between hippocampal volume in each hemisphere. The random effects structure of the final model included by-participant random intercepts and slopes for time with correlation parameters set to zero. No significant predictors or interactions were found.

Discrimination

This model tested whether hippocampal volume predicted overnight change on non-native discrimination performance. Fixed effects included time (deviation coded: immediate posttest = $-.5$, next-day posttest = $.5$), the interaction of hemisphere and hippocampal volume and their interactions (hemisphere was deviation coded as before), and total intracranial volume to account for head size (lmer syntax: $\text{dprime} \sim \text{time} * (\text{hemisphere} : \text{hippocampal volume}) + \text{total intracranial volume}$). There was a significant difference in the two

time points, $\beta = .17$ (95% CI $[.10, .25]$), $SE = .04$, $t = 4.27$, $p < .001$. Because hippocampal volume was mean-centred, this suggests that individuals with average hippocampal volume (at least average in this sample) improved on the discrimination task after the overnight interval. Hippocampal volume did not predict d' scores; however, there was an interaction between hippocampal volume in the right hemisphere and time, $\beta = .14$ (95% CI $[.03, .25]$), $SE = .06$, $t = 2.43$, $p = .02$. An effect in the same direction was found for the left hemisphere, but this interaction did not reach significance, $\beta = .09$ (95% CI $[-.02, .20]$), $SE = .06$, $t = 1.52$, $p = .13$. This suggests the relationship between hippocampal volume in the right hemisphere and d' scores was stronger at the next-day posttest than the immediate posttest. In other words, right hippocampal volume positively predicted overnight improvement on the discrimination task (see Figure 6).

Gyrification

The local gyrification index is a ratio with a minimum value of one, so in order to make the intercept more interpretable (i.e. interpret it as zero), we subtracted one from each participant's local gyrification index in each hemisphere. Models testing the relationship between gyrification of the transverse temporal gyri and non-native speech sound learning behavioural tasks are described below.

Identification

This analysis tested the relationship between gyrification measures of the transverse temporal gyri and non-native identification performance. Fixed factors included time (deviation coded as before), the interaction of the local gyrification index of the transverse temporal gyrus and hemisphere (deviation coded as before), and their interactions ($\text{accuracy} \sim \text{time} * (\text{hemisphere} : \text{local gyrification index})$). Random effects in the final model included by-

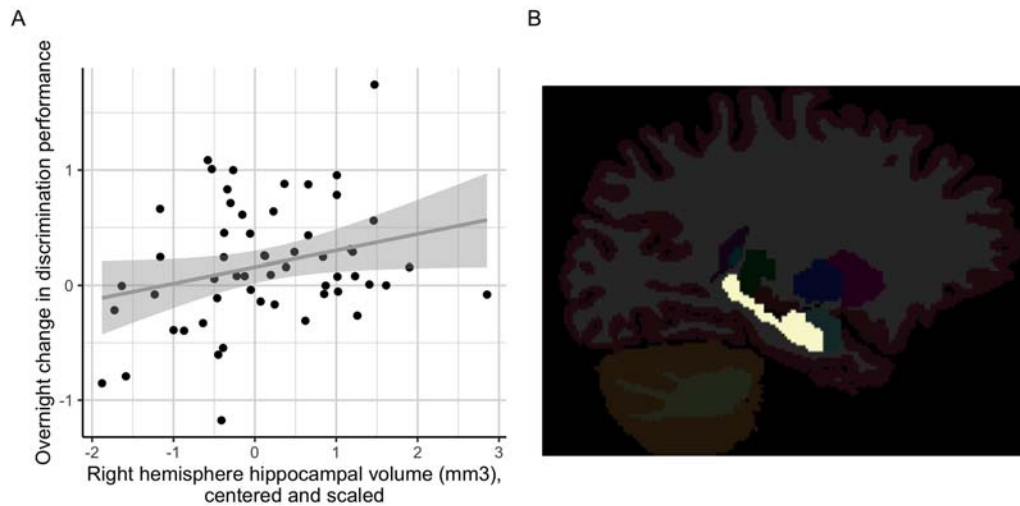


Figure 6. A. Volume of the right hippocampus positively predicts overnight change in discrimination performance. Note that overnight change is plotted as a difference score for ease of visualisation, but the interaction between hippocampal volume and time point was tested in the statistical model. B. Right hippocampus region of interest.

participant random intercepts and slopes for time with correlations of random effects set to zero. The model revealed no significant effects or interactions, suggesting that gyrification of the transverse temporal gyri was not a predictor of identification performance.

Discrimination

This model tested the relationship between gyrification of the transverse temporal gyri and non-native discrimination. Fixed effects included time (backwards difference coded as before), the interaction of the local gyrification index of the transverse temporal gyrus and hemisphere (deviation coded as before), and their interactions ($d'_{\text{prime}} \sim \text{time} * (\text{hemisphere} : \text{local gyrification index})$). We found a difference between the immediate posttest and the next-day posttest, $\beta = -2.08$ (95% CI $[-3.73, -.43]$), $SE = .85$, $t = -2.45$, $p = .01$. We additionally found an interaction of the local gyrification index in the

left hemisphere and the difference between the immediate posttest and the next-day posttest, $\beta = .59$ (95% CI $[-.16, 1.01]$), $SE = .22$, $t = 2.66$, $p = .008$; and the same interaction in the right hemisphere, $\beta = .58$ (95% CI $[-.16, 1.00]$), $SE = .22$, $t = 2.66$, $p = .008$, suggesting that gyrification in the bilateral transverse temporal gyri was positively related to overnight change in discrimination performance (see Figure 7A). However, after visually inspecting the data in Figure 7B, it looked as if the difference in slopes indicated by the interactions found in this analysis were a result of the negative relationship between gyrification and discrimination performance at the immediate posttest going away at the next-day posttest. We tested this by nesting the interaction between the local gyrification index and hemisphere within time to get simple effects of gyrification in each hemisphere at each time point. This model showed no significant relationships between gyrification and

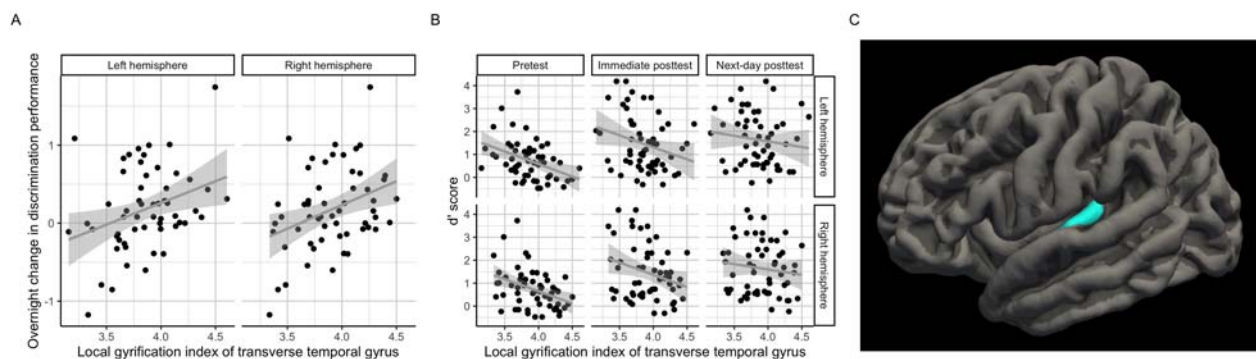


Figure 7. A. Local gyrification index of the bilateral transverse temporal gyri predicts overnight change in discrimination performance. B. Relationship between the local gyrification index of the transverse temporal gyri and d' scores at each time point. C. Transverse temporal gyrus region of interest.

discrimination performance at any time point; however, the direction of the relationships between gyrification and discrimination performance were negative at the pretest and immediate posttest and close to zero at the next-day posttest.

Discussion

In the current study, we tested whether individual variability in brain structure is related to behavioural performance on non-native speech sound learning tasks. Previous studies have examined relationships between brain structure and non-native speech sound *learning*, and the goal of the current study was to test whether individual differences in brain structure also predict *retention* of non-native speech sounds after an interval of sleep. In the present study, we used surface-based analysis to derive structural measures of surface area, cortical thickness, volume, and gyrification.

Structure of the left superior temporal gyrus predicts non-native speech sound learning

In a region of interest analysis, we found that individual differences in brain structure of areas related to speech perception predicted non-native speech sound learning. Specifically, we found that surface area and volume of the left superior temporal gyrus interacted with time in statistical models predicting discrimination scores, suggesting that surface area and volume of this region were more strongly related to discrimination scores after training than before. We interpret this to mean that surface area and volume of the left superior temporal gyrus predicted learning of the non-native speech sounds. The (left) superior temporal gyrus (or adjacent regions in the superior temporal sulcus) is often thought to underlie native speech category representations (e.g. Chang et al., 2010; Myers, 2007), and several fMRI studies have shown clusters of activation in this region in response to learned speech categories. For example, Desai et al. (2008) found that listeners who learned sine-wave variants of native-language speech sounds showed more activation in the left superior temporal gyrus/sulcus after learning to perceive these sounds as speech, and the degree of that activation was related to how categorically (or speech-like) listeners perceived the sine-wave speech sounds. In a non-native speech sound learning study, Golestani and Zatorre (2004) found activation of the superior temporal gyrus in response to non-native speech sounds after training. Our results extend these previous findings to suggest that individual differences in brain structure in

the left superior temporal gyrus predicts non-native speech sound learning.

These findings should be interpreted with some caution, however. The superior temporal gyrus is a large region, and earlier studies have suggested that different parts of it may be related to different functions. For example, the posterior portion of the superior temporal gyrus has been found to support categorical perception of speech (Chang et al., 2010), while more anterior portions are thought to underlie representations of semantic concepts (Ralph et al., 2010). The relationship between structural measures of specific subregions of the superior temporal gyrus and non-native speech sound learning should be investigated in future work.

Structure of the transverse temporal gyrus predicts retention of non-native speech sounds

Over the long term, successful learning of non-native speech sounds requires maintenance and retention of these sound representation over time. For identification data, surface area of the left transverse temporal gyrus interacted with time, such that the relationship between surface area and identification performance was stronger at the *immediate* posttest than the next-day posttest. This suggests that surface area of this region is inversely related to overnight improvement on the identification task. The direction of this relationship was unexpected, and visual inspection of the scatter plot in Figure 4A suggests the relationship may be weak. If this relationship is real, we speculate that it is related to the task demands. For example, the identification task we used requires the listener to categorise the sounds and match them with a picture. It is possible that performance on this task relies less on sensory and perceptual processes than a discrimination task does, in which a listener has to compare subtle acoustic differences between tokens (e.g. Guenther et al., 1999). The transverse temporal gyri are often associated with the processing of more fine-grained acoustic detail rather than speech category-level information (e.g. Binder et al., 2000; Okada et al., 2010). It is therefore possible that the processing of fine-grained acoustic detail is not as helpful in the long run for learning to identify new speech categories, a situation where listeners ultimately need to collapse many acoustic exemplars into a single category. Finally, we did not have a pre-training measure of identification (like we did with discrimination) because listeners cannot perform an identification task without being familiarised with the sound-word pairings beforehand. Therefore, we did not have a contrast (i.e. posttest-pretest) to test for identification

improvement after training (learning). It is possible that we would have seen similar brain-behaviour relationships for identification learning as we saw for discrimination if we had a way to measure changes from pretest to immediate posttest in the identification task.

We additionally tested whether gyrification of the bilateral transverse temporal gyri predicted behavioural measures of non-native discrimination or identification. Based on previous findings (e.g. Golestani et al., 2006, 2011; Turker et al., 2017), we predicted that more gyrification of the transverse temporal gyrus in either hemisphere would predict better performance on non-native speech measures. We first found that a participant's local gyrification index was positively related to overnight change in non-native discrimination performance, but we did not find any relationship between gyrification and identification performance.

At first glance, the positive relationship between gyrification and overnight improvement on the non-native discrimination task seems intuitive; however, the reason for this relationship was unexpected. Visual inspection of the plotted data in Figure 7B suggests that gyrification negatively predicted performance on the pretest and immediate posttest, but by the next-day posttest, the relationship was attenuated. This suggests that a greater amount of gyrification in these regions predicted poorer naive discrimination of the sounds, which does not seem entirely consistent with prior studies showing more instances of split or duplicate transverse temporal gyri in expert phoneticians (Golestani et al., 2011), faster learners of the Hindi dental/retroflex contrast (Golestani et al., 2006), or more accurate imitation (production) of the Hindi dental/retroflex sounds (Turker et al., 2017). Our findings seem to contradict previous findings; however, previous studies used different tasks and measures of gyrification. Specifically, all three of the previous studies discussed measured the actual number of split or duplicate transverse temporal gyri, and it is possible that the local gyrification index used in the current study is capturing something different than the morphological differences observed in previous studies. Ultimately, future research will need to test whether a larger local gyrification index is related to split or duplicate gyri.

The current finding that transverse temporal gyrification is related to overnight change in discrimination could be indicative of different learning strategies between good and poor perceivers. For example, the good perceivers may have been able to rely on auditory acuity throughout the entire training and testing process, whereas poor perceivers may rely more on consolidation processes to catch up after a period of offline consolidation. In other words, the good perceivers may

have been able to rely on acoustic differences of the stimuli both before and after training in order to discriminate them, while the poor perceivers may have tapped into category learning strategies which helped them more after a period of offline consolidation. The finding that surface area and volume of the left superior temporal gyrus predicted learning is also consistent with this interpretation, namely that perceptual regions play a role early in speech sound learning. It is also possible we are seeing hints of structural variation that has been seen in phonological dyslexia as in the study by Leonard et al. (2001). In that study, they found that having multiple transverse temporal gyri was associated with phonological dyslexia. Perhaps the relationship between structural variation and discrimination performance seen in the current study is related to subtle differences in reading and language skills in typical learners. In the end, however, the relationship between gyrification and non-native discrimination disappears after a delay, so it seems that any early disadvantages associated with more gyrification in these areas were not long-lasting.

Hippocampal volume and retention of non-native sound contrasts

Prior work has raised the possibility that individual differences in non-native speech perception may be partially attributable to individual differences in memory consolidation of learned phonetic information (Earle et al., 2017). While we found that there was no significant improvement on discrimination or identification of the non-native speech sound contrast after an overnight delay, when we added measurements of hippocampal volume to a model predicting non-native discrimination scores, we found that learners with average hippocampal volume (at least in our sample) showed overnight improvement. We additionally found that volume of the right hippocampus positively predicted overnight change in discrimination performance, which supports the idea that domain-general memory consolidation processes are at least partially involved in the learning trajectory of non-native speech sounds. This is consistent with some findings in the memory consolidation literature, in which a larger hippocampus is related to better delayed or sometimes even immediate recall in patients with Alzheimer's disease (Köhler et al., 1998) or healthy young adults (Pohlack et al., 2014; but see Van Petten, 2004). Learning non-native speech sounds is an interesting problem of learning and memory because of the difficulty and large amount of individual variability seen in adult learners. The current results suggest that hippocampal volume may predict the

degree to which a learner can take advantage of memory consolidation processes, at least after a short delay. Hippocampal volume has been found to predict other types of learning, such as statistical learning in children (Finn et al., 2019) and it has been found to increase as a result of language learning (Bellander et al., 2016; Mårtensson et al., 2012). Our findings suggest that we may want to look to memory processes in the future to explain individual differences in longer-term outcomes of non-native speech sound learning.

Conclusions

The current study tested relationships between individual variation in brain structure and non-native speech sound learning and retention after an overnight, approximate 12-hour delay. Surface area and volume of the left superior temporal gyrus positively predicted learning of a non-native speech contrast as measured by a discrimination task, while surface area of the left transverse temporal gyrus negatively predicted retention as measured by an identification task. Volume of the right hippocampus positively predicted behavioural improvement after the overnight interval on the discrimination task. Writ large, these results underscore the importance of two separate systems supporting non-native speech sound learning, namely memory and perception. First, the fact that variability in hippocampal volume predicts learning suggests that domain-general memory processes may partially account for individual differences in non-native learning. From a practical perspective, second language instructors may find that increasing the opportunities to practice and consolidate speech category information over successive days may help this information be better retained in memory. Second, the finding that regions rather early in the neural processing hierarchy (especially the transverse temporal and superior temporal gyri) are predictive of learning suggests that individual differences in early sensory/perceptual processes may play a role in learning. What is less clear is whether greater perceptual acuity (leading to finer-grained detection of acoustic differences between categories), or less perceptual acuity (allowing listeners to ignore unimportant acoustic differences between categories) is the more optimal pattern. Other work (e.g. McCandliss et al., 2002), shows that learners may learn better when they initially hear exaggerated acoustic differences between speech categories. Ultimately, these patterns underscore the importance of both perception and memory processes for non-native speech sound learning and suggest that individual differences in brain structure in areas related to these processes predict learning and retention of non-native speech sounds.

Notes

1. For linear mixed effects models, normality of the residuals was verified by visually inspecting qq plots using the `qqnorm()` function in R. Logistic regression does not make this assumption, so this diagnostic was not performed for those models.
2. We acknowledge that there are other regions of interest that could have been selected based on the literature, but out of concern for statistical power, we did not want to have too many predictors in our models. We ultimately chose the regions whose structure or function had been shown in previous literature to be predictive specifically of *individual differences* in non-native speech sound learning, rather than just non-native speech sound learning at the group level. The whole-brain exploratory analysis addresses this limitation: Any robust relationships between behaviour and brain structures that were not included in the region of interest analyses should be captured in the whole-brain analysis.
3. We use the term deviation coding to describe sum coding in the narrower sense that the factor levels are coded as $-.5$ and $.5$ rather than -1 and 1 . We find this more intuitive to interpret for factors with two levels because the coefficient represents the difference between the two levels.

Acknowledgments

The authors are grateful to Dr. Peter Molfese for help with the gyrification preprocessing. This material is based upon work supported in part by the National Science Foundation under grant DGE-1747486 to the University of Connecticut, and NSF BCS 1554810 to EBM. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by National Science Foundation [grant numbers BCS 1554810,DGE-1747486].

References

- Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., Krivitsky, P. N., & Bolker, M. B. (2015). Package 'lme4'. *Convergence*, 12(1), 2. <https://doi.org/10.18637/jss.v067.i01>
- Bellander, M., Berggren, R., Mårtensson, J., Brehmer, Y., Wenger, E., Li, T.-Q., Bodammer, N. C., Shing, Y.-L., Werkle-Bergner, M., & Lövdén, M. (2016). Behavioral correlates of changes in hippocampal gray matter structure during acquisition of

- foreign vocabulary. *Neuroimage*, 131, 205–213. <https://doi.org/10.1016/j.neuroimage.2015.10.020>
- Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America*, 109(2), 775–794. <https://doi.org/10.1121/1.1332378>
- Binder, J., Frost, J., Hammeke, T., Bellgowan, P., Springer, J., Kaufman, J., & Possing, E. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, 10(5), 512–528. <https://doi.org/10.1093/cercor/10.5.512>
- Boersma, P., & Weenink, D. (2013). Praat: doing phonetics by computer [computer program]. version 5.3. 51. 2. Online: <http://www.praat.org>
- Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics*, 61(5), 977–985. <https://doi.org/10.3758/BF03206911>
- Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, 101(4), 2299–2310. <https://doi.org/10.1121/1.418276>
- Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365–376. <https://doi.org/10.1038/nrn3475>
- Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, 13(11), 1428–1432. <https://doi.org/10.1038/nn.2641>
- Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage*, 9(2), 179–194. <https://doi.org/10.1006/nimg.1998.0395>
- Davis, M. H., Di Betta, A. M., Macdonald, M. J., & Gaskell, M. G. (2009). Learning and consolidation of novel spoken words. *Journal of Cognitive Neuroscience*, 21(4), 803–820. <https://doi.org/10.1162/jocn.2009.21059>
- Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: Neural and behavioural evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1536), 3773–3800. <https://doi.org/10.1098/rstb.2009.0111>
- Desai, R., Liebenthal, E., Waldron, E., & Binder, J. R. (2008). Left posterior temporal regions are sensitive to auditory categorization. *Journal of Cognitive Neuroscience*, 20(7), 1174–1188. <https://doi.org/10.1162/jocn.2008.20081>
- Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3), 968–980. <https://doi.org/10.1016/j.neuroimage.2006.01.021>
- Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage*, 53(1), 1–15. <https://doi.org/10.1016/j.neuroimage.2010.06.010>
- Earle, F. S., & Arthur, D. T. (2017). Native phonological processing abilities predict post-consolidation nonnative contrast learning in adults. *The Journal of the Acoustical Society of America*, 142(6), EL525–EL531. <https://doi.org/10.1121/1.5013141>
- Earle, F. S., Landi, N., & Myers, E. B. (2017). Sleep duration predicts behavioral and neural differences in adult speech sound learning. *Neuroscience Letters*, 636, 77–82. <https://doi.org/10.1016/j.neulet.2016.10.044>
- Earle, F. S., & Myers, E. B. (2014). Building phonetic categories: An argument for the role of sleep. *Frontiers in Psychology*, 5, 1192. <https://doi.org/10.3389/fpsyg.2014.01192>
- Earle, F. S., & Myers, E. B. (2015). Sleep and native language interference affect non-native speech sound learning. *Journal of Experimental Psychology: Human Perception and Performance*, 41(6), 1680–1695. <https://doi.org/10.1037/xhp0000113>
- Finn, A. S., Kharitonova, M., Holtby, N., & Sheridan, M. A. (2019). Prefrontal and hippocampal structure predict statistical learning ability in early childhood. *Journal of Cognitive Neuroscience*, 31(1), 126–137. https://doi.org/10.1162/jocn_a_01342
- Fischl, B. (2012). Freesurfer. *Neuroimage*, 62(2), 774–781. <https://doi.org/10.1016/j.neuroimage.2012.01.021>
- Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., & Dale, A. M. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3), 341–355. [https://doi.org/10.1016/S0896-6273\(02\)00569-X](https://doi.org/10.1016/S0896-6273(02)00569-X)
- Flege, J. E., Munro, M. J., & MacKay, I. R. (1995). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America*, 97(5), 3125–3134. <https://doi.org/10.1121/1.413041>
- Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second-language acquisition. *Journal of Memory and Language*, 41(1), 78–104. <https://doi.org/10.1006/jmla.1999.2638>
- Fuhrmeister, P., & Myers, E. B. (2020). Desirable and undesirable difficulties: Influences of variability, training schedule, and aptitude on nonnative phonetic learning. *Attention, Perception, & Psychophysics*, 82(4), 2049–2065. <https://doi.org/10.3758/s13414-019-01925-y>
- Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, 9(6), 641–651. <https://doi.org/10.1177/1745691614551642>
- Golestani, N. (2014). Brain structural correlates of individual differences at low-to high-levels of the language processing hierarchy: A review of new approaches to imaging research. *International Journal of Bilingualism*, 18(1), 6–34. <https://doi.org/10.1177/1367006912456585>
- Golestani, N., Molko, N., Dehaene, S., LeBihan, D., & Pallier, C. (2006). Brain structure predicts the learning of foreign speech sounds. *Cerebral Cortex*, 17(3), 575–582. <https://doi.org/10.1093/cercor/bhk001>
- Golestani, N., Paus, T., & Zatorre, R. J. (2002). Anatomical correlates of learning novel speech sounds. *Neuron*, 35(5), 997–1010. [https://doi.org/10.1016/S0896-6273\(02\)00862-0](https://doi.org/10.1016/S0896-6273(02)00862-0)
- Golestani, N., Price, C. J., & Scott, S. K. (2011). Born with an ear for dialects? Structural plasticity in the expert phonetician

- brain. *Journal of Neuroscience*, 31(11), 4213–4220. <https://doi.org/10.1523/JNEUROSCI.3891-10.2011>
- Golestani, N., & Zatorre, R. J. (2004). Learning new sounds of speech: Reallocation of neural substrates. *Neuroimage*, 21(2), 494–506. <https://doi.org/10.1016/j.neuroimage.2003.09.071>
- 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(3), 311–343. <https://doi.org/10.1177/0267658312461497>
- Greve, D. N., & Fischl, B. (2018). False positive rates in surface-based anatomical analysis. *NeuroImage*, 171, 6–14. <https://doi.org/10.1016/j.neuroimage.2017.12.072>
- Guenther, F. H., Husain, F. T., Cohen, M. A., & Shinn-Cunningham, B. G. (1999). Effects of categorization and discrimination training on auditory perceptual space. *The Journal of the Acoustical Society of America*, 106(5), 2900–2912. <https://doi.org/10.1121/1.428112>
- Köhler, S., Black, S., Sinden, M., Szekely, C., Kidron, D., Parker, J., Foster, J. K., Moscovitch, M., Wincour, G., Szalai, J. P., & Bronskill, M. J. (1998). Memory impairments associated with hippocampal versus parahippocampal-gyrus atrophy: An mr volumetry study in Alzheimer's disease. *Neuropsychologia*, 36(9), 901–914. [https://doi.org/10.1016/S0028-3932\(98\)00017-7](https://doi.org/10.1016/S0028-3932(98)00017-7)
- Lee, Y.-S., Turkeltaub, P., Granger, R., & Raizada, R. D. (2012). Categorical speech processing in broca's area: An fmri study using multivariate pattern-based analysis. *Journal of Neuroscience*, 32(11), 3942–3948. <https://doi.org/10.1523/JNEUROSCI.3814-11.2012>
- Leonard, C. M., Eckert, M. A., Lombardino, L. J., Oakland, T., Kranzler, J., Mohr, C. M., King, W. M., & Freeman, A. (2001). Anatomical risk factors for phonological dyslexia. *Cerebral Cortex*, 11(2), 148–157. <https://doi.org/10.1093/cercor/11.2.148>
- Lim, S.-J., & Holt, L. L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive Science*, 35(7), 1390–1405. <https://doi.org/10.1111/j.1551-6709.2011.01192.x>
- Luthra, S., Fuhrmeister, P., Molfese, P. J., Guediche, S., Blumstein, S. E., & Myers, E. B. (2019). Brain-behavior relationships in incidental learning of non-native phonetic categories. *Brain and Language*, 198, 104692. <https://doi.org/10.1016/j.bandl.2019.104692>
- Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology press.
- Marie, D., Maingault, S., Crivello, F., Mazoyer, B., & Tzourio-Mazoyer, N. (2016). Surface-based morphometry of cortical thickness and surface area associated with heschl's gyri duplications in 430 healthy volunteers. *Frontiers in Human Neuroscience*, 10, 69. <https://doi.org/10.3389/fnhum.2016.00069>
- Mårtensson, J., Eriksson, J., Bodammer, N. C., Lindgren, M., Johansson, M., Nyberg, L., & Lövdén, M. (2012). Growth of language-related brain areas after foreign language learning. *NeuroImage*, 63(1), 240–244. <https://doi.org/10.1016/j.neuroimage.2012.06.043>
- Mathôt, S., Schreij, D., & Theeuwes, J. (2012). Opensesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, 44(2), 314–324. <https://doi.org/10.3758/s13428-011-0168-7>
- Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language*, 94, 305–315. <https://doi.org/10.1016/j.jml.2017.01.001>
- McClelland, B. D., Fiez, J. A., Protopapas, A., Conway, M., & McClelland, J. L. (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, & Behavioral Neuroscience*, 2(2), 89–108. <https://doi.org/10.3758/CABN.2.2.89>
- McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3), 419–457. <https://doi.org/10.1037/0033-295X.102.3.419>
- Myers, E. B. (2007). Dissociable effects of phonetic competition and category typicality in a phonetic categorization task: An fmri investigation. *Neuropsychologia*, 45(7), 1463–1473. <https://doi.org/10.1016/j.neuropsychologia.2006.11.005>
- Myers, E. B., Blumstein, S. E., Walsh, E., & Eliassen, J. (2009). Inferior frontal regions underlie the perception of phonetic category invariance. *Psychological Science*, 20(7), 895–903. <https://doi.org/10.1111/j.1467-9280.2009.02380.x>
- Myers, E. B., & Swan, K. (2012). Effects of category learning on neural sensitivity to non-native phonetic categories. *Journal of Cognitive Neuroscience*, 24(8), 1695–1708. https://doi.org/10.1162/jocn_a_00243
- Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I.-H., Saberi, K., Serences, J. T., & Hickok, G. (2010). Hierarchical organization of human auditory cortex: Evidence from acoustic invariance in the response to intelligible speech. *Cerebral Cortex*, 20(10), 2486–2495. <https://doi.org/10.1093/cercor/bhp318>
- Perrachione, T. K., Lee, J., Ha, L. Y., & Wong, P. C. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America*, 130(1), 461–472. <https://doi.org/10.1121/1.3593366>
- Piske, T., MacKay, I. R., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics*, 29(2), 191–215. <https://doi.org/10.1006/jpho.2001.0134>
- Pohlack, S. T., Meyer, P., Cacciaglia, R., Liebscher, C., Ridder, S., & Flor, H. (2014). Bigger is better! hippocampal volume and declarative memory performance in healthy young men. *Brain Structure and Function*, 219(1), 255–267. <https://doi.org/10.1007/s00429-012-0497-z>
- Qin, Z., & Zhang, C. (2019). The effect of overnight consolidation in the perceptual learning of non-native tonal contrasts. *PLoS one*, 14(12). <https://doi.org/10.1371/journal.pone.0221498>
- Rakic, P. (2000, January). Radial unit hypothesis of neocortical expansion. In *Novartis Foundation symposium* (pp. 30–52). John Wiley.
- Ralph, M. A. L., Sage, K., Jones, R. W., & Mayberry, E. J. (2010). Coherent concepts are computed in the anterior temporal lobes. *Proceedings of the National Academy of Sciences*, 107(6), 2717–2722. <https://doi.org/10.1073/pnas.0907307107>
- R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing. <https://www.R-project.org/>

- Rodriguez, S. M., Archila-Suerte, P., Vaughn, K. A., Chiarello, C., & Hernandez, A. E. (2018). Anterior insular thickness predicts speech sound learning ability in bilinguals. *Neuroimage*, 165, 278–284. <https://doi.org/10.1016/j.neuroimage.2017.10.038>
- Sadakata, M., & McQueen, J. M. (2014). Individual aptitude in mandarin lexical tone perception predicts effectiveness of high-variability training. *Frontiers in Psychology*, 5, 1318. <https://doi.org/10.3389/fpsyg.2014.01318>
- Schaer, M., Cuadra, M. B., Schmansky, N., Fischl, B., Thiran, J.-P., & Eliez, S. (2012). How to measure cortical folding from mr images: A step-by-step tutorial to compute local gyrification index. *JoVE (Journal of Visualized Experiments)*, 59, e3417. <https://doi.org/10.3791/3417>
- Sebastián-Gallés, N., Soriano-Mas, C., Baus, C., Díaz, B., Ressel, V., Pallier, C., Costa, A., & Pujol, J. (2012). Neuroanatomical markers of individual differences in native and non-native vowel perception. *Journal of Neurolinguistics*, 25(3), 150–162. <https://doi.org/10.1016/j.jneuroling.2011.11.001>
- Serniclaes, W., Van Heghe, S., Mousty, P., Carré, R., & Sprenger-Charolles, L. (2004). Allophonic mode of speech perception in dyslexia. *Journal of Experimental Child Psychology*, 87(4), 336–361. <https://doi.org/10.1016/j.jecp.2004.02.001>
- Singmann, H., Bolker, B., & Westfall, J. (2019). *Aust. f. afex: Analysis of factorial experiments. R package version 0.23–0*.
- Turker, S., Reiterer, S. M., Seither-Preisler, A., & Schneider, P. (2017). “When music speaks”: Auditory cortex morphology as a neuroanatomical marker of language aptitude and musicality. *Frontiers in Psychology*, 8, 2096. <https://doi.org/10.3389/fpsyg.2017.02096>
- Van Petten, C. (2004). Relationship between hippocampal volume and memory ability in healthy individuals across the lifespan: Review and meta-analysis. *Neuropsychologia*, 42(10), 1394–1413. <https://doi.org/10.1016/j.neuropsychologia.2004.04.006>
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.
- White, T., Su, S., Schmidt, M., Kao, C.-Y., & Sapiro, G. (2010). The development of gyrification in childhood and adolescence. *Brain and Cognition*, 72(1), 36–45. <https://doi.org/10.1016/j.bandc.2009.10.009>
- Wierenga, L. M., Langen, M., Oranje, B., & Durston, S. (2014). Unique developmental trajectories of cortical thickness and surface area. *Neuroimage*, 87, 120–126. <https://doi.org/10.1016/j.neuroimage.2013.11.010>
- Winkler, A. M., Kochunov, P., Blangero, J., Almasy, L., Zilles, K., Fox, P. T., Duggirala, R., & Glahn, D. C. (2010). Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies. *Neuroimage*, 53(3), 1135–1146. <https://doi.org/10.1016/j.neuroimage.2009.12.028>
- Wong, P. C., Warrier, C. M., Penhune, V. B., Roy, A. K., Sadehh, A., Parrish, T. B., & Zatorre, R. J. (2008). Volume of left heschl’s gyrus and linguistic pitch learning. *Cerebral Cortex*, 18(4), 828–836. <https://doi.org/10.1093/cercor/bhm115>
- Yi, H.-G., Maddox, W. T., Mumford, J. A., & Chandrasekaran, B. (2016). The role of corticostriatal systems in speech category learning. *Cerebral Cortex*, 26(4), 1409–1420. <https://doi.org/10.1093/cercor/bhu236>