Article accepted for publication in *Studies in Second Language Acquisition*

doi:10.1017/S027226312000039X

Even in the best-case scenario L2 learners have persistent difficulty perceiving and utilizing tones in Mandarin: Findings from behavioral and ERP experiments

Eric Pelzl[1,2], Ellen F. Lau[1], Taomei Guo[3], & Robert DeKeyser[1]

1 University of Maryland, College Park
2 The Pennsylvania State University
3 Beijing Normal University

Abstract:

Lexical tones are widely believed to be a formidable learning challenge for adult speakers of non-tonal languages. While difficulties—as well as rapid improvements—are well documented for beginning second language (L2) learners, research with more advanced learners is needed to understand how tone perception difficulties impact word recognition once learners have a substantial vocabulary. The present study narrows in on difficulties suggested in previous work, which found a dissociation in advanced L2 learners between highly accurate tone identification and largely inaccurate lexical decision for tone words. We investigate a 'best-case scenario' for advanced L2 tone word processing by testing performance in nearly ideal listening conditions—with words spoken clearly and in isolation. Under such conditions, do learners still have difficulty in lexical decision for tone words? If so, is it driven by the quality of lexical representations or by L2 processing routines? Advanced L2 and native Chinese listeners made lexical decisions while EEG was recorded. Nonwords had a first syllable with either a vowel or

tone that differed from that of a common disyllabic word. As a group, L2 learners performed less accurately when tones were manipulated than when vowels were manipulated. Subsequent analyses showed that this was the case even in the subset of items for which learners showed correct and confident tone identification in an offline written vocabulary test. ERP results indicated N400 effects for both nonword conditions in L1, but only vowel N400 effects in L2, with tone responses intermediate between those of real words and vowel nonwords. These results are evidence of the persistent difficulty most L2 learners have in using tones for online word recognition, and indicate it is driven by a confluence of factors related to both L2 lexical representations and processing routines. We suggest that this tone nonword difficulty has real world implications for learners: it may result in many toneless word representations in their mental lexicons, and is likely to affect the efficiency with which they can learn new tone words.

Introduction

People often struggle to learn the unfamiliar speech sounds of a new language. Unsurprisingly, difficulty distinguishing speech sounds often leads to difficulty distinguishing words that contain them (e.g., Broersma & Cutler, 2011). In some case, even when second language (L2) learners have mastered novel speech sounds, they may still have difficulty using them to recognize words (Darcy et al., 2013; Díaz et al., 2012).

This latter pattern applies to L2 learning of lexical tones in Mandarin Chinese. In a previous study (Pelzl et al., 2019), we found that a group of advanced L2 Mandarin learners (native speakers of English with an average of ten years learning/using Mandarin) identified tones on single syllables with near-native accuracy, but performed below chance on a lexical decision task that required using tones to reject disyllabic nonwords.

The present study narrows in on these L2 tone word recognition difficulties. We focus on two general classes of explanation for L2 phonological and lexical difficulties (or *phonolexical* difficulties, cf. Chrabaszcz & Gor, 2014). The first attributes difficulties primarily to weaknesses in the quality of L2 *lexical representations* (Cook et al., 2016; Cook & Gor, 2015; Darcy et al., 2013; Gor, 2018; Melnik & Peperkamp, 2019). In the case of tones, this would mean that frequent errors occur in L2 tone word recognition because the representations that are being activated either lack tones or have low quality (uncertain) tone information. A second class of explanations attributes L2 difficulties to the influence of L1 *processing biases* (Chang, 2018; MacWhinney & Bates, 1989; Strange, 2011). In this case, the problem for L2 learners of tonal languages is that they focus perceptual attention (cf. Chang, 2018) on segmental cues to the exclusion of relevant tonal cues. This routine is successful in the L1, but in a tonal L2 leads to spurious activation of words with mismatching tones.

Using a lexical decision task with concurrent EEG, and an offline test of explicit lexical and tonal knowledge, we aim to see to what extent the representational and processing accounts can shed light on outcomes in advanced L2 learners of Mandarin.

Second language learning of Mandarin tones

In lexical tone languages pitch differentiates words from one another. For example, in Mandarin Chinese the syllable /ma/ spoken with a high pitch is ‘mom’, but with a low pitch is ‘horse’. For speakers of non-tonal languages the very idea that words could work this way can be hard to fathom. Perhaps for this reason, people often take for granted that learning L2 tones will be difficult, though—as we review below—research indicates that the difficulty of L2 tone learning is not absolute.

The primary acoustic cue for tone is fundamental frequency (F0) (Ho, 1976; Howie, 1974), the lowest frequency component of a sound wave, which humans perceive as its pitch. F0 is used in all languages in some form (intonation, stress), so it is not novel in and of itself. What sets tone languages apart is the functional use of F0 as a *lexical* cue.

Non-tonal language speakers need to learn at least two qualitatively novel things related to F0 (Figure 1). First, they must learn to treat F0 patterns as discrete *tone categories*. For example an L2 Mandarin speaker (or "learner") must learn that there are four tones: a high tone (Tone 1, or T1), a rising tone (T2), a low tone (T3), and a falling tone (T4). This implies learners must be able to hear the differences between tones (auditory perception). However, knowing these tone categories is not enough.

Tone must also be integrated as a necessary (abstract) feature in a word's phonological form. We will refer to this as learning *tone words*. Tone words are often illustrated with the syllable *ma* /mɑ/, which is a different word with each of the four tones (*ma1* 'mom'; *ma2* 'hemp'; *ma3* 'horse'; *ma4* 'scold'). While not every syllable in Mandarin occurs with all four of the tones every syllable of a word requires a tonal feature to be complete (though in some cases, the required feature is a lack of a tone, as in the case of the morpheme *me* 么). So then, to

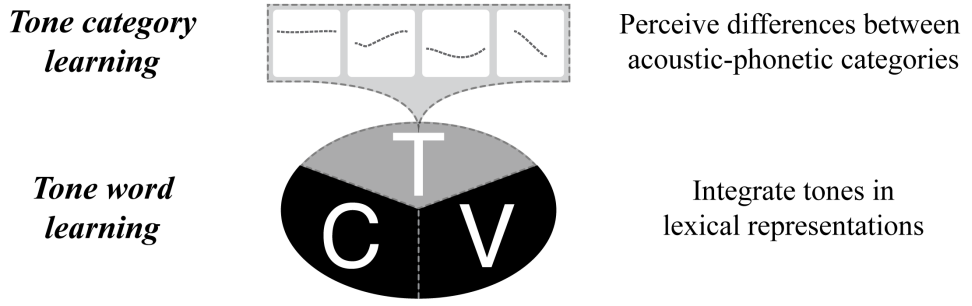

FIGURE 1. Two distinct challenges of second language lexical tone learning, learning tone categories, and learning tone words (C=consonant; V=vowel; T=tone)

successfully learn tone words, learners must be able not only to perceive tone categories, but to encode them (abstractly) in long-term memory for *lexical representations*, and to retrieve words during real-time *lexical processing* using tones.

Tone category and tone word learning in naïve or novice L2 learners

Previous research suggests that most people, given enough time and training, can learn to hear differences among tone categories and to identify them. As we might expect, people with no previous tone language experience make many errors identifying or discriminating tones (e.g., Alexander et al., 2005; Bent et al., 2006; Broselow et al., 1987; Gottfried, 2007; Y.-S. Lee et al., 1996; So & Best, 2010). But their errors are perhaps less surprising than their accuracy. Naïve participants generally perform well above chance, indicating they are not just guessing. For classrooms learners with only slightly more experience, accuracy in tone identification is often at or above 80% (e.g., C.-Y. Lee et al., 2009; Wang et al., 1999; Zhang, 2011; though accuracy may decline if more difficult tasks are used, e.g., Wiener et al., 2019). These patterns contrast with some truly difficult L2 speech sounds. For instance, Japanese speakers who have attained advanced proficiency in English may still perform at or below chance distinguishing /r/ and /l/ (Brown, 1998). Similarly, English speakers who have achieved advanced proficiency in Russian typically display pronounced difficulties discriminating certain hard/soft consonant distinctions (Chrabaszcz & Gor, 2014). At least compared to such cases, basic auditory perception of tones appears less challenging (cf. Antoniou & Wong, 2016 for a comparison when training Mandarin tones and Hindi stops, suggesting tones are an easier learning target).

A number of L2 training studies have combined tone category and tone word learning in a single training routine, pairing pictures with small sets of words that differ only by tones. Such

studies typically find clear improvements after training, but individual differences (musical expertise, pitch perception) can have a strong impact on outcomes (e.g., Bowles et al., 2016; Chandrasekaran et al., 2010; Dong et al., 2019; M. Li & DeKeyser, 2017; Perrachione et al., 2011; Sadakata & McQueen, 2014; Wong & Perrachione, 2007). For example, in Wong & Perrachione (2007) almost half of participants failed to reach even 60% accuracy in matching 18 tone words to pictures (6 sets of three-way tone contrasts), even after 10 or more training sessions.

Two training studies have demonstrated the separability of tone category and word learning, by first training tone category identification and then tone words. Ingvalson et al. (2013) found that participants with less aptitude showed improved outcomes by first engaging in tone category learning, and only then proceeding to tone word learning. Cooper & Wang (2013) found a similar pattern for non-musicians trained first with Cantonese tones categories and then tone words.

A potential limitation on the generalizability of the tone word training results reviewed above is that the studies have relied on very small sets of tone word stimuli. In order to make tones as salient as possible, each word contrasts with two or three others. This may prove to be optimal for tone training (though the long-term benefits are still unclear), but necessarily fails to capture the complexity of a real tone language lexicon, especially when words longer than a syllable are considered (for an example of tone word training with a larger number of stimuli that more realistically reflect the statistical properties of Mandarin, though only with single syllables, see Wiener et al., 2018; for a set of studies that includes disyllabic tone words, see Bowles et al., 2016; Chang & Bowles, 2015).

In the case of Mandarin, there are major differences in the qualities of monosyllabic and disyllabic words. One crucial difference is the likelihood of a word having tone neighbors—that is, words that share all phonological features except for a tone (e.g., *tang1* /tɑŋ/ 'soup' and *tang2* 'sugar') or tones (e.g., *you1yu4* /jou1y4/ 'melancholy' and *you2yu2* 'squid'). Tone neighbors are the norm for monosyllabic words, but much less common for disyllabic or multi-syllabic words (Table 1). Importantly, there are many more disyllabic than monosyllabic words, which means most words learners encounter do not have tone neighbors. At the same time, even though monosyllabic words do have tone neighbors, many of these words are among the earliest to be learned and are encountered with extreme (token) frequency (Tao, 2015), so that interlocutors can typically intuit intentions, even when words are mispronounced. This sets up a scenario where, early on, learners encounter few consequential tone neighbors and little pressure to avoid confusion. As their vocabularies grow, they will become familiar with more and more words with tone neighbors and will need to discuss topics that require less frequent words where tones may become more critical cues for listeners. This delay in experiencing the communicative value of tones may lead to a large backlog of L2 words with incorrect or missing tone features, which could have major impacts on whether and how L2 learners use tones in real time word recognition.

To understand how tone category and tone word learning may break down when thousands of words are known, we need to examine outcomes from 'training' with the full

TABLE 1. Word counts according to word length (syllables) in SUBTLEX-CH (Cai & Brysbaert, 2010) for most frequent (*logW*) 10,000 words.

	total words	number of tone neighbors
monosyllables	2021	4.13 (max 32)
disyllables	7118	.10 (max 5)
trisyllables	717	.01 (max 1)

complexity of a real language lexicon. This means we need studies with advanced L2 tone language learners.

Advanced L2 Chinese research

Research with advanced L2 learners of tone languages is still rather rare, but a handful of studies provide some indication of what typical long-term outcomes for tone category and tone word learning look like.

At the level of tone category learning, previous studies with advanced L2 Chinese leaners (English L1) have found that they can achieve near-native performance on identification of tones on isolated monosyllables (C.-Y. Lee et al., 2009; Pelzl et al., 2019; Zhang, 2011; for related work with Dutch L1 speakers, see Zou et al., 2016). In Pelzl et al. (2019), we found that even when tone identification was challenging (syllables clipped from continuous speech), advanced L2 participants performed nearly identically to native Mandarin participants, with a clear difference appearing only for T2. Along the same lines, Shen and Froud (2016) found that behavioral identification and discrimination performance (using tone continua) for advanced L2 learners was near-native. Interestingly, when Shen and Froud (2018) tested the same participants using ERPs, they found their MMN and P300 responses during passive listening were distinct from native patterns (for similar results in L2 learners with varied non-tonal L1s, see Yu et al., 2019). These disjunctive results suggest that advanced L2 learners can develop tone categories, but that the categories are in some way distinct from those of native Mandarin speakers.

Given advanced learners' ability to achieve high performance at tone identification, a key question that arises is whether this perceptual capacity will translate into high performance in online tone word recognition. Our previous study (Pelzl et al., 2019) also examined this question.

The same advanced L2 participants that performed at near-native levels on tone identification completed a lexical decision task with disyllabic words, and nonwords that mismatched real words by a tonal or segmental contrast. Tonal nonwords differed from real words only with respect to the tone of the first syllable (e.g., nonword *fang4zi* /fɑŋ4tsɻ/ derived from real word *fang2zi* 'house'). Segmental nonwords differed from real words with respect to the rhyme of the first syllable (e.g., nonword *feng2zi* /fəŋ2tsɻ/ derived from real word *fang2zi*). As in the tone identification, the stimuli were clipped from continuous speech in sentences. Compared to native speakers, L2 learners performed significantly less accurately on both types of nonword, but the difference in accuracy between the segmental and tonal conditions was particularly striking. For segmental nonwords, mean L2 accuracy was 84% (compared to 96% for L1), while for tonal nonwords it was 35% (L1: 91%). Performance did not appear to be due to lack of word knowledge, as most L2 participants knew upwards of 95% of the critical vocabulary, and (with just one exception) participants failed to reach native-speaker levels for rejection of tonal nonwords even when they performed near ceiling on an offline test of tone knowledge for the critical vocabulary. Summarizing the results, our previous study suggested that tone identification data indicated L2 learners can achieve strong auditory perception of tone categories, but that lexical decision data suggested they have persistent difficulty representing and/or processing tone words.

The present study

The current study was designed to narrow in on the causes that drive difficulty in learning tone words, focusing on the issues of lexical representation and processing raised previously (Pelzl et al., 2019), and highlighted in our introduction. We investigate a 'best-case scenario' for

advanced L2 tone word processing by testing performance in nearly ideal listening conditions—with words spoken clearly and in isolation. Under such conditions, do learners still have difficulty in lexical decision for tone words? If so, is it driven by the quality of lexical representations or by L2 processing routines?

One possible explanation for the low L2 accuracy in rejecting tone nonwords observed by Pelzl et al. (2019) was that the challenging stimuli—requiring listeners to process multiple syllables at a naturalistic pace—induced a processing bottleneck, so that listeners did not have enough time to utilize the routines they used so successfully in tone identification (cf. phonetic and phonological modes in Strange, 2011). If this were the case, performance might recover if words were pronounced more slowly. To address this possibility, the present study will test whether differences between tonal and segmental nonwords persist with more slowly and clearly pronounced stimuli. This will answer the first research question: (1) *Are L2 listeners equally accurate in rejection of isolated disyllabic nonwords that differ from real words only with respect to either a vowel or a tone?*

A second possible explanation for low L2 accuracy in rejecting tone nonwords is that it might have been due to a lack of certainty about the phonological form of relevant real words on the part of learners. Cook and Gor (cf. Cook & Gor, 2015; Gor, 2018; Gor & Cook, 2018) have posited that L2 learners' subjective familiarity with words can provide an explanation for why they might be more permissive in accepting phonologically similar words compared with L1 listeners. In this case, the hypothesis is that less familiar words have lower quality phonological representations and are more likely to be incorrectly accepted, while more familiar words have higher quality representations and are more likely to be correctly rejected. While we did measure offline knowledge of words and tones in Pelzl et al. (2019), we did not attempt to measure

confidence for the meanings or tones of the associated words. By measuring subjective confidence, the current study will account more thoroughly for the role of L2 familiarity in lexical decision outcomes, answering the second research question: (2) *Does lexical familiarity impact L2 behavioral responses to tone nonwords?*

Finally, the current study will take advantage of event-related potentials (ERPs) as a measure of continuous online responses to gain fuller insight into the L2 tone word recognition process. The behavioral outcome in a lexical decision task only reflects the final decision point for each trial, leaving the process leading up to that decision unexamined. In this sense, the difference we found in the lexical decision task in Pelzl et al. (2019) was only quantitative, not qualitative.[1] It is possible that, despite lower accuracy overall, L2 learners nevertheless display qualitatively equivalent responses to both vowel and tone word mismatches.

To address this possibility, the current study will use ERPs to assess the word recognition process as it unfolds during each trial. ERPs are particularly valuable because they can capture qualitative aspects of word recognition processes, namely whether responses occur within the same time window, and whether the magnitude of responses in different conditions is comparable.

The present study will focus on the N400, which is particularly useful in examination of lexical recognition processes. The N400 is a negative-going ERP response that peaks approximately 400 ms after stimulus onset and can be used as an index of the ease or difficulty a listener has in accessing lexical targets (Kutas & Federmeier, 2000; Kutas & Hillyard, 1980, 1984; Lau et al., 2008). Several previous studies have found the N400 in native Chinese speakers to be sensitive to lexical tone mismatches in contextually expected words (*in sentences*: Brown-Schmidt & Canseco-Gonzalez, 2004; Li, Yang, & Hagoort, 2008; Pelzl et al., 2019; Schirmer,

Tang, Penney, Gunter, & Chen, 2005; *with picture cues*: Malins & Joanisse, 2012; J. Zhao, Guo, Zhou, & Shu, 2011). However, no previous research has investigated advanced L2 neural sensitivity to tone mismatches in isolated disyllabic words. By examining L2 ERPs to nonwords, we will have a continuous measure of L2 tone processing, allowing us to answer a third research question: (3) *Are L2 listeners equally sensitive to vowel and tone mismatches (as indexed by the N400)?* Importantly, we will only be examining trials with *correct* rejections of nonwords. For correct rejections, the N400 amplitude should be more negative than that of real words ('the N400 effect'), indicating the difficulty the listener has accessing a word. If we see similar N400 effects for tone and vowel nonwords, this will indicate that the same process attains for both. If we find smaller N400 responses for tones, this will indicate that *even when nonwords are correctly rejected*, L2 sensitivity to tones is diminished. This might occur if, for example, L2 listeners rely on slow, explicit judgments to arrive at correct rejections, rather than on the faster and more automatic processes indexed by the N400.

Participants

We recruited 19 native English speakers who had achieved relatively advanced proficiency in spoken Mandarin Chinese. One participant was excluded due to early onset of learning (age 7) and possible tone language exposure in the family home. This left 18 advanced L2 participants. Table 2 summarizes their general learning characteristics, as well as scores on the screening measures, and results on a tone identification task (for details, see supplementary materials). This study used the same screening measures (vocabulary, Can-do self-assessment) and criteria as in Pelzl et al. (2019), in order to maintain at least a lower bound of comparability with the population tested in that study (one L2 participant scored a bit lower (65.7) than

TABLE 2. Background information, screening measures, and tone identification scores for L2 participants (*n*=18)

	mean (sd)	range
Age at testing	25.7 (4.8)	18-38
Age of onset	17.5 (3.9)	11-25
Semesters of formal study	8.9 (4.9)	3-20
Years in immersion	3.4 (2.6)	0.7-9
Total years learning	8.2 (3.7)	3-19
Can-do self-assessment (%)	82.9 (7.5)	72.8-96.8
Vocabulary self-assessment (%)	88.2 (9.2)	65.7-100
Tone identification accuracy (%)	85.3 (7.8)	71.8-99.2

criterion (70) on the vocabulary test, but was accepted nonetheless as advanced L2 participants were difficult to find). Twenty-four native Chinese speakers also completed the experiment (average age = 26.1). Four were excluded due to excessive EEG artifacts, leaving twenty for all analyses presented below.

All participants gave informed consent and were compensated for their time.

Stimuli design and production

We selected 96 disyllabic real words (e.g., *fa1yin1* /fɑ1in1/ 'method') to be used in an auditory lexical decision task. All were high frequency nouns. On the basis of the real words, two types of nonwords were created, differing from real words only with respect to a tone or vowel (Figure 2). For the tone mismatch condition, the tone of the first syllable was changed producing a nonword (e.g., *fa**2**yin1*). We will refer to these items as *tone nonwords*. For the vowel mismatch condition, the vowel (and only the vowel) on the first syllable was changed producing a nonword (e.g. *fu1yin1* /f**u**1in1/), i.e., *vowel nonwords*. (Additional details of procedures for selection and quality control of stimuli, the approach applied for T3 sandhi, as well as the complete list of all stimuli, can be found in supplementary materials online.)

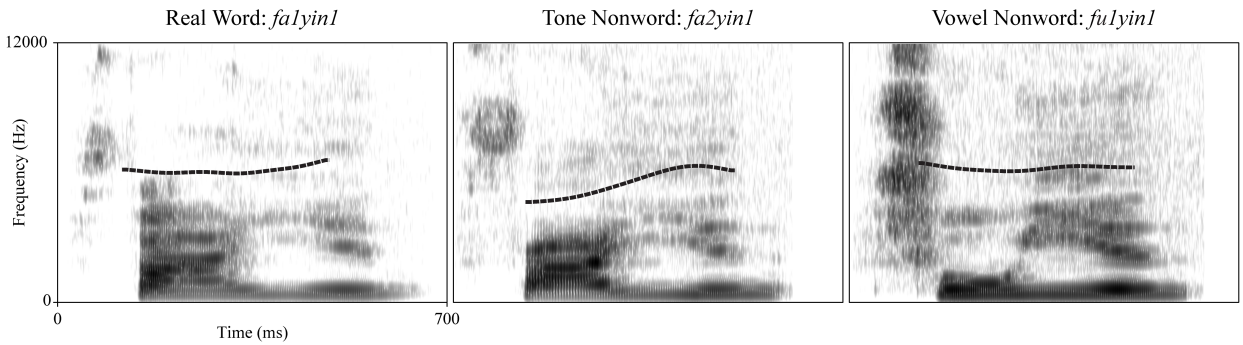

FIGURE 2. Illustration of pitch contours and spectrograms for the Real Word *fa1yin1* 'pronunciation', Tone Nonword *fa2yin1*, and Vowel Nonword *fu1yin1*

These stimuli improve on those of Pelzl et al. (2019) in several ways. First, all tones are balanced across real words, and tone changes are balanced across tone nonwords—that is, T1 becomes T2, T3, and T4 an equal number of times, and similarly for other first syllable tones. Second, whereas Pelzl et al. (2019) swapped out entire syllable rhymes, including syllable final /n/ and /ŋ/ (e.g., *xiang3fa3* /ɕiɑŋ3fɑ3/ 'thought' became the nonword *xu3fa3* /ɕy3fɑ3/), the current stimuli limited changes to vowels.[2] Third, in order to prevent listeners from rejecting nonwords before the onset of the second syllable, we avoided creating syllables that never occur in Mandarin (e.g., *fai* /fɑi/) or are very rare (e.g., *cen* /tsʰən/).

As noted above, the current study aims to explore advanced L2 tone perception in a best-case scenario. To this end, we recorded a native Chinese speaker (female) from northern China who spoke with a standard Mandarin accent. She produced the stimuli in isolation, speaking a with a clear voice, at a comfortable, but relaxed, speech rate. Using *Praat*, all stimuli were cut out of the original audio files to create individual .wav files. The average intensity of each file was scaled to 70 dB, and 200 ms of silence were appended at the end of each file. This resulted in 96 triplets consisting of a real word and its vowel and tone nonword counterparts. Stimuli were divided into three lists, each containing 32 real words, 32 vowel nonwords, and 32 tone nonwords. Additionally, the 32 disyllabic real word filler trials were included in each list to

balance the proportion of correct 'yes' answers across the experiment. Importantly, no item was repeated in both its real and nonword forms for the same participant, as such repetition might lead to undesirable strategizing.

Vocabulary Test

We also constructed an offline vocabulary test. The format is illustrated in Figure 3. For each L2 participant, the test included all real word counterparts for vowel and tone nonwords encountered during the lexical decision task (64 words). Each item provided Chinese characters and toneless Pinyin. Participants supplied tones (numbers 1-4 for each syllable), an English definition, and a confidence rating from 0-3 for both the tones and the definition of each item. Participants were informed that the 0-3 scale had the following meaning: *0 = I don't recognize this word; 1 = I recognize this word, but am very uncertain of the tones/meaning; 2 = I recognize this word, but am a bit uncertain of the tones/meaning*; *3 = I recognize this word, and am certain of the tones/meaning*. This scale remained visible as a reference throughout the test. For any tones or definitions they did not know, participants were told to leave the answer blank and supply "0" for confidence.

Procedures

CHINESE	PINYIN	TONES	CONFIDENCE RATING (0-3)	DEFINITION	CONFIDENCE RATING (0-3)
律师	lüshi	40	2	lawyer	3
办法	banfa	43	3	method	3
牛排	niupai	23	1	beef ribs	2

FIGURE 3. Format of items for the offline vocabulary knowledge test

We used an auditory lexical decision task. Participants heard a single disyllabic Mandarin word or nonword and decided whether it was a real word or not. EEG was recorded along with the behavioral response for each trial. After the experiment, L2 participants completed an offline vocabulary knowledge test of the real word counterparts of all nonwords they heard in the lexical decision task.

Thirty-six participants (24 L1 and 12 L2) were tested in the lab at Beijing Normal University (BNU). Seven additional L2 participants were tested under conditions as similar as possible in the lab at the University of Maryland (UMD). Each participant was seated in front of a computer monitor and fit with an EEG cap. Auditory stimuli were presented using a single high quality audio monitor (JBL LSR305) placed centrally above the computer monitor.

For the lexical decision task, instructions presented onscreen included an illustrative example of each type of nonword: "*zhong1guo2* is a real word, but *zhang1guo2* and *zhong4guo2* are not real words in Mandarin." Instructions were presented in English for L2 participants, and in Chinese for L1 participants. Instructions were followed by ten practice items with stimuli not included in the experiment. Participants then completed 128 lexical decision trials. Trials were divided into seven blocks (roughly 20 in each) with self-paced breaks between each block. Stimuli were counterbalanced across three lists, and each list was given four unique pseudo-random orders so that stimuli of a single condition type was never repeated more than three times in a row, and strings of expected yes/no answers never extended beyond three items in a row. Timing parameters are shown in Figure 4.

After the ERP experiment was finished, L2 participants completed the offline vocabulary test.

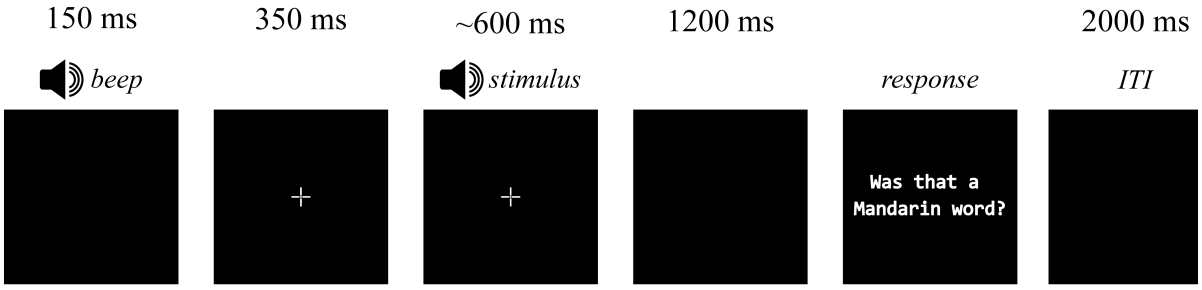

FIGURE 4. Trial structure and timing parameters of the ERP experiment

EEG recording

Raw EEG was recorded continuously at a sampling rate of 1000 Hz using a Neuroscan SynAmps data acquisition system and an electrode cap (BNU: Quik-CapEEG; UMD: Electrocap International) mounted with 29 AgCl electrodes at the following sites: *midline*: Fz, FCz, Cz, CPz, Pz, Oz; *lateral*: FP1, F3/4, F7/8 FC3/4, FT7/8, C3/4, T7/8, CP3/4, TP7/8, P4/5, P7/8, and O1/2 (UMD*:* had FP2, but *no* Oz). Recordings were referenced online to the right mastoid and re-referenced offline to averaged left and right mastoids. The electro-oculogram (EOG) was recorded at four electrode sites: vertical EOG was recorded from electrodes placed above and below the left eye; horizontal EOG was recorded from electrodes situated at the outer canthus of each eye. Electrode impedances were kept below 5kΩ. The EEG and EOG recordings were amplified and digitized online at 1 kHz with a bandpass filter of 0.1-100 Hz.

EEG data processing

All trials were visually inspected and evaluated individually for artifacts using EEGLAB v10.2.5.8b (Delorme & Makeig, 2004) and ERPLAB v3.0.2.1 (Lopez-Calderon & Luck, 2014) running under MATLAB R2013b (MathWorks, 2013). Data from four L1 participants were excluded due to having more than 40% artifacts on experimental trials. After excluding these

participants, artifact rejection affected 8.45% of experimental trials (L1 8.08%; L2 8.86%). Trial-level data for each subject baselined to the mean of the 100 ms preceding the onset of the auditory stimulus was exported for further processing in *R* (R Core Team, 2019). A single average amplitude was obtained for each trial for each electrode for each subject in a slightly delayed auditory N400 window (400-900 ms). This window was chosen on the basis of two criteria. First, the average duration of stimuli was approximately 600 ms. Listeners could only notice a nonword sometime *after the onset of the second syllable*, suggesting any time earlier than 300 ms would be inappropriate. Second visual inspection of grand average waveforms across all scalp electrodes suggested 900 ms was a reasonable endpoint to capture N400 effects, and is sufficiently generous so that it does not underestimate potentially slower L2 responses.

Data from fifteen central electrodes (F3, Fz, F4, FC3, FCz, FC4, C3, Cz, C4, CP3, CPz, CP4, P3, Pz, P4) were chosen for final analysis as visual inspection of grand average waveforms suggested these electrodes had strong and consistent N400 peaks across conditions, and we had no theoretical motivation for positing that ERP responses would vary across regions. To reduce some mild non-normality in the data, any trial with an absolute value greater than 50μV was removed prior to final data analysis. Finally, only trials that elicited correct behavioral responses (correct acceptance or correct rejection) were retained for final analysis. After all of these steps, the final EEG dataset contained 43,567 data points (80.0% out a of total possible 54,720 data points: L1=88.1%; L2=70.2%). We note that the loss of data disproportionately affects L2 data, which reduces power for finding effects (an alternative analysis retaining all trials is included in online supplementary materials, substantive results are the same as those reported here).

Behavioral lexical decision task results and statistical analysis

Reliability for the lexical decision task data was high for all three lists (list A: α=.94; list B: α=.93.; list C: α =.93). Descriptive results are shown in Table 3. The L1 group displayed high accuracy across all conditions, while the L2 group had noticeably lower accuracy overall, with tone nonwords registering the lowest. D-prime (*d′*) was also calculated for each participant, contrasting vowel nonwords and real words, and tone nonwords and real words, using Laplace smoothing to correct for infinite values (Barrios et al., 2016; Jurafsky & Martin, 2009). As with accuracy, *d′* results suggest overall higher sensitivity to nonwords for L1 listeners with little difference between nonword conditions (vowel *d′*=3.78, *sd*=.46; tone *d′*=3.81, *sd*=.46). In contrast, L2 has less sensitivity overall and a larger difference between conditions that suggests vowel nonwords are detected more readily than tone nonwords (vowel *d′*= 2.31, *sd*=.55; tone *d′*=1.59, *sd*=.78). When considered individually (Figure 5), all but one L2 participant had a lower *d′* for tone than vowel nonwords. All but three scored below the lowest L1 *d′* for tone nonwords, while for vowel nonwords eight learners were in the range of l1 scores. Only one L2 participant performed near the level of the average L1 scores overall.

All statistical analyses reported below were conducted in *R* (version 3.6.1, R Core Team, 2019). Mixed-effects models were fit using the *lme4* package (version 1.1.21, Bates, Mächler, Bolker, & Walker, 2015). Effects coding was applied using the *mixed* function in *afex* (Singmann et al., 2017).

TABLE 3. Descriptive accuracy results for the Lexical Decision Task

group	cond	mean acc % (sd)
L1 (*n*=20)	real	98.1 (13.6)
	vowel	95.2 (21.5)
	tone	95.3 (21.2)
L2 (*n*=18)	real	85.6 (35.2)
	vowel	84.9 (35.8)
	tone	61.5 (48.7)

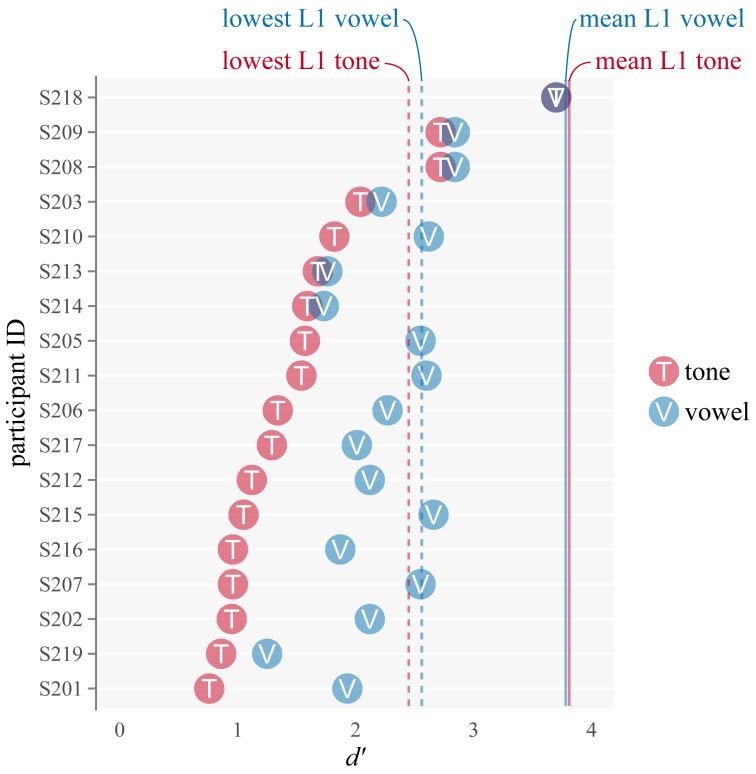

FIGURE 5. L2 participants d-prime scores for vowel and tone nonwords. Lowest individual L1 scores are indicated by dashed lines. Mean L1 scores are indicated with solid lines.

Accuracy results were submitted to a generalized linear mixed-effects model (using the *bobyqa* optimizer) with crossed random effects for subjects and items. The dependent variable was accuracy (1, 0). Fixed effects included the factors *condition* (real word, tone nonword, vowel nonword), and *group* (L1, L2), and their interaction. The maximal random effects model was fit first (Barr et al., 2013; Bates, Kliegl, et al., 2015). Model convergence difficulties were addressed by suppressing correlations in random effects (using "expand_re = TRUE" in the *mixed* function). The best fitting model was determined by model comparison conducted through likelihood ratio tests, building from the maximal model (which was rejected due to convergence issues) to progressively less complex models. Inclusion of the nuisance factor *list* (with subjects nested under lists) did not improve model fit, and so was not retained in the final model. The

final model included by-subject random intercepts and slopes for the effect of condition, and by-item random intercepts and slopes for condition and group, but not their interaction *(glmer* model formula*:* accuracy ~ condition * group + (condition || subject) + (condition + group || item)). Results are depicted graphically in Figure 6.

Table 4 reports main effects and interactions. P-values were obtained using the likelihood ratio test ("LRT") method. The effects of condition and group were both statistically significant. There was also a significant interaction between condition and group.

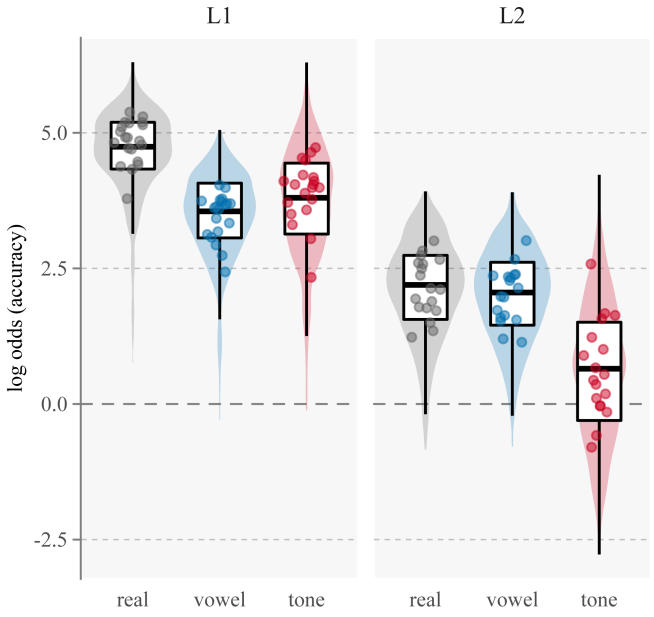

FIGURE 6. Boxplots of model estimated log odds of an accurate response for Lexical Decision Task. Shaded areas indicate the distribution of responses. Each circle indicates an individual participant's model estimated mean score. The dashed line indicates chance performance.

TABLE 4. Mixed Model ANOVA Table for accuracy results (Type 3 tests, LRT-method)

Effect	Df	Chisq.	Chi Df	Pr(>Chisq)	
condition	11	25.59	2	<.001	***
group	12	53.84	1	<.001	***
condition × group	11	8.43	2	.015	*

Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1

TABLE 5. Planned comparisons for accuracy results in the lexical decision task (p-values adjusted by Holm method)

						95% CI	
Comparison	Estimate	SE	z value	Pr(>|z|)		*lower*	*upper*
L1 Vowel vs. Tone	0.26	0.44	0.58	.582		-0.78	1.29
L2 Vowel vs. Tone	1.78	0.37	4.87	<.001	***	0.93	2.64
L1 V-T vs. L2 V-T	-1.52	0.50	-3.03	.005	**	-2.70	-0.35

Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1

Critical planned comparisons are reported in Table 5. The Holm method was used to correct for multiple comparisons. Though we are primarily interested in testing accuracy in correct rejection of vowel and tone nonwords in L2, implicit in this comparison is that there is a difference between the differences in accuracy for vowel and tone nonwords for L1 and L2. This is borne out in our comparisons. There was no significant difference in L1 accuracy of correct rejections for vowel and tone nonwords, whereas for the L2 group accuracy for correct rejection of nonwords differed significantly for vowels and tones. L2 listeners were about two and a half times more likely to incorrectly accept tone nonwords than vowel nonwords (38.5*/15.1=2.55*). Finally, the difference between L2 vowel and tone was significantly larger than the difference between L1 vowel and tone.

ERP results and statistical analysis of correct trials only

N400 average amplitudes for trials that received a correct response in the lexical decision task are shown in Table 6 and depicted visually as grand average waveforms in Figure 7. Across all midline and central electrodes, L1 displays strong N400 effects to both vowel and tone nonwords. In contrast L2 shows attenuated N400 effects overall, and visually different magnitudes of N400 for vowel and tone nonwords, with tone nonword responses diverging less strongly from real word responses.

TABLE 6. Mean amplitude (μV) of ERP responses (correct trials only)

Group	Condition	meanAmp	SE
L1	real	-1.80	.098
L1	vowel	-4.86	.101
L1	tone	-5.17	.098
L2	real	0.06	.107
L2	vowel	-1.75	.107
L2	tone	-1.22	.118

Averaged N400 amplitudes from the 400-900ms window were submitted to a linear mixed-effects model with crossed random effects for subjects and items. Models included fixed effects for *condition* (real word, vowel nonword, tone nonword) and *group* (L1, L2) and their interactions. Convergence difficulties were addressed by specifying uncorrelated random effects. Effects coding was used, and p-values were obtained using Satterthwaite's method. The maximal

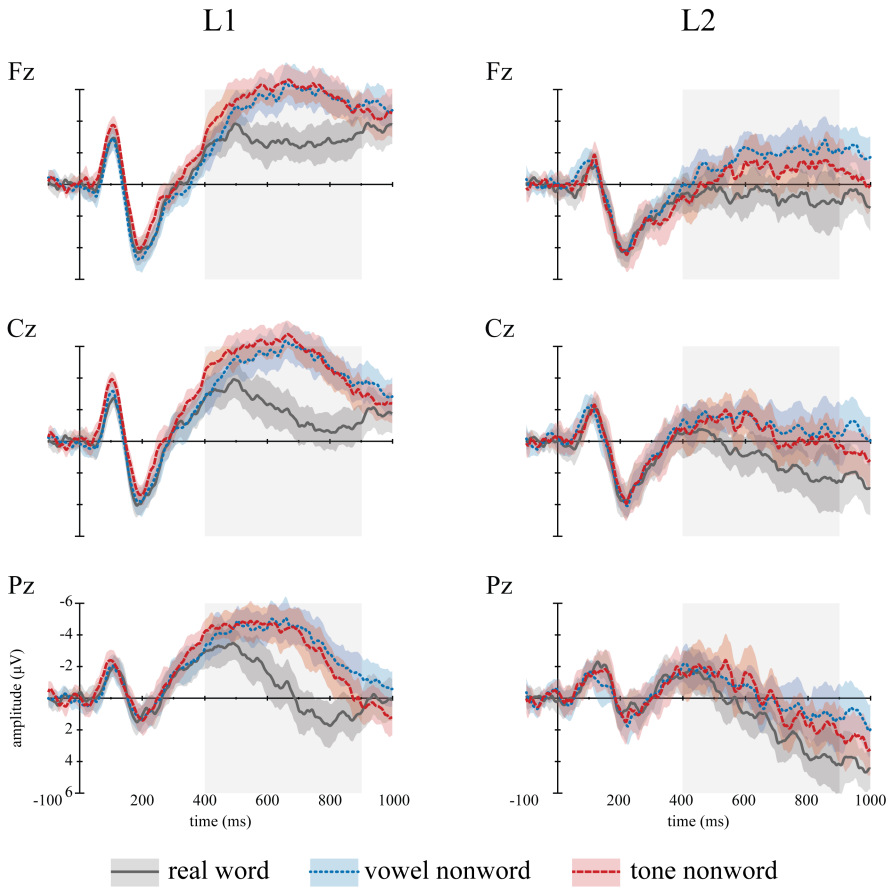

FIGURE 7. Grand average waveforms for lexical decision task (40 Hz low pass filter), only correct trials are included. Shaded areas around lines represent 95% within-subjects confidence intervals.

TABLE 7. Mixed Model ANOVA Table for N400 results (Type 3 tests, LRT-method)

Effect	numer Df	denom Df	F	Pr(>F)	
condition	2	112.91	18.10	<.001	***
group	1	44.23	12.95	<.001	***
condition × group	2	105.11	3.03	.053	.

Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1

model that successfully converged was fit first and was then compared to less complex models to test random effects. The final model included random intercepts for subjects and items, and by-item random slopes for the effect of group (*lmer* model formula: amplitude ~ condition * group + (1 | subject / electrode) + (1 + group || item)).

Model results are reported in Table 7. The main effects of group, and condition, and their interaction were statistically significant.

Planned comparisons are reported in Table 8. Model estimates in planned comparison can be interpreted as amplitude differences (in μV). For L1 listeners, real words evoked significantly more positive amplitudes than either vowel nonwords or tone nonwords, while there was no statistically significant difference between vowel and tone nonword responses. For L2 listeners, real words evoked a significantly more positive response than vowel nonwords, while there was no significant difference between tone nonwords and either real words or vowel nonwords. Finally, the difference of differences between L1 and L2 tone and vowel nonwords was not

TABLE 8. Planned comparisons for ERP results of lexical decision task (p-values adjusted by Holm method)

							95% CI	
Group	Comparison	Estimate	SE	z value	Pr(>|z|)		*lower*	*upper*
L1	Real vs. Vowel	2.86	0.51	22.79	<.001	***	1.53	4.18
	Real vs. Tone	3.26	0.81	25.16	<.001	***	1.15	5.38
	Vowel vs. Tone	0.41	0.75	2.17	1.000		-1.56	2.37
L2	Real vs. Vowel	1.48	0.52	11.68	.020	*	0.13	2.83
	Real vs. Tone	0.72	0.83	6.34	1.000		-1.46	2.90
	Vowel vs. Tone	0.76	0.78	4.35	1.000		-1.27	2.79
L1 vs. L2	Vowel vs. Tone	1.17	1.05	4.74	1.000		-1.58	3.91

statistically significant. Visual depiction of model estimated results are shown in boxplots in Figure 8.

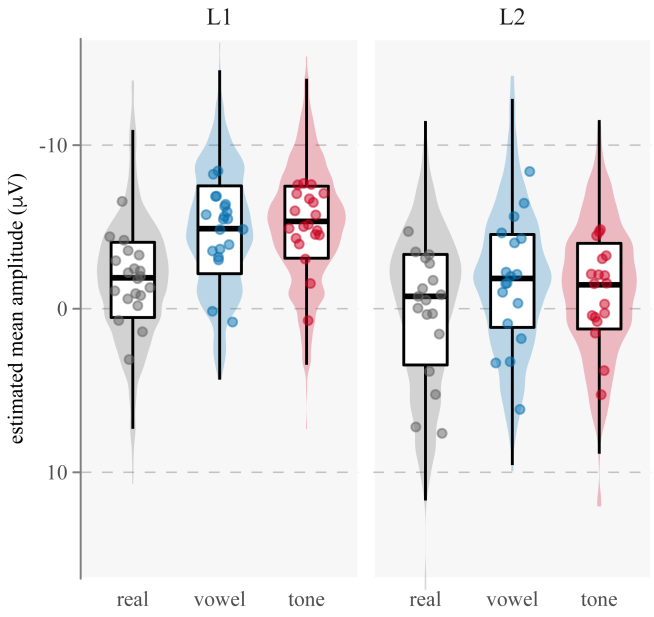

FIGURE 8. Boxplots of model estimated N400 amplitudes (400-900 ms window) for correct trials in the lexical decision task. Shaded areas indicate the distribution of responses. Each circle indicates a single participant's model estimated mean amplitude.

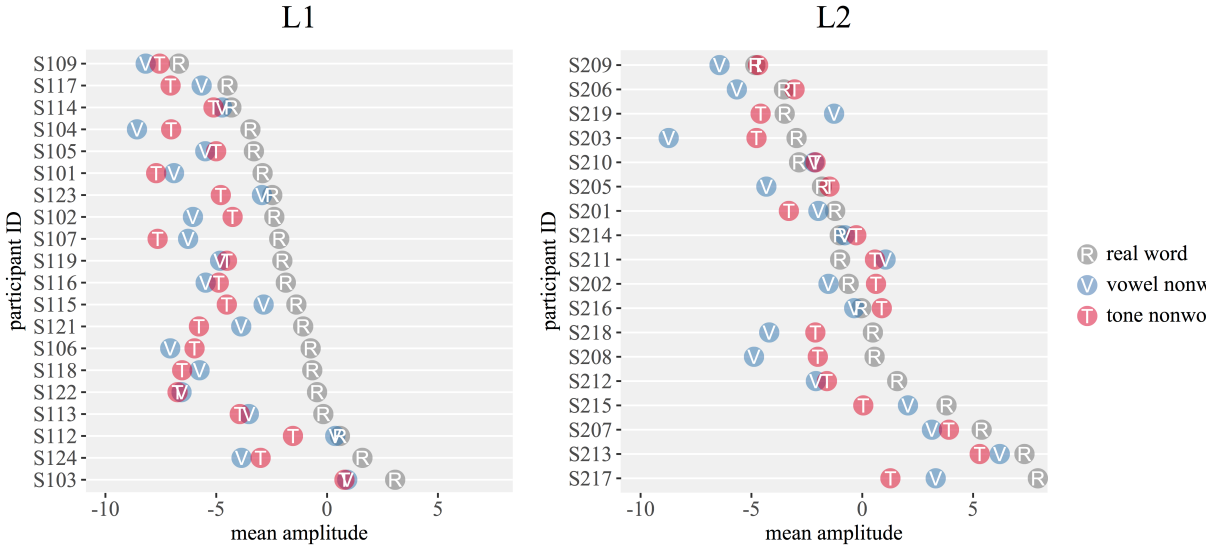

FIGURE 9. Individual participant's mean amplitude for each condition (correct trials only). Participant IDs are on the y-axis, amplitude is on the x-axis. L1 results are displayed on the left, L2 on the right. Each circle represents the mean for one condition. Participants are ordered according to mean amplitude in the real word condition.

In order to capture patterns at the individual level, we plotted each participant's mean amplitude for the three conditions (Figure 9). Although L1 participants varied as to whether tone or vowel nonwords elicited stronger negativity, all L1 participants display greater negativity for nonwords than real words. In contrast, L2 participants display much less consistency. While some participants display clear nonword responses, many participants' N400 effects are small or non-existent. Ten L2 participants' tone N400s are smaller than their vowel N400s, though five individuals show the opposite pattern, and three display no nonword N400 effects at all.

In summary, for trials with correct responses, the L1 group displayed significant and strong N400 effects for both vowel and tone nonwords, and this was consistent across all L1 participants. The L2 group displayed significant N400 effects only for vowel nonwords, with weaker N400 effects for tone nonwords, intermediate between vowel nonwords and real words. This was reflected at the level of individuals by inconsistent N400 effects, with tone nonwords overall less likely to elicit N400s than vowel nonwords.

Offline vocabulary test data processing

The offline vocabulary data are used to consider how familiarity with words and tones impacts lexical decisions, and to evaluate the general quality of L2 tone word knowledge. The test produced four data points for each nonword that an L2 participant encountered: an accuracy score for the tones and definition they supplied, and a confidence rating for each. For example, if the word was *fa1yin1*, and the participant provided 11 as the answer for tones, this would be scored as 1, while any other set of two numbers would result in a score of 0 for the tone on that item. Note that this scoring method counted tones on both syllables, whereas the nonwords only ever mismatched real words with respect to tones on the first syllable. In that sense, this scoring

approach is strict. Definitions were also scored 1 for correct, or 0 for incorrect. For both of these scores, there was also an accompanying confidence rating, ranging from 0 to 3. One participant's vocabulary test data was lost due to a coding error.

Overall, L2 learners supplied correct tones for about 74% of the items (807 out of 1088 total responses), and correct definitions for about 91% of the items overall (990 out 1088 total responses).

Items given a confidence score of 0 for either tones or vowels were discarded before further analyses (a total of 40 trials), and four trials were missing data (i.e., unanswered). This left a total of 1044 items (90.6% of all L2 nonword trials) that had data for all four cells (i.e., tone and definition accuracy, and tone and definition confidence ratings).

Offline vocabulary test results

Table 9 presents vocabulary results for tone responses. This data can give us insight into the quality of L2 tone representations for known words, as well as its relation to performance in the lexical decision task. Results are listed according confidence ratings. For example, for real word counterparts to vowel nonwords, participants assigned a rating of 3 'high' to the tones they supplied for 377 items. Table 9 also lists the accuracy of the supplied tones, and the accuracy of lexical decisions for those items. Even for high confidence items (*"I recognize this word, and am certain of the tones")*, tone answers were inaccurate more than 10% of the time, and lexical decision accuracy was lower for tone nonwords than vowel nonwords. For mid and low confidence items, tone and lexical decision accuracy fell even further.

TABLE 9. Results of L2 offline vocabulary test requiring participants to supply tones and tone confidence ratings for real word counterparts of critical nonwords. Tone accuracy indicates whether supplied tones were correct. Lexical decision task (LDT) accuracy indicates whether the related nonwords were correctly rejected in the lexical decision task.

Tone confidence ratings and accuracy of L2 supplied tones				
Condition	conf. rating	k (items)	tone acc. %	LDT acc. %
Vowel nonword counterparts	3 (high)	377	87	84
	2 (mid)	132	56	86
	1 (low)	16	62	88
Tone nonword counterparts	3 (high)	385	85	66
	2 (mid)	130	52	52
	1 (low)	4	25	0

Table 10 provides parallel results for vocabulary definitions, allowing us to separately evaluate the quality of lexical-semantic knowledge. In contrast to sometimes questionable confidence in tone knowledge, L2 participants' confidence about their knowledge of definitions seems quite accurate—high confidence items were correctly defined 98% of the time. In other words, they know which words they know, and which they do not. There is not a clear relationship between this knowledge and performance on the lexical decision task, which follows insofar as the lexical decision task only tested word form recognition, not semantic knowledge.

In sum, results of the vocabulary knowledge test suggest L2 participants have substantial difficulty encoding tones in lexical representations. Even when explicit knowledge is fully available and words are confidently recognized, L2 tone knowledge was still inaccurate over

TABLE 10. Results of L2 offline vocabulary test requiring participants to supply definitions and definition confidence ratings for real word counterparts of critical nonwords. Def. accuracy indicates whether supplied definitions were correct. Lexical decision task (LDT) accuracy indicates whether the related nonwords were correctly rejected in the lexical decision task.

Definition confidence ratings and accuracy of L2 supplied definitions				
Condition	conf. rating	k (items)	def. acc. %	LDT acc. %
Vowel nonword counterparts	3 (high)	462	98	85
	2 (mid)	49	65	76
	1 (low)	8	62	94
Tone nonword counterparts	3 (high)	458	98	63
	2 (mid)	50	80	51
	1 (low)	17	59	62

10% of the time. (For complete by-item vocabulary test results, see the online supplementary materials.)

Does lexical familiarity impact L2 behavioral responses? ("the best-case scenario")

Next, we used the offline vocabulary results to evaluate the extent to which lexical decision errors reflect deficits of offline vocabulary and tone knowledge. To this end, we reanalyzed lexical decision results for the subset of trials characterized by accurate and confident L2 knowledge for both tones and meanings (i.e., 3s for all four response categories on the vocabulary test). This comprised 301 tone nonword and 303 vowel nonword trials (604 total, 55% of total nonword trial data). By testing this data, we get a 'best-case scenario' for L2 participants: *When lexical knowledge is highly accurate and confident, do L2 learners reject vowel and tone nonwords with equal accuracy?*

Table 11 presents descriptive accuracy results for the two nonword conditions in the best-case scenario data for the lexical decision task. The accuracy results were submitted to a generalized linear mixed-effects model following the same procedures as outlined for previous analyses. The model included the fixed effect of nonword condition. The maximal model was fit, and included random intercepts for subjects and items, and random slopes for the by-subject and by-item effects of condition (*lmer* model formula*:* accuracy ~ condition + (condition || subject) + (condition || item).

Table 11. Descriptive accuracy results for the 'best-case scenario' analysis of the lexical decision task

group	cond	mean acc. % (sd)
L2 (n=17)	vowel	85 (35)
	tone	67 (47)

Table 12. Comparison of conditions for accuracy results in the 'best-case scenario' analysis of the lexical decision task (Type 3 tests, LRT-method)

						95% CI	
Comparison	b	SE	z	p		*lower*	*upper*
Tone vs vowel nonword	-1.31	0.36	-3.64	<.001	***	-2.02	-0.61

Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1

Results are displayed in Table 12. There was a statistically significant difference in accuracy for vowel and tone nonwords. So then, even in the best-case scenario—with near perfect word and tone knowledge—L2 participants still display a more limited ability to reject tone nonwords than vowel nonwords.

Does lexical familiarity impact ERP responses?

Due to limited power, statistical modeling of the best-case scenario for ERP data was not possible. However, as the ERP analysis was conducted on only those trials that resulted in correct decisions, it is possible to consider the quality of offline knowledge associated with those decisions in order to examine whether insufficient explicit knowledge of tones contributed to ERP differences. That is, even though L2 participants ultimately made the correct decision on these trials, they still may have been guessing or using other strategies (e.g., they might know that a specific tone is *not* correct, even though they do not know explicitly what the correct tone actually is).

For these trials, L2 knowledge of definitions for the real word counterparts of nonwords was very accurate (vowel nonwords: *mean*=97%; tone nonwords: *mean*=96%). L2 knowledge of tones, however, was not nearly so high (tone nonwords 80%), and varied rather extremely across participants, with the lowest mean average being 31%, and the highest 100%. The extreme low score was somewhat atypical of the group overall. Only two participants scored below 50%.

Nevertheless, these results suggest that, insofar as we can equate online and offline word knowledge, even for correctly rejected tone nonword trials, L2 participants did not have accurate explicit knowledge of the appropriate tones for target words 20% of the time. This might have further reduced the amplitude of tone nonword responses.

General Discussion

We conducted a lexical decision study with ERP recordings in advanced L2 Mandarin learners whose L1 was English, in order to determine whether and why processing lexical tone is selectively difficult for learners.

Our first research question asked whether L2 listeners were equally accurate in rejection of isolated disyllabic vowel and tone nonwords. Here we found a clear answer. Across all analyses, the L2 group showed consistently weaker performance for tone nonwords than vowel nonwords. With only one exceptional individual, L2 participants showed weaker sensitivity (*d'*) for tones than vowels in lexical decision, and gaps between the lowest L1 score and L2 scores were more common and larger for tones than vowels. These data replicate the same tone-vowel discrepancy in lexical decision accuracy we found for words extracted from continuous speech in Pelzl et al. (2019), and show that the selective deficit in tone word processing extends to more slowly and clearly produced stimuli.

Digging a bit deeper, our second question asked whether a learner's familiarity with the critical words might moderate their performance on tone nonwords. Looking only at trials for which learners had accurate and confident knowledge of the critical words and their appropriate tones, we saw only a slight improvement in accuracy for rejection of tone nonwords. Overall, the disadvantage for tone versus vowel nonwords persisted.

While L2 learners made more errors on lexical decisions to tone nonwords than vowel nonwords, the deficit should not be exaggerated. The L2 group still made accurate decisions in the majority of tone nonword trials. So then, we can also ask about these trials, when their responses were correct. Our third research question was whether ERPs (for correct trials only) might reveal equal L2 sensitivity to vowel and tone mismatches, as these were the trials when listeners achieved successful responses (i.e., correct rejections of nonwords). Our results were consistent with weaker L2 sensitivity for tones compared to vowels in this case as well. While the L2 group displayed statistically significant N400s to vowel nonwords, the tone response was intermediate between real word and vowel nonword responses. Furthermore, measures of offline L2 knowledge for correctly rejected tone nonwords suggest that, for approximately 20% of those trials, the correct response was not necessarily indicative of correct tone knowledge.

We now consider these results in light of the two broad accounts of L2 phonological and lexical difficulty highlighted in our introduction.

Missing and incorrect L2 tone word representations

Perhaps the most straightforward explanation for L2 difficulty in lexical decision tasks requiring tone knowledge is simply that this lexical tone knowledge was never accurately encoded in the learner's long-term memory for many words. A number of scholars (Cook & Gor, 2015; Gor, 2018; Gor & Cook, 2018; Diependaele, Lemhöfer, & Brysbaert, 2013; Veivo & Järvikivi, 2013) have argued that L2 knowledge for less familiar words (usually lower frequency items) is characterized by low-quality, or 'fuzzy', phonological representations. Learners cannot display sensitivity to lexical cues they do not remember or remember incorrectly. In the current study, explicit vocabulary test results point to ongoing weaknesses in advanced L2 explicit

knowledge of tones: for the group, 25% of supplied tones in the offline test were incorrect, and even when learners indicated the highest level of confidence in their tone knowledge, they were still in error more than 10% of the time. If this scales up to a vocabulary of thousands of words, many advanced L2 Mandarin speakers may (confidently) misremember tones for hundreds or thousands of words.

Despite the potentially large scale of (explicit) L2 tone knowledge deficits, missing or misremembered tone knowledge cannot provide a full account of L2 tone word performance in our lexical decision task, as inaccurate decisions were observed even for words for which learners showed confident and accurate tone knowledge in the offline task.

Uncertainty in L2 tone word representations

Apart from the accuracy of encoded L2 tone knowledge itself, another qualitative aspect of that knowledge which could affect lexical decision performance is learners' *(un)certainty* about their own knowledge of the relevant real words (cf. Cook & Gor, 2015; Gor, 2018; Gor & Cook, 2018; Veivo & Järvikivi, 2013). In the present case, even if L2 listeners accurately perceive a tone nonword (e.g., they know they heard *fa**2***yin1*), perhaps they still accept it because they are not *confident* of the tones of the real word counterpart (e.g., *fa**1**yin1).* This uncertainty would make them more permissive in the decision process.

Again, this is not a fully sufficient explanation for present results. While uncertainty may play some role in L2 performance, the best-case scenario analysis suggests that L2 tone nonword inaccuracy is not due solely to such uncertainty. Even when participants had fully accurate *and confident* explicit tone knowledge for real words, they responded incorrectly in the lexical decision task about one-third of the time.

Format of L2 tone word representations

A third possibility is that aspects of L2 tone word knowledge could be represented in a qualitatively different way from L1, such that L2 listeners retrieve tone knowledge more slowly and less accurately under time pressure. For example, L2 learners might encode tone word information as declarative (explicit) knowledge, rather than automatic (implicit) knowledge (DeKeyser, 2007; DeKeyser, 2003). This knowledge might even be encoded in representational modalities other than phonological form, such as visual encoding of tone diacritics or full orthographic (Pinyin romanization) representations of words (cf. Bassetti et al., 2015, and related articles).

Our analysis of ERPs from correct trials only was intended as an initial exploration of this possibility: if lexical tone is encoded in a qualitatively different way in L2 learners, then we might expect the retrieval of this representation to manifest differently in the neural response, *even on trials where retrieval was successful.* For example, if L2 learners retrieve explicit knowledge of tone at a slower timescale, then we might expect that they could fail to show native-like responses to tone nonwords early in processing, and still successfully reject tone nonwords using the slower pathway by the end of the trial. Although not conclusive, our ERP results are consistent with this possibility: unlike native speakers, L2 learners showed a smaller N400 response to tone nonwords than vowel nonwords on trials in which their behavioral response was correct rejection. For example, on the implicit/explicit account this pattern could arise if L2 learners initially retrieve a tone-less form of a real word representation on both real word and tone nonword trials, and only later in the trial retrieve the explicitly encoded tone information that distinguishes the words from the nonwords.

If correct, this interpretation of the ERP results would have strong implications for L2 learners' tone processing capacities in real world situations, as it would suggest their access of lexical tone information could often be too slow to impact processing of continuous speech. In other words, if they often succeed in comprehension of Mandarin speech, it will be despite ineffective tone processing. However, we believe further ERP work is needed before drawing these strong conclusions. Though we are inclined to believe present results accurately reflect L2 tone ERP responses, we must acknowledge that the weaker tone N400 effects in our present analysis could be due simply to lack of sufficient data. After removing incorrect responses, there were nearly 25% fewer trials available for L2 tone than vowel nonwords. Additionally, offline vocabulary results suggest that, for some participants the available data contained a substantial number of guesses. Extending this reasoning then, it is possible that given more data in the tone word condition and less noise in the offline knowledge estimates (i.e., limiting analysis to trials where the participant truly had perfect tone word knowledge), we would discover that L2 tone nonword N400 effects on correct responses were equivalent to those of vowel nonwords. Future replications of present results and refined methods for examining the nature of L2 tone representations will be necessary to fully remove doubts along these lines.

Processing biases could drive L2 tone word errors

A second class of explanations for L2 tone word errors in the lexical decision task is that they reflect differences in the processing biases used by L1 and L2 listeners, rather than or in addition to differences in their stored lexical representations. This class of explanation can straightforwardly account for the discrepancy between offline and online accuracy in retrieving tone word knowledge: both tasks would in principle draw on the same, intact tone word

knowledge, and it is the L2 processing routine for the lexical decision task that is driving the errors. Given their lifetime of experience attending primarily to segmental cues in word recognition, non-tonal L1 speakers default to the same processing routine for tone words, with F0 cues playing little role in accessing lexical candidates.

The Automatic Selective Perception model (ASP) (Strange, 2011; see also recent discussion of perceptual attention by Chang, 2018) posits that task demands will play a key role in determining when L2 learners are able to successfully attend to novel L2 sounds. When task demands are low (as in many identification or discrimination tasks), L2 learners are able to direct attention to acoustic-phonetic cues that are not used or are given much lower weight in their L1. As task demands increase (recognizing words, interpreting semantic content), learners fall back on L1 perceptual routines, often leading to lower levels of performance. Zou et al. (2016) suggest the ASP correctly predicted the outcomes of their study. They found that, as L2 Mandarin proficiency increased, native Dutch speakers showed greater reliance on tones in a challenging AXB task, indicating convergence on appropriate weighting of Mandarin F0 cues. Paired with results from Pelzl et al. (2019), our present results also fit well with the ASP model. When task demands decreased (relative to Pelzl et al. 2019), L2 learners showed an increasing ability to rely on F0 cues to successfully reject many nonwords. However, of the two tasks, that of Pelzl et al. (2019) is likely more similar to the speech learners most often encounter, and so more likely to reflect *typical* L2 use of tone cues.

Wiener and colleagues (Liu & Wiener, 2020; Wiener, 2019; Wiener et al., 2018, 2019; Wiener & Lee, 2020) present a slightly different view of L2 tone learning, though in many ways it seems complimentary to the ASP model. They frame L2 tone-learning under the umbrella of *dimension-based statistical learning* (Idemaru & Holt, 2011, 2014), drawing a distinction

between signal-based and knowledge-based (or probability-based) processing. In the first case, listeners rely on the low-level acoustic-phonetic input of the speech signal itself to recognize words; in the latter, they rely on knowledge of the statistical properties of specific syllables or words in their L2 experience. For example, listeners may know from experience that some syllable + tone combinations are either highly probably or very unlikely to occur. This knowledge may guide their initial processing of relevant syllables, especially under more difficult listening conditions (e.g., multiple talkers, noise). When listening conditions are easier (e.g., a familiar voice in clear speech), listeners can rely more heavily on the acoustic-phonetic signal. Wiener and colleagues have so far not addressed how such processes might play out beyond a single syllable. Given how rare tone neighbors are for disyllabic words, L2 listeners may typically recognize disyllabic words even if they ignore tone (F0) cues. This may lead them to rely on a knowledge-based processing strategy that attends to segmentally defined disyllablic sequences, while disregarding tones.

Though not focused on processing *per se*, recent work by Chan & Leung (2019) is also amenable to processing accounts in that it is deals with statistical learning mechanisms. They examined Cantonese and English L1 participants' abilities to pick up on co-occurrence patterns of syllable-initial segmental cues and specific tones (i.e., syllables beginning with aspirated stop consonants always had rising tones; syllables beginning with an approximant always had falling tones). Whereas Cantonese participants showed some ability to generalize the (implicit) pattern after training, English participants failed to show evidence of learning. Chan and Leung framed their study as an examination of the phonological level of L2 tone learning (cf. the phonetic-phonological-lexical continuum described by Wong & Perrachione, 2007), and suggest that L2 tone learning might be particularly difficult when it comes to the formation of implicit and

abstract phonological tone representations. Similar to Wiener and colleagues, Chan and Leung focus on the level of single syllables, and it is unclear what they would expect for multi-syllable strings, except that it is unlikely to be that L2 performance will increase when confronted with such stimuli.

Finding ways to investigate L2 tone learning in the context of syllables and words of different lengths, while also addressing fundamental differences in the statistical and acoustic properties of single and multi-syllable strings will be required to more fully address these issues.

Practical implications

Despite the tone word difficulty reported above, it is not necessarily the case that this set of L2 learners has many *practical* difficulties as a result of tone word misperception. All of our L2 participants could be characterized as successful language learners. After years of classroom study, they were using Mandarin to communicate on a regular basis in their daily lives, often at a professional level. So then, does the tendency to incorrectly accept nonwords have any bearing on real world L2 Mandarin learners?

We tentatively suggest that it does. Though nonwords by definition do not occur in native Mandarin speech, words with incorrect or missing tones clearly do exist in the vocabulary of many L2 learners. The inability to differentiate these mistaken words from their real word tone neighbors will prevent learners from recognizing their own incorrect tone knowledge. Similarly, whenever a learner encounters a spoken word they have not previously learned, the inability to recognize that word's tones in real time will prevent them from acquiring a fully accurate representation of the word. Even if these difficulties do not cause consistent lexical confusion for the L2 learner as listener, they may still cause difficulties for those who listen to the learner.

Gaps in tone knowledge will lead to production of tone errors that are potentially confusing or misleading for listeners who do process tones—as we see from the strong N400 responses to tone nonwords by L1 listeners in the current study (see also Pelzl et al., 2020; and Pelzl et al., in press, for investigations of the impacts of L2 tone errors on native listeners).

Conclusion

The present study extends our understanding of L2 tone word difficulties by demonstrating that, even in fairly ideal circumstances, L2 learners have considerable difficulty recognizing words on the basis of tones. Our results suggest that both representational and processing issues are at play in these difficulties. L2 learners seem able to function at a high level despite these tone difficulties, but the difficulties nevertheless pose a considerable learning challenge that may have real impacts on the efficiency with L2 learners can expand their Mandarin lexicon.

Notes

1. Though Pelzl et al. (2019) conducted an ERP experiment as well, it targeted responses to words embedded in sentences, and L2 results were difficult to interpret.

2. One reviewer noted that there may be directional effects in perception of monophthong-to-diphthong vs. diphthong-to-monophthong changes. We made no attempt to control such effects. Our expectation would be that, even if they were controlled, vowel changes would remain easier for L2 learners than tone changes.

References

Alexander, J. A., Wong, P. C., & Bradlow, A. R. (2005). Lexical tone perception in musicians and non-musicians. *Interspeech*, 397–400. http://groups.linguistics.northwestern.edu/speech_comm_group/publications/2005/Alexander-Wong-Bradlow-2005.pdf

Antoniou, M., & Wong, P. C. (2016). Varying irrelevant phonetic features hinders learning of the feature being trained. *The Journal of the Acoustical Society of America*, *139*(1), 271–278.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barrios, S. L., Namyst, A. M., Lau, E. F., Feldman, N. H., & Idsardi, W. J. (2016). Establishing New Mappings between Familiar Phones: Neural and Behavioral Evidence for Early Automatic Processing of Nonnative Contrasts. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00995

Bassetti, B., Escudero, P., & Hayes-Harb, R. (2015). Second language phonology at the interface between acoustic and orthographic input. *Applied Psycholinguistics*, *36*(1), 1–6. https://doi.org/10.1017/S0142716414000393

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious Mixed Models. *ArXiv:1506.04967 [Stat]*. http://arxiv.org/abs/1506.04967

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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*(1), 97–103. https://doi.org/10.1037/0096-1523.32.1.97

Bowles, A. R., Chang, C. B., & Karuzis, V. P. (2016). Pitch ability as an aptitude for tone learning. *Language Learning*, *66*(4), 774–808. https://doi.org/10.1111/lang.12159

Broersma, M., & Cutler, A. (2011). Competition dynamics of second-language listening. *The Quarterly Journal of Experimental Psychology*, *64*(1), 74–95.

Broselow, E., Hurtig, R. R., & Ringen, C. (1987). The Perception of second language prosody. In G. Ioup & S. H. Weinberger (Eds.), *Interlanguage Phonology: The Acquisition of a Second Language Sound System* (pp. 350–364). Newbury House Publishers.

Brown, C. A. (1998). The role of the L1 grammar in the L2 acquisition of segmental structure. *Second Language Research*, *14*(2), 136–193.

Brown-Schmidt, S., & Canseco-Gonzalez, E. (2004). Who do you love, your mother or your horse? An event-related brain potential analysis of tone processing in Mandarin Chinese. *Journal of Psycholinguistic Research*, *33*(2), 103–135.

Chan, R. K., & Leung, J. H. (2019). Why are lexical tones difficult to learn? Insights from the incidental learning of tone-segment connections. *Studies in Second Language Acquisition*, 1–27. https://doi.org/10.1017/S0272263119000482

Chandrasekaran, B., Sampath, P. D., & Wong, P. C. M. (2010). Individual variability in cue-weighting and lexical tone learning. *The Journal of the Acoustical Society of America*, *128*(1), 456–465.

Chang, C. B. (2018). Perceptual attention as the locus of transfer to nonnative speech perception. *Journal of Phonetics*, *68*, 85–102. https://doi.org/10.1016/j.wocn.2018.03.003

Chang, C. B., & Bowles, A. R. (2015). Context effects on second-language learning of tonal contrasts. *Journal of the Acoustical Society of America*, *136*(6), 3703–3716.

Chrabaszcz, A., & Gor, K. (2014). Context effects in the processing of phonolexical ambiguity in L2: Context effects in processing of L2. *Language Learning*, *64*(3), 415–455. https://doi.org/10.1111/lang.12063

Cook, S. V., & Gor, K. (2015). Lexical access in L2: Representational deficit or processing constraint? *The Mental Lexicon*, *10*(2), 247–270. https://doi.org/10.1075/ml.10.2.04coo

Cook, S. V., Pandža, N. B., Lancaster, A. K., & Gor, K. (2016). Fuzzy nonnative phonolexical representations lead to fuzzy form-to-meaning mappings. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01345

Cooper, A., & Wang, Y. (2013). Effects of tone training on Cantonese tone-word learning. *The Journal of the Acoustical Society of America*, *134*(2), EL133–EL139.

Darcy, I., Daidone, D., & Kojima, C. (2013). Asymmetric lexical access and fuzzy lexical representations in second language learners. *The Mental Lexicon*, *8*(3), 372–420. https://doi.org/10.1075/ml.8.3.06dar

DeKeyser, R. (2007). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in Second Language Acquisition* (pp. 97–113). Lawrence Erlbaum Associates.

DeKeyser, R. M. (2003). Implicit and explicit learning. In C. J. Doughty & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 313–348). Blackwell Publishing.

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*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

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*(6), 680–689. https://doi.org/10.1016/j.lindif.2012.05.005

Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency effect in first- and second-language word recognition: A lexical entrenchment account. *The Quarterly Journal of Experimental Psychology*, *66*(5), 843–863. https://doi.org/10.1080/17470218.2012.720994

Dong, H., Clayards, M., Brown, H., & Wonnacott, E. (2019). The effects of high versus low talker variability and individual aptitude on phonetic training of Mandarin lexical tones. *PeerJ*, *7*, e7191. https://doi.org/10.7717/peerj.7191

Gor, K. (2018). Phonological priming and the role of phonology in nonnative word recognition. *Bilingualism: Language and Cognition*, *21*(03), 437–442. https://doi.org/10.1017/S1366728918000056

Gor, K., & Cook, S. V. (2018). A mare in a pub? Nonnative facilitation in phonological priming. *Second Language Research*, 18.

Gottfried, T. L. (2007). Music and language learning: Effects of musical training on learning L2 speech contrasts. In O.-S. Bohn & M. J. Munro (Eds.), *Language Experience in Second Language Speech Learning* (pp. 221–237). John Benjamins.

Ho, A. T. (1976). The acoustic variation of Mandarin tones. *Phonetica*, *33*, 353–367.

Howie, J. (1974). On the domain of tone in Mandarin. *Phonetica*, *30*(3), 129–148.

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1939–1956. https://doi.org/10.1037/a0025641

Idemaru, K., & Holt, L. L. (2014). Specificity of dimension-based statistical learning in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1009–1021. https://doi.org/10.1037/a0035269

Ingvalson, E. M., Barr, A. M., & Wong, P. C. M. (2013). Poorer phonetic perceivers show greater benefit in phonetic-phonological speech learning. *Journal of Speech Language and Hearing Research*, *56*(3), 1045. https://doi.org/10.1044/1092-4388(2012/12-0024)

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463–470.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potential reflect semantic incongruity. *Science*, *207*, 203–205.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*, 161–163.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933. https://doi.org/10.1038/nrn2532

Lee, C.-Y., Tao, L., & Bond, Z. S. (2009). Speaker variability and context in the identification of fragmented Mandarin tones by native and non-native listeners. *Journal of Phonetics*, *37*(1), 1–15. https://doi.org/10.1016/j.wocn.2008.08.001

Lee, Y.-S., Vakoch, D. A., & Wurm, L. H. (1996). Tone perception in Cantonese and Mandarin: A cross-linguistic comparison. *Journal of Pscyholinguistic Research*, *25*(5), 527–542.

Li, M., & DeKeyser, R. (2017). Perception practice, production practice, and musical ability in L2 Mandarin tone-word learning. *Studies in Second Language Acquisition*, 1–28. https://doi.org/10.1017/S0272263116000358

Li, X., Yang, Y., & Hagoort, P. (2008). Pitch accent and lexical tone processing in Chinese discourse comprehension: An ERP study. *Brain Research*, *1222*, 192–200.

Liu, J., & Wiener, S. (2020). Homophones facilitate lexical development in a second language. *System*, *91*, 102249. https://doi.org/10.1016/j.system.2020.102249

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00213

MacWhinney, B., & Bates, E. (Eds.). (1989). *The Cross-linguistic Study of Sentence Processing*. Cambridge University Press.

Malins, J. G., & Joanisse, M. F. (2012). Setting the tone: An ERP investigation of the influences of phonological similarity on spoken word recognition in Mandarin Chinese. *Neuropsychologia*, *50*(8), 2032–2043. https://doi.org/10.1016/j.neuropsychologia.2012.05.002

Melnik, G. A., & Peperkamp, S. (2019). Perceptual deletion and asymmetric lexical access in second language learners. *The Journal of the Acoustical Society of America*, *145*(1), EL13–EL18. https://doi.org/10.1121/1.5085648

Pelzl, E., Carlson, M. T., Guo, T., Jackson, C. N., & van Hell, J. G. (2020). Tuning out tone errors? Native listeners do not down-weight tones when hearing unsystematic tone errors in foreign-accented Mandarin. *Bilingualism: Language and Cognition*, 1–8. https://doi.org/10.1017/S1366728920000280

Pelzl, E., Lau, E. F., Guo, T., & DeKeyser, R. (2019). Advanced second language learners' perception of lexical tone contrasts. *Studies in Second Language Acquisition*, *41*(1), 59–86. https://doi.org/10.1017/S0272263117000444

Pelzl, E., Lau, E. F., Guo, T., Jackson, S. R., & Gor, K. (in press). Behavioral and neural responses ton tone errors in foreign-accented Mandarin. *Language Learning*.

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*(1), 461–472.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Sadakata, M., & McQueen, J. M. (2014). Individual aptitude in Mandarin lexical tone perception predicts effectiveness of high-variability training. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01318

Schirmer, A., Tang, S.-L., Penney, T. B., Gunter, T. C., & Chen, H.-C. (2005). Brain responses to segmentally and tonally induced semantic violations in Cantonese. *Journal of Cognitive Neuroscience*, *17*(1), 1–12.

Shen, G., & Froud, K. (2016). Categorical perception of lexical tones by English learners of Mandarin Chinese. *The Journal of the Acoustical Society of America*, *140*(6), 4396–4403. https://doi.org/10.1121/1.4971765

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2017). *Afex: Analysis of factorial experiments* (R package version 0.17-8) [Computer software]. http://cran.r-project.org/package=afex

So, C. K., & Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Language and Speech*, *53*(2), 273–293.

Strange, W. (2011). Automatic selective perception (ASP) of first and second language speech: A working model. *Journal of Phonetics*, *39*(4), 456–466. https://doi.org/10.1016/j.wocn.2010.09.001

Tao, H. (2015). Profiling the Mandarin spoken vocabulary based on corpora. In W. S.-Y. Wang & C. Sun (Eds.), *The Oxford Handbook of Chinese Linguistics*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199856336.013.0031

Veivo, O., & Järvikivi, J. (2013). Proficiency modulates early orthographic and phonological processing in L2 spoken word recognition. *Bilingualism: Language and Cognition*, *16*(04), 864–883. https://doi.org/10.1017/S1366728912000600

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, *106*, 3649.

Wiener, S. (2019). Second language learners develop non-native lexical processing biases. *Bilingualism: Language and Cognition*, 1–12. https://doi.org/10.1017/S1366728918001165

Wiener, S., Ito, K., & Speer, S. R. (2018). Early L2 spoken word recognition combines input-based and knowledge-based processing. *Language and Speech*, 002383091876176. https://doi.org/10.1177/0023830918761762

Wiener, S., Lee, C., & Tao, L. (2019). Statistical regularities affect the perception of second language speech: Evidence from adult classroom learners of Mandarin Chinese. *Language Learning*, *69*(3), 527–558. https://doi.org/10.1111/lang.12342

Wiener, S., & Lee, C.-Y. (2020). Multi-talker speech promotes greater knowledge-based spoken Mandarin word recognition in first and second language listeners. *Frontiers in Psychology*, *11*, 214. https://doi.org/10.3389/fpsyg.2020.00214

Wong, P. C. M., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics*, *28*(04), 565–585.

Yu, K., Li, L., Chen, Y., Zhou, Y., Wang, R., Zhang, Y., & Li, P. (2019). Effects of native language experience on Mandarin lexical tone processing in proficient second language learners. *Psychophysiology*, *56*(11). https://doi.org/10.1111/psyp.13448

Zhang, L. (2011). Meiguo liuxuesheng Hanyu shengdiaode yinwei he shengxue xinxi jiagong. *Shijie Hanyu Jiaoxue (Chinese Teaching in the World)*, *25*(2), 268–275.

Zhao, J., Guo, J., Zhou, F., & Shu, H. (2011). Time course of Chinese monosyllabic spoken word recognition: Evidence from ERP analyses. *Neuropsychologia*, *49*(7), 1761–1770. https://doi.org/10.1016/j.neuropsychologia.2011.02.054

Zou, T., Chen, Y., & Caspers, J. (2016). The developmental trajectories of attention distribution and segment-tone integration in Dutch learners of Mandarin tones. *Bilingualism: Language and Cognition*, 1–13. https://doi.org/10.1017/S1366728916000791