Studies in Second Language Acquisition, 2018, page 1 of 28.
doi:10.1017/S0272263117000444

Research Article

ADVANCED SECOND LANGUAGE LEARNERS' PERCEPTION OF LEXICAL TONE CONTRASTS

Eric Pelzl*

University of Maryland, College Park

Ellen F. Lau

University of Maryland, College Park

Taomei Guo

Beijing Normal University

Robert DeKeyser

University of Maryland, College Park

Abstract

It is commonly believed that second language (L2) acquisition of lexical tones presents a major challenge for learners from nontonal language backgrounds. This belief is somewhat at odds with research that consistently shows beginning learners making quick gains through focused tone training, as well as research showing advanced learners achieving near-native performance in tone identification tasks. However, other long-term difficulties related to L2 tone perception may persist, given the additional demands of word recognition and the effects of context. In the current study, we used behavioral and event-related potential (ERP) experiments to test whether perception of Mandarin tones is difficult for advanced L2 learners in isolated syllables, disyllabic words in isolation, and disyllabic words in sentences. Stimuli were more naturalistic and challenging than in previous research. While L2 learners excelled at tone identification in isolated syllables, they performed with very low accuracy in rejecting disyllabic tonal nonwords in isolation and in sentences. We also report ERP data from critical mismatching words in sentences; while L2 listeners showed no significant differences in responses in any condition, trends were not

Thanks to Junjie Wu, Di Lu, Yongben Fu, and Chunyan Kang for help running participants in Beijing, to Man Li for help proofreading initial rounds of Chinese stimuli, to Anna Chrabaszcz for helpful advice and Matlab scripts, and to Brendan Cone for help editing many, many sound files. This research was supported in part by NSF-IGERT grant 0801465 and NSF-EAPSI grant 1514936.

*Correspondence concerning this article should be addressed to Eric Pelzl, Second Language Acquisition, University of Maryland, College Park, MD 20742, USA. E-mail: pelzlea@gmail.com

inconsistent with the overall pattern in behavioral results of less sensitivity to tone mismatches than to semantic or segmental mismatches. We interpret these results as evidence that Mandarin tones are in fact difficult for advanced L2 learners. However, the difficulty is not due primarily to an inability to perceive tones phonetically, but instead is driven by the need to process tones *lexically*, especially in *multisyllable* words.

INTRODUCTION

It is commonly believed that second language (L2) acquisition of lexical tones presents a major challenge for learners from nontonal language backgrounds (e.g., Showalter & Hayes-Harb, 2013; Wang, 2013). This belief appears somewhat at odds with research that consistently shows beginning learners making quick gains in tone identification accuracy through focused training (e.g., Wang, Spence, Jongman, & Sereno, 1999). Similarly, while there are relatively few studies that examine more advanced L2 learners of tone languages, several suggest such learners can obtain a high degree of mastery in perceiving tones (Lee, Tao, & Bond, 2010; Zhang, 2011; Zou, Chen, & Caspers, 2016). However, these results do not necessarily rule out long-term difficulties related to L2 tone perception. Research with novice learners has focused on short-term outcomes and may not generalize well to real-world learning conditions. Research with advanced learners has focused almost exclusively on tone perception in nonmeaningful isolated syllables, leaving the effects of context and the demands of word recognition unexamined. Even if advanced learners can master low-level phonetic perception of tones, these additional factors might still create long-term difficulties.

In the current study we used behavioral and event-related potential (ERP) measures to examine the lexical tone perception abilities of advanced L2 learners of Mandarin Chinese in isolated syllables, isolated disyllabic words, and disyllabic words in sentential contexts. Our goal was to determine whether L2 tones are truly difficult to learn and, if so, to begin exploring reasons for that difficulty. Specifically, we wished to examine whether difficulties in tone perception appear in phonetic tasks (i.e., listener's ability to identify tones nonlexically), lexical tasks, or sentence-level tasks. To ensure that any differences across tasks were not attributable to the acoustics of the tones, we used a novel design that presented word and syllable stimuli extracted from the same sentence context, thus increasing the difficulty of tasks. This study makes a novel and important contribution to research on lexical tones and to the broader discussion of the role of difficult speech sounds in L2 lexical recognition.

BACKGROUND

DIFFICULT SOUNDS IN L2 SPEECH PERCEPTION

It is well known that unfamiliar L2 speech sounds can cause major difficulties for language learners, especially those who start learning after early childhood (e.g. Abrahamsson, 2012; Flege, Munro, & MacKay, 1995). Recent research has specifically explored ways in which such sounds can induce *lexical* confusion (Broersma, 2012; Broersma & Cutler, 2008, 2011; Chrabaszcz & Gor, 2014; Darcy,

Daidone, & Kojima, 2013; Díaz, Mitterer, Broersma, & Sebastián-Gallés, 2012; Lukianchenko, 2014; Pallier, Colomé, & Sebastián-Gallés, 2001; Sebastián-Gallés & Díaz, 2012). When two or more sounds in an L2 are difficult for learners to distinguish, this can lead to so-called phonolexical ambiguity, that is, *phonologically* induced *lexical* ambiguity (cf. Chrabaszcz & Gor, 2014). In such cases, words sound alike to L2 listeners, even though they are easily distinguishable to native (L1) listeners (for further discussion, see especially Broersma & Cutler, 2011; Cook & Gor, 2015; Darcy et al., 2013).

Though increased proficiency may alleviate difficulties in lower-level phoneme perception tasks, it appears that for much of the learning process, the ability to categorize L2 sounds phonetically does not *necessarily* predict success in lexical tasks (cf. Sebastián-Gallés & Díaz, 2012). Examining Dutch learners of English, Díaz et al. (2012) found that some learners who performed within the range of native speakers on a phoneme categorization task targeting English /æ/ and /ɛ/, nevertheless had significant difficulties with lexical tasks that targeted the same phonological distinction. Sebastián-Gallés & Díaz (2012) refer to this discontinuous pattern as "graded learning." There are now several examples of such discontinuities in the L2 speech learning literature (e.g., Darcy et al., 2013; Díaz et al., 2012). These studies show that even when L2 listeners excel at tasks that involve individual *sounds* (i.e., identification, categorization, or discrimination), they may nevertheless have significant difficulty in tasks that require utilizing those sounds for the recognition of *words*.

In light of such discontinuities, in the following text we will make a consistent distinction in our use of the words *phonetic* and *phonological* (categorization, perception, etc.). *Phonetic* implies learners can distinguish and label L2 phonemes, but does not require that they are able to encode these phoneme categories lexically. *Phonological* implies that learners can establish phonetic categories and *encode them* for lexical recognition. Importantly, this distinction applies to both segmental phonemes and suprasegmental phonemes, such as lexical tones.

L2 PERCEPTION OF LEXICAL TONES: ADVANCED LEARNERS AND LEXICAL RECOGNITION

Lexical tone languages use contrasts in pitch (*f*0) *height* (high, low) and/or *contour* (rising, falling) to distinguish lexical meanings. In the case of Mandarin Chinese, there are four contrastive tones, conventionally numbered one to four. The tones can be illustrated with the syllable *tang* (/taŋ/). When spoken with a high, level pitch (T1) *tāng* means "soup"; with a rising pitch (T2) *táng* means "candy"; with a low or dipping pitch (T3) *tǎng* means "lie (down)"; and with a falling pitch (T4) *tàng* means "scald" (the diacritics indicate high ¯, rising ´, low/dipping ˇ, and falling ` tones, respectively). Because nearly every syllable of every Mandarin word bears a tone, any weaknesses in tone perception have the potential to induce large amounts of phonolexical ambiguity. For instance, to a hypothetical L2 learner who was completely "tone deaf," all four *tang* words would sound the same.

At present, the handful of studies that have investigated more experienced L2 Mandarin learners present a somewhat mixed picture of their tone perception abilities. Experienced L2 listeners appear to perform comparably to native listeners on simple monosyllabic tone identification tasks (e.g., Lee, Tao, & Bond, 2009). The picture is less

clear for disyllabic or multisyllabic stimuli. Zhang (2011) examined native English speakers studying Mandarin in China. Learners were classified as either "novice" (7–9 months of study in China) or "intermediate" (20–25 months). While L2 accuracy scores did not differ significantly from those of natives on a monosyllabic tone identification task, learners nevertheless had difficulty on a repeated "word" recognition task. Here, participants were briefly trained to recognize a set of disyllabic nonword stimuli, and afterward tested on their ability to correctly reject untrained nonwords that differed from trained items by a single tone (e.g., trained /nákū/ vs. untrained /nàkū/). Relative to native speakers, learners both at novice and intermediate levels were significantly more likely to accept untrained items. Hao (2012) also examined more experienced L2 learners. She compared the acquisition of Mandarin tones by native speakers of English and Cantonese (English speakers mean length of study was 2.68 years). Stimuli included monosyllabic real words and disyllabic nonwords. Participants completed an identification task during which they heard the stimuli one at a time and wrote the corresponding tone symbol (¯ˊˇˋ). Results for English speakers indicated fairly high accuracy rates for T1 (approaching 90%) and T4 (> 90%), but much lower accuracy on T2 and T3 (roughly 70% for both). Additionally, identifying tones in monosyllabic stimuli was easier than in disyllabic stimuli, and for disyllabic stimuli, tones on initial syllables were more difficult than on final syllables (for similar outcomes, cf. Sun, 1998).

While the studies in the preceding text suggest that experienced L2 learners have relatively good phonetic tone perception in monosyllables and some continued difficulties with disyllabic stimuli, there are some significant limitations. First, L2 proficiency was based only on length of study, so it is not clear that the "advanced" participants were indeed highly proficient. Second, there were no conditions testing segmental contrasts (e.g., Zhang, 2011, did not test nákū vs. níkū). This makes it impossible to tease apart tone-related effects from task-related effects, such as memory constraints, which are certain to play a role as stimuli increase in length. Third, the studies did not address how difficulty perceiving tones might impact lexical recognition; disyllabic stimuli were, for the most part, not even real words.

In naïve learners, lexical recognition ability has been tested in a number of tone training studies (see especially the work of Patrick Wong and his colleagues, e.g., Wong & Perrachione, 2007), where participants must learn to associate tonal words with pictures, rather than tone labels. Generally, naïve learners can make significant progress on such tasks, though individual differences play a large role, at least in the short time span of such studies (Chandrasekaran, Sampath, & Wong, 2010; Perrachione, Lee, Ha, & Wong, 2011). However, participants generally learn only a couple dozen words, all of which form minimal tone contrasts with other known words in the training vocabulary (e.g., *mā* vs. *má* vs. *mà*), thereby enhancing the salience of tones as a lexical feature. These studies thus are limited in their ability to reflect the realities of the much larger L2 Mandarin lexicon that experienced learners have acquired.

In real-life L2 acquisition, minimal tone contrasts between any two words—let alone between highly imageable nouns—are likely to be rather rare early on, and will become common only as the learner's vocabulary increases to hundreds or thousands of words. Even when such contrasts do occur, there are almost always large asymmetries in the frequency with which its members appear (e.g., *mā* "mom" is much more frequent than *má* "hemp"). Furthermore, except for Chang and Bowles (2015) (see also Bowles,

Chang, & Karuzis, 2016), previous tone training studies have used only monosyllabic target words, while the actual Mandarin lexicon is largely disyllabic (Duanmu, 2007). Disyllabic minimal tone contrasts are naturally much less frequent than monosyllabic ones, but this does not necessarily mean they are less problematic (for discussion of ways *nonword* neighbors can lead to word recognition problems, see Broersma & Cutler, 2011), and, at least for novice learners, disyllabic tone words do appear to be much more difficult to recognize (Chang & Bowles, 2015). In short, the actual impacts of tone confusion on L2 lexical recognition cannot be understood unless we examine learners who have already achieved a sizable vocabulary and can comprehend connected speech.

Lexical recognition has been a central issue in work examining L1 perception of tones, both for isolated words and in sentential contexts. Several studies have provided evidence of native speakers' early sensitivity to tonal errors *in connected speech* and seem to indicate that native speakers process tonal and segmental cues equally fast in context. For example, Brown-Schmidt and Canseco-Gonzalez (2004) used electroencephalography (EEG) to measure ERPs while participants listened to spoken Mandarin sentences, and examined the N400, a negative-going ERP component that peaks around 400 ms after stimulus onset (Kutas & Hillyard, 1980). The amplitude of the N400 can be modulated by manipulating the probability of target words occurring in context, typically operationalized as *cloze probability*, that is, the probability of test takers providing a given word when prompted with the preceding context. As cloze probability increases, the amplitude of the N400 decreases, reflecting the ease or difficulty with which listeners access or integrate words during online processing (Kutas & Federmeier, 2000; Kutas & Hillyard, 1984; Lau, Phillips, & Poeppel, 2008). By comparing listeners' ERPs to expected sentence endings with mispronunciations of tones, segments, or both, Brown-Schmidt and Canseco-Gonzalez were able to show that in L1 speakers the onset and amplitude of the increased N400 responses for unexpected segmental and tonal information were not significantly different. Other ERP studies have shown similar results, demonstrating that native Chinese speakers can quickly and efficiently use lexical tones to identify words both in isolation (Malins & Joanisse, 2012; Zhao, Guo, Zhou, & Shu, 2011) and in connected speech (Schirmer, Tang, Penney, Gunter, & Chen, 2005).

THE PRESENT STUDY

The present study investigated whether and why tones might be difficult for advanced L2 learners by examining their performance at the phonetic level (isolated syllables) and the phonological level (isolated disyllabic words and disyllabic words in sentences). To this end, we carried out three experiments, using behavioral (Experiments 1–3) and ERP (Experiment 3) methods.[1] As noted previously, to control acoustic differences across experiments and to increase the difficulty of isolated word and syllable tasks, all stimuli were extracted from the same set of sentence recordings (see online supplementary materials for details and lists of all stimuli).

PARTICIPANTS

L2 participants were 16 (seven female) native speakers of English who had achieved high levels of proficiency in Mandarin Chinese. They were recruited in Beijing, China (n = 9)

or in Maryland (n = 7), through posters, e-mail, and word of mouth. Two screening measures were used to ensure selection of highly proficient L2 learners (details in online supplementary materials). General biographical and language learning information was gathered with a brief questionnaire, and is shown for the 16 L2 participants, along with scores from the two screening tests, in Table 1.

Twenty L1 speakers of Mandarin (14 female, average age 22.95) served to establish a baseline for comparison of L2 results. All self-identified as native speakers of Mandarin (*Putonghua*). All L1 participants were recruited and tested at Beijing Normal University (BNU).

All procedures were approved by the University of Maryland Institutional Review Board (IRB) and BNU prior to the involvement of any participants. All participants were right-handed, provided informed consent, and were compensated for their time in the lab.

EXPERIMENT 1: TONE IDENTIFICATION TASK—TONES IN ISOLATED SYLLABLES

Experiment 1: Research Questions and Task Design

Experiment 1 aimed to answer the questions: *Are advanced L2 learners of Mandarin able to identify tones on isolated syllables? How does their performance compare to that of L1 speakers?* To answer these questions, we administered a tone identification task (Tone ID). In the Tone ID, participants heard isolated monosyllabic Mandarin words and were asked to identify the tones by pressing a corresponding number key (e.g., 1 for T1). Stimuli were balanced so that there were 15 instances of each tone. No two stimuli shared the same segmental structure. All monosyllabic stimuli were extracted from disyllabic words (Experiment 2) that had been extracted from sentences (Experiment 3). This design has the merit of allowing relatively fair comparisons of acoustic characteristics of stimuli across experiments.

Another motivation for the present design was to increase difficulty and avoid ceiling performance. One factor that could contribute to ceiling performance is stimuli with unnaturally long durations, which can occur when they are produced in isolation. In many previous studies, stimuli had average syllable durations between 300 and 500 ms (or even longer). Durations in the current study are more naturalistic: *T1*: m = 224 ms (*range* = 180–313 ms); *T2*: m = 208 ms (152–284 ms); *T3*: m = 221 ms (168–265 ms); *T4*: m = 217 ms (146–294 ms). This should make tone identification more challenging.

TABLE 1. Background information and screening measure scores for L2 participants (n = 16)

	Mean (*SD*)	Range
Age at testing	29.4 (7.6)	22–49
Age of onset	18.6 (1.8)	16–24
Semesters of formal study	8.2 (5.0)	2–21
Years in immersion	2.3 (1.6)	0.4–6
Total years learning	10.8 (7.7)	4–30
Can-do self-assessment (%)	86.3 (8.5)	67.2–97.6
Vocabulary self-assessment (%)	85.3 (12.1)	60.9–100

Additionally, as documented by Xu (1997), when tones are produced in context, pitch onsets and offsets shift to accommodate preceding or following tones. These contextual influences are likely to further increase the difficulty of our Tone ID.

Though we are calling the Tone ID a "syllable" level task, we acknowledge that the syllables used can all be mapped to real, monosyllabic Mandarin words—though they were not produced as such. For example, gāo ("tall") was extracted from the word *gāozhōng* ("high school"). Crucially, the Tone ID does not *require* word recognition and can be accomplished successfully even if participants do not recognize a syllable as a word.

Experiment 1: Procedures

The Tone ID was administered to participants while they were seated comfortably in a quiet room in the Beijing or Maryland lab. The experiment was presented using PsychoPy version 1.82.01 on a MacBook Air (1.7 Ghz Intel Core i5) running OSX 10.9.5. Participants wore earphones and were instructed to adjust the sound to a comfortable volume. Each trial began with a 350 ms beep followed by a 200 ms silence before presentation of the target item. Participants were instructed to respond as quickly and accurately as possible, but there was no time limit on responses. A 1,000 ms silence followed each response before presentation of the next item. Participants first completed eight practice items with feedback (indicating only whether a response was correct/incorrect) and then completed a total of 60 experimental trials without feedback. Experimental trials were presented in a random order different for each participant. The entire task lasted about five minutes.

Experiment 1: Analysis and Results

Tone ID reliability was acceptable ($\alpha = .75$), and the task appears to have effectively eliminated across-the-board ceiling performance, even for the native Mandarin speakers. Mean accuracy appears comparable for both groups (Table 2) on T1, T3, and T4, with accuracy for T1 being higher than for T3 or T4. In the case of T2, there seems to be a considerable difference in accuracy between groups.[2]

Accuracy results were submitted to a generalized linear mixed-effects model with crossed random effects for subjects and items. This and all subsequent statistical analyses were conducted in *R* (version 3.3.3, R Core Team, 2017). Models were run using the *lme4* package (version 1.1-12, Bates, Mächler, Bolker, & Walker, 2015). For the Tone ID model, the dependent variable was accuracy (1, 0). Fixed effects included *tone condition* (T1, T2, T3, T4) and *group* (L1 or L2) and their interactions. As the monosyllabic Tone ID stimuli all mapped to real words, word *frequency* (log-transformed and z-scored) was included in model comparisons as a nuisance factor along with all two- and three-way interactions. Frequency was computed as the logW-CD value taken from the SUBTLEX-CH corpus (Cai & Brysbaert, 2010) for the most commonly occurring form of each monosyllabic Tone ID stimulus. Model comparisons here, and subsequently, were carried out with a "maximal" approach using backward selection to arrive at the best-fitting and most parsimonious model (Barr, Levy, Scheepers, & Tily, 2013). For Tone ID, the final model included fixed effects of tone condition and group, and by-subject and by-item

TABLE 2. Means and standard deviations for accuracy by tone condition in the Tone ID

Condition	Group	Mean	*SD*
T1	L1	0.92	0.27
	L2	0.91	0.29
T2	L1	0.93	0.25
	L2	0.77	0.42
T3	L1	0.79	0.41
	L2	0.77	0.42
T4	L1	0.85	0.36
	L2	0.85	0.36

random intercepts, by-subject random slopes for the effect of condition, and by-item random slopes for the effect of group. The factor *frequency* was dropped as it did not significantly improve model fit or substantively change results. The model reported in the following text and all subsequent generalized linear mixed-effects models were fit using the *bobyqa* optimizer in *lme4*. Default model output here reports simple effects and should be interpreted with reference to the baseline (T1 for the L1 group), that is, *b* estimates indicate difference (in logits) relative to the likelihood of correct identification of T1 by the L1 group. These estimates indicate the size and direction of effects (cf. Jaeger, 2008). Additional comparisons of interest not provided in the default output from *lme4* were examined using the *multcomp* package (Hothorn, Bretz, & Westfall, 2008) and are labeled "additional comparisons" in each table. No corrections for multiple comparisons were applied to mixed-effects model results.

Results are shown in Table 3. The intercept indicates that the L1 group was more likely than not to correctly identify T1 items. T2 and T4 did not differ significantly from T1. There was a marginal effect for T3 suggesting the L1 group was less likely to correctly identify T3 than T1 (b = -1.69, SE = .87, p = .052). Transforming the log-odds (exp(b)), we see that L1 listeners were about five times (1/.19 = 5.39) less likely to respond correctly to T3 items than T1 items.

Model output and additional comparisons revealed no significant differences between groups for T1 (b = -.89, SE = .59, p = .131), T3 (b = .44, SE = .53, p = .403), or T4 (b = .57, SE = .51, p = .265). However, L2 was significantly less likely to correctly identify T2 items (b = 2.14, SE = .49, $p < .001$). The size of the b estimate indicates that the L2 group was almost nine times ($exp(b)$ = 8.53) less likely to respond correctly to T2 than the L1 group.

Experiment 1: Discussion

The results of the Tone ID task are consistent with the view that advanced L2 learners can successfully identify tones on isolated syllables. Except for T2, advanced L2 learners performed comparably to native speakers, despite the fact that the stimuli used here were relatively naturalistic (and therefore more challenging). These results are in line with previous studies that found strong L2 performance on monosyllabic tone identification (Lee et al., 2010; Zhang, 2011), as well as somewhat lower performance for T2 and T3 (Hao, 2012; Sun, 1998). Whereas T2 seems uniquely difficult for L2 learners, T3 appears to be more difficult for both L1 and L2 listeners.

TABLE 3. Results of generalized linear mixed-effects modeling for Tone ID

*model formula (glmer): accuracy ~ condition * group + (1 + condition | subject) + (1 + group | item)*

	Fixed Effects						Random Effects	
	b	*exp(b)*	*SE*	*z*	*p*		subjects *SD*	items *SD*
Intercept (*L1/T1*)	3.84	46.71	0.68	5.65	<.001	*	0.86	1.95
T2	-0.26	0.77	0.88	-0.29	.769		0.44	
T3	-1.69	0.19	0.87	-1.94	.052	†	0.76	
T4	-0.96	0.38	0.89	-1.08	.279		0.74	
L2	-0.89	0.41	0.59	-1.51	.131			1.00
L2 × T2	-1.25	0.29	0.67	-1.88	.061	†		
L2 × T3	0.45	1.57	0.68	0.66	.507			
L2 × T4	0.32	1.38	0.70	0.46	.647			
Additional comparisons								
L1 T2 – L2 T2	2.14	8.53	0.49	4.33	<.001	*		
L1 T3 – L2 T3	0.44	1.55	0.53	0.84	.403			
L1 T4 – L2 T4	0.57	1.77	0.51	1.12	.265			

*= statistically significant; † = marginal

Why is T2 uniquely difficult in L2? Some researchers suggest that rising tones are marked, and therefore naturally more difficult for humans to perceive than other tones (cf. Zhang, 2016). A second factor, specific to Mandarin, might be the tendency for traditional pedagogy to emphasize the dipping nature of T3 (rather than its more common contextual realization as a low tone, i.e., nondipping), which could lead to confusion because T2 can also have a slight dip in its contour. There is some indication from production research that such an emphasis may lead to confusion between T2 and T3 (e.g., Zhang, 2014). This might also explain the lower L2 accuracy for T3, though it may also be that, as is likely the case for L1 listeners, confusion between T3 and T2 is driven by T3 sandhi processes that change T3 to a rising tone (T2) when it occurs before another T3 (cf. Huang & Johnson, 2010).

EXPERIMENT 2: TONES IN ISOLATED WORDS—AUDITORY LEXICAL DECISION TASK

Experiment 2: Research Questions and Task Design

As discussed in the preceding text, even with excellent performance on tasks targeting phonetic categorization, L2 learners may still have difficulty when the same sound categories are involved in a lexical task. Thus, in our second experiment, we aimed to answer the questions: *Are advanced L2 learners of Mandarin Chinese able to use tonal and/or segmental cues to distinguish isolated real words from nonwords? How does their performance compare to that of L1 speakers?*

To examine listeners' abilities to utilize tonal distinctions when processing isolated words, an auditory lexical decision task (LDT) was carried out using individually presented words and nonwords. In the LDT, participants heard a single disyllabic stimulus item and immediately decided whether it was a word. For each real word,

corresponding nonwords were created that differed from the real words in either rhyme (vowels, occasionally including a syllable final /n/ or /ŋ/) or tone. For example, the real word *fángzi* ("house") had the corresponding nonwords *féngzi* (segmental nonword) and *fàngzi* (tonal nonword). Disyllabic stimuli were extracted from continuous speech in sentences. Four presentation lists of 120 items were created. Each list contained 60 real words, 30 segmental nonwords, and 30 tonal nonwords. Lists were balanced such that, across participants, all items occurred in all conditions. It was expected that the L2 group would excel at accepting real words and rejecting segmental nonwords, but that they might have considerable difficulty rejecting tonal nonwords.

Experiment 2: Procedures

Equipment and lab conditions were the same as for Experiment 1. Each trial began with a 350 ms beep, followed by a 200 ms silence and then presentation of the target item. Participants were instructed to respond by pressing "J" for *yes* and "F" for *no* (on a QWERTY keyboard) as quickly and accurately as possible, but there was no time limit on responses. An 800 ms interval followed each response before the next trial began. Participants first completed 12 practice items with feedback (correct/incorrect). There was no feedback during the experiment. Each participant completed 120 total trials delivered in two blocks of 60 with a break provided after the first block. Conditions were balanced so that an even number of real words and nonwords (balanced between segmental and tonal manipulations) occurred in each block. Presentation of items within the blocks was generated in a random order for each participant. The entire task took about ten minutes.

Experiment 2: Analysis and Results

Reliability for the four lists of the LDT was consistently high (list 1: $\alpha = .93$; list 2: $\alpha = .96$; list 3: $\alpha = .95$; list 4: $\alpha = .91$). Descriptive statistics for the LDT are shown in Table 4.[3] Both L1 and L2 groups were quite accurate in correctly accepting real words. The L1 group performed with similarly high accuracy in rejecting segmental nonwords, while L2 was somewhat less accurate. For tonal nonwords, the L1 group performed with slightly lower accuracy than in other conditions, while L2 accuracy was much lower.

As in Experiment 1, results were submitted to a generalized linear mixed-effects model with crossed random effects for subjects and items. The dependent variable was accuracy

TABLE 4. Means and standard deviations for accuracy by condition in the LDT

Condition	Group	Mean	*SD*
Real words	L1	0.94	0.24
	L2	0.91	0.29
Segmental nonwords	L1	0.96	0.20
	L2	0.84	0.36
Tonal nonwords	L1	0.91	0.29
	L2	0.35	0.48

(1, 0). Fixed effects included *condition* (real word, segmental nonword, tonal nonword) and *group* (L1 or L2) and their interaction. The final model included by-subject and by-item random intercepts, and by-subject random slopes for the effect of *condition* and by-item random slopes for the effect of *group* (the maximal model with by-item random slopes for condition and group was tested, but failed to converge; other similar models did not provide improvement over the final model).

Though the default *lme4* output for our model is reported in Table 5, including comparisons with performance on real words, we will focus here especially on the critical comparisons between nonword conditions and between groups. For L1 listeners there was a small but significant difference in the likelihood of correct rejections between segmental and tonal nonwords ($b = 1.00$, $SE = .45$, $p = .026$). L1 listeners were about three times less likely to respond correctly to tonal nonwords than segmental nonwords ($exp(b) = 2.73$). L2 listeners were also less likely to correctly reject tonal nonwords than segmental nonwords ($b = 3.24$, $SE = .33$, $p < .001$). This effect was very large, with L2 listeners about 26 times less likely to correctly reject tonal nonwords than segmental nonwords ($exp(b) = 25.56$). Compared to L1, L2 was less likely to correctly reject segmental ($b = 2.32$, $SE = .50$, $p < .001$) and tonal nonwords ($b = 4.56$, $SE = .49$, $p < .001$). The effect was again very large for tonal nonwords. L2 listeners were about 96 times less likely to correctly reject tonal nonwords than L1 listeners ($exp(b) = 95.69$).

Experiment 2: Discussion

The results of Experiment 2 show a dramatic difference between the L1 and L2 groups in their ability to reject nonwords on the basis of tones. This might indicate a disconnect between L2 abilities to categorize tones as phonetic objects and abilities to utilize those categories as lexical cues. Alternatively (or additionally), we might think that the task shows

TABLE 5. Results of generalized linear mixed effects modeling for LDT

*model formula (glmer): accuracy ~ condition * group + (1 + condition | subject) + (1 + group | item)*

	Fixed Effects						Random Effects	
	b	*exp(b)*	*SE*	*z*	*p*		subjects *SD*	items *SD*
Intercept (*L1/ real words*)	4.15	63.56	0.35	11.92	< .001	*	0.59	1.86
segmental nonwords	0.50	1.64	0.48	1.04	.298		1.06	
tonal nonwords	-0.51	0.60	0.47	-1.09	.277		1.27	
L2	-1.14	0.32	0.39	-2.88	.004	*		1.44
L2 × segmental nonwords	-1.19	0.31	0.55	-2.15	.031	*		
L2 × segmental nonwords	-3.42	0.03	0.59	-5.80	< .001	*		
Additional comparisons								
L1 segmental – L1 tonal	1.00	2.73	0.45	2.23	.026	*		
L2 segmental – L2 tonal	3.24	25.56	0.33	9.80	< .001	*		
L1 segmental – L2 segmental	2.32	10.22	0.50	4.68	< .001	*		
L1 tonal – L2 tonal	4.56	95.69	0.49	9.34	< .001	*		

* = statistically significant

the difficulty of identifying tones in more complex (disyllabic) stimuli as opposed to monosyllables. Finally, we might wonder if the results reflect something about L2 vocabulary knowledge rather than tone perception *per se*—though their much stronger performance in the segmental nonword condition suggests this alone is not a sufficient explanation for the results. We will delay fuller discussion of these issues until we have examined results from Experiment 3.

EXPERIMENT 3: TONES IN SENTENCES—SENTENCE JUDGMENT AND EVENT-RELATED POTENTIAL EXPERIMENT

Experiment 3: Research Questions and Task Design

The (in)ability to perceive tones phonologically in isolated words might not speak directly to the ability to handle lexical tone cues during comprehension of continuous speech, where constraining context might make tone errors more (or less) salient. To this end, in Experiment 3 we aimed to answer the following questions: *Do advanced L2 learners of Mandarin Chinese display evidence of early sensitivity to tonal and segmental errors in sentential context (as indexed by the N400)? Do advanced L2 learners display differences in the amplitude of N400 effects in response to tonal errors compared to segmental errors? Are the N400 patterns of L2 learners different from those of L1 speakers?* To address these questions, we used a sentence judgment task (SJT) along with an ERP experiment testing L1 and L2 listeners' abilities to utilize lexical tones to disambiguate words in sentences. The SJT is a coarse behavioral measure of listener ability to detect critical segmental and tonal manipulations. ERPs go a step further and allow us to capture online sensitivity to specific tones and segments *in context*. This approach parallels previous L1 Mandarin research (reviewed previously).

Materials consisted of 120 target sentences, each occurring in four versions with different critical word conditions: expected, semantic mismatch, segmental mismatch, and tonal mismatch. An example sentence is shown in Figure 1. In the expected condition, the critical word is the highly predictable *xiōngdì* ("brothers"), which is expected to elicit a large reduction in the N400 response. In the semantic mismatch condition, the target word *càidān* ("menu") is expected to elicit an increased (i.e., more negative) N400 amplitude. This condition serves to verify that listeners display the increased N400 responses expected in normal language processing. In the segmental mismatch condition, the target nonword *xuēdì* differs from the expected word with respect to the rhyme (*ue* vs. *iong*) and should elicit an increased N400 amplitude. As with the semantic mismatch, the segmental mismatch is expected to be quite clear to both L1 and L2 listeners. In the critical tone mismatch condition, the target is the nonword *xióngdì* which differs from the expected word in a single tone (rising instead of high). If listeners are sensitive to tonal mismatches, this should elicit an increase in N400 amplitude similar to that found in the semantic and segmental mismatch conditions.

Four experimental lists were constructed, each consisting of 120 sentences, 30 in each condition. Lists were balanced across participants so that no one heard the same sentence twice, but all sentences in all four conditions were used repeatedly in the experiment overall. Additionally, 60 filler items (well-formed sentences presented with no manipulations) were added to each list to maintain a balance of acceptable and unacceptable items.

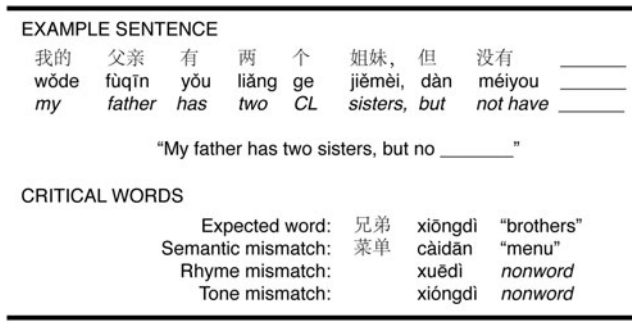

FIGURE 1. Example of sentence stimuli used in the sentence judgment task/ERP experiment.

Experiment 3: Procedures

For all L1 participants and nine L2 participants, Experiment 3 was conducted in the lab at Beijing Normal University (BNU). An additional seven L2 participants were tested in the lab at the University of Maryland (UMD). Where conditions at UMD differed from those at BNU, notes are distinguished using "UMD:".

Participants were seated comfortably in a quiet room in front of a computer monitor. Before the experiment began, participants were shown their EEG output and familiarized with the potential effects of movement and blinking. Audio was delivered by two open speakers (Edifier R1600TIII) placed on either side of the display monitor (UMD: a single audio monitor [JBL LSR305] placed centrally over the display computer monitor). Instructions were given orally in Chinese to all participants, and presented visually on the monitor (in Chinese text for L1 participants, in English text for L2 participants). It was explained that, after hearing a sentence, participants would be required to answer the question "Did this sentence sound OK to you?" (Chinese: *Zhege juzi tingqilai shifou zhengque?*) by pressing "J" for *yes* and "F" for *no*. To make clear what "OK" (*zhengque*) meant, participants completed 10 practice items with feedback. Feedback indicated the correct answer with a brief explanation of the problem if the answer was "no." The researcher also gave oral instructions to make it clear that there were only three ways in which a sentence could sound not "okay": A word could be *mispronounced* either (a) in *segments* (*fayin*) or (b) *tones* (*yudiao/shengdiao*), or (c) a word could be *inappropriate in context* (*bu he yujing*). After practice, the experiment began. Each trial was preceded by a 150 ms beep, at which point a fixation-cross appeared. After a 1,000 ms pause, the sentence was played. When the sentence finished, the fixation-cross disappeared and there was a 2,000 ms silence before the question prompt appeared on the screen. After the participant responded, there was a 3,000 ms silence before the beep cuing the next trial. After every 30 trials, participants were allowed to rest. Not including electrode cap preparation, the entire experiment lasted approximately 50 minutes. During the experiment, the researcher monitored for eye-blinks and other artifacts, reminding participants or adjusting as necessary. The sentences were delivered using Matlab R2013b (MathWorks, 2013) and the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997).

Experiment 3: EEG Recording

Raw EEG was recorded continuously at a 1,000 Hz sampling rate using a Neuroscan SynAmps data acquisition system and an electrode cap (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: FP2 and *no* Oz). Recordings were referenced online to the right mastoid (UMD: left 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 5 kΩ. The EEG and EOG recordings were amplified and digitized online at 1 kHz with a bandpass filter of 0.1-100 Hz.

Experiment 3: SJT Analysis and Results

Before considering the ERP results, we address the results of the accompanying behavioral task, the SJT. In the SJT both the expected and semantic mismatch conditions were expected to be relatively easy, that is, accepting well-formed sentences and rejecting sentences with a semantically ill-fitting word should be easy, assuming listeners comprehend the sentences. In contrast, the two phonological conditions could be quite challenging as listeners had to identify a single mispronounced word occurring at an unpredictable location within a sentence.

SJT reliability was high for all lists (list 1: $\alpha = .96$; list 2: $\alpha = .97$; list 3: $\alpha = .96$; list 4: $\alpha = .90$). Descriptive results of the SJT (Table 6) suggest that L1 listeners generally performed well in all conditions, with some drop in accuracy for sentences in the tonal mismatch condition. In contrast, the L2 group appeared to struggle, even with sentences in the expected word and semantic mismatch conditions. This suggests that comprehending the sentences presented a considerable challenge to at least some of the L2 participants. L2 scores were similarly low on the segmental mismatch, and much lower for the tonal mismatch condition.

Accuracy results were submitted to a generalized linear mixed-effects model with crossed random effects for subjects and items. The dependent variable was accuracy (1, 0). Fixed effects included *condition* (expected word, semantic mismatch,

TABLE 6. Means and standard deviations for accuracy by condition in the SJT

Condition	Group	Mean	*SD*
Expected word	L1	0.92	0.27
	L2	0.71	0.45
Semantic mismatch	L1	0.91	0.28
	L2	0.76	0.43
Segmental mismatch	L1	0.95	0.22
	L2	0.72	0.45
Tonal mismatch	L1	0.84	0.36
	L2	0.40	0.49

segmental mismatch, tonal mismatch) and *group* (L1 or L2) and their interaction. The final model included by-subject and by-item random intercepts, as well as by-subject and by-item random slopes for the effect of condition. A more complex model with by-item random slopes for both condition and group (as in later ERP models) failed to converge.

Model output is reported in Table 7. Here we focus on the critical comparisons between groups and between mismatch conditions, though it is notable that, even for well-formed sentences, L2 was significantly less accurate than L1 (b = -1.66, SE = 0.26, $p < .001$). For mismatch conditions, which required participants to reject sentences, L1 was significantly less likely to respond correctly to tonal mismatches compared to both semantic (b = 0.86, SE = 0.34, p = .011) and segmental mismatches (b = 1.16, SE = 0.31, $p < .001$). Compared to tonal mismatches, semantic mismatches were roughly twice as likely to be correctly rejected (*exp(b)* = 2.36), and segmental mismatches about three times as likely (*exp(b)* = 3.18).

For L2 listeners, correct responses were marginally less likely for segmental than for semantic mismatch sentences (b = 0.39, SE = 0.23, p = .097), and the likelihood of correct rejection for tonal mismatch sentences was significantly lower than both mismatch conditions (L2 semantic vs. L2 tonal: b = 2.14, SE = 0.30, $p < .001$; L2 segmental vs. L2 tonal: b = 1.76, SE = 0.24, $p < .001$). Compared to tonal mismatches, semantic mismatches were roughly nine times as likely to be correctly rejected (*exp(b)* = 8.53), and segmental mismatches about six times as likely (*exp(b)* = 5.79).

TABLE 7. Results of generalized linear mixed effects modeling for SJT

model formula (glmer): accuracy ~ condition * group + (1 + condition | subject) + (1 + condition | item)

	Fixed Effects						Random Effects	
	b	*exp(b)*	*SE*	*z*	*p*		subjects *SD*	items *SD*
Intercept (*expected word/L1*)	2.75	15.61	0.23	12.18	<.001		0.52	0.74
semantic mismatch	0.26	1.29	0.39	0.66	.509		1.13	1.19
rhyme mismatch	0.55	1.74	0.38	1.47	.141		1.03	0.80
tone mismatch	-0.60	0.55	0.33	-1.84	.066	†	0.86	1.36
L2	-1.66	0.19	0.26	-6.31	<.001	*		
L2 × semantic mismatch	0.26	1.30	0.48	0.55	.580			
L2 × rhyme mismatch	-0.42	0.65	0.46	-0.91	.362			
L2 × tone mismatch	-1.02	0.36	0.40	-2.55	.011	*		
Additional comparisons								
L1 semantic – L1 segmental	-0.30	0.74	0.33	-0.92	.359			
L1 semantic – L1 tonal	0.86	2.36	0.34	2.54	.011	*		
L1 segmental – L1 tonal	1.16	3.18	0.31	3.7	<.001	*		
L2 semantic – L2 segmental	0.39	1.47	0.23	1.66	.097	†		
L2 semantic – L2 tonal	2.14	8.53	0.30	7.08	<.001	*		
L2 segmental – L2 tonal	1.76	5.79	0.24	7.25	<.001	*		
L1 semantic – L2 semantic	1.40	4.05	0.37	3.81	<.001	*		
L1 segmental – L2 segmental	2.09	8.05	0.37	5.68	<.001	*		
L1 tonal – L2 tonal	2.68	14.65	0.30	9.08	<.001	*		

* = statistically significant; † = marginal

Between-groups comparisons for mismatch conditions show that L2 was significantly less likely than L1 to correctly reject sentences in semantic ($b = 1.40$, $SE = 0.37$, $p < .001$), segmental ($b = 2.09$, $SE = 0.37$, $p = .011$), and tonal ($b = 2.68$, $SE = 0.30$, $p < .001$) mismatch conditions. These effects were larger for tonal mismatches. Compared to L1 listeners, L2 listeners were about four times less likely to correctly reject semantic mismatch sentences ($exp(b) = 4.05$), about eight times less likely for segmental mismatches ($exp(b) = 8.05$), and about 15 times less likely for tonal mismatches ($exp(b) = 14.65$).

Experiment 3: SJT Discussion

The behavioral results of the SJT suggest that, regardless of group, tones are a subtler cue than rhymes in sentences, or at least that listeners have some difficulty deciding to reject sentences on the basis of a single word with a mismatching tone. While this effect was seen in both groups, its size was about twice as large for L2 listeners, suggesting it is quite difficult for them to correctly reject such sentences. Additionally, the relatively poor performance of the L2 group on well-formed sentences suggests that, as a group, they were performing the task with some difficulty, perhaps due to a lack of sufficient comprehension of target sentences (e.g., indicating that "My father has two sisters, but no *menu*." is *okay*). In fact, after the experiments, a few L2 participants mentioned that they had difficulty understanding "some sentences." Importantly, despite the apparent difficulty of the task, we nevertheless see a L2 pattern consistent with that found in the LDT, namely, significantly worse performance on rejection of tonal mismatches than segmental mismatches.

Experiment 3: EEG Data Processing

All trials were evaluated individually for artifacts using EEGLAB v. 10.2.5.8b (Delorme & Makeig, 2004) and ERPLAB v3.0.2.1 (Lopez-Calderon & Luck, 2014) running under MATLAB R2013b (MathWorks, 2013). Data from two L1 participants and one L2 participant were excluded due to having greater than 50% artifacts on experimental trials (either overall or in a single condition). After exclusion of these participants, artifact rejection affected 12.8% of experimental trials for the L1 group (n = 18) and 17.1% for the L2 group (n = 15). Trial-level data for each subject baselined to the mean of the 100 ms preceding the onset of the critical word was exported for further processing in *R* (R Core Team, 2017). A single average amplitude was obtained for each trial for each subject in the N400 window (300–500 ms) and, additionally, a later window (550–800 ms) for the late positive component (LPC) as clear effects were seen in the L1 responses in this window. These time windows were chosen on the basis of visual inspection of grand average waveforms and consideration of N400 and LPC (P600) effects in previous research. Any trial with an absolute value greater than 50μV was removed. Data from nine central-posterior electrodes (C3, Cz, C4, CP3, CPz, CP4, P3, Pz, P4) were chosen for final analysis as these are electrodes where N400 and LPC effects are commonly observed. After all these steps, the final dataset contained 30,275 data points for the N400 (84.9% out of total possible 35,640 data points), and 30,231 data points for the LPC (84.8%).

Experiment 3: ERP Analysis and Results

Grand average waveforms for all trials are illustrated in Figure 2 from electrode CPz for the L1 and L2 groups. Visually, they suggest strong N400 effects for all mismatch conditions for the L1 group with large LPCs for rhyme and tone mismatch conditions. For the L2 group, there is only slight differentiation between conditions, with the semantic mismatch consistently more negative than the other conditions. Expected word and tone mismatch responses appear largely comparable, with responses to the rhyme mismatch perhaps a bit more negative in the 550–800 ms time window.

Mean amplitude results from the N400 and LPC windows were submitted to linear mixed-effects models with crossed random effects for subjects and items. Both the N400 and LPC models were constructed with the same procedures. The dependent variable was mean amplitude. Fixed effects included *condition* (expected word, semantic mismatch, segmental mismatch, tonal mismatch) and *group* (L1, L2) and their interaction. Final models included by-subject and by-item random intercepts, by-subject random slopes for the effect of *condition*, and by-item random slopes for the effect of *condition*. Restricted maximum likelihood estimation was used for the models reported in the following text. As the best method of obtaining p-values from t-distributions in linear mixed-effects models remains controversial, we treat absolute t-values equal to or greater than 2 as significant, and greater than 1.65 as marginal (Gelman & Hill, 2007). In the N400 and LPC models, *b* estimates represent mean amplitude in μV.

Model results for the N400 (Table 8) indicate that L1 responses to critical words in the expected word condition were positive in amplitude ($b = 1.71$). Relative to this baseline, L1 responses to critical words in mismatch conditions were all significantly more negative (semantic: $b = -2.48$, $SE = 0.79$, $t = -3.13$; rhyme: $b = -2.19$, $SE = 0.86$, $t = -2.54$; tone: $b = -2.09$, $SE = 0.83$, $t = -2.51$). There was no significant difference between rhyme and tone mismatch conditions ($b = -0.10$, $SE = 0.75$, $t = -0.13$). For L2, there were no significant differences between responses to expected words and responses to any of the mismatch conditions (expected vs. semantic: $b = .75$, $SE = 0.73$, $t = -1.02$; expected vs. rhyme: $b = 0.62$, $SE = 0.86$, $t = -0.73$; expected vs. tone: $b = 0.05$, $SE = 0.94$, $t = -0.05$). There was also no difference between rhyme and tone mismatch conditions ($b = 0.05$, $SE = 0.94$, $t = -0.05$).

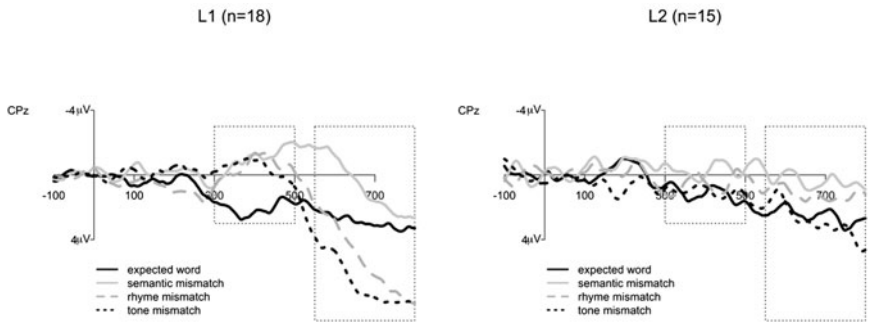

FIGURE 2. Grand average waveforms for electrode CPz. Responses of the L1 group are on the left; responses of L2 group are on the right. All waveforms were generated from the raw EEG data after applying a low-pass filter of 20Hz. The critical 300-500ms window and 550-800ms are indicated in dashed boxes.

TABLE 8. Results of linear mixed-effects model for ERP experiment: 300–500 ms

model formula (lmer): mean.amp ~ 1 + word * group + (1 + word | subject + (1 + word * group | item)

	Fixed Effects				Random Effects	
	b	*Std. Error*	*t-value*		*subjects SD*	*items SD*
Intercept (*expected word/L1*)	1.71	0.64	2.65		1.87	4.77
semantic mismatch	-2.48	0.79	-3.13	*	1.87	6.79
rhyme mismatch	-2.19	0.86	-2.54	*	2.52	6.29
tone mismatch	-2.09	0.83	-2.51	*	2.58	5.64
L2	-0.54	0.94	-0.58		—	6.82
L2 × semantic mismatch	1.73	1.12	1.54		—	9.29
L2 × rhyme mismatch	1.57	1.22	1.28		—	8.44
L2 × tone mismatch	2.04	1.25	1.63		—	8.70
Additional comparisons						
L1 rhyme – L1 tone	-0.10	0.75	-0.13			
L2 expected – L2 semantic	0.75	0.73	-1.02			
L2 expected – L2 rhyme	0.62	0.86	-0.73			
L2 expected – L2 tone	0.05	0.94	-0.05			
L2 rhyme – L2 tone	-0.57	0.79	0.47			

* = statistically significant (|t| > 2.00)

The intercept for LPC model results (Table 9) indicates that the amplitude of L1 responses to expected words in the later time window tended to be positive (b = 2.30). Relative to this response, responses to semantic mismatches were somewhat more negative (b = -1.42, SE = 0.99, t = -1.44), while responses to both rhyme and tone mismatches were significantly more positive (rhyme: b = 2.26, SE = 1.01, t = 2.23; tonal: b = 3.51, SE = 0.90, t = 3.90). There was no significant difference between rhyme and tone mismatch conditions (b = -1.25, SE = 0.91, t = -1.37). For L2, there were no significant differences between responses to expected words and words in any mismatch conditions (semantic: b = 1.32, SE = 0.96, t = 1.38; rhyme: b = 1.30, SE = 0.99, t = 1.32; tone: b = -0.65, SE = 1.00, t = -0.66). However, compared to tone mismatches, responses to rhyme mismatches were significantly more negative (b = -1.97, SE = 1.03, t = -1.92), and semantic mismatches were marginally more negative (b = -1.96, SE = 0.87, t = -2.26).[4]

Experiment 3: Discussion

As expected based on previous research with native Mandarin speakers, the L1 group displayed N400 effects in all mismatch conditions. Additionally, they displayed strong positive LPCs for rhyme and tone mismatches. Though less often discussed, this latter result is consistent with patterns discernable in previous research (cf. central and posterior electrodes in Figure 1 in Malins & Joanisse, 2012) and may be in part attributable to the judgment task used in the present experiment. Though interpretation of the L1 responses is not a priority of the current study, we briefly offer the following explanations. First, the pronounced negative response to the semantic mismatches is

TABLE 9. Results of linear mixed-effects model for ERP experiment: 550–800 ms

model formula (lmer): mean.amp ~ 1 + word * group + (1 + word | subject) + (1 + word * group | item)

	Fixed Effects				Random Effects	
	b	*Std. Error*	*t-value*		*subjects SD*	*items SD*
Intercept (*expected word/L1*)	2.30	0.66	3.48		1.94	4.82
semantic mismatch	-1.42	0.99	-1.44		2.80	7.47
rhyme mismatch	2.26	1.01	2.23	*	2.99	7.32
tone mismatch	3.51	0.90	3.90	*	2.83	5.87
L2	-0.22	0.98	-0.22		—	7.12
L2 × semantic mismatch	0.10	1.39	0.07		—	9.87
L2 × rhyme mismatch	-3.57	1.43	-2.47	*	—	9.91
L2 × tone mismatch	-2.85	1.32	-2.15	*	—	8.54
Additional comparisons						
L1 rhyme – L1 tone	-1.25	0.91	-1.37			
L2 expected – L2 semantic	1.32	0.96	1.38			
L2 expected – L2 rhyme	1.30	0.99	1.32			
L2 expected – L2 tone	-0.65	1.00	-0.66			
L2 rhyme – L2 tone	-1.96	0.87	-2.26	*		
L2 semantic – L2 tone	-1.97	1.03	-1.92	†		

* = statistically significant (|t| > 2.00); † = marginal (|t| > 1.65)

consistent with the behavior of the N400 and indicates more effortful lexical-semantic access caused by the semantically unexpected words (e.g., Kutas & Federmeier, 2000). The negative-to-positive amplitude pattern for both the rhyme and tone mismatches suggests a similar N400 response of surprisal to the unexpected mispronunciations, along with a later attempt to repair the misfitting phonological material (cf. Gibson, Bergen, & Piantadosi, 2013).

In contrast, L2 listeners showed no clear N400 or LPC effects. There were no significant differences between responses to expected words and responses to mismatching words in any of the conditions, though, relative to tones, responses to mismatching rhyme and semantic words did tend to be more negative overall. Interpreted liberally, this tendency might suggest that some L2 listeners were sensitive to rhyme and semantic mismatches some of the time. While we do not wish to base any strong claims on these results, we do note that the overall pattern is not inconsistent with what was found for the same words in the LDT and the behavioral responses in the SJT. In other words, the ERP results present no evidence that L2 listeners have sensitivity to tones in sentence contexts. At the same time, it seems reasonable to conclude that present results reflect the difficulty some participants had comprehending some of the target sentences, thus making word-level ERP effects from our mismatch conditions hard to evaluate.

GENERAL DISCUSSION

The present study set out to test whether perception of Mandarin lexical tones, which is generally thought to be difficult for L2 learners to master, remains difficult for relatively

advanced learners. Given a task that targeted phonetic categorization of isolated tones (Tone ID), these results suggest that most learners can achieve high levels of perceptual ability. In contrast, for tasks that targeted perception of tones in disyllabic words (the LDT and SJT), most learners appear to struggle—even after many years of experience with Mandarin. The discontinuities we observed between lower and higher levels of speech perception are consistent with findings in broader L2 speech perception research (cf. Sebastián-Gallés & Díaz, 2012), and extend those findings to the suprasegmental level.

Before we consider *why* learners might have these difficulties, we first consider individual performance across tasks, and whether performance is related to a learner's vocabulary knowledge.

NEAR-NATIVE L2 PERFORMANCE

To investigate individual performance patterns across tasks, we considered how many L2 learners scored within the range of native listeners on each task, that is, with what we might call near-native performance. We operationalized near-native performance with a generous criterion, setting the lowest individual L1 mean score for each condition in each task as the lower bound. Performance at or equal to this criterion score was considered near-native. Boxplots depicting individual L1 and L2 performance in the Tone ID, LDT, and SJT are depicted in Figure 3.

In the Tone ID, the majority of L2 learners performed at or above the level of the lowest L1 score for each condition, with somewhat weaker performance for T2 (T1 = 16/16; T2 = 10/16; T3 = 14/16; T4 = 16/16). This suggests that, for the current sample of advanced L2 learners, near-native performance on the Tone ID was *the norm*, though T2 may pose long-term difficulties for some learners.[5]

In the LDT, 15 L2 participants scored within the native range for correct acceptance of real words, 11 for rejection of segmental nonwords, but only one for rejection of tonal nonwords with 80% correct rejection. Three other L2 learners correctly rejected more than 50% of the tonal nonwords, with all others below 50% accuracy. In the SJT, seven L2 participants scored within the native range in the expected condition, correctly accepting normally produced sentences. Ten were within the native range for rejecting semantic mismatches, eight for segmental mismatches, and just two for tonal mismatches. In other words, it appears that, even for the best L2 learners in our sample, perception of tones in disyllabic words and sentential contexts remains a major challenge. Of our 16 learners, only one performed at near-native levels in the LDT, and just two in the SJT—despite a very generous near-native criterion.

EXPLICIT VOCABULARY KNOWLEDGE

To successfully reject tonal nonwords in the LDT and SJT, participants needed to know the real words that were manipulated in these tasks, as well as the tones that belong to those words. We attempted to control for effects of L2 knowledge by selecting only highly frequent vocabulary for our stimuli, but we also administered an Explicit Vocabulary Knowledge Test (EVKT) to all L2 participants, after Experiment 3. The EVKT tested explicit knowledge of tones and meanings for the 60 real word counterparts

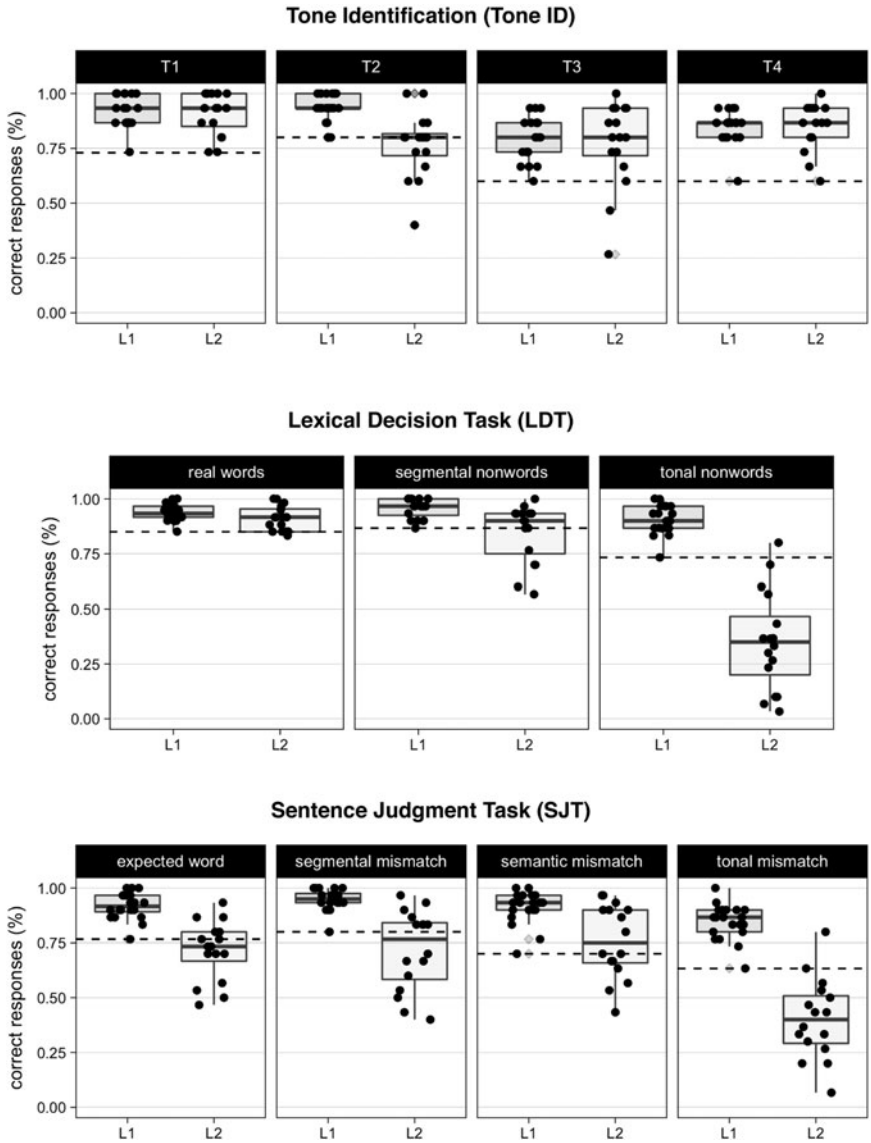

FIGURE 3. Boxplots for the Tone ID (top), LDT (middle), and SJT (bottom). Each black dot indicates an individual's mean performance in the task. Dashed lines indicate the lowest L1 mean score(s). Gray diamonds indicate outliers.

to the tonal nonwords that a participant had encountered in Experiments 2 and 3 (i.e., from the participant's assigned experimental list). The test provided characters and toneless romanization ("Pinyin") of Mandarin, and required participants to provide tones and an English definition for each word (examples in Figure 4). For each syllable of a disyllabic word, a correct tone was awarded one point and no points were awarded for incorrect tones (120 points possible). The average score for tones was 84.7% (sd = 11;

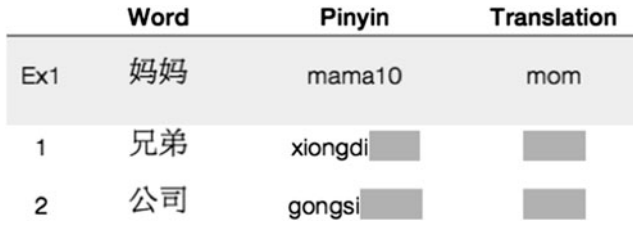

FIGURE 4. Example of test items from the explicit vocabulary knowledge test (EVKT).

range = 54.1–97.5). Three participants knew less than 80% of the tones, but most knew more than 85%. English definitions were also scored one point per correct translation (60 points total). The average score for definitions was 95.2% (*sd* = 6.7; *range* = 78.3–100), with all but three participants knowing 95% (57 out of 60 words) or more of the tested vocabulary. In other words, while there was some inconsistency, in general *participants had strong knowledge of both tones and meanings for target vocabulary*. This indicates that it was not a lack of vocabulary knowledge that drove L2 results in the LDT (though vocabulary difficulties may have played a larger role in comprehension for the sentence level experiment). Importantly, there were no statistical outliers in the L2 group for the LDT or SJT (cf. Figure 3), and the majority of scores in the L2 distribution fell below 50% accuracy, so it is not the case that effects were driven by a few individuals with low vocabulary knowledge.

The fact that even individuals with highly accurate explicit knowledge of tones and vocabulary did not consistently perform at near-native levels in the LDT and SJT suggests that explicit tone and vocabulary knowledge are not enough to master perception of Mandarin tones. (A table with all individual results is available in the supplemental materials online.) We would also suggest that the fact that some advanced learners seemed to have difficulty remembering the appropriate tones for words in the EVKT is consistent with the idea that *encoding and retrieving* tones *lexically* may be a long-term L2 challenge.

WHAT MAKES TONES DIFFICULT?

We believe the results presented here can be interpreted as strong evidence that Mandarin tones are in fact difficult for advanced L2 learners to perceive. However, the nature of the difficulty needs to be specified. It is *not* the case that tones are difficult for learners *to hear* or *categorize as phonetic objects*. Rather, it seems that, given enough time and experience, highly accurate phonetic categorization of tones is within the grasp of most L2 learners. This makes sense insofar as pitch serves as a prosodic feature of languages, even when it is not a lexical cue. It would seem to follow that most normal, healthy individuals will be able to perceive pitch, though they may vary greatly in how well they can learn to *repurpose it as a lexical cue*. We suggest it is in this latter respect that tones are truly difficult for L2 learners. Even more specifically, we suggest that this difficulty is particularly prominent *when phonological tone perception is required of disyllabic (or multisyllabic) words.*

The LDT and SJT examined only disyllabic words, as most monosyllables would not satisfy the word/nonword neighborhood properties required by our design. So, we cannot make a direct comparison of the effects of lexicality for monosyllabic and multisyllabic words based on our results. However, previous research is suggestive. Chang and Bowles (2015) found that disyllabic tone words were much more difficult for their *naïve* participants to learn than monosyllabic words. Similarly, Hao (2012) and Sun (1998) both found that their learners were less accurate in tone identification with disyllabic and multisyllabic stimuli, suggesting some amount of phonetic perception difficulty in multisyllabic contexts, even for more advanced learners. While it might be tempting to conclude from these results that longer always equals harder, the contrast between segmental and tonal nonwords in the current LDT results argues against the idea that disyllabic words are just generally more difficult for advanced learners. Rather, we believe that syllable count may play an oversized role in the difficulty associated with tonal word recognition in L2 Mandarin.

As just noted, Chang and Bowles (2015) specifically examined the role of syllable count in L2 tone perception. They attributed the difficulty they observed to the warping of neighboring tone contours that occurs in disyllables; in other words, they argued for difficulties in what we are calling phonetic perception of tones. We suggest that while phonetic perception may play a strong role for beginners, highly experienced learners are less likely to struggle due to tonal coarticulation. In the current study's Tone ID task, isolated tones were recognized consistently despite the tone-warping effects that came from extracting syllables out of words in sentences. Therefore, we wish to finish with a more speculative discussion about the nature of L1 and L2 tone and word representations, with particular consideration given to the role additional syllables play in the development and utilization of these representations.

First, we consider tone categories. We posit that it is relatively easy for learners to form phonetic categories for the Mandarin tones, but that these tone categories are *not* necessarily easy to encode lexically, especially for multisyllabic words. Figure 5 illustrates how encounters with individual words reinforce abstract tone categories, but that this will not lead to equally robust entries for those words. Each time a word is heard, the learner has the opportunity to reinforce both that word and the relevant tone category (or categories) associated with it. As there are only a handful of tone distinctions, the tone categories become rather robust quite quickly. In contrast, strengthening the connection between a given word and its tone category is a much slower process, highly dependent on word frequency. Thus, as we found in the current study, tone categories can become quite robust without necessarily leading to equally robust word recognition. Figure 5 suggests that tone categories represent single tones, which might be expected to constrain how efficiently they operate for multisyllabic words, which need connections to multiple tone categories. Alternatively, we could entertain the possibility of "compound tone categories" that map multiple tones to multisyllabic words (e.g., the category [T1+T1] vs. [T1+T2] and so on). The increased number of possible tone combinations will necessarily lead to slower reinforcement of these compound categories, and here the role of tonal coarticulation might play a significant role in adding to the difficulty of forming robust categories. In other words, at the level of category formation, there may be reasons for multisyllabic

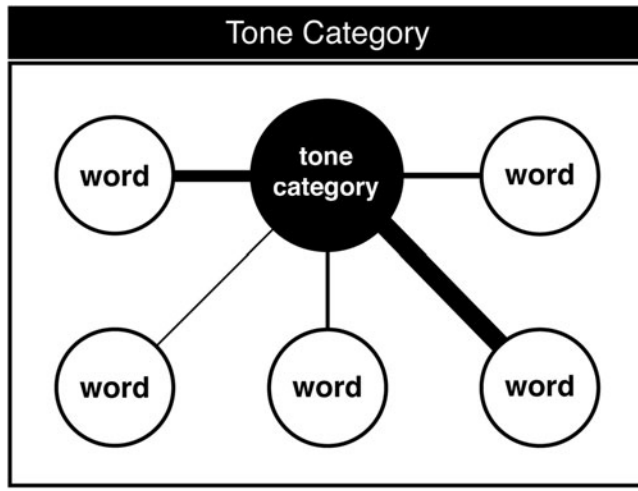

FIGURE 5. Tone category representation. Many words link to a single tone category. The category itself becomes quite robust, but the links between the category and individual words vary in strength depending on word frequency.

tonal words to present learners with more than a simple linear increase in difficulty over monosyllabic words.

A second way to think about tone difficulty focuses on the processing biases of L2 listeners. We assume that during spoken word recognition, individual phonetic cues are given different weights by the listener. For a nontonal language listener, pitch differences receive relatively minimal weight as lexical cues. In fact, even for native Mandarin listeners, tone cues may be relatively uninformative due to the small number of possible tone contrasts (as well as the large number of monosyllabic homophones, cf. Wiener & Ito, 2016). In other words, even for native listeners, once the segmental structure of a syllable is recognized, semantic representations can often be accessed *regardless of tone*. Critically, tone categories become even less informative as words increase in number of syllables, as there is much lower likelihood of tonal minimal pairs occurring. These distributional properties of Mandarin do little to push back against a native bias to attend to segmental cues and dismiss pitch cues. This may make it exceedingly difficult for L2 listeners to learn to reweight tone cues during Mandarin word recognition.

Both the representational and processing accounts in the preceding text predict that, while not impossible, tones will be exceedingly difficult for L2 learners to acquire at the level of automatized (or perhaps implicit) knowledge, and that, all else being equal, tones on multisyllable words will be persistently more difficult than on monosyllabic ones. In this regard, it is notable that, in the present study, accurate explicit knowledge of tones was apparently enough for success in the Tone ID, but was insufficient in the LDT, even though nothing in the task disallowed its use.

Our accounts do not align easily with previous research of L2 tone categories, which has been framed in the context of models such as PAM L2 (e.g., So & Best, 2010, 2014, attempt to account for the influences of nontonal L1 intonation categories on perceptual assimilation patterns for Mandarin tones), or through application of Optimality Theory (OT) (Zhang,

2016). While those approaches have the potential to help us understand the formation of specific L2 tone categories, in their current form, they are a bit orthogonal to the questions we discussed previously, as they do not address the distinction we highlight between phonetic and phonological perception. That is, while PAM or OT might give us a way to understand why, for example, T2 is particularly difficult for L2 learners to identify in the Tone ID, they do not seem to help us understand the LDT results where tone *in general*—but not any specific tone—is difficult to apply in L2 *word recognition*. As laid out in the preceding text, we believe L2 tone category formation and the difficulties associated with it cannot be understood without careful consideration of the character of the L2 lexicon. So far, this has not been a part of the PAM or OT accounts of L2 tone category formation.

Future research should aim to flesh out and formalize the preceding accounts and test their predictions.

LIMITATIONS

An obvious limitation of the present study is the relatively small number of L2 participants, which necessarily constrains the force of generalizations made with the current data. At the same time, we note that the size of observed effects in the LDT and SJT suggests the key patterns of tone-related results would be unlikely to change with additional participants.

Additionally, it must be conceded that the current sample may be biased by self-selection, that is, it may be that only (or mostly) participants who initially excel at tones will continue to study until they reach advanced levels. This limits how confidently we can generalize present Tone ID results to the broader population of L2 learners, particularly in light of research with naïve learners that shows the strong influence individual differences can play in initial tone learning (e.g., Bowles et al., 2016; Li & DeKeyser, 2017; Perrachione et al., 2011). However, this limitation only strengthens the force of results for our lexical and sentential tasks. If *even the most successful learners* find tones persistently difficult in word recognition, then learners who fail to perceive tones phonetically will certainly find this to be a major challenge.

Finally, though we have described L2 participants in the current study as "advanced" learners, we are not claiming that they have reached ultimate attainment, or that stronger L2 tone performance is not possible. However, we do believe these participants represent a normal range of outcomes for L2 learners of Mandarin after years of concerted effort.

CONCLUSION

The results reported here strongly suggest that lexical tones are difficult even for advanced L2 learners of Mandarin. Importantly, this difficulty does not seem to be due to phonetic perception of tone categories, as nearly all participants were near-native in their ability to categorize tones, despite the challenge presented by our naturalistic stimuli. Instead, L2 tone difficulty appears related to the lexical encoding and retrieval of tones in multisyllabic words. This type of difficulty persists, even when learners have highly accurate explicit knowledge of tones and vocabulary. In this sense, tone perception appears to present a considerable challenge for L2 learners.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0272263117000444

NOTES

[1]*Note on order of experiments:* Though tone identification is presented first for ease of exposition, the order of the experiments in the lab was (a) ERP, (b) lexical decision, then (c) tone identification, followed by the explicit vocabulary knowledge test.

[2]Response times were recorded, but are not reported here due to space limitations.

[3]Response times were recorded. However, as we expected a high error rate for L2 tonal nonwords, we did not plan to analyze response times and they are not reported here.

[4]One reviewer recommended examination of only correct trials. As this would result in a loss of more than 30% of L2 data, with 60% of L2 data lost for the critical tone mismatch condition, we have not pursued it here. However, for interested readers, we have tried to address this issue with an additional analysis in an online supplement (Appendix A5). Despite added complexity, the additional analysis does not reveal substantial differences with the models reported here.

[5]Two reviewers expressed some concern about the fixed order of experiments, with Tone ID always occurring last. They suggested participants might have grown more attuned to tones as the study progressed, thus performing more strongly on the Tone ID. While it is impossible to entirely rule this out, we do not feel it is likely. Whereas the sentence experiment lasted nearly an hour, the LDT took less than 10 minutes. If there was a learning effect for Tone ID, why not also for the LDT?

REFERENCES

Abrahamsson, N. (2012). Age of onset and nativelike L2 ultimate attainment of morphosyntactic and phonetic intuition. *Studies in Second Language Acquisition*, *34*, 187–214.

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*, 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Bowles, A. R., Chang, C. B., & Karuzis, V. P. (2016). Pitch ability as an aptitude for tone learning. *Language Learning*, *66*, 774–808.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 443–446.

Broersma, M. (2012). Increased lexical activation and reduced competition in second language listening. *Language and Cognitive Processes*, *27*, 1205–1224.

Broersma, M., & Cutler, A. (2008). Phantom word activation in L2. *System*, *36*, 22–34.

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

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*, 103–135.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS One*, *5*, e10729.

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*, 456–65.

Chang, C. B., & Bowles, A. R. (2015). Context effects on second-language learning of tonal contrasts. *Journal of the Acoustical Society of America*, *136*, 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*, 415–455.

Cook, S. V., & Gor, K. (2015). Lexical access in L2: Representational deficit or processing constraint? *The Mental Lexicon*, *10*, 247–270.

Darcy, I., Daidone, D., & Kojima, C. (2013). Asymmetric lexical access and fuzzy lexical representations in second language learners. *The Mental Lexicon*, *8*, 372–420.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*, 9–21.

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

Duanmu, S. (2007). *The phonology of standard Chinese* (2nd ed.). New York, NY: Oxford University Press.

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America*, *97*, 3125–3134.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*, 8051–8056.

Hao, Y.-C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, *40*, 269–279.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*, 346–363.

Huang, T., & Johnson, K. (2010). Language specificity in speech perception: Perception of Mandarin tones by native and nonnative listeners. *Phonetica*, *67*, 243–267.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*, 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*, 920–933.

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*, 1–15.

Lee, C.-Y., Tao, L., & Bond, Z. S. (2010). Identification of acoustically modified Mandarin tones by non-native listeners. *Language and Speech*, *53*, 217–243.

Li, M., & DeKeyser, R. (2017). Perception practice, production practice, and musical ability in L2 Mandarin tone-word learning. *Studies in Second Language Acquisition*, *39*, 563–620.

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*, 213.

Lukianchenko, A. (2014). *From sound to meaning: Quantifying contextual effects in resolution of L2 phonolexical ambiguity*. College Park: University of Maryland.

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*, 2032–2043.

Mathworks, USA. (2013). *MATLAB and statistics release 2013a*. Natick, MA: The MathWorks, Inc.

Pallier, C., Colomé, A., & Sebastián-Gallés, N. (2001). The influence of native-language phonology on lexical access: Exemplar-based versus abstract lexical entries. *Psychological Science*, *12*, 445–449.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *Journal of the Acoustical Society of America*, *130*, 461–472.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

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*, 1–12.

Sebastián-Gallés, N., & Díaz, B. (2012). First and second language speech perception: Graded learning. *Language Learning*, *62*, 131–147.

Showalter, C. E., & Hayes-Harb, R. (2013). Unfamiliar orthographic information and second language word learning: A novel lexicon study. *Second Language Research*, *29*, 185–200.

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*, 273–293.

So, C. K., & Best, C. T. (2014). Phonetic influences on English and French listeners' assimilation of Mandarin tones to native prosodic categories. *Studies in Second Language Acquisition*, *36*, 195–221.

Sun, S. H. (1998). *The development of a lexical tone phonology in American adult learners of standard Mandarin Chinese*. Honolulu, HI: Second Language Teaching & Curriculum Center.

Wang, X. (2013). Perception of Mandarin tones: The effect of L1 background and training. *The Modern Language Journal*, *97*, 144–160.

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., & Ito, K. (2016). Impoverished acoustic input triggers probability-based tone processing in mono-dialectal Mandarin listeners. *Journal of Phonetics*, *56*, 38–51.

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

Xu, Y. (1997). Contextual tonal variation in Mandarin. *Journal of Phonetics*, *25*, 61–83.

Zhang, H. (2014). The third tone: Allophones, sandhi rules and pedagogy. *Journal of the Chinese Language Teachers Association*, *49*, 117–145.

Zhang, H. (2016). Dissimilation in the second language acquisition of Mandarin Chinese tones. *Second Language Research*, *32*, 427–451.

Zhang, L. (2011). Meiguo liuxuesheng Hanyu shengdiaode yinwei he shengxue xinxi jiagong. *Shijie Hanyu Jiaoxue Chinese Teaching in the World*, *25*, 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*, 1761–1770.

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*, *20*, 1017–1029.