ABSTRACT

Title of Dissertation: SECOND LANGUAGE LEXICAL REPRESENTATION AND PROCESSING OF MANDARIN CHINESE TONES

Eric A. Pelzl, Doctor of Philosophy, 2018

Dissertation directed by: Professor Robert DeKeyser
Second Language Acquisition

This dissertation investigates second language (L2) speech learning challenges by testing advanced L2 Mandarin Chinese learners' tone and word knowledge. We consider L2 speech learning under the scope of three general hypotheses. (1) *The Tone Perception Hypothesis:* Tones may be difficult for L2 listeners to *perceive* auditorily. (2) *The Tone Representation Hypothesis:* Tones may be difficult for L2 listeners to *represent* effectively. (3) *The Tone Processing Hypothesis:* Tones may be difficult for L2 listeners to *process* efficiently.

Experiments 1 and 2 test tone perception and representation using tone identification tasks with monosyllabic and disyllabic stimuli with L1 and advanced L2 Mandarin listeners. Results suggest that both groups are highly accurate in identification of tones on isolated monosyllables; however, L2 learners have some difficulty in disyllabic contexts. This suggests that low-level auditory perception of tones presents L2 learners with persistent long-term challenges. Results also shed

light on tone representations, showing that both L1 and L2 listeners are able to form abstract representations of third tone allotones.

Experiments 3 and 4 test tone representation and processing through the use of online (behavioral and ERP) and offline measures of tone word recognition. Offline results suggest weaknesses in L2 learners' long-term memory of tones for specific vocabulary. However, even when we consider only trials for which learners had correct and confident explicit knowledge of tones and words, we still see significant differences in accuracy for rejection of tone compared to vowel nonwords in lexical recognition tasks. Using a lexical decision task, ERP measures in Experiment 3 reveal consistent L1 sensitivity to tones and vowels in isolated word recognition, and individual differences among L2 listeners. While some are sensitive to both tone and vowel mismatches, others are only sensitive to vowels or not at all. Experiment 4 utilized picture cues to test neural responses tied directly to tone and vowel mismatches. Results suggest strong L1 sensitivity to vowel mismatches. No other significant results were found.

The final chapter considers how the three hypotheses shed light on the results as a whole, and how they relate to the broader context of L2 speech learning.

SECOND LANGUAGE LEXICAL REPRESENTATION AND PROCESSING OF MANDARIN CHINESE TONES

by

Eric A. Pelzl

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Robert DeKeyser, Chair
Professor Ellen Lau
Professor Kira Gor
Professor Scott Jackson
Professor Colin Phillips, Dean's Representative

Dedication

To family and friends—especially those friends I've made along the way while learning and teaching Chinese.

Acknowledgements

Six years ago I arrived at UMD intent on doing research in Chinese as a second language. At the time, I vowed there were two issues I would never research: tones and characters. The list of people I need to thank for making it possible for me to break (half) my vow is long.

Thanks, first of all, to my committee: My advisor Robert DeKeyser who has been unfailingly supportive, patient, and constructive—qualities that I do not take for granted. Your guidance and example have always made me a better scholar. Ellen Lau who went far beyond the expected duties of an IGERT mentor and enriched my intellectual experience at UMD in every way. I will greatly miss our weekly meetings. Kira Gor whose enthusiasm for research and all things phonolexical created an exciting environment in which to pursue tone research. Scott Jackson who always managed to find time to answer my almost endless stream of questions about R and mixed-effects models. Finally, Colin Phillips whose passion for language science and running got me more involved in both. Though not on my committee, Bill Idsardi gets special mention for willingly spending a semester teaching me phonology one-on-one and humoring me by making tones the main topic of the class.

Many classmates/friends also played significant roles in making my years here enjoyable and intellectually rewarding. Chief among them is Alia Lancaster who I was privileged to spend time with in classes, CASL offices, DIT offices, and at board game parties. Alix, Chris, Peter, Peter, Rachel, Zoe and some other cats who joined or upended (Bruce!) board game nights and were always a pleasure to spend time with. Along with Zoe, the other best officemates I ever had were Anton and Lara. Along

with Alia the other best co-workers I ever had were Martyn and Megan. I will miss lunches and happy hours with all of you. Other friends who made the years better by their presence include Payman, Yuichi, Stephen, Başak, Nick, Natalia, Allyson, Man Li, Nick, Chia-Hsuan, Paulina, and Bradford.

Thanks to my friends and colleagues in Beijing: Guo Taomei at Beijing Normal University for generously opening her lab to my visits and making sure I had all the support necessary to complete research projects in limited windows of time. Wu Junjie, Chen Mo, Fu Yongben, Kang Chunyan, Lu Di, Li Shuhua, Zhang Zhaoqi, and Zhang Liang for technical help with EEG, Matlab, and participant recruitment, and more nuanced assistance with navigation of local bureaucracy (yuck!) and restaurants (yum!), as well as some memorable game and trivia nights. Thanks also to Zhang Kai, Kevin Fedewa, Shianna Fairbanks, Chrissy Stouder, John Wendland, Christine Swanson, Glen Thompson, Ma Yinqiu, all my EAPSI friends and many others who helped make time in Beijing fly by far too quickly.

I am also grateful to the National Science Foundation for financial support that enriched my education and research, and made trips to Beijing possible (NSF-IGERT grant 0801465, NSF-EAPSI grant 1514936, and Ling-DDRI grant 1728851).

Living for six years in a house shared with four other people resulted in some of the best friendships I've had, and I'm very grateful for all the good times that we shared. Kate shared a home with me for as long as anyone outside my family ever has and I'm forever in her debt for the fine drinks and even finer conversations that took place in the kitchen and living room on Harvard Road. Those good times were only

made better when Joe, Stephen, and/or Chris joined in. Thanks to all of you for keeping me sane.

Sanity was also preserved by near-annual escapes to the best board-gaming destination in southern Minnesota at Don Zimmerman's home, as well as side trips to the Ebeling and Kuehn homes in Milwaukee. Visits with old friends from Middlebury—Di Ruihe, He Xiaoman, An Long, Diao Wenhao, Bai Jianhua—were much too rare, but always wonderful.

Finally, profound thanks and love to my family, especially my parents, who have never hesitated in loving and supporting me as I travelled across oceans and made dramatic course changes in my life.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Overview

Mandarin Chinese is a *lexical tone language*. This means that in Mandarin, along with vowels and consonants, pitch (F0) contrasts are one of the basic phonological elements of words. For a word's phonological form to be complete, it is not enough for a Mandarin speaker to encode consonants (C) and vowels (V) in long-term memory, tones (T) must also be specified. For example, for the word meaning 'mom', the CV representation /ma/ is not a complete phonological form, the representation must include a tone (CVT): /ma1/. Said another way, all words in Mandarin are *tone words*, that is, they include tone as an essential feature. Even if the tone does not distinguish a given word from others, that word must still have a tone.

This dissertation takes the case of Mandarin tone as a lens through which to consider second language (L2) speech learning challenges. Importantly, in the case of non-tonal language learners such as those who will be the focus of this dissertation, a potential contrast between tone learning and other instances of difficult L2 phoneme learning is that tones may be best understood as a natural class of sounds that lie outside the learners native language (L1) phonological system. This means that not only do L2 listeners need to learn to categorize these sounds in ways they previously did not, they must also reorganize their phonological system to utilize tones as essential *lexical* cues.

By testing the tone and word knowledge of advanced L2 Mandarin learners, this dissertation aims to shed new light on L2 speech learning, and to refine the

discussion of L2 Mandarin tone learning by constraining the space of discussion under the scope of three general hypotheses. These hypotheses are:

(1) *The Tone Perception Hypothesis:* Tones may be difficult for L2 listeners to *perceive* auditorily, that is, hearing differences between linguistic tones is difficult.

(2) *The Tone Representation Hypothesis:* Tones may be difficult for L2 listeners to *represent* in an effective way, that is, L2 tone categories may be poorly formed or not encoded directly in the phonological form of mental lexical entries.

(3) *The Tone Processing Hypothesis:* Tones may be difficult for L2 listeners to *process* efficiently, that is, native language (L1) processing biases may impede L2 ability to fully utilize tone cues in real time.

We will consider each of these in some depth later on, but it is worth noting up front that these hypotheses are *not* mutually exclusive. So while this dissertation hopes to provide some limitations on the scope of individual hypotheses, and perhaps bolster evidence that favors the explanatory power of one over another, it would be too ambitious to expect that a series of four experiments will allow us to conclusively exclude any one of them from further consideration. Such an aim is unrealistic in any case. L2 learning is complex, and at different points in tone learning for different individuals, different difficulties may play a larger or smaller role. Consequently, this dissertation will primarily view these hypotheses as a useful framework that can sharpen our discussion of tone learning difficulties. To this end, I will consider how the evidence presented in experimental results below might influence us to favor one

or another account, but I will also suggest ways in which multiple hypotheses could play a role.

Taking previous research (Pelzl, Lau, Guo, & DeKeyser, 2018) as a starting point, this dissertation will present a series of four experiments. Experiment 1 and 2 (Chapter 2) will address the Tone Perception Hypothesis through the use of tone identification tasks using monosyllabic and disyllabic stimuli with L1 and advanced L2 Mandarin listeners. As we will review below, previous research provides evidence that isolated monosyllabic tone identification is not a serious challenge for advanced L2 learners, but it remains possible that *contextual* tone identification (in disyllabic or longer strings) may present considerable difficulties. Experiments 1 and 2 will also address the Tone Representation Hypothesis through examination of allotonic variation of the Mandarin third tone. Here the assumption is that if L2 learners have acquired nativelike tone representations, then they should represent allotones abstractly as a single tone category. As this assumption has not been previously tested in L1 perception, Experiments 1 and 2 will also explore allotone perception in native Mandarin speakers.

Chapters 3 and 4 will target the Tone Representation and Processing Hypotheses through the use of online and offline measures of Mandarin tone word recognition. The aim of these experiments will be to test the quality of advanced L2 learners' offline lexical tone knowledge, and to examine how that knowledge is or is not deployed in real time processing. These experiments will use both behavioral (accuracy) and neural (ERP) measures to provide insight into the online responses of L1 and L2 listeners to nonwords that differ from real words in a single tone.

Additionally, by examining the role of individual tones in nonword confusions, we will shed further light on the Tone Perception Hypothesis.

Finally, in Chapter 5 we will step back once again to consider how the three guiding hypotheses help us to interpret the results of all the experiments as a whole, both in the context of L2 Mandarin learning, as well as in the broader context of L2 speech learning.

1.2 The four tones of Mandarin Chinese

Mandarin Chinese is a lexical tone language in which *pitch distinctions* that accompany syllables indicate *lexical distinctions*. These pitch distinctions are typically characterized along two dimensions: height (high or low) and contour (rising, falling, or some combination). Additional features that can help distinguish lexical tones include duration, intensity, and voice quality (i.e., creakiness, cf. R. Yang, 2015), though pitch is usually considered the primary cue (Duanmu, 2007).

In modern standard Mandarin, there are four canonical tones, by convention labeled with the numbers one through four (Figure 1.1). Tone 1 (T1) is a *high-level* tone, tone 2 (T2) is a *rising* tone, tone 3 (T3) is a *low* tone, tone 4 (T4) is a *falling* tone.[1] The pitch contour of T3 can vary depending on contextual factors, an issue I will revisit in detail in Chapter 2. Like vowel and consonant distinctions, lexical tones are *phonemic* (or *tonemic*), that is, they can critically distinguish one word from another (e.g., *ma1* 'mom', *ma2* 'hemp', *ma3* 'horse', *ma4* 'to scold'). In addition to the four full tones, there is a *neutral tone* that can occur on unstressed syllables

[1] Below, I will use italicized Pinyin romanization with numbers to indicate the tone of a given syllable (e.g., *ma1*).

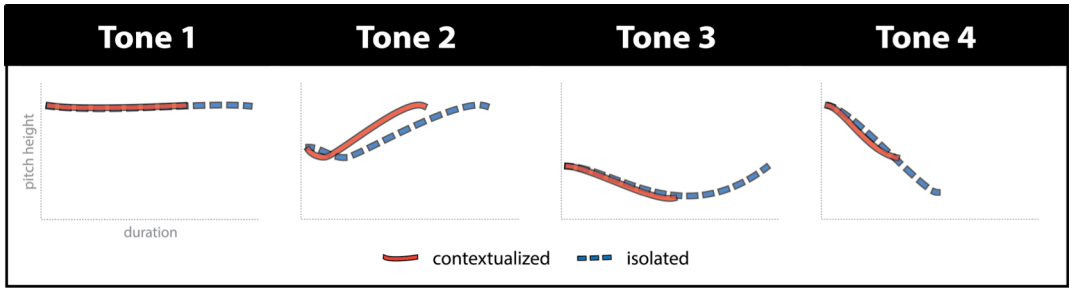


Figure 1.1 The four tones of Mandarin Chinese

(Chao, 1968). The pitch of the neutral tone varies depending on its context, with its specific pitch being determined largely by the pitch of the preceding syllable (Chen & Xu, 2006; Lee & Zee, 2008).

1.2.1 Tone learning

Perceiving, representing, and processing lexical tones comes naturally (after childhood acquisition) to L1 Mandarin speakers. For them, tone distinctions are simply a fact of word recognition. For L2 learners, however, each new word of Mandarin requires the learning of a feature that was not originally necessary in word recognition. In this dissertation I will often use the term *tone word* to highlight the role of tone in L2 word learning.[2] The contention is that, if we exclude tones from our metrics, most L2 Mandarin word learning is successful. Furthermore, in practice, it will often be enough for the learner to encode the segments of Mandairn words without encoding tones. As has been pointed out by others (Wiener & Ito, 2015, 2016), many syllables in Mandarin only ever occur with one of the four tones (e.g.,

[2] Note that I am not using this term in the sense of 'tone-word' found in Cooper & Wang (2013). For them, a tone-word must form a minimal tone contrast with another word. In my usage, this contrastive aspect is not necessary.

neng2 ‘can’). In that sense, the tone will never differentiate those words from other real words. However, that does not mean those words can be produced with other tones without at least sounding odd. If non-distinctive tones are perhaps the exception among monosyllables, for disyllabic Mandarin words, relatively few are even minimally differentiated by tones. Nevertheless, for the L2 learner, completely successful learning of any Mandarin word requires learning its tone(s). Thus, it is tone word learning.

1.2.2 General factors that influence L2 perception of Mandarin tones

There has been considerable research on L2 (often completely naïve) perception of Mandarin tones, with studies using both observational and experimental designs (see Tables A1.1 and A1.2 in Appendix A). Here I highlight a few key findings to position us in the broader context before we consider advanced L2 tone learning.

First, results across studies are consistent in showing the pervasive influence of L1 (particularly non-tonal L1) on L2 perception of tones (Hallé, Chang, & Best, 2004; Leather, 1987; L. Lee & Nusbaum, 1993; Y.-S. Lee, Vakoch, & Wurm, 1996; Repp & Lin, 1990; Schaefer & Darcy, 2014; Stagray & Downs, 1993; Y. Wang, Jongman, & Sereno, 2001). Second, individual differences in aptitude for pitch-contour perception and musical experience can lead to strikingly different outcomes in tone training studies with naïve learners (*aptitude:* Ingvalson, Barr, & Wong, 2013; Perrachione, Lee, Ha, & Wong, 2011; Wong & Perrachione, 2007; *music experience:* Alexander, Wong, & Bradlow, 2005; Gottfried, 2007; Wong, Skoe, Russo, Dees, & Kraus, 2007; Zhao & Kuhl, 2015, *and many others*). Importantly, these advantages

lead to better tone word learning outcomes (e.g., Bowles, Chang, & Karuzis, 2016), at least for minimal vocabularies comprising a few dozen words. However, the long-term impacts of these individual differences are still unclear, and some training studies suggest a relatively small amount of initial *non-lexical* tone identification training may help non-musicians and learners with lower pitch aptitude to catch up (*Mandarin:* Ingvalson et al., 2013; *Cantonese:* Cooper & Wang, 2013). Finally, while most of the studies above investigate overt tone perception or learning, some recent research also suggests that L2 learners are able to track probabilistic distributions of tones over individual syllables in Mandarin-like artificial languages (Potter, Wang, & Saffran, 2016; Wiener, Ito, & Speer, 2016a, 2016b, 2018), though not all studies of this type have resulted in equal success for learners (Caldwell-Harris, Lancaster, Ladd, Dediu, & Christiansen, 2015; Ong, Burnham, & Escudero, 2015; Wang & Saffran, 2014), with previous L2 (tone) language experience appearing to play a significant role in outcomes (Potter et al., 2016; Wiener et al., 2016b).

1.3 Difficult sounds in L2 word learning

In the past two decades a growing body of research has examined cases of difficult L2 phonemes that have strong effects on learners' abilities to efficiently recognize words in the L2 (Amengual, 2016; Barrios, Jiang, & Idsardi, 2016; Barrios, Namyst, Lau, Feldman, & Idsardi, 2016; Broersma & Cutler, 2008, 2011, 2011; Chrabaszcz & Gor, 2014; Cutler, Weber, & Otake, 2006; Darcy, Daidone, & Kojima, 2013; Darcy et al., 2012; Díaz, Mitterer, Broersma, & Sebastián-Gallés, 2012; Escudero, Hayes-Harb, & Mitterer, 2008; Lukianchenko, 2014; Ota, Hartsuiker, & Haywood, 2009; Pallier, Colomé, & Sebastián-Gallés, 2001; Weber & Cutler, 2004).

For example, native speakers of standard Dutch were found to have difficulty distinguishing the vowels /ɛ/ and /æ/ in L2 English (Cutler, Weber, Smits, & Cooper, 2004). Further research has shown that difficulties with these vowels at the *phonetic* level regularly lead to various types of confusion at the *lexical* level (Sebastián-Gallés & Díaz, 2012), ranging from confusion of minimal pairs (e.g., Díaz, Mitterer, Broersma, & Sebastián-Gallés, 2012), to more subtle but potentially more pervasive effects in the activation of spurious competitors during lexical competition (e.g., Broersma & Cutler, 2011; Weber & Cutler, 2004). Similar issues have been examined with Spanish-Catalan bilinguals' perception of Catalan vowels /e/ and /ɛ/ (e.g., Pallier et al., 2001), Japanese speakers' perception of English /l/ and /r/ (e.g., Cutler et al., 2006), English speakers' perception of German rounded vowels and Japanese long and short consonant distinctions (Darcy et al., 2013), and English speakers' perception of Russian hard and soft consonant contrasts (e.g., Chrabaszcz & Gor, 2014; Lukianchenko, 2014).

The L2 challenge posed by the target phonemes in these studies ranges from fairly mild (e.g., Darcy et al., 2013), to a near-complete inability of learners to distinguish sounds (e.g., Brown, 1998). In general, it seems reasonable to assume that the severity of the effects a phonemic distinction has on lexical recognition will correspond to a large degree with the difficulty listeners have in auditory perception of that distinction. However, in this sense the L2 acquisition of Mandarin tones may be interestingly different from the cases mentioned above. While tones are certainly unfamiliar to English speakers, they may not always present the same degree of low-level (i.e., acoustic/phonetic) perceptual challenge as difficult phonemes in other L2s.

Nevertheless, they still appear to create significant challenges for L2 lexical encoding and speech processing (Pelzl et al., 2018). In other words, it may be that the most persistent difficulties in L2 Mandarin tone learning occur at the intersection of *phonological learning* (i.e., acquiring the sound categories of the L2), and *word learning*—at what is sometimes called the 'phonolexical' level (Chrabaszcz & Gor, 2014; Cook & Gor, 2015). Whereas *perceiving* pitch differences is not necessarily difficult for L2 listeners, particularly more experienced ones, L2 tone acquisition is nevertheless persistently difficult due to the need to *repurpose* pitch as a lexical cue. This repurposing (or its limitations) impacts lexical encoding and lexical retrieval—and the impact is substantial since *all Mandarin words are tone words*.

1.4 Evidence that L2 phonetic perception of tones may be *relatively* easy

It is generally taken for granted that perception of lexical tones poses a severe difficulty for most L2 learners. In previous research, I have argued, contrary to this belief, that perception of phonetic properties of lexical tones is *relatively easy* for most L2 learners (Pelzl et al., 2018). This claim is intended to be understood in two ways. First, relative to other documented L2 speech learning challenges, Mandarin tones appear only moderately difficult. Second, relative to the difficulty L2 learners have in repurposing Mandarin tones for *lexical* representation and processing, merely perceiving phonetic differences between tones may not be so challenging.

Evidence that supports the first point comes from two main sources. First, there is a logical argument. As pitch is a universal feature of language for conveying emotion, and, in the case of many languages (e.g., English) functions as a prosodic cue for stress on words and in phrases, normal-hearing individuals are expected to be

capable of perceiving pitch differences. This contention receives some support in research with infants that finds, unlike many other phonological distinctions to which infants seem to become less sensitive as they hone in on their native language(s), pitch perception appears to remain relatively robust, though it does vary depending on the salience of specific tone contrasts (L. Liu & Kager, 2014; R. Shi, Santos, Gao, & Li, 2017).

The second source of evidence is empirical, drawn from the large body of previous research on non-native tone perception. This research suggests that even completely naïve English listeners regularly perform well above chance in tone identification and discrimination tasks (Table 1.1), and display strong sensitivity to tone contrasts (cf. the relatively strong d-prime scores in Huang & Johnson, 2010). Similarly, a large body of training studies demonstrate substantial improvement in tone identification after relatively minimal training (e.g., Chang & Bowles, 2015;

Table 1.1 Accuracy rates reported for naïve listeners in discrimination and identification tasks. Participant L1 is English. Percentages rounded to nearest whole number.

Study	L2 Experience	Task	Target	Accuracy	Chance	> Chnc
Alexander, Wong, & Bradlow (2005)	naïve (musicians)	2AFC	overall	89%	50%	Yes
	naïve (non-musicians)	2AFC	overall	69%	50%	Yes
	naïve (musicians)	AX	overall	87%	50%	Yes
	naïve (non-musicians)	AX	overall	71%	50%	Yes
Broselow, Hurtig, & Ringen (1987)	naïve	4AFC	T1	81%	25%	Yes
		4AFC	T2	67%	25%	Yes
		4AFC	T3	78%	25%	Yes
		4AFC	T4	94%	25%	Yes
Bent, Bradlow, & Wright (2006)	naïve	4AFC	overall	59%	25%	Yes
Gottfried (2007)	naïve (musicians)	4AFC	overall	48%	25%	Yes
	naïve (non-musicians)	4AFC	overall	39%	25%	Yes
So & Best (2010)	naïve (non-musicians)	4AFC	T1	69%	25%	Yes
		4AFC	T2	52%	25%	Yes
		4AFC	T3	60%	25%	Yes
		4AFC	T4	19%	25%	No
Lee, Vakoch, & Wurm (1996)	naïve	AX	same	96%	50%	Yes
		AX	different	74%	50%	Yes

AX = same/different discrimination task; 2AFC = two alternative forced-choice identification task; 4AFC = four alternative forced-choice identification task

Wang, Spence, Jongman, & Sereno, 1999). Research with experienced learners provides further evidence that simple tone identification is not particularly difficult. Lee, Tao, & Bond (2010) report that third-year learners achieved near-perfect results on a tone identification task, with a notable exception for T2 (four-alternative forced choice (4AFC): T1= 97%; T2=75%; T3=92%; T4=100%). L. Zhang (2011) similarly reports that learners who had spent between 20-25 months studying in China achieved roughly 95% accuracy on a tone identification task (4AFC), and even learners who had spent much less time in China (7-9 months) achieved roughly 94% accuracy on this task. The performance of both L2 groups was not statistically significantly different from that of native speakers, though their response times were slower. Perhaps the strongest evidence to date comes from Pelzl et al. (2018). Using fast, co-articulated monosyllabic stimuli clipped from disyllabic words, we found that advanced learners' L2 tone identification accuracy was largely similar to that of native speakers (who were notably also not at ceiling), with only accuracy for T2 being significantly lower than that of native speakers (Figure 1; T1: L1 *mean*=92%, L2 *mean* = 91%; T2: L1 *mean* =93%, L2 *mean* =77%; T3: L1 *mean* =79%, L2 *mean* =77%; T4: L1 *mean* =85%, L2 *mean* =85%). Again, this contrasts with what is found in most cases of difficult phonemes in the L2 literature, where even advanced learners still make a large number of errors (e.g., Brown, 1998; Cutler et al., 2004).

In short, Mandarin tones do no seem to present learners with a low-level perceptual challenge of the same magnitude as difficult phonemes in other reported L2 cases, such as English /r/ and /l/ for Japanese learners (e.g., Brown, 1998), English /æ/ and /ɛ/ for Dutch learners (e.g., Díaz et al., 2012), or Russian palatal consonants

for English learners (e.g. Chrabaszcz & Gor, 2014). Considering this, I have argued that the difficulty of Mandarin tones instead is primarily related to difficulties in encoding of L2 tone categories in lexical representations and the utilization of those representations during real-time speech processing.

1.5 Research on L2 tone category learning

To encode lexical tones, it is necessary for L2 learners to form tone categories, that is, abstract representations that treat what are in fact continuous pitch differences as if they belong to discrete categories. These categories then also need to be encoded as lexically distinctive cues. Accounting for L2 category formation and its limits is a major goal of prominent L2 models of speech learning (*perceptual assimilation model (PAM):* Best, 1994; Best & Tyler, 2007; *speech learning model:* Flege, 1995). Just how such models apply to tones for learners from non-tonal L1s remains an open question. Can non-tonal L2 learners form tone categories 'from scratch' or do they adapt L1 pitch features for new purposes? So and Best (2010, 2014) propose that listeners adapt lexical tones to their native language 'prosodic categories'—but this raises a number of questions that have yet to be answered, for example, how categorical are prosodic categories (cf. Ladd & Morton, 1997; Wagner & Watson, 2010), and how does L1 transfer occur between, e.g., word stress and lexical tones as compared to sentence intonation and lexical tones?

While it seems clear that L1 intonation can impact pitch perception *in general* (Braun, Galts, & Kabak, 2014; Braun & Johnson, 2011; So & Best, 2010, 2014), it is far from clear that such issues relate to tone in a *categorical* fashion such that, if you expect pitch to fall at the end of sentences, this will create general biases for

perception of lexical tones in all contexts. Even proponents of PAM show some uncertainty as to whether non-tonal L1 listeners are expected to treat tones as linguistic sounds that lie outside the their current phoneme space (i.e., *uncategorizable*) or as non-linguistic sounds (i.e., *non-assimilable*, like Zulu clicks) (cf. discussion in So & Best, 2010). So far, results conducted in the PAM framework, suggest that even if L1 prosodic categories play some role in shaping L2 perception of tones in certain prosodic positions (for example, English phrase-final falling intonation may warp perception of phrase final pitch), the phonetic properties of the tones themselves are likely to play a larger role in explaining tone confusion patterns overall (So & Best, 2010, 2014).

Several recent studies (Ling, Schafer, & Grüter, 2016; G. Shen & Froud, 2016, 2018) have taken a different approach to understanding L2 tone categorization by applying variations of classic categorical perception paradigms (e.g., Lisker & Abramson, 1964). These studies have found intriguing results with experienced L2 learners (as opposed to naïve listeners). Ling, Schafer, and Grüter (2016) found that L2 learners with previous classroom exposure to Mandarin resembled native speakers in tone categorization when performing an identification task (with tone continua), but not when performing a discrimination task targeting the same stimuli. Along the same lines, G. Shen and Froud (2016) found that behavioral identification and discrimination performance (again with tone continua) for 'advanced' L2 learners was near-native, while ERP responses (MMN, P300) for these stimuli during passive listening were distinct from native patterns (G. Shen & Froud, 2018). Such inconsistent results suggest that L2 learners can to some degree develop tone

categories, but that—at least in the absence of massive L2 exposure—the categories are not fully nativelike.

1.6 Lexical difficulties and L2 tones

To fully evaluate the development of L2 tone categorization, research must consider lexical representations and processing. Categorization that is only successful in phonetic tasks will be of limited use for L2 listeners. Surprisingly, though the lexical nature of tones and their relationship to semantic distinctions has been the focus of many training studies with *naïve* learners (e.g., Chandrasekaran, Sampath, & Wong, 2010; Chang & Bowles, 2015; Cooper & Wang, 2013; Ingvalson et al., 2013; Perrachione et al., 2011; Wong et al., 2007), lexical issues have been largely overlooked when dealing with *experienced L2* learners. This is crucial, as, even when learners have strong auditory tone perception abilities, this does not necessarily translate into robust lexical representations or real-time processing. Practically speaking, this is also what actually matters to learners, as it has the potential to affect their success in understanding and conveying meaning.

In previous research (Pelzl et al., 2018), we found that even advanced L2 learners who excelled at tone identification (as reported above) performed quite poorly when tone cues were needed to reject nonwords in a lexical decision task. We used a paradigm that pitted recognition of tonal nonwords against recognition of segmental nonwords. All stimuli were disyllabic. Tonal nonwords differed from real words only with respect to the tone of the first syllable (e.g., nonword *fang**4**zi* /fɑŋ4ts/ derived from real word *fang**2**zi* /fɑŋ2ts/ ‘house’). Segmental nonwords differed from real words with respect to the rhyme of the first syllable (e.g., nonword *f**e**ng2zi*

/fəŋ2ts/ derived from real word ***fa****ng2zi*). 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%). This performance did not appear to be due to lack of word knowledge, as most participants knew upwards of 95% of the critical vocabulary, and (with just one exception) even participants who performed near ceiling on an offline test of tone knowledge for the critical vocabulary failed to reach native-speaker levels for rejection of tonal nonwords. While there are several aspects of this research that can be improved upon, the results strongly suggest that L2 learners have particular difficulty encoding and/or processing tone information during spoken word recognition.

1.7 Three Hypotheses to explain L2 tone difficulties

As mentioned above, this dissertation will focus on three general hypotheses that posit different sources of difficulties for L2 tone learning. It should be clear by now that I believe a full understanding of L2 tone learning difficulties cannot be achieved outside the context of lexical representations and word recognition processes. So then, before we turn our attention to the three hypotheses, it will be useful to briefly consider L1 Mandarin word recognition.

1.7.1 The role of tones in L1 Mandarin word recognition

Here I outline a generic model of L1 Mandarin word recognition that roughly reflects the characteristics of most current models of lexical recognition (e.g., Cutler,

2012; Luce & Pisoni, 1998; Marslen-Wilson, 1987; McClelland & Elman, 1986; Norris, 1994; Norris & McQueen, 2008), but critically addresses the role of tones (cf. Shuai & Malins, 2017). This model does not address whether and to what extent episodic details (e.g., speaker identity, emotion) are encoded in the lexical entries themselves (for recent discussions see, Kazanina, Bowers, & Idsardi, 2017; Pierrehumbert, 2016), and, at least for the moment, we will not consider additional complications that might apply in Mandarin word recognition, such as the role of syllable+tone co-occurrence frequencies and probabilities (Wiener & Ito, 2015, 2016).

Figure 1.2 depicts the process of L1 Mandarin spoken word recognition. The process begins with perception of a word as acoustic input. That input includes fine acoustic-phonetic detail, all of which is perceived, but not all of which will be relevant in a give instance of word recognition. For example, information about the identity or emotions of the speaker, while available in the input, is not expected to play a strong role in word recognition in most cases. Instead, for Mandarin listeners, it will be the segmental (C, V) and suprasegmental (T) cues that drive lexical access. Figure 1.2 indicates features that are utilized for word recognition by surrounding them with a box and an arrow pointing to the matching lexical target. In this case, consonant, vowel and tone cues are all utilized, while cues to speaker identify (S) and emotion (E) are not.

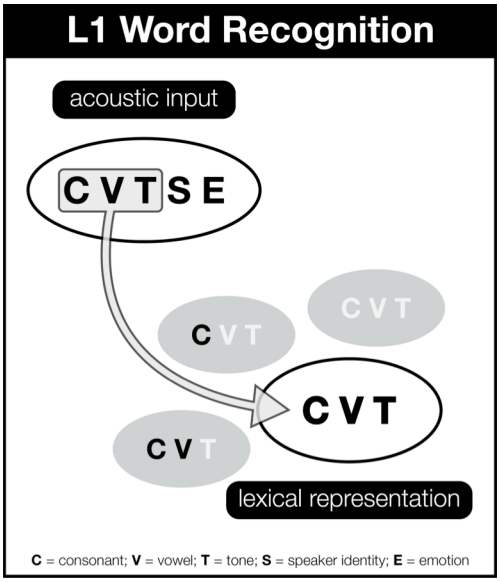


Figure 1.2. Model of L1 Mandarin word recognition, with acoustic features of the input activating phonological entries in the mental lexicon.

Word recognition in this model assumes lexical competition whereby the unfolding acoustic signal drives activation of all plausibly matching lexical candidates. As more of the signal becomes available, mismatching candidates are eliminated. These eliminated competitor words are indicated in Figure 1.2 in gray circles, with the features that match the input in black, and those that do not match the input in light gray. The features of the final word that gets selected are all in black, indicating a complete match between the input and the phonological form encoded in the mental representation.

This basic model of L1 Mandarin word recognition will serve as a counterpart for consideration of ways in which L2 Mandarin word recognition might break down due to tones.

1.7.2 The Tone Perception Hypothesis

The Tone Perception Hypothesis: Tones may be difficult for L2 listeners to perceive auditorily, that is, hearing differences between linguistic tones is difficult.

As reviewed above, L2 speech learning often involves low-level difficulty due to inaccurate auditory perception of unfamiliar sounds. Though I have argued that evidence so far suggests this is not such a formidable long-term problem in L2 learning of tones, it is too early to completely dismiss such difficulties. For one, they certainly can play a significant role for many learners at the early stages of Mandarin learning, and thus may have lingering effects at later stages. Additionally, the influences of a wide range of complicating factors have still been minimally researched (e.g., coarticulation, speech rate, noise). For now, then, it is still worth considering the Tone Perception Hypothesis as a potentially significant cause of difficulties in more advanced learners.

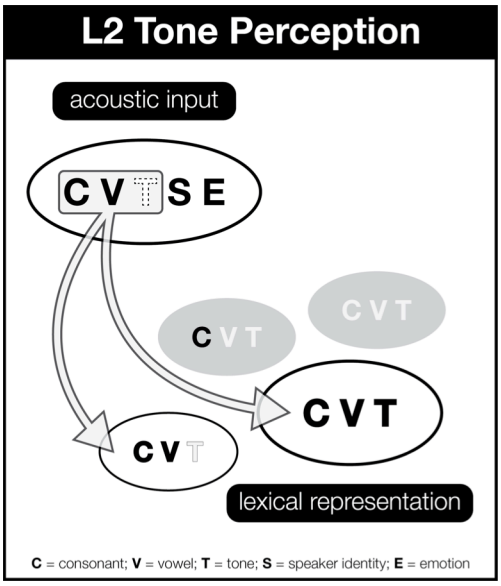


Figure 1.3. Illustration of breakdown in L2 word recognition due to difficulty in perception of lexical tones.

Figure 1.3 illustrates how tone perception difficulties could negatively impact (isolated) Mandarin word recognition for learners who have already established a reasonably sizable L2 lexicon. In this case, even though both segmental and suprasegmental features are engaged in lexical access, the tone feature is misperceived (indicated by the dotted white T in Figure 1.3). This misperception is such that some or all tonal distinctions in the acoustic input are ambiguous. Consequently, any L2 lexical representations that match the segmental portion of the input, and are plausible matches for the ambiguous tone, will be equally valid competitors. In other words, even if tones are correctly represented in long-term memory, and are actively processed during lexical recognition, weaknesses in tone perception could lead to spurious lexical competition, and, ultimately, 'recognition' of an incorrect word.

In the extreme case, where accurate tone perception was never possible at all, it would seem unlikely that any useful tone representations could be encoded by the learner. But, as we reviewed above, this extreme case seems highly unlikely for most L2 learners. A more reasonable possibility is that *some* tones in *some* contexts may be difficult to perceive and consequently could induce word recognition difficulties of the type depicted in Figure 1.3. This latter possibility will be addressed in Chapter 2, where we investigate L2 tone perception in isolated monosyllables and in disyllabic contexts.

1.7.3 The Tone Representation Hypothesis

> *The Tone Representation Hypothesis:* Tones may be difficult for L2 listeners to *represent* in an effective way, that is, L2 tone categories may be poorly

formed or not encoded directly in the phonological form of mental lexical entries.

Figure 1.4 illustrates how tone representation problems could once again lead to spurious lexical competition for the L2 learner. While all segmental and suprasegmental cues in the input are accurately perceived and fully utilized for lexical access, when contacting the lexical representations, there is ambiguity induced by a poor quality tone feature (cf. Cook & Gor, 2015; Diependaele, Lemhöfer, & Brysbaert, 2013; Perfetti, 2007). "Poor quality" here is used as a catchall for different possible problems. A tone feature could be *missing*, *incorrect*, or too *permissive.* In the latter case, this means that, even though the encoded tone is "correct", this abstract tone representation is not distinctive enough (compared to tone features in other lexical representations) to either spur on correct activation or inhibit spurious activation. For now, we make the simplifying assumption that all of these problems have the same result, namely, making lexical competition less efficient due to

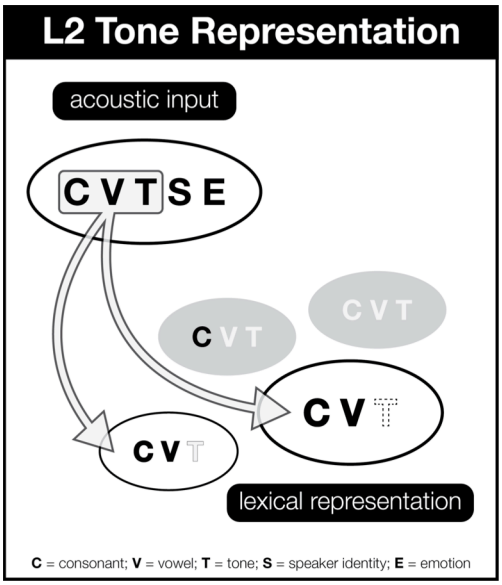


Figure 1.4. Illustration of breakdown in L2 word recognition due to difficulty in representation of lexical tones.

spurious activation of tonal competitors. Experiments 3 and 4 will explore this type of problem by using nonword neighbors of real words that are distinguished from those real words by either a vowel or tone feature. In this way we will test whether L2 listeners are able to use these features to effectively inhibit lexical activation. Successful rejection of tone nonwords indicates that acoustic cues are perceived and utilized to inhibit incorrect lexical selection, thus implying that representations are fully specified for tones.

1.7.4 The Tone Processing Hypothesis

> *The Tone Processing Hypothesis:* Tones may be difficult for L2 listeners to *process* efficiently, that is, native language (L1) processing biases may impede L2 ability to fully utilize tone cues in real time.

This final hypothesis posits that, even if perception and representation of tones are unproblematic, there might still be difficulty in lexical recognition due to ingrained L1 processing biases (MacWhinney & Bates, 1989; Strange, 2011; Zou, Chen, & Caspers, 2016). In this case, as depicted in Figure 1.5, tone features are not automatically involved in lexical recognition. Instead the L1 bias for segmental cues drives L2 lexical access such that any potential match for the segmental form of the input becomes a lexical competitor. Thus, upon hearing '*ma1*', all words that match the segmental form of /ma/ are accessed.

The Processing Hypothesis will be examined in Experiments 3 and 4, where we measure event-related potentials to test real-time neural responses to tone and vowel nonwords. If we find evidence of neural sensitivity to tones, this will indicate that L2 listeners can successfully process tone cues in real time. Additionally, we will

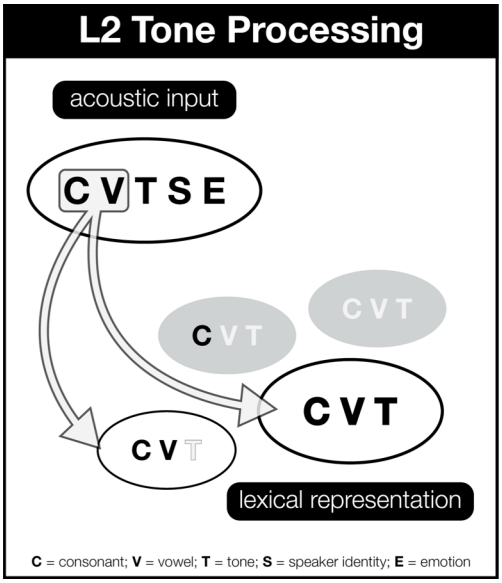


Figure 1.5. Illustration of breakdown in L2 word recognition due to difficulty in processing of lexical tones.

contrast lexical recognition in complete isolation (Experiment 3) and when contextualizing cues (pictures) create strong expectations about the phonological form of a word (Experiment 4). If we find success in lexical recognition in the latter experiment, but not in the former, this would suggest that pre-activation of the lexical representations allows for successful processing of tones that might not occur otherwise.

Chapter 2: Perception of tones and allotones in isolation and in context

2.1 Introduction

For those who grow up speaking non-tonal L1s, a necessary step towards learning to recognize tone words is forming discrete phonetic categories out of the varying, but patterned pitch contours that accompany Mandarin syllables. As reviewed in Chapter 1, previous research shows that—at the level of phonetic perception—this task seems tractable, if not always easy. However, a serious limitation on much of the research is its reliance on isolated monosyllabic (MS) stimuli. Out of the 47 observational studies listed in Tables A1 (Appendix A1) 36 used only MS stimuli. Out of the 31 training studies listed in A2 (Appendix A1) only five included disyllabic (DS) stimuli. Due to the predominant use of MS stimuli in tone research, there are still many limitations on what we know about L2 abilities to learn phonetic tone categories beyond this simplest context.

Over-reliance on tasks that narrowly target phonetic perception without considering phonological or lexical aspects of tone learning further limits what we can say about the quality of the tone categories L2 leaners acquire. That is, we do not typically know whether the type of phonetic tone categories targeted can be usefully encoded in mental lexical representations, and whether they can serve for tone word recognition across a variety of contexts, and not just tones in isolation. As noted in Chapter 1, one way in which some recent research has attempted to target the *phonological* representation of tones is to test beginning learners' tone category perception in the context of words (e.g., Wong & Perrachione, 2007). Another approach, reviewed above, it to use test the phonological warping of tone perception

by probing tone category boundaries using identification and discrimination of tone continua (Ling et al., 2016; G. Shen & Froud, 2016, 2018). One more method, which to my knowledge has not been previously explored for L2 tones, is to test phonological abstraction by examining tone identification of tonal variants (allotones). Allotones provide a useful test case for the abstractness of tone representations, as successful allotone acquisition means that sounds with surface differences in pitch contour will be perceived as a single tone category.

2.1.1 Goal of the chapter

In this chapter, I present tone identification experiments meant to address the limitations noted above. This chapter aims to shed light on L1 and L2 perception of tones in isolated monosyllables (MS) and in disyllabic (DS) context, with special focus on an underexplored allotone of Mandarin T3. The goal is to determine whether context affects L1 and advanced L2 tone perception similarly, and whether L1 and L2 listeners are able to dismiss phonetic differences in the surface form of separate T3 allotones by identifying both with the same category label. We will also examine L1 and L2 error patterns for T3 allotones to further understand how T3 phonetic realization influences perception.

2.1.2 Phonetic perception of tones in advanced L2 learners

As noted already, previous research has strongly suggested that advanced L2 learners (from non-tonal languages such as English and Dutch) can generally achieve high accuracy in tone identification, particularly for tones on isolated monosyllables (C.-Y. Lee, Tao, & Bond, 2009; Pelzl et al., 2018; L. Zhang, 2011). Nevertheless, not

all Mandarin tones are equally easy. T2 and T3 seem to cause relatively more difficulty for learners, even after several years of classroom study (Hao, 2012; Sun, 1998). At the same time, some level of difficulty identifying T3 is also typical of L1 listeners (T. Huang & Johnson, 2010; Pelzl et al., 2018). In Pelzl et al. (2018), we found that when tone identification was challenging, L1 and advanced L2 listeners performed nearly identically, both showing a considerable drop in accuracy for T3. However, performance was not identical across the board. For T2 the L1 group was significantly more accurate than L2. Such results suggest that, for the most part, with respect to MS tone identification, L2 difficulty at the level of phonetic perception is not strongly different than what is experienced by *all* tone language listeners. In this respect, a simplistic version of the Tone Perception Hypothesis that posits general difficulty with auditory perception of tones does not offer a satisfactory explanation for advanced L2 listeners' tone difficulties.

In addition to potential L2 weakness for T2, a second experiment left open the possibility that L2 tone perception difficulties may be uniquely severe in *disyllabic contexts*. When the exact same acoustic MS tokens used in the tone identification task were presented in their original DS context (words), we found that advanced L2 listeners had great difficulty rejecting nonwords on the basis of tones. Compared both to L1 listeners and to their own performance on nonwords with mismatching rhymes, L2 performed with very low accuracy (*mean* = .35). Since the lexical decision task differed from the tone identification task both in the length of its stimuli (two syllables vs. one) and in the nature of the task (lexical decision vs. tone identification), the role of stimulus length on its own remains ambiguous. To address

this issue, Experiment 2 below directly contrasts advanced L2 tone identification in MS and DS contexts, i.e. using the same task for both.

2.1.3 Mandarin allotones

Besides the obvious difference in the number of tones involved in MS and DS contexts, tones in these contexts also vary with respect to their *surface realization* (W.-S. Lee & Zee, 2014; Xu, 1997). This is particularly so for T3. In carefully produced citation form, T3 is typically a low *dipping* tone—which I will call *T3D*. T3D is notable not only for its dipping quality, but also for its duration; it is considerably longer than the other tones. Importantly, in contextualized speech T3 is rarely realized as a dipping tone (typically only when it is given prominence in pre-pausal position). Instead, contextualized T3 is most often realized as a low or low-falling tone with a relatively flat contour (Duanmu, 2007; J. Zhang & Lai, 2010). I will refer to this form as *T3F* (i.e., *falling*). In contrast to T3D, the duration of T3F syllables is comparable to that of the other three tones (all approximately 200ms in natural speech, cf. Duanmu, 2007).

T3 also undergoes a well-known process of contextually determined tone change whereby it ends up resembling a different tone entirely. This process is typically called tone *sandhi*. Under the application of T3 sandhi, T3 becomes T2 when it precedes another T3. That is, despite the relevant syllable having an *underlying* T3, it will be realized with the surface form of a rising tone, which is generally judged to be indistinguishable from T2 (Duanmu, 2007; Speer, Shih, & Slowiaczek, 1989).

Previous research suggests the degree of difference between realizations of T3F and T3D varies across contexts and speakers (cf. Duanmu, 2007; T. Huang & Johnson, 2010; F. Shi, 2009). Specific T3 allotone realizations may also be partially sociolinguistic in nature (e.g., indexing age or regional affiliation), and might produce some emotional or rhetorical effects when exaggerated. *Despite these differences*, T3F and T3D are still perceived as a single T3 tone category and native speakers without linguistic training are unlikely to be aware of these changes. While T3 sandhi is almost 'famous' in Mandarin linguistic circles, even linguistically savvy native speakers may be unaware of the distinction between T3F and T3D. This does not mean that these distinctions are necessarily hard to perceive, and native speakers are unlikely to have difficulty noticing the relevant acoustic differences if their attention is drawn to them. However, to my knowledge, no research has yet examined perception of T3D and T3F allotones in native Mandarin listeners (Gårding, Kratochvil, Svantesson, & Zhang (1986) used synthesized stimuli to test perception of T3 and T4 in context, specifically examining the role of inflection point and creaky voice, but did not contrast T3D and T3F specifically).

In contrast to T3, allotones of the other citation tones are not commonly discussed, though T4 is known to have a somewhat shorter fall in context than in pre-pausal position (see Chao, 1968, for a less common and somewhat debated T2 variant). Still, it is worth making a short note here about the contour of T2. While it is correct to describe T2 as a rising tone, in careful acoustic analysis (e.g., Xu, 1997), it can be seen that T2 often also has a slight dip at its onset before it rises, even when produced in isolation. This might be one cue that would lead it to be perceived as

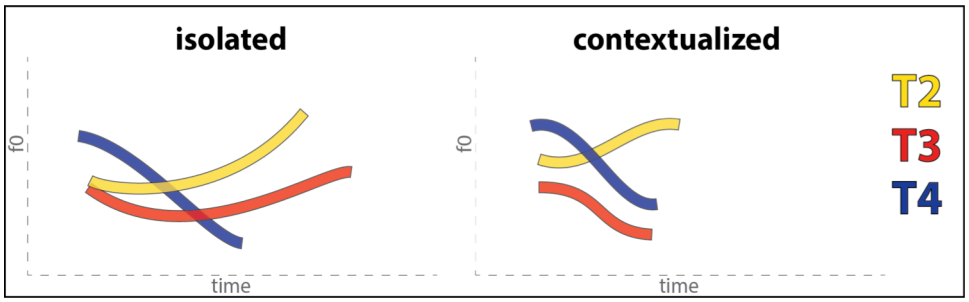


Figure 2.1. Comparison of F0 contours of T2, T3, and T4 when produced in isolated monosyllabic and in disyllabic context. Contour based on production from a single male native speaker.

similar to T3, especially T3D, and in fact several studies have manipulated this and other features of T2 and T3 contours to test native Mandarin tone perception (Blicher, Diehl, & Cohen, 1990; Moore & Jongman, 1997; J. Shen, Deutsch, & Rayner, 2013; X. S. Shen & Lin, 1991).

2.2 Experiment 1: L1 tone identification in monosyllables and disyllables

2.2.1 Experiment 1: Motivation

As noted above, I am unaware of previous research on L1 perception of T3F and T3D allotones (for production, see J. Zhang & Lai, 2010). To fill that gap, and to provide direction for the design of an experiment targeting advanced L2 perception, we conducted a tone identification experiment meant to tease apart phonetic and phonological aspects of T3 allotone perception in L1 listeners.

As highlighted above, the surface F0 differences between T3D and T3F suggest that perceived similarities with other tones will depend on which variant is under discussion. The contrasting surface F0 patterns of T3D and T3F, as well as their similarities to other tones are illustrated in Figure 2.1, in slightly idealized form based on the productions of a single male speaker. Figure 2.1 highlights how these similarities are impacted by tone production in isolation and in context. T3D, due to

its dipping contour, may resemble T2, as both tones can have an initial fall followed by a rise. In contrast, for T3F, the falling contour suggests similarity with T4, with the main difference between T3F and T4 being pitch onset (though T4 in context may also have a higher offset). These patterns of similarity suggest that in a tone identification task, likely error patterns would have T3D misidentified as T2 and T3F as T4.

2.2.2 Experiment 1: Research questions

In Experiment 1, a tone identification task was conducted using three types of tone stimuli: monosyllables produced in isolation, disyllables produced in isolation, and single syllables clipped from the disyllabic stimuli. In each trial of the tone identification task, participants heard an auditory stimulus and were asked to judge its tone, or, in the case of disyllabic stimuli, the tone of its first syllable. While stimuli included all tones, the critical results pertain the T3 allotones. Specifically, the experiment aimed to answer the following questions:

(1.1) *Does L1 accuracy for T3 allotones vary according to whether syllables are presented in isolation or in disyllabic context?*

(1.2) *Does the direction of L1 errors for T3 vary according to the phonetic form of T3 (T3D vs. T3F)?*

Question 1.1 explores whether there are limits on the robustness of tone category abstraction for T3F. As *contextualized* T3F is pervasive in natural Mandarin speech, it is expected that listeners will consistently identify it as T3. However, when T3F is presented *in isolation*, we expect the lack of context will induce some amount of misperception among listeners due to the strong phonetic similarity with T4, and

the rarity with which isolated T3F occurs in natural language. In other words, although it is generally the case that the most strongly weighted cue Mandarin listeners use to identify tones is F0 (Gandour, 1983; Howie, 1976), in this case, we expect that isolated F0 will be insufficient to fully disambiguate T3F from T4. This would indicate the important role of *relative F0* in L1 tone recognition (cf. J. Huang & Holt, 2009).

Question 1.2 aims to further test the effects of phonetic similarity by examining whether T3 confusion patterns play out as expected, namely, that when T3D is misidentified, it will most often be as T2. In contrast, when T3F is misidentified, it will most often be as T4. If these error patterns were to be realized, this would provide some evidence that phonetic influences are primarily responsible for L1 tone confusions of T3. We will return to this point in the discussion for Experiment 1.

2.2.2 Experiment 1: Participants

Participants were 36 native Chinese (25 female, average age 21) living in Beijing, China. All were current or former graduate or undergraduate students at local universities. All participants identified themselves as native speakers of Mandarin (*Putonghua*). Screening procedures further asked them to verify that the tones and pronunciation of any local Chinese language they might speak were the same as Mandarin. Participants were predominantly from northern regions (e.g., Beijing, Heilongjiang, Shandong), though a small number of them were accepted because they insisted that that their mother tongue was Mandarin, despite growing up in areas where other regional varieties of Chinese are prevalent (e.g., Guizhou, Zhejiang,

Jiangsu). While it is possible that such varied experience could impact outcomes[3], in the current instance, we were primarily interested in establishing a general pattern of native Mandarin tone perception as a baseline for comparison with L2 learners. Future work might endeavor to control regional influences more strongly. All participants gave informed consent and were compensated for their time. Procedures were approved by the Institutional Review Board of the University of Maryland (UMD) and the local equivalent at Beijing Normal University (BNU).

2.2.3 Experiment 1: Stimuli

Stimuli consisted of 192 unique audio tokens, 64 for each of three target types: monosyllables (MS), disyllables (DS), and clipped syllables (CS). They were based on four nonword monosyllables that were chosen from accidental gaps in the Mandarin lexicon (*bou* /pəu/*, chei* /tʂʰəi/*, fai* /fai/*, tiu* /tʰiəu/). The syllables were intended to be easy for native Mandarin speakers to pronounce as each one had at least one real word neighbor that differed in only a single consonantal feature (e.g., aspiration in the real word *pou1* /pʰəu/ as in 'cut open' vs. the unaspirated form in nonword *bou* /pəu/). These four nonword syllables were paired with the four standard Mandarin tones to create a set of sixteen monosyllabic stimuli. Additionally, disyllabic stimuli were created by adding the syllable *ba* /pa/ after each of the sixteen monosyllables. This extra syllable always bore a so-called "neutral" (i.e., unstressed) tone in the nonword, and was selected because it is often neutralized in standard

[3] For example some southern regional dialects have no rising tone category. Speakers from such regions might then be largely unable to differentiate T2 from other tones.

Mandarin vocabulary, and because the closure in production of /p/ allows for cleanly clipping off the final syllable. The F0 patterns of neutral tones are contextually conditioned. In final position on an isolated disyllable, the F0 of the neutral tone will be strongly dependent on the preceding tone (W.-S. Lee & Zee, 2008, 2014). In the present case, neutral tones were chosen to avoid creating strong coarticulatory influences on the tones of the first syllables. Additional practical considerations were to make it less likely that participants in the tone identification task would attempt to identify the second syllable of DS targets, and to limit the overall number of possible disyllable tone configurations.

When testing native listeners and advanced L2 learners, monosyllabic tone identification stimuli run the risk of producing strong ceiling effects, rendering experimental results uninformative. Thus, several steps were taken to ensure the relative difficulty of the stimuli in the present experiment. First, nonwords were created such that they all contained diphthong or triphthong vowels (/ei/, /ou/, /ai/, /iou/), as diphthongs have been found to make tone identification more difficult (B. Yang, 2012). The use of nonwords serves a similar purpose and also limits the likelihood of unwanted lexical or distributional influences on tone identification (Fox & Unkefer, 1985; Wiener & Ito, 2016).

One final method to increase the difficulty of the current task was to introduce variability by having multiple speakers produce the stimuli. Four native Mandarin speakers, two male and two female, each produced the full set of 32 stimuli for recording, along with additional nonword practice stimuli for the nonwords *diang* and *diangba*. Each speaker read the stimuli off of cue cards presented in a randomized

order, first completing two runs without recording in order to get them comfortable reading Pinyin romanization (which most Chinese in the PRC are familiar with, but not accustomed to reading in the absence of Chinese characters). Then the entire set of stimuli was recorded three times for each speaker.

After recording, all sound files were labeled and clipped from the original recording using *Praat* (Boersma & Weenink, 2010) to create the individual stimulus items. The third set of recordings from each speaker was always used, with individual items replaced from the first or second recording if necessary. For disyllabic items, two forms were created. A full form that included both syllables, and a clipped form that included only the first syllable, with the second syllable cut off at the last zero-crossing prior to the closure of the /p/ in *ba*. This process created 48 stimuli from each of the four speakers for a total of 192 unique stimuli.

The F0 contours of the first syllables of all stimuli are depicted in Figure 2.2. It can be seen that overall tone height varies across speakers, especially female vs. male. Additionally, clear differences in the contour of T3 can be seen for its isolated and contextualized forms. Whereas the isolated form is much longer and tends to have a dipping contour, the contextualized form is shorter and typically consists only of a falling contour without a later rise.

The stimuli were organized into three sets of 64 items according to their type (MS, DS, CS). For each set, two blocks of 32 items were formed, balanced so that an equal sample of each nonword, speaker, and tone occurred in each block. Six block orders were created, with both blocks of one type being presented together, but so that

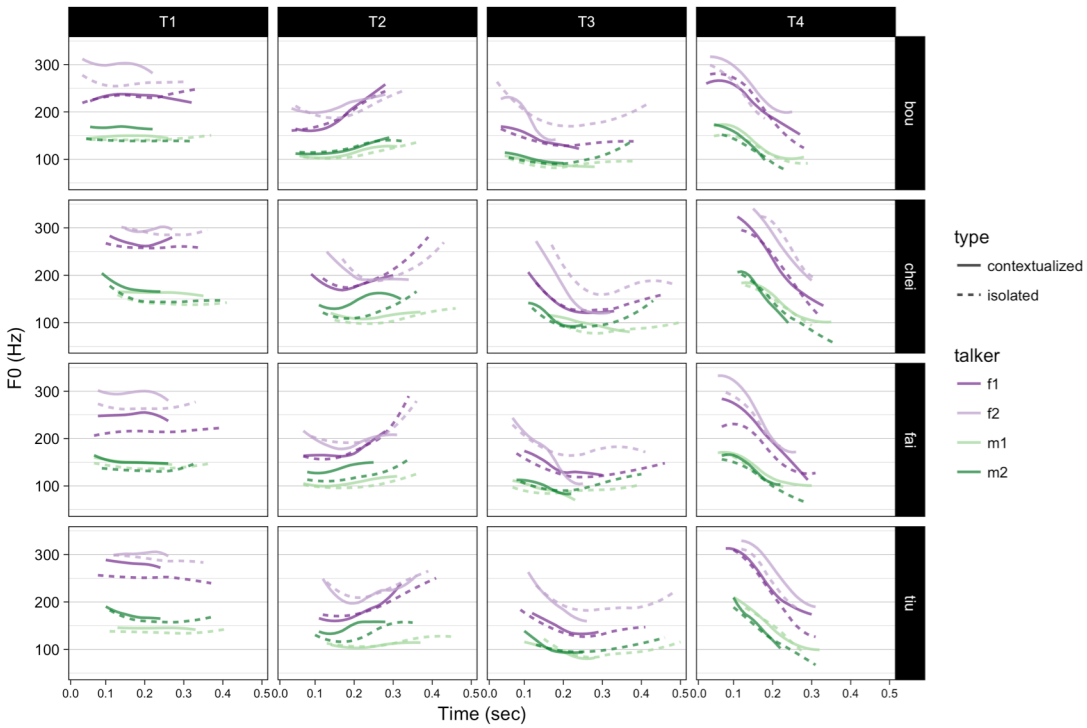


Figure 2.2. Smoothed F0 contours of first syllables of stimuli used for tone identification (Experiments 1 and 2). Dashed lines represent productions of isolated monosyllables. Solid lines represent productions of contextualized tones (i.e., first syllable of disyllabic items). Stimuli were produced by two female (f1, f2) and two male (m1, m2) native Mandarin speakers.

the order of types presented first, second, and third was balanced across participants (MS-DS-CS; MS-CS-DS; DS-MS-CS; DS-CS-MS; CS-MS-DS; CS-DS-MS).

2.2.3 Experiment 1: Procedures

After giving informed consent, and completing a brief background survey, participants were seated with a laptop and headphones in a quiet lab room at BNU. The experiment was run using *PsychoPy* (Peirce, 2007). Instructions were presented on the screen in Chinese. On each trial they heard a single stimulus and responded by pressing a number key corresponding to the Mandarin tone they believed they heard. The index and middle fingers of the right and left hands were used for responses, with one finger placed on each number key (1-4). For blocks of disyllabic stimuli, they were instructed only to judge the first syllable. Participants first completed eight

practice items in an MS block, then completed eight more in a DS block. No feedback was provided during practice, as the intention was to familiarize people with the task rather than train responses. After completing the practice trials, each participant completed six blocks of 36 trials, with trials in each block presented in a random order unique to that participant. The entire task took about fifteen minutes. Both accuracy and decision times were recorded, though only accuracy will be reported here.

2.2.3 Experiment 1: Accuracy results and analysis

Reliability for Experiment 1 data was high (α=.94). Descriptive results (Table 2.1) of mean accuracy suggest consistently strong performance for T1 and T4, regardless of stimulus type. For T2, accuracy is highest for MS (*mean*=94%), and drops somewhat for DS (*mean*=87%) and CS (*mean*=84%). For T3, there appears to be variation across contexts with mean accuracy lower for MS (*mean*=77%), very high for DS (*mean*=94%), and extremely low for CS (*mean*=32%). Figure 2.3 depicts results in box plots.

Accuracy results were submitted to a generalized linear mixed-effects model with crossed random effects for subjects, talkers, and items. Statistical analyses were conducted in *R* (version 3.3.3, R Core Team, 2017), and models were fit using the *lme4* package (version 1.1-12, Bates, Mächler, Bolker, & Walker, 2015) with the *bobyqa* optimizer. Effects coding was applied using the *mixed* function in *afex* (Singmann, Bolker, Westfall, & Aust, 2017), and p-values were obtained using the likelihood ratio test ("LRT") method. Results are reported for Chi-square tests in ANOVA tables for this and subsequent generalized linear mixed-effects models. The

Table 2.1. Mean accuracy and std. dev. results for the tone identification task (Experiment 1)

condition	tone	accuracy % (sd)
MS	T1	96 (20)
	T2	94 (25)
	T3	77 (42)
	T4	97 (18)
DS	T1	94 (25)
	T2	87 (34)
	T3	94 (24)
	T4	97 (18)
CS	T1	94 (25)
	T2	84 (36)
	T3	32 (47)
	T4	94 (23)

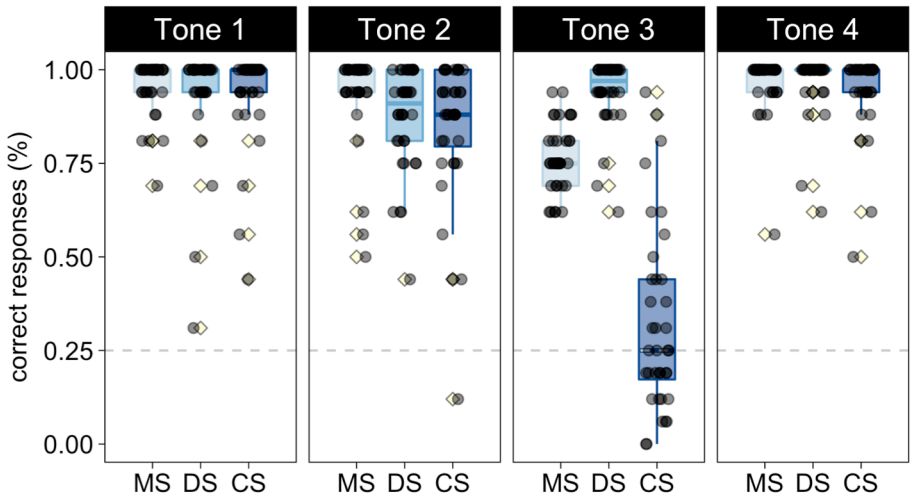


Figure 2.3. Boxplots of accuracy results for tone identification (Experiment 1). Each circle indicates an individual participant's mean score. Diamonds indicate that scores at that level are outliers. The dashed line indicates the level that would be equivalent to chance performance.

dependent variable was accuracy (1, 0). Fixed effects included *tone* (T1, T2, T3, T4), *context type* (MS, DS, CS), and their interaction. The structure of items in the experiment is rather complex, with each item crossed for the four nonwords (*bou, chei, fai, tiu*), the four tones, and the four talkers. To address this, both nonword and talker random effects were included. The maximal model was fit first (Barr, Levy,

Table 2.2. Mixed Model ANOVA Table (Type 3 tests, LRT-method) (Experiment 1)

Effect	df	Chisq.	p-value	
context	46	7.30	.026	*
tone	45	14.22	.003	**
context × tone	42	16.98	.009	**
Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1				
model formula: accuracy ~ context * tone + (context * tone ll subject) + (context * tone ll talker) + (context * tone ll nonword)				

Scheepers, & Tily, 2013; Bates, Kliegl, Vasishth, & Baayen, 2015). Model convergence difficulties were addressed by suppressing correlations in random effects (using "expand_re = TRUE" in the *mixed* function). The maximal model was compared to several simpler model structures, but in the end was retained due to providing significantly better fit as determined by model comparison conducted through likelihood ratio tests. List (order of blocks) was added to the maximal model as a nuisance factor (with subjects nested under lists), but did not significantly improve model fit and thus was dropped from the final analysis. The final model included by-subject, by-talker, and by-item (nonword) random intercepts and slopes for the effects of context, tone, and their interaction.

Main effects and interactions are reported in Table 2.2. There were significant main effects for context (χ^2=7.30, p = .026) and tone (χ^2=14.22, p = .003), and a significant context-by-tone interaction (χ^2=16.98, p = .009).

Table 2.3 reports planned comparisons to test the effects of T3 across contexts. The Holm method was applied for correction of multiple comparisons. Model estimates (b) are reported as log odds and indicate the size and direction of effects (cf. Jaeger, 2008). The difference between T3 in MS and DS context was marginally significant (b = -1.42, SE = -.73, z= -1.94, p = .053). On the basis of

Table 2.3. Planned comparisons for tone identification (Experiment 1)

Comparison	b	SE	z value	Pr(>|z|)	
MS T3 – DS T3	-1.42	-.73	-1.94	.053	.
MS T3 – CS T3	3.11	1.22	2.55	.022	*
DS T3 – CS T3	4.53	0.95	4.75	<.001	**
Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1					

descriptive results (error rates), we can say that errors for T3 in DS were about six times more likely than for T3 in MS (.23/.4=5.75). The difference between T3 in MS and CS was statistically significant (b = 3.11, SE = 1.22, z= 2.55, p = .022). Errors for T3 were about three times more likely in CS than in MS (.68/.23=2.96). Finally, the difference between T3 in DS and CS was also statistically significant (b = 4.53, SE = .95, z= 4.75, p < .001), with errors about 17 times more likely for T3 in CS compared to DS (.68/.4=17). In summary, there was a strong and significant effect for T3 in CS compared to DS, while the effect for T3 in MS compared to DS was smaller and marginally significant.

The specific pattern of T3-by-context interactions depicted by the mean scores and supported by the statistical analyses (CS < MS ≤ DS) can be observed in the raw data for 30 out of 36 participants. In only three cases did participants perform more accurately in CS than MS, and all three still showed less accurate performance than in DS. Three participants were less accurate in MS than in DS, but less accurate still in CS. In other words, despite some amount of variation in the interaction pattern, no participant was ever more accurate for T3 in CS than in DS, and strong performance on CS was exceedingly rare—only one participant (an outlier) scored above 90% accuracy.

Table 2.4. Counts of correct and incorrect responses by tone type for each tone in the tone identification task (Experiment 1). Trials for each tone are indicated in separate tables. Response type (tone) is indicated in rows, with context type indicated in columns. The row of correct responses for a given tone type is highlighted in gray.

Tone 1 trials			
	MS	**DS**	**CS**
T1	551	539	539
T2	23	21	24
T3	2	13	7
T4	0	3	6

Tone 2 trials			
	MS	**DS**	**CS**
T1	21	48	31
T2	539	501	486
T3	13	23	43
T4	3	4	16

Tone 3 trials			
T3	**MS**	**DS**	**CS**
T1	9	9	25
T2	123	13	21
T3	441	541	187
T4	3	13	343

Tone 4 trials			
	MS	**DS**	**CS**
T1	3	0	3
T2	1	5	4
T3	15	14	25
T4	557	557	544

2.2.4 Experiment 1: Error pattern results

Error patterns for all tones are shown in Table 2.4. Our analysis will focus only on T3. It can be observed that in MS, out of 135 total T3 errors, 123 (91%) were T2 responses. For DS there were just 35 total errors (6.1% of trials), and no clear confusion pattern emerges, with 9 T1, 13 T2 and 13 T4 errors. Finally, for CS, out of 389 T3 errors, 343 (88.2%) were T4—and in fact the majority of all responses for T3 targets in CS context were T4 (59.5%).

2.2.5 Experiment 1: Discussion

There are three main findings from Experiment 1. First, as expected, accuracy across MS, DS, and CS varied, with rather dramatic effects for T3 in CS. Second,

these results were highly consistent, with 30 participants showing the same pattern of effects for mean accuracy in MS, and all participants displaying a drop in accuracy for CS compared to DS. Finally, at the group level, error patterns for T3 show clear trends of misidentification as T2 in MS, correct identification in DS, and misidentification as T4 in CS. In short, results in this experiment are largely in line with our expectations—though the size of the CS effect on T3 identification was much larger than expected. These results suggest that while L1 phonological perception of T3 allotones is fairly robust for MS and DS, the surface phonetics of T3 do have the potential to be misleading. In MS, this induces some confusion with T2; in CS this induces very strong confusion with T4.

As noted earlier, the confusion of isolated T3D (MS) with T2 is consistent with previous studies (e.g., Huang & Johnson, 2010). While the set-up of the current study highlights the potential *phonetic* sources of that confusion, we cannot rule out alternative sources. The perceived similarity between T3 and T2 could be shaped (in part) by L1 experience with T3 sandhi rather than purely phonetic considerations. Nevertheless, this confusion is apparently strongly contextually conditioned. In DS the misperception of T3 as T2 all but disappears, showing that L1 listeners have abstracted the T3 category away from its surface form in DS context.

That there is a conditioning effect of context on L1 Mandarin tone identification is not a novel finding *per se* (Braun & Johnson, 2011; Fox & Qi, 1990; J. Huang & Holt, 2009; X. S. Shen & Lin, 1991). However, examination of this effect in the case of T3F (DS vs. CS), and the notable size of the effect, make this a unique contribution to L1 Mandarin tone research, and thus require a bit more discussion.

There are several factors in the current experiment that may have enhanced the strength of the T3F effect. First and foremost, the CS condition in the present experiment is by definition not natural. While T3F occurs with tremendous frequency in context, it is likely to be relatively rare in isolation in natural speech, and listeners rarely if ever need to identify isolated T3D for communicative ends. Additionally, the use of nonword stimuli prevented listeners from relying on lexical knowledge, such as word or syllable frequency, when identifying tones—a factor that is known to impact the way L1 listeners utilize tone cues (Fox & Unkefer, 1985; Wiener & Ito, 2015, 2016). Finally, the use of multiple speakers may have prevented listeners from quickly forming generalizations about the F0 range of any given speaker, thus making relative pitch judgments more difficult. Nevertheless, it seems that these factors—which applied to all stimuli—had a uniquely strong impact on T3F identification.

The results of Experiment 1 motivated further examination of T3 allotones in a second experiment conducted with advanced L2 leaners.

2.3 Experiment 2: Mandarin tone confusions and the effects of T3 allotones on advanced L2 Mandarin learners

2.3.1 Experiment 2: Motivation and research questions

Experiment 2 uses two of the conditions (MS and DS) from Experiment 1, and extends the scope of investigation to the tone perception of advanced L2 learners.[4] We aim to answer the following questions:

[4] In order to maximize efficiency, this second experiment did not include the CS condition, anticipating that effects would largely mirror those of Experiment 1. As

(2.1) *Are advanced L2 listeners equally as accurate as L1 listeners for tones in isolation and in context?*

(2.2) *Do L1 and L2 accuracy for T3 allotones vary similarly for isolated T3D and contextualized T3F?*

(2.3) *Do L1 and L2 listeners make the same patterns of errors for isolated T3D and contextualized T3F?*

Question 2.1 addresses the Tone Perception Hypothesis presented in Chapter 1, and follows up on the results from Pelzl et al. (2018), which found advanced L2 learners to perform near-natively for isolated MS stimuli. By adding the DS context, we can test whether, compared to L1 listeners, L2 learners show significant difficulties in tone identification between MS and DS contexts. If so, the Tone Perception Hypothesis may be a useful framing for our understanding of some aspects of L2 tone difficulties.

Questions 2.2 and 2.3 address the Tone Representation Hypothesis. In Experiment 1 we saw that L1 listeners abstract away from surface phonetics when identifying T3 in context, performing with very high accuracy for T3F is DS context. Experiment 2 will test whether advanced L2 learners have similarly learned to abstract away from surface phonetics when identifying T3 in DS contexts. By comparing error patterns for L1 and L2, we will potentially gain insight into what is driving T3 confusions.

---

even L1 listeners do not treat T3F as an allotone in isolation, this condition is somewhat orthogonal to the main interests of this study, namely abstraction of L2 T3 perception.

Table 2.5. Background information and screening measure 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

Finally, Experiment 2 aims to avoid some potential confounds that have been common in previous tone identification experiments. Specifically, this experiment avoided the use of potentially meaningfully words or syllables that, even if not processed lexically, might allow statistical processing mechanisms to guide tone identification (Wiener & Ito, 2015, 2016; Wiener et al., 2018). Additionally, regardless of stimulus type (MS, DS), only one answer was required for each trial. Previous disyllabic tone identification experiments have required participants to identify multiple tones for DS stimuli in a single trial. This allows a role for memory limitations and potential confusion when multiple labels need to be provided at the same time. Finally, as noted above in the description of stimuli for Experiment 1, several steps were taken to avoid ceiling effects that may have obscured results in some previous studies.

2.3.2 Experiment 2: Participants

L2 participants were 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 (9 female) advanced L2 participants who will be the critical L2

sample for the remainder of this dissertation. Table 2.5 summarizes these 18 L2 participants' general learning characteristics, as well as scores on the screening measures.[5] This study used the same criteria for participation as in previous work with advanced L2 Mandarin learners (Pelzl et al., 2018), thus aiming to maintain at least a lower bound of comparability with the population tested in that study.[6] Twenty-four native Chinese speakers (14 female, average age = 26.1) also completed the experiment. All participants gave informed consent and were compensated for their time. Procedures were approved by the Institutional Review Board of the University of Maryland (UMD) and the local equivalent at Beijing Normal University (BNU).

2.3.3 Experiment 2: Task and stimulus design

As in Experiment 1, all participants completed a tone identification task (Tone ID). In each trial of the Tone ID participants heard a single Mandarin nonword and responded by pressing a number to represent the tone of the first syllable.

---

[5] Additional details for the screening measures, are available online at: https://www.cambridge.org/core/journals/studies-in-second-language-acquisition/article/advanced-second-language-learners-perception-of-lexical-tone-contrasts/BA813543C79288DF6B92E929442CB4BC - fndtn-supplementary-materials.

[6] One L2 participant scored a bit lower (65.7) than criterion (70) on the vocabulary test, but was accepted nonetheless as L2 participants were difficult to come by.

Stimuli were the same as Experiment 1, except that clipped syllables were not included. There were a total of 128 unique stimuli, 64 MS and 64 DS. Two blocks of 32 stimuli were created for both MS and DS, maintaining a balance of tones, nonwords, and talkers. Four block orders were created in total, alternating blocks of 32 MS and 32 DS stimuli (MS1-DS1-MS2-DS2; MS2-DS2-MS1-DS1; DS1-MS1-DS2-MS2; DS2-MS2-DS1-MS1). The order of blocks was balanced across participants.

2.3.4 Experiment 2: Procedures

The Tone ID was conducted in a single experimental session following three ERP experiments (reported in later chapters). Participants were seated in a quiet lab room at BNU or at UMD. The Tone ID was presented on a laptop running *PsychoPy* (Peirce, 2007) and audio was played using a single high quality audio monitor (JBL LSR305) placed centrally in front of and above the participant. Instructions were presented on the screen in English for L2 participants, and in Chinese for L1 participants. Participants were instructed to place the four fingers of their right hand on the numbers 1-4 on the keypad. On each trial they heard a single stimulus and responded by pressing a number key corresponding to the Mandarin tone they believed they heard. For blocks of disyllabic stimuli, they were instructed only to judge the first syllable. Participants first completed four practice items in an MS block, and then four practice items in a DS block. No feedback was given during practice. After practice trials, each participant completed four blocks of 36 trials, with trials in each block presented in a random order unique to each participant. The entire

task took about ten minutes. Accuracy and decision times were recorded, though current analyses will only examine accuracy.

2.3.5 Experiment 2: Accuracy results and analysis

Reliability of the data for the Tone ID was high (α = .91). As seen in Table 2.6, L1 and L2 performed quite strongly across tones in MS, with the exception of T3 (L1: *mean*=81%; L2: *mean*=71%). For DS however, there appear to be differences in accuracy for L1 and L2. L1 appears to largely maintain accuracy for T1, T2, and T4, while accuracy for T3 (*mean*=92%) is higher than it was in MS. In contrast, L2 performs with somewhat lower accuracy in DS for T1 (*mean*=81%) and T2 (*mean*= 80%), continues to show lower accuracy for T3 (*mean*=68%), but maintains high accuracy for T4 (*mean*= 97%). Descriptive results are depicted visually in boxplots in Figure 2.4.

Table 2.6. Mean accuracy and standard deviation for tone identification (Experiment 2)

Group	Context	Tone	Accuracy % (sd)
L1	MS	T1	94 (24)
		T2	93 (25)
		T3	81 (39)
		T4	97 (18)
	DS	T1	92 (27)
		T2	89 (32)
		T3	92 (26)
		T4	96 (19)
L2	MS	T1	99 (12)
		T2	90 (31)
		T3	71 (46)
		T4	98 (13)
	DS	T1	81 (39)
		T2	80 (40)
		T3	68 (47)
		T4	97 (18)

The same analysis procedures were followed as for Experiment 1. Accuracy results were submitted to a generalized linear mixed-effect model. The dependent variable was accuracy (1, 0). Fixed effects included *tone* (T1, T2, T3, T4), *context* (MS, DS, CS), and *group* (L1, L2), and their interactions. Model convergence difficulties were again addressed by suppressing correlations in random effects. As in Experiment 1, after model comparison, the maximal model was retained as the best fitting model. This final model included by-subject random intercepts and slopes for the effects of context and tone and their interaction, as well as by-talker, and by-nonword random intercepts and slopes for the effects of context, tone, and group and their interactions.

Main effects and interactions are presented in Table 2.7. There were significant main effects of context (χ^2=7.04, p = .008) and tone (χ^2=10.12, p = .017),

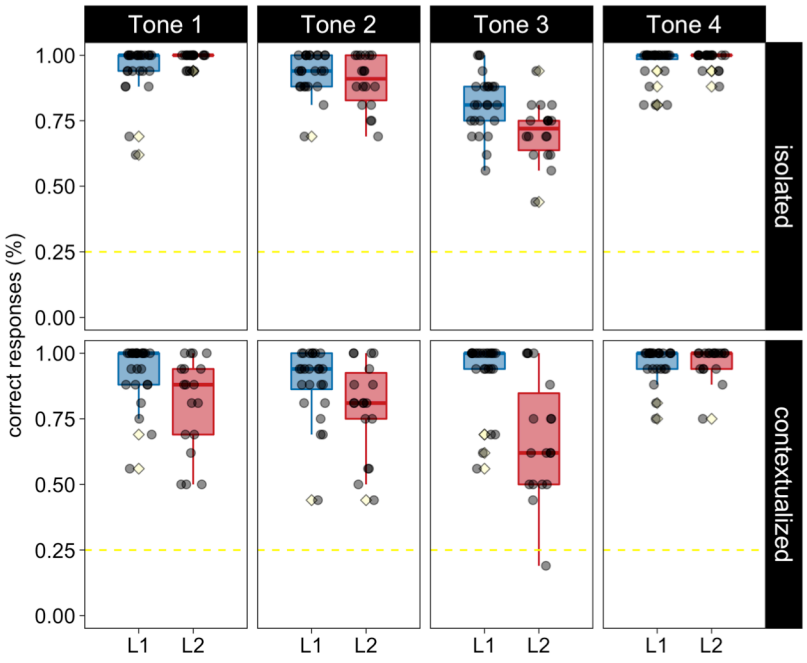


Figure 2.4. Boxplots of accuracy results for tone identification (Experiment 2). Each circle indicates an individual participant's mean score. Diamonds indicate that scores at that level are outliers. The dashed line indicates the level which would be equivalent to chance performance.

Table 2.7. Mixed Model ANOVA Table (Type 3 tests, LRT-method) (Experiment 2)

Effect	Df	Chisq.	Chi Df	Pr(>Chisq)	
context	55	7.04	1	.008	**
tone	53	10.12	3	.017	*
group	55	1.23	1	.267	
context × tone	53	8.44	3	.038	*
context × group	53	6.71	1	.010	**
tone × group	53	10.46	3	.015	*
context × tone × group	53	6.88	3	.076	.
Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1					
model formula: accuracy ~ context * tone * group + (context * tone || subject) + (context * tone * group || talker) + (context * tone * group || nonword)					

and significant two-way interactions for context-by-tone (χ^2=8.44, p = .038), context-by-group (χ^2=6.71, p = .010), and tone-by-group (χ^2=10.46, p = .015). The three-way interaction of context, tone, and group was marginally significant (χ^2= 6.8, p = .076).

Table 2.8 summarizes planned comparisons meant to address our research questions. We first consider interactions of context and group. The difference between L1 MS and L2 MS was not significant (b = -0.17, SE = 0.43, z= -0.37, p = .692). The difference between L1 and L2 in DS was statistically significant (b = 1.03, SE = 0.40, z= 2.57, p = .010). Compared to L1 participants, L2 participants were about two times more likely to incorrectly identify tones in DS (.19/.8=2.38).

We next consider comparisons of accuracy for T3 by context and by group. Family-wise correction was applied using the Holm method. Despite the apparent difference in raw means, the difference between T3 in MS and DS contexts for L1 was not statistically significant. As expected from descriptive statistics, there were also no significant differences for L2 accuracy for T3 according to context, nor was there a significant difference between L1 and L2 accuracy for T3 in MS. For T3 in DS context, however, there was a significant difference in accuracy for L1 and L2

Table 2.8. Planned comparisons for tone identification (Experiment 2)

Comparison	Estimate	SE	z value	Pr(>|z|)	
context × group					
L1 MS vs. L2 MS	-0.17	0.43	-0.37	.692	
L1 DS vs. L2 DS	1.03	0.40	2.57	.010	*
T3 specific comparisons					
L1 MS T3 vs. L1 DS T3	-1.07	0.83	-1.28	.399	
L2 MS T3 vs. L2 DS T3	0.25	0.81	0.32	.751	
L1 MS T3 vs. L2 MS T3	0.96	0.55	1.75	.239	
L1 DS T3 vs. L2 DS T3	2.28	0.57	4.00	<.001	***
Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1					

groups (b = 2.28, SE = 0.57, z= 4.00, p < .001). L2 was about four times more likely than L1 to make errors when identifying T3 in DS (.32/.8=4).

2.3.6 Experiment 2: Error patterns

Error patterns for all tones are shown in Table 2.9. As in Experiment 1, we will focus only on results for T3. In MS stimuli, L1 listeners' errors for T3 were predominantly misidentifications as T2 (65 out of 72 errors, 90%). While L2 listeners made more errors overall, the pattern seems to largely reflect what is seen for L1 with 81 out of 84 errors (96%) being T2. For DS, L1 listeners made relatively few T3 errors. The majority, 72% (21 out of 29), were T2 errors, with 6 T4 errors (21%), and 2 T1 errors (7%). As already seen, the L2 group was significantly less accurate for T3 in DS, but proportionally, the number of errors roughly corresponds to what we find for the L1 group—62 out of 92 errors are T2 (67%), 23 are T4 (25%), and 7 are T1 (8%). In other words, at the group level, despite lower accuracy overall, L2 errors for T3 in DS appear proportionately quite similar to L1 errors.

Table 2.9. Counts of correct and incorrect responses by tone type for each tone in the tone identification task (Experiment 2). Trials for each tone are indicated in separate tables. Response type (tone) is indicated in rows, with context type indicated in columns. The row of correct responses for a given tone type is highlighted in gray.

L1

Tone 1 trials		
	MS	**DS**
T1	361	354
T2	22	23
T3	1	5
T4	0	2

Tone 2 trials		
	MS	**DS**
T1	7	26
T2	358	340
T3	19	17
T4	0	1

Tone 3 trials		
	MS	**DS**
T1	4	2
T2	65	21
T3	312	355
T4	3	6

Tone 4 trials		
	MS	**DS**
T1	2	2
T2	6	2
T3	5	11
T4	371	369

L2

Tone 1 trials		
	MS	**DS**
T1	284	233
T2	3	9
T3	0	4
T4	1	42

Tone 2 trials		
	MS	**DS**
T1	9	23
T2	258	229
T3	19	27
T4	2	9

Tone 3 trials		
	MS	**DS**
T1	1	7
T2	81	62
T3	204	196
T4	2	23

Tone 4 trials		
	MS	**DS**
T1	2	3
T2	1	5
T3	2	2
T4	283	278

2.3.7 Experiment 2: Discussion

Results from experiment 2 can be summarized as follows. While L1 and L2 were comparably accurate for tone identification in MS, they differed significantly in accuracy for initial syllables in DS contexts. For the specific case of T3 allotones, we find that whereas both L1 and L2 show some inaccuracy for T3 in MS, in DS L1 is highly accurate for T3 while L2 remains somewhat inaccurate. Nevertheless, the pattern of L2 T3 identification errors does not appear obviously different than that of the L1 group. We will discuss each of these issues in more depth in the general discussion.

2.4 Experiments 1 & 2: General Discussion

Experiments 1 and 2 set out to explore L1 and L2 tone perception across MS and DS contexts, and to examine the quality of listeners' abstract tone representations by testing categorization of T3 allotones. In this way we aimed to test advanced L2 tone recognition under the scope of the Tone Perception and Tone Representation Hypotheses. As L1 issues were addressed at some length in the discussion for Experiment 1, discussion here will focus primarily on results of Experiment 2 and implications for L2 tone perception.

2.4.1 Tone perception in isolated monosyllables and disyllabic context

The first question we aimed to answer in Experiment 2 was whether advanced L2 listeners are equally as accurate as L1 listeners for tones in isolation and in context. Results show that when contrasting MS with DS perception, for L1 an ambiguous MS tone (T3) became clearer in DS, while for L2 clear MS tones became more ambiguous in DS. This result is largely consistent with previous L2 studies in suggesting that even rather advanced L2 learners tend to have greater difficulty with tone identification in multi-syllable contexts—especially on the initial syllables of words (Broselow, Hurtig, & Ringen, 1987; Hao, 2012, 2018; Sun, 1998). However, the notably different effect for L1 places it in an interesting light, showing that, compared to L1 listeners, advanced L2 learners seem less able to capitalize on relative pitch cues to identify contextualized tones.

The size of this negative effect should not be exaggerated. Compared to the L1 group, the L2 group was twice as likely to make errors in DS overall—but this still represents a moderately high level of accuracy (overall DS *mean*=81%). In other

words, this is evidence of L2 tone perception weakness, but certainly not tone deafness. It is also important to note the large variability in L2 performance. Figure 2.4 suggests that, even though three or four of the L2 listeners were dramatically impacted by DS context for T1 and T2, most of them remained quite accurate, that is, within the normal range of L1 performance. A quick test of L2 participants who fall within 1 standard deviation of the L1 mean for each tone-by-context combination shows that 14 out of 18 L2 participants meet this criterion in all cases except T3 in DS (where only eight L2 participants meet this criterion). Finally, it is also worth keeping in mind that this task was designed to be challenging. The stimuli were nonwords, produced by multiple talkers, and even the L1 group did not perform at ceiling in most cases (except MS T4, where L2 was also at ceiling). In this light, the advanced L2 learners performed quite well—far above chance in all cases with only one exception for one participant (cf. T3 DS where one L2 participant was below chance).

As noted above, one of the main motivations for the current experiment was to separate tone perception from tone word recognition, as the latter might introduce additional difficulties. In that light, the current results do suggest that, though they may be relatively mild (at least for *isolated* MS and DS perception), perceptual difficulties do persist for advanced L2 learners and are exacerbated by context. However, the reason for DS difficulty remains a bit unclear.

One potential explanation would be that, in the case of initial syllables in DS words, difficulty increases because syllable duration is significantly decreased—or, said another way, *speech rate* is increased. Examination of stimuli used in

Experiment 2 finds that average duration of MS items was 374 ms, which is almost 100 ms longer than the duration of initial syllables in DS targets (294 ms). Perhaps the faster speech rate of the target tones in DS stimuli induced L2 inaccuracy. However, this is not a completely satisfying explanation. In our previous study (Pelzl et al., 2018) with a similarly advanced sample of L2 learners, stimuli durations were even shorter (an average of around 220 ms per syllable), nevertheless, L2 accuracy appears to have been slightly higher overall for that experiment.

The same goes for the impact of tone coarticulation on the surface F0 of target syllables. While it is true that initial syllable tones in our DS stimuli underwent some shifts to accommodate the following syllable, this was also true of the stimuli in our previous study—except that in that case, clipping syllables out of DS context removed the potentially helpful contextualizing cue of the following syllable. Additionally, acoustic studies of Mandarin (e.g., Xu, 1997) suggest that the magnitude of coarticulatory effects is typically greater on following tones (carry-over effects) than on preceding tones (anticipatory effects), so it is less likely that initial syllable tones would be more distorted by coarticulation than final syllables—especially as the current stimuli used unstressed second syllables. However, results consistently show that perception of tones on initial syllables is more challenging for L2 learners (Broselow et al., 1987; Hao, 2012, 2018; Sun, 1998).

This leaves the possibility that the increased difficulty L2 listeners experience for DS tone perception may be *due to the presence of a second syllable*. This might indicate that L2 listeners interpret the *holistic* pitch for a bundle of syllables, rather than discrete tones for each syllable. This does not seem entirely likely, as previous

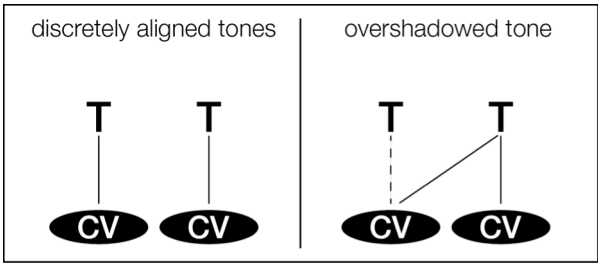


Figure 2.5. Depiction of clear perception of tones aligned discretely with syllables (on the left), and a more recent tone overshadowing a previous tone (on the right).

DS tone identification studies have required labeling of both tones, and this approach does not appear to make the task particularly difficult. Additionally, it does not account for stronger performance on final syllables observed in previous studies (Broselow et al., 1987; Hao, 2012, 2018; Sun, 1998).

Another possibility is that memory constraints might come into play (e.g. the phonological loop, Baddeley, 1968). In Mandarin, the dominance of the syllable as a discrete functional unit (in opposition to words or phonemes, cf. discussion of syllables and phonemes in production planning in C. Li, Wang, & Idsardi, 2015, and O'Seaghdha, Chen, & Chen, 2010) might have unexpectedly strong effects on serial memory. One hypothesis might be that, all else being equal, in longer strings of tones it will always be the earliest tones/syllables that will be misidentified.

Finally, it may be that there is some asymmetrical perceptual interference such that the character of earlier tones can be colored by more recent tones (Figure 2.5). This would explain why tones in isolation are unaffected, and examination of specific cases might suggest how the overshadowing tone might distort the earlier tone. For example, in the present case with DS T3 stimuli, the neutral tone is expected to be somewhat higher than the offset of T3F. This might make T3F sound more like a rising tone. This idea would need to be further fleshed out before it can be fully tested.

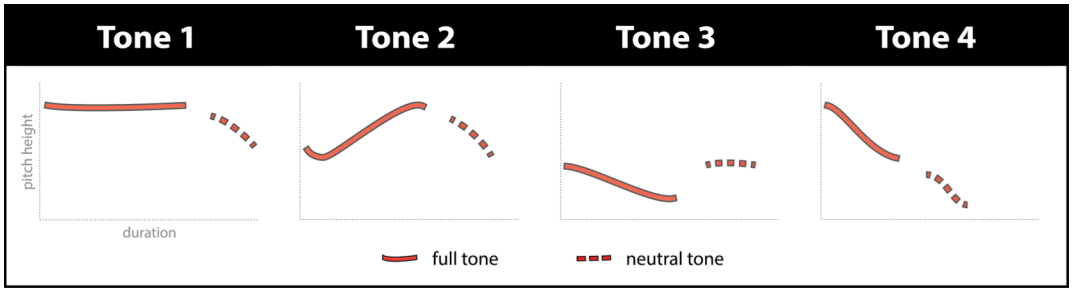


Figure 2.6. Depiction of the four tones followed by a neutral tone in word final position.

While the apparent difficulties encountered for T1, T2, and T3 need to be addressed, the apparent ease with which T4 is consistently identified in DS contexts is also in need of explanation. One possibility is that this high performance in the current situation is due to the relatively stable shape of T4 in both isolated MS and contextualized DS forms. As can be observed in the previously presented Figure 2.2, along with the major contextual changes undergone by T3, some T1 and T2 tokens also appear to vary a bit more widely from the citation form of that tone. In contrast, while the duration of T4 varies, its contour is very consistent. Closer analysis of which stimuli were related to T1 and T2 errors might reveal that it was a subset of the stimuli that were responsible for those errors in DS context.

Another possibility is that the form of the neutral tone served as a distinctive cue for T4. Though acoustic analyses have not been carried out on those syllables in the current stimuli, based on previous research (W.-S. Lee & Zee, 2014), we can speculate as to the likely character of the neutral tones in the present stimuli. Figure 2.6 presents speculative versions of the neutral tone following each of the full tones. It can be seen that the neutral tone following T4 has a somewhat lower pitch than other realizations of the neutral tone, and that it continues the trajectory of the preceding pitch contour (i.e., continues to fall). In contrast, all other forms of the neutral tone are a bit higher in overall pitch (thus less mutually distinctive) and move

away from the contour of the preceding full tone. This distinctive neutral tone following T4 might help listeners identify the tone more accurately. Additional acoustic analyses might reveal whether such a pattern is apparent for the current stimuli.

The role of multiple talkers in the present results deserves a brief comment. Previous L2 research has suggested that talker variation does not disproportionately impact tone identification accuracy for L2 compared to L1 listeners (C.-Y. Lee et al., 2009; C.-Y. Lee, Tao, & Bond, 2013). However, those studies presented only MS targets (sometimes occurring finally in a carrier phrase). It might be the case that there is some sort of interaction involved with talker variability in the current study, such that it has stronger impacts on L2 listeners for initial syllables in disyllabic contexts. It is also possible that, randomly, the specific talkers used in the current study were more challenging than talkers in previous studies.

All of these above uncertainties emphasize the need for more research to contrast tone perception using single and multi-syllable stimuli. While use of MS stimuli is understandable for many practical reasons, it also clearly limits our understanding of L1 and L2 tone perception processes. Future work can attempt to address some of the open questions raised above.

2.4.2 Tone 3 allotones

Shifting discussion away from broad contextual effects to the level of abstract representations, results paint a fairly encouraging picture for L2 tone perception. While L2 listeners identified T3 less accurately overall in disyllabic context than L1 listeners did, the error patterns they made suggest that, as a group, L2 learners tend to

make the same types of confusions as L1 listeners—just more of them. In short, it seems that L2 has successfully abstracted allotonic variation of T3 to be part of a single category. This suggests that successful formation of abstract phonological tone representations does occur, and adds one more type of evidence to bolster results from categorical perception studies reviewed above (Ling et al., 2016; G. Shen & Froud, 2016, 2018). At the same time, just as those studies suggest some incompleteness in L2 categories, the lower accuracy for T3 DS in Experiment 2, also suggests some differences for L2 categories. In this case, as discussed above—that they are less robust than L1 to the effects of context.

While the allotone results can be taken as evidence of abstraction for the T3 *category*, it is important to point out that this does not mean this category can *necessarily* be encoded lexically in a native-like fashion, i.e., automatically and implicitly. If L2 learners can only retrieve categories through explicit and effortful processes, then they will be minimally useful for lexical representations.

2.4.4 Remaining questions: L2 tone pedagogy

One motivation for examining T3 allotones was that L2 Mandarin teaching is somewhat divided over how best to present T3 to learners. While traditional pedagogical practice has emphasized T3D at the expense of T3F, recent discussions have consistently argued for a more prominent role for T3F (cf., Lin, 1985; J. Shi, 2007; Sparvoli, 2017; H. Zhang, 2014). Present results unfortunately do not provide much support for any particular position, but they do suggest that, at least for advanced learners, T3 allotones are often being correctly encoded. Whether this occurred due to specific instructional practices, incidental provision of explicit

information (e.g., in online discussions), or simply through the accumulation of L2 tone experience remains unclear. Future work might use the present allotonic paradigm as a way to shed light on developmental issues in L2 tone perception.

Chapter 3: Investigation of tones in L2 lexical recognition of words in isolation

3.1 Introduction

Experiment 2 utilized a tone identification task to investigate perception and representation of tones in advanced L2 Mandarin learners. While useful, tone identification is limited to providing information about tone categories themselves, and cannot speak directly to processes of tone word recognition. The next two chapters utilize lexical tasks to examine lexical processes in advanced L2 learners.

Chapter 3 lays out the motivation, logic, and results of a lexical decision task (LDT) meant to shed light on the role of tones in online word recognition in advanced L2 Mandarin listeners. The chapter will also report results from offline vocabulary and tone knowledge tests targeting vocabulary used in the experiment, and use those results to further explore the nature of online responses. All of these results will be discussed with reference to the three hypotheses outlined in Chapter 1.

3.1.1 Background and Motivation

As reviewed above, in Pelzl et al. (2018) we used DS nonwords in a lexical decision task contrasting tonal and segmental mismatches with real words. We found that advanced L2 learners were largely unable to utilize tone cues to reject these nonwords; as a group they performed below chance (35% correct rejection). This suggested a general tendency to accept (non)words without reference to tone cues. This effect was particularly striking when contrasted with results for segmental nonwords, where the L2 group achieved 84% accuracy overall. Our statistical model

suggested that L2 learners were about 26 times less likely to reject a tone nonword than a segmental nonword.

The findings in Pelzl et al. (2018) provide convincing evidence of persistent tone difficulty for advanced L2 learners. The results, however, allowed for several potential explanations of that difficulty, which Experiment 3 will attempt to address.

First, the low L2 accuracy for tonal nonwords 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 than L1 listeners. In this case, the hypothesis is that less familiar words have lower quality ('fuzzy') 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. Though we measured offline knowledge of tones in our previous study, we did not attempt to measure L2 confidence for the tones or meanings of the associated words. In Experiment 3, by measuring confidence in tones and definitions, the current study will attempt to account more thoroughly for the role of L2 familiarity in LDT outcomes. Assuming Cook and Gor are correct, this can provide some insight into the quality of L2 tone representations. If high quality (i.e., correct and confident) tone representations lead to more accuracy in rejection of tone nonwords, this suggests that the Tone Representation Hypothesis can account for L2 word recognition difficulties.

3.1.2 Benefits of ERPs

Another limitation of our previous study is that accuracy results in a task such as a LDT only reflect the final decision point for each trial. This leaves open the possibility that sensitivity to tones could be present during the word recognition process, but for some reason (perhaps lack of confidence) did not result in a correct rejection of tonal nonwords. To address this limitation, the current study will use event-related potentials (ERPs) to assess the listener's word recognition process as it unfolds during each trial.

Because ERPs allow us to examine the listener's response to words (or nonwords) as it unfolds over time, they are more sensitive than accuracy or response time measures, and potentially able to capture implicit evidence of learned representations that would be unobservable from overt behavior (e.g., McLaughlin, Osterhout, & Kim, 2004). The ability to examine both neural and concurrent behavioral responses makes ERPs highly useful for examining word recognition processes.

The N400 component is particularly useful in examination of lexical recognition processes, and will be the main ERP outcome of interest for the LDT in Experiment 3. 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, Phillips, & Poeppel, 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., 2018; 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 words.

3.1.3 Disyllabic word recognition in Mandarin

In addition to extending L2 tone research, the present study also expands on previous Mandarin ERP research more generally by examining the role of tones and vowels in isolated DS word recognition. Most previous ERP research has examined L1 sensitivity to tone mismatches in MS words in predictive contexts. To my knowledge, only one previous ERP study has examined spoken disyllabic Mandarin word recognition (cf. Liu, Shu, & Wei, 2006), however, that study did not specifically consider tonal effects.

There are good reasons to conduct additional research on disyllabic spoken word recognition in Mandarin. First, although MS Mandarin words—with their many homophones and tone neighbors—are quite novel and tend to attract most of the attention, the majority of words in Mandarin are DS (Duanmu, 2007). As already argued above in Chapter 2, by restricting research to MS word recognition, we run the risk of severely misrepresenting the process. Unlike MS words, DS (and longer)

Table 3.1. Word counts according to word length (syllables) in SUBTLEX-CH (Cai & Brysbaert, 2010) for most frequent 10000 words.

	total words	ton neighbors	Homophones
monosyllables	2021	5.13 (max 33)	2.02 (max 15)
disyllables	7118	1.10 (max 6)	1.02 (max 4)
trisyllables	717	1.01 (max 2)	1.01 (max 2)

A value of 1 for tone neighbors or homophones would indicate that no tone neighbors or homophones exist.

words have very few tone neighbors or homophones (Table 3.1). This makes the process of DS word recognition possibly very different from the process of MS word recognition. While neither type of word should be neglected, at least in ERP research, DS words require much more attention than they have received so far.

The differences between MS and DS words also give DS words some desirable properties for examining word recognition processes, and make it easy to avoid some common pitfalls in MS word recognition paradigms.[7] Specifically, DS words allow for the relatively straightforward creation of nonword phonological neighbors for real words. This allows for phonological mismatches to be determined solely on the basis of whether a target is or is not a word. MS words, in contrast, are not so flexible and nonwords are harder to come by. Consequently, when working with MS words, it is usually necessary to provide some type of constraining context to test the critical manipulation (e.g., a tone or vowel mismatch). This is not problematic, *per se*, but does tie responses to a larger context of expectations. In contrast, while DS words can be used in contextual word recognition paradigms, they are also usable in complete isolation.

The present LDT will take advantage of DS words to test whether L1 and advanced L2 listeners show sensitivity to phonological mismatches during isolated spoken word recognition. By examining both behavioral and ERP responses, we will

[7] One common pitfall is treating bound morphemes as independent MS words. This is problematic in that N400 effects might be elicited due to the oddness of hearing such morphemes on their own, rather than due to the phonological properties of interest.

test whether listeners are equally sensitive to tone and vowel mismatches during spoken word recognition.

3.2 Experiment 3: Lexical decision for words in isolation

3.2.1 Experiment 3: research questions and hypotheses

Experiment 3 will utilize a LDT while recording EEG in order to address the following research questions regarding advanced L2 learners' behavioral and neural responses to tones and vowels in nonwords.

(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?*

(2) *Are L2 listeners equally sensitive to vowel and tone mismatches in isolated disyllabic words (as indexed by the N400)?*

Question (1) investigates whether advanced L2 listeners show different levels of behavioral accuracy depending on the nature of phonological mismatches with real words. Based on our previous results (Pelzl et al., 2018), we expect that L2 listeners will be less accurate in rejection of tone nonwords compared to vowel nonwords, demonstrating less ability to use tone cues than segmental cues in online word recognition. However, as explained below, the current task is intentionally less difficult than the previous study, and thus is it possible that we could find highly accurate L2 performance for tone mismatches.

Question (2) asks whether neural responses demonstrate N400 effects to both vowel and tone nonwords. The most straightforward outcome would be that, if behavioral results show less accuracy for tone than vowel mismatches, N400 responses would similarly reflect less robust responses for tones compared to vowels.

However, it is possible that we could see different patterns. First, ERP responses to mismatching tones could be evident despite poor behavioral accuracy for those trials. This would indicate implicit tone knowledge that is not evident in the behavioral response. Alternatively, we could find that, despite high accuracy in behavioral performance, neural responses to tones (and vowels) are weak or non-existent within the N400 window. 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.

For the L1 group, we expect high accuracy and strong N400 effects for both vowel and tone nonwords. As the response will be tied to lexical recognition in isolation, and thus not related to confounding expectations about phonological form, we expect no differences between nonword conditions. That is, responses in both conditions should equally reflect difficulty in accessing a real word.

Although direct comparison of L1 and L2 groups is not a major concern in the current study, because the native response pattern is implicitly assumed in evaluation of L2 results, the two groups will be compared in statistical analyses. The L1 group also serves as a test of the experimental materials to show whether they effectively induce the expected nonword effects.

Questions (3-5), below, are of a more exploratory nature and will be pursued by considering the relationship of L2 learners' explicit knowledge and subjective confidence (measured by an offline vocabulary test) to online accuracy and ERP results. All of these questions assume we will find less accurate L2 performance for tone cues compared to vowels. Specifically:

(3) *Does lexical familiarity impact L2 behavioral responses?*

That is, will we find that correct explicit knowledge paired with high subjective confidence leads to higher accuracy in the rejection of tone and vowel nonwords?

(4) *Do specific tone confusions impact L2 behavioral responses?*

This question examines whether low-level perceptual difficulties (i.e., specific difficult tone contrasts) might be partially or wholly responsible for lower accuracy in rejection of tone nonwords for L2 learners.

(5) *Does lexical familiarity impact ERP responses?*

This question asks whether L2 N400 effects are potentially modulated by explicit knowledge.

3.2.2 Experiment 3: Task

Experiment 3 used an auditory lexical decision task (LDT) without priming. Participants heard a single disyllabic Mandarin word or nonword and had to decide whether it was a real word or not. ERPs and behavioral accuracy were recorded for each trial. After the experiment, participants were also given an offline vocabulary knowledge test targeting the real word counterparts of all nonwords they heard in the LDT.

3.2.3 Experiment 3: stimuli design and production

Stimuli selection began with a set of 96 DS real words (e.g., *fang1fa3* /fɑŋ1fɑ3/ 'method'). All real words were high frequency nouns, mostly selected from a widely used Chinese textbook series (*Intergrated Chinese*, 2008). Six critical real

words were chosen for each tone combination (T1T1, T1T2, T1T3, T1T4, T2T1, etc.), avoiding words with neutral tones or *erhua* (a syllable final "-r" [ɚ]) on the second syllable. An additional 32 words were chosen as fillers, following the same guidelines as for critical stimuli, maintaining the balance of tone types across real words. Where the textbook wordlist proved insufficient, additional high frequency words were selected from the SUBTLEX-CH corpus (Cai & Brysbaert, 2010) relying on the author's intuition to select words likely to be known.

On the basis of the real words, two types of (pronounceable) nonwords were created, differing from real words only with respect to a tone or vowel. For the tone mismatch condition, the tone of the first syllable was changed producing a nonword (e.g., *fang**2**fa3* vs. real word *fang**1**fa3*). I 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. *f**e**ng1fa3* /fəŋ1fɑ3/ vs. *f**a**ng1fa3)*, i.e., *vowel nonwords*.

The stimuli in the present design display several improvements over those used in our previous study (Pelzl et al., 2018). First, all tones are balanced across real words, and tone changes are also balanced across tone nonwords—that is, T1 becomes T2, T3, and T4 an equal number of times, and similarly for other first syllable tones.[8] Second, whereas Pelzl et al. (2018) swapped out entire syllable

[8] In the case where T2 or T3 changes might have been confused with T3 sandhi, the relevant change type was avoided. T3T3 words never became T2T3 nonwords; T2T3 words never became T3T3 nonwords. Instead T1 and T4 changes were balanced across those items.

rhymes, including syllable final /n/ and /ŋ/ (e.g., *xiang3fa3* ‘thought’ /ɕiɑŋ3fɑ3/ became the nonword *xu3fa3* /ɕy3fɑ3/), the current stimuli limited changes to vowels. Some effort was also made to minimize the ‘magnitude’ of vowel changes—though this was largely based on intuition rather than empirical evidence or theory (e.g., vowel features). These steps made the segmental nonwords in the current design a bit more challenging and were meant to allow for a fairer comparison between nonword types. Third, in order to prevent listeners from rejecting nonwords before onset of the second syllable, syllable gaps (e.g., *fai* /fai/) and very rare syllables (e.g., *cen* /tsʰən/) were avoided, and the first syllables of all nonwords were checked against the most frequent 5000 words in the SUBTLEX-CH corpus to be sure there were viable lexical competitors. In a small number of difficult cases, this restriction was waved because there were competitor words L2 learners were likely to be familiar even though they were not in the most frequent 5000 (e.g., the syllable *shao3* in vowel nonword *shao3du2* occurs in the word *shao3shu4* ‘minority’, a word L2 are likely to know, despite being somewhat less frequent). Nonwords were checked against several large comprehensive Mandarin dictionaries using the *Pleco* Chinese dictionary app. Finally, care was taken that the initial syllable of critical stimuli was never repeated within a list (if fillers repeated a syllable, the filler always occurred second).

Due to the way the stimuli in Pelzl et al. (2018) were constructed—by clipping (non)words out of fluently produced sentences—that LDT was very challenging. In some sense those results can be viewed as a worst-case scenario for L2 tone performance. The current study aimed to explore tone perception in less extreme circumstances using somewhat easier stimuli. To this end, all stimuli were

produced in isolation by a native Chinese (female) speaking at a comfortable rate. The program WaveSurfer (Sjolander, 2000) was used for recording, which was conducted in a sound booth at UMD using the internal microphone of a laptop computer. Audio was recorded at a 48,000 Hz sampling rate (eleven items were originally recorded at 44,100 Hz and later resampled to 48,000Hz). Each word or nonword was presented to the speaker in Pinyin (Mandarin romanization) in a random order using a presentation script in *Praat*. Any items that were judged to be mispronounced were later re-recorded by the same speaker under the same conditions (except for eleven items that were re-recorded in a sound booth at BNU). Using *Praat*, all individual stimuli were cut out of the original audio files to create individual .wav files for each item. The average intensity of each file was scaled to 70dB, and 200 ms of silence were appended at the end of each file.

Average duration of stimuli was examined using *Praat* and is shown in Table 3.2. There were some slight differences in duration between real words (and fillers) and nonwords. While the differences were not statistically significant ($F_{(3,316)}$ =1.644, p =.179), this does not prove they are not practically significant for listeners. However, given the diversity of initial syllables involved, the fact that none of them were repeated across items in a list, and that a given real word and its nonword counterparts never occurred in the same list, it seems quite unlikely that duration

Table 3.2. Average durations of auditory stimuli for the Lexical Decision Task (Experiment 3)

condition	avg. dur. (sd)
real words	600 (83)
vowel nonwords	621 (70)
tone nonwords	615 (73)
fillers	600 (73)

alone would be a useful cue of differences between conditions.

The above process resulted in 96 triplets consisting of a real word and its vowel and tone nonword counterparts. The stimuli were divided into three balanced 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.

An offline vocabulary test was also constructed. The format is illustrated in Figure 3.1. For each L2 participant, the test included all real word counterparts for vowel and tone nonwords encountered during the LDT. Each item provided Chinese characters and toneless Pinyin and required participants to supply 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 has 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

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	2	beef ribs	2

Figure 3.1. Example of format for offline vocabulary knowledge test

blank and supply “0” for confidence.

3.2.4 Experiment 3: Procedures

Thirty-six participants (24 L1, 12 L2) were tested in the lab at Beijing Normal University (BNU). Seven additional L2 participants were tested under highly comparable conditions 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 LDT, instructions presented on screen critically 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 balanced 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.

The beginning of each trial was signaled with a 150 ms ‘beep’ and the appearance of a fixation cross. After 350 ms, the auditory stimulus played. 1200 ms after the end of the auditory target, the fixation cross disappeared and a question prompt asked for participants to decide whether what they heard was or was not a

Mandarin word. After their decision was made, there was a 2 sec pause before the next trial began. The entire lexical decision experiment lasted approximately 15 minutes.

After all the ERP and listening experiments were finished, L2 participants completed the offline vocabulary test to establish their knowledge and subjective confidence in that knowledge for the real word counterparts of vowel and tone nonwords that had occurred in the LDT.

3.2.5 Experiment 3: 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.

3.2.6 Experiment 3: 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 100ms preceding the onset of the auditory stimulus was exported for further processing in *R* (R Core Team, 2017). A single average amplitude was obtained for each trial for each electrode for each subject in an auditory N400 window (400-900ms). 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 suggests 900 ms is a reasonable end point 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 L1 grand average waveforms suggested these electrodes had strong and consistent N400s, nor was there any 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.

Table 3.3. Descriptive accuracy results for the Lexical Decision Task (Experiment 3)

group	cond	mean acc % (sd)
L1 (*n*=24)	real	98 (15)
	vowel	94 (24)
	tone	95 (22)
L2 (*n*=18)	real	86 (35)
	vowel	85 (36)
	tone	63 (48)

Finally, only trials that elicited correct behavioral response (correct acceptance or correct rejection) were retained for final analysis. After all of these steps, the final dataset contained 43,567 data points (80.0% out a of total possible 54,720 data points: L1=88.1%; L2=70.2%).

3.2.7 Experiment 3: Behavioral LDT results and statistical analysis

Reliability for the LDT data was high for all three lists (list A: α=.94; list B: α=.92.; list C: α =.93). Descriptive behavioral results from the lexical decision task are shown in Table 3.3. L1 displayed high accuracy across all conditions, while L2

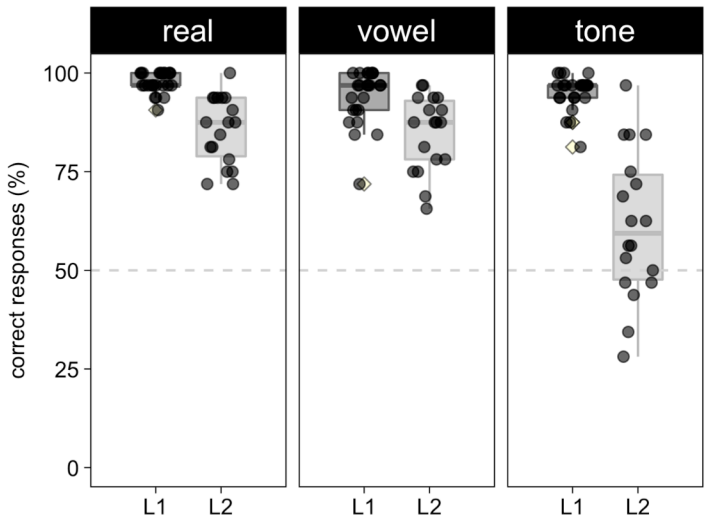


Figure 3.2. Boxplot of accuracy results for Lexical Decision Task (Experiment 3). Each circle indicates an individual participant's mean score. Diamonds indicate that scores at that level are outliers. The dashed line indicates the level which would be equivalent to chance performance.

had noticeably lower accuracy overall, with tone nonwords registering the lowest accuracy. To capture sensitivity for these results d-prime 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, Namyst, et al., 2016; Jurafsky & Martin, 2009). As with accuracy, d-prime results suggest overall higher sensitivity to nonwords for L1 listeners with little difference between nonword conditions (vowel *d'*=3.63(*sd*=.55); tone *d'*=3.67(*sd*=.45)). In contrast, L2 has less sensitivity overall and a larger difference between conditions that might suggest vowel nonwords are detected more readily than tone nonwords (vowel *d'*= 2.35(*sd*=.67); tone *d'*=1.69(*sd*=.90)).

Behavioral results were submitted to a generalized linear mixed-effects model, with the factors *condition* (real word, tone nonword, vowel nonword), and *group* (L1, L2), and their interaction. Model fitting procedures were the same as for experiments 1 and 2. Model convergence difficulties were addressed by suppressing correlations in random effects. Inclusion of the nuisance factor *list* (with subjects nested under lists) did not improve model fit, and so it was not retained in the final model. The fully specified model included by-subject random intercepts and slopes for the effect of condition, and by-item random intercepts and slopes for condition and group and their interaction.

Table 3.4 reports main effects and interactions for accuracy in Experiment 3. The effects of condition (χ^2=29.04, p<.001) and group (χ^2=52.20, p<.001) were both statistically significant. Critically, there was also a significant interaction between condition and group (χ^2=11.24, p=.004).

Table 3.4. Mixed Model ANOVA Table for accuracy results (Type 3 tests, LRT-method) (Experiment 3)

Effect	Df	Chisq.	Chi Df	Pr(>Chisq)	
condition	13	29.04	2	<.001	***
group	14	52.20	1	<.001	***
condition × group	13	11.24	2	.004	**
Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1					
model formula: accuracy ~ condition * group + (condition || subject) + (condition * group || item)					

Table 3.5. Planned comparisons for accuracy of Lexical Decision Task (Experiment 3)

Comparison	Estimate	SE	z value	Pr(>|z|)	
L1 Vowel vs. Tone	0.11	0.41	0.27	.789	
L2 Vowel vs. Tone	1.85	0.37	5.03	<.001	***
L1 V-T vs. L2 V-T	-1.74	0.50	-3.45	.001	**
Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1					

Planned comparisons are reported Table 3.5. 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 L1 does *not* display a similar difference. This is in fact born 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 (b= 1.85, SE = .37, z = 5.03, p <.001). L2 listeners were about two and a half times more likely to incorrectly accept tone nonwords than vowel nonwords (*.27/.15=2.6*). Finally, the difference between L2 vowel and tone was significantly larger than the difference between L1 vowel and tone (b= -1.74, SE = .50, z = -3.45, p <.001).

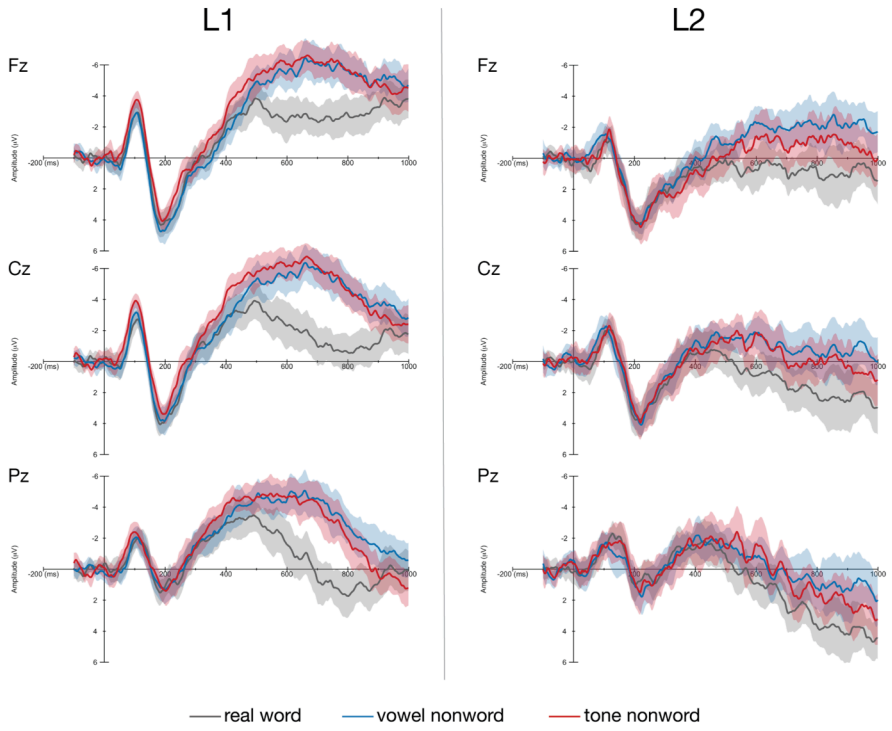


Figure 3.3. Grand average waveforms for LDT (Experiment 3), only correct trials are included (40 Hz low pass filter). Shaded areas around lines represent 95% within-subjects confidence intervals.

3.2.8 Experiment 3: ERP results and statistical analysis

N400 amplitudes for the LDT (correct trials only) are depicted visually as grand average waveforms in Figure 3.3. Across all midline and central electrodes, L1 appears to show strong N400 effects to both vowel and tone nonwords. In contrast L2 appears to show attenuated N400 effects overall, and different magnitudes of N400 for vowel and tone nonwords, with tone nonword responses diverging less strongly, from real word responses.

Averaged N400 amplitudes from the 400-900ms window were submitted to a linear mixed-effects model with crossed random effects for subjects and items, and with electrodes nested under subjects. The nesting of electrodes reflects the assumption that amplitude variation is strongly related across electrodes for each

specific subject. This approach does not entirely account for relatedness between electrodes, as it makes no distinction between electrodes closer or further away from one another. However, this seems like an acceptable first approximation as we lack clear hypotheses about differential effects for ERP amplitudes in different regions of interest (e.g., anterior vs. posterior regions). Models included fixed factors for *condition* (real word, vowel nonword, tone nonword) and *group* (L1, L2) and their interactions. All analyses were once again conducted in *R* (version 3.3.3, R Core Team, 2017), and models were fit using the *lme4* package (version 1.1-12, Bates, Mächler, Bolker, & Walker, 2015), in conjunction with the *mixed* function in *afex* (Singmann et al., 2017). Convergence difficulties were addressed by specifying uncorrelated random effects. Effects coding was used, and p-values were obtained using Satterthwaite's method. Despite remaining controversy about the proper way to calculate degrees of freedom for linear mixed-effect models, recent advice is that using p-values is nevertheless more conservative (reduces Type I error) than using absolute t-values >2 (Luke, 2017). Results are reported for *F*-tests in ANOVA tables for this and subsequent linear mixed-effects models. The maximal model was fit first and was then compared to less complex models to test random effects (Barr et al., 2013). The maximal model was retained and fit using REML. This model included random intercepts for subjects and items, by-subject random slopes for condition, and by-item random slopes for condition and group and their interaction.

Model results are reported in Table 3.6. There were statistically significant main effects of condition ($F_{2,\ 113.299}$=18.04, p <.001) and group ($F_{1,\ 44.248}$=12.95,

Table 3.6. Mixed Model ANOVA Table for ERP results (Type 3 tests, LRT-method) (Experiment 3)

Effect	numer Df	denom Df	F	Pr(>F)	
condition	2	113.299	18.04	<.001	***
group	1	44.258	12.95	<.001	***
condition × group	2	105.509	3.03	.053	.
Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1					
model formula: amplitude ~ condition * group + (condition || subject / electrode) + (condition * group || item)					

Table 3.7. Planned comparisons for ERP results of Lexical Decision Task (Experiment 3)

Group	Comparison	Estimate	SE	z value	Pr(>|z|)	
L1	Real vs. Vowel	2.86	0.51	5.64	<.001	***
	Real vs. Tone	3.27	0.81	4.04	<.001	***
	Vowel vs. Tone	0.41	0.75	0.54	.586	
L2	Real vs. Vowel	1.48	0.52	2.87	.013	*
	Real vs. Tone	0.72	0.83	0.87	.659	
	Vowel vs. Tone	-0.76	0.78	-0.98	.659	

p<.001), and the interaction of condition and group was marginally significant ($F_{2, 105.509}$=3.03, p=.053).

Planned comparisons (with Holm adjustments for p-values) are reported in Table 3.7. In the ERP models reported below and in later sections, *b* estimates in planned comparison can be taken to represent amplitude differences in μV. For L1 listeners, real words evoked significantly more positive amplitudes than either vowel nonwords (b= 2.86, SE = .51, z = 5.64, p <.001) or tone nonwords (b= 3.27, SE = .81, z = 4.04, p <.001), while there was no significant difference between vowel and tone nonword responses. For L2 listeners, real words evoked a significantly more positive response than vowel nonwords (b= 1.48, SE = .52, z = 2.87, p=.013), while there was no significant difference between tone nonwords and either real words or vowel nonwords. Visual depiction of model results are shown in violin plots in Figure 3.4—

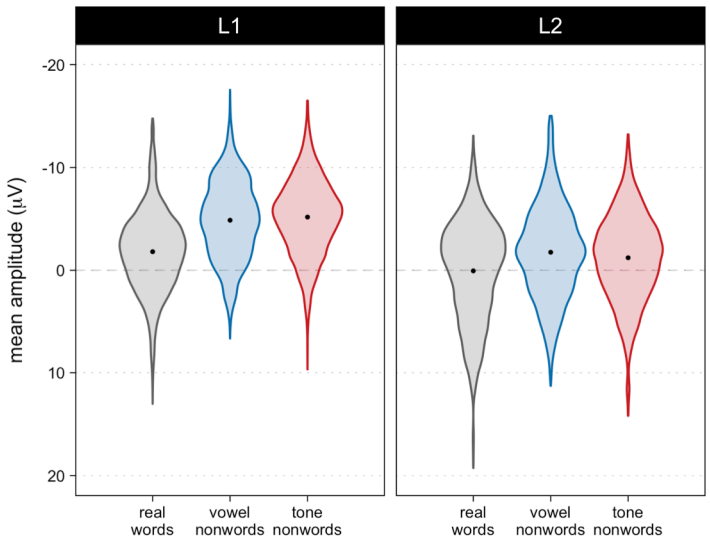


Figure 3.4. Violin plots for model estimates of N400 amplitudes (400-900ms) from Lexical Decision Task (Experiment 3). The black dots represent the model estimated group means for each condition, with shaded areas representing the distribution of estimated responses.

note that, in contrast to waveform visualizations, these results are based on model-generated estimates rather than descriptive statistics.

In summary, the L1 group displayed significant and similarly strong N400 effects for both vowel and tone nonwords. In contrast, the L2 group displayed significant N400 effects only for vowel nonwords, while tone nonwords elicited N400 responses that were intermediate between vowel nonwords and real words.

3.2.9 Experiment 3: Offline vocabulary test data processing

The offline vocabulary test produced four data points for each nonword that an L2 participant encountered. For each word they received an accuracy score for the tones and definition they supplied. For example, if the word was *lü4shi1* 'lawyer', and the participant provided 41 as the answer for tones, this would be scored as 1, while an any other set of two numbers would result in a score of 0 for the tone on that item. Note that this scoring counted tones on both syllables, whereas the LDT

nonwords only ever mismatched real words with respect to tones on the first syllable. In that sense, this scoring approach is rather 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 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 items 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).

3.2.10 Experiment 3: Offline vocabulary test results

Descriptive results for tones in the offline test are displayed in Table 3.8, along with related accuracy for those items in the LDT. It can be seen that, even for

Table 3.8. Results of L2 offline vocabulary test requiring participants to supply tones and tone confidence ratings for nonwords. Tone accuracy indicates whether supplied tones were correct. LDT accuracy indicates whether the related nonwords were correctly rejected in the LDT.

Confidence ratings and accuracy of L2 supplied tones				
Condition	conf. rating	k (items)	tone acc. %	LDT acc. %
Vowel	3 (high)	377	87	84
	2 (mid)	132	56	86
	1 (low)	16	62	88
Tone	3 (high)	385	85	66
	2 (mid)	130	52	52
	1 (low)	4	25	00

Table 3.9. Results of L2 offline vocabulary test requiring participants to supply definitions and definition confidence ratings for nonwords. Def. accuracy indicates whether supplied definitions were correct. LDT accuracy indicates whether the related nonwords were correctly rejected in the LDT.

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

high confidence items, explicit tone knowledge was often somewhat inaccurate (*mean*=85%), and as confidence decreased, tone accuracy tended to decrease as well. As we would expect, tone accuracy and confidence only appear to impact LDT accuracy for tone nonwords, and not vowel nonwords.

Results for definitions are displayed in Table 3.9. It can be seen that L2 participants' subjective confidence about their knowledge of definitions seems quite accurate, as high confidence items were correctly defined 98% of the time. There does not appear to be a straightforward relationship between definition confidence and accuracy in the LDT. This makes sense insofar as the LDT did not test semantic knowledge, but only word form recognition.

In sum, we find that L2 offline knowledge suggests some difficulties in accurate encoding of tones in lexical representations. Even when explicit knowledge is fully available and words are confidently recognized, L2 tone knowledge is still inaccurate over 10% of the time. Obviously, such limitations could impact online responses in the LDT.

3.2.11 Experiment 3: Exploratory "Best Case Scenario" analysis

3.2.11.1 Overall Accuracy

As an attempt to clean up some of the noise introduced to the LDT by insufficient L2 word knowledge, an exploratory 'Best Case Scenario' analysis was conducted. In this analysis, we retain only the subset of trials that targeted items (real word counterparts of nonwords) for which an L2 participant had indicated correct and confident knowledge (confidence rating = 3) of both tones and definitions. This comprised 301 tone nonword and 303 vowel nonword trials (604 total, 55% of total nonword trial data).

Table 3.10 presents descriptive accuracy results for the two nonword conditions in the Best Case Scenario data for the LDT. 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.

Table 3.10. Descriptive accuracy results for the 'Best Case Scenario' analysis of the LDT

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

Table 3.11. Comparison of conditions for accuracy results in the 'Best Case Scenario' analysis of the LDT (Type 3 tests, LRT-method) (Experiment 3)

Effect	b	SE	z	p	
Tone vs vowel nonword	-1.29	0.42	-3.08	.002	**
Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1					
model formula: accuracy ~ condition + (condition | subject) + (condition | item)					

Results are displayed in Table 3.11. There was a significant difference in accuracy for tone and vowel nonwords (*b*=-1.29, *SE* = .42, *z* = -3.08, *p*=.002).

In summary, even after accounting for limits on (offline) L2 word knowledge and subjective confidence of that knowledge, L2 still displays a more limited ability to reject tone nonwords than vowel nonwords.

3.2.11.2 Accuracy by tone manipulation

Another consideration that can be explored in the LDT data is whether there was any evidence of differential impacts according to the type of tone switch between the real word and its tone nonword form. That is, whether changing T1 to T2, T1 to T3, T1 to T4, and so on were equal in the difficulty they induced in listeners. Table 3.10 shows the accuracy of LDT decisions for tone nonwords according to the tone switch that created the nonword. Two points stand out. First, tone switches involving tones that have been found to be highly confusable with one another (e.g., T2-T3, T1-T4) seem related to lower overall accuracy (T2 to T3: *mean*=55%; T3 to T2: *mean*=30%; T1 to T4: *mean*= 55%; T4 to T1: *mean*=68%). Similarly, what we would expect to be easy tone distinctions (high vs. low pitch height, rising vs. falling contours) in general do appear to have higher accuracy (T1 to T3: *mean*=91%; T3 to T1: *mean*=72%; T2 to T4: 81%; T4 to T2: 84%). At the same time, apart from T1 to T3, any given switch produces a drop of 15% or more in accuracy. In other words, while some specific tones may be harder than others, almost all tones appear difficult to some degree—even when offline knowledge suggests learners have correct and highly confident knowledge of which tones belong to which word.

Table 3.12 L2 LDT accuracy by tone switch in the Best Case Scenario analysis (Experiment 3)

switch type	total data points	accuracy %
T1 to T2	29	62
T1 to T3	23	91
T1 to T4	33	55
T2 to T1	22	55
T2 to T3	20	55
T2 to T4	21	81
T3 to T1	18	72
T3 to T2	20	30
T3 to T4	30	73
T4 to T1	34	68
T4 to T2	25	84
T4 to T3	26	81
	total: 301	*67*

3.2.11.3 Quality of L2 knowledge for correct trials in ERP data

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. For these trials, L2 knowledge of definitions for 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.

3.3 Experiment 3: General Discussion

The results of Experiment 3 can be summarized briefly as follows. Advanced L2 listeners were significantly less accurate at rejecting tone nonwords than vowel nonwords. This accuracy difference persisted even when we examined only those trials where learners had accurate and confident offline knowledge of the tones and words that were being manipulated. While some of the difficulties seem attributable to specific tone confusions (e.g., T2 vs. T3), it is nevertheless the case that almost all tones created some level of difficulty. For ERPs, the L2 group displayed significant N400 effects for vowel nonwords that were correctly rejected, but the N400 response for correctly rejected tone nonwords was intermediate between real word and vowel nonword responses. Measures of L2 offline 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. Finally, L1 responses demonstrated strong and equivalent N400 effects for both tone and vowel nonwords.

We will consider each of these results in more detail below.

3.3.1 L2 accuracy for tone nonwords

As in our previous study (Pelzl et al., 2018), we found that advanced L2 learners are significantly less accurate at detecting tone mismatches in DS words compared to segmental (vowel/rhyme) mismatches. At the same time, the tone effects found in the present study are not nearly as strong as in the previous LDT. As suggested earlier, stimuli in the present study were expected to be easier than in our previous study. They can be considered easier in at least two ways. First, they were produced in isolation rather than clipped from sentences, thus reducing coarticulation

of tones. Second, they were produced at a slower rate (due to being produced in isolation). In other words, while it is unsurprising that the effect of tone nonwords is weaker here than in Pelzl et al. (2018), the fact that it persists demonstrates that the difficulties we observed previously were not merely an artifact of those stimuli. Advanced L2 listeners do in fact have more difficulty with Mandarin tone distinctions than they do with segmental distinctions.

More importantly, the significant difference in L2 accuracy for tone and vowel nonwords persisted even when we attempted to account for the accuracy and strength of offline knowledge associated with the critical vocabulary. As noted above, one possible explanation for previous LDT results was that, upon hearing tone nonwords, learners may have accurately *perceived* them, but proceeded to accept them as real words due to some *uncertainty* about their own knowledge of the relevant real words (Cook & Gor, 2015; Gor, 2018; Gor & Cook, 2018; Veivo & Järvikivi, 2013). For example, if they accurately perceived the nonword *fang**2***fa3*, perhaps they nevertheless accepted it because they were not confident that they knew the correct tones for the real word counterpart (*fang**1**fa3),* and so were more permissive in their decision process. This explanation would fall under the scope of the Tone Representation Hypothesis, thus favoring a representational account of L2 tone difficulty. While not providing a definitive answer, the Best Case Scenario analysis suggests that, if there is uncertainty involved in the rejection of tone nonwords, it cannot be reduced to uncertainty due to subjective confidence in the learners' *offline* knowledge about tones.

Another possible explanation for the lower accuracy of L2 tone performance is that it is driven by specific tones that are particularly difficult for L2 learners. This explanation would fall under the scope of the Tone Perception Hypothesis, and, more generally, would place *specific* tone contrasts in the category of 'difficult L2 sounds.' Again, the Best Case Scenario analysis suggests this is not the case. Although LDT accuracy for words involving T2 and T3 switches were decidedly lower than other tone switches, there was nevertheless a consistent drop in accuracy across almost all tone switch types. In other words, if the L2 difficulty is due to tone perception, it is tone perception *in general*, and not exclusively perception *of difficult tone contrasts*.

3.3.2 L2 ERP results

As noted earlier, a potential advantage of ERPs over simpler behavioral measures is the opportunity to examine the unfolding response to words as it occurs. In this respect, ERP results for the LDT provide some evidence that L2 behavioral decisions reflect typical lexical recognition processes. That is, for a simple LDT such as used here, we expect that stronger N400 effects will occur when listeners fail to access a word, potentially due to activation spreading to phonological neighbors of the targeted (non)word (Winsler, Midgley, Grainger, & Holcomb, 2018). This failure in lexical access should most often result in a rejection ('no' response) for that trial. Thus, we should expect overall that the N400 amplitude for correct rejections will be more negative than for correct acceptances. Present results for vowel nonwords are consistent with this account, showing significant N400 effects, though with an overall smaller magnitude than similar effects in L1 N400s.

For tone nonwords, however, L2 responses were not significantly different from either real word or vowel nonword responses, even though the analysis included only correctly rejected trials. There are several ways to account for these results. First, we might simply believe that the weak N400 effects are due to lack of power. This is not unreasonable, as there were nearly 25% fewer trials available for L2 tone nonwords than vowel nonwords. Additionally, offline vocabulary results suggest that, for some participants the available data contained a substantial number of guesses. Thus, we might believe that, given enough data, we would find N400 effects for all L2 participants. While possible, this does not seem particularly likely since it is clear that L2 participants have difficulty with tones behaviorally.

An alternative interpretation is that L2 N400 responses for tones failed to be different from real words because—for most L2 participants—there was a lack of consistent sensitivity to tone cues. Said another way, the intermediate nature of the L2 N400 for tone nonwords is due to averaging across participants. Consideration of individual ERP results lends some support to this contention. Figure 3.6 depicts individual participants' N400 responses (nonword minus real word amplitudes) across electrodes, with tone N400 effects on the right and vowel N400 effects on the left. Whereas all L1 participants display rather clear N400 effects for both conditions, fewer L2 participants appear to display N400 effects for the tone nonword condition, though, importantly, some show very clear N400 effects. Additionally, it seems to be the case that if a given L2 participant showed a tone N400 effect, they also displayed a vowel N400 effect, though the opposite is not necessarily true (though S219 appears to be an exception). If we believe this latter explanation for reduced L2 tone nonword

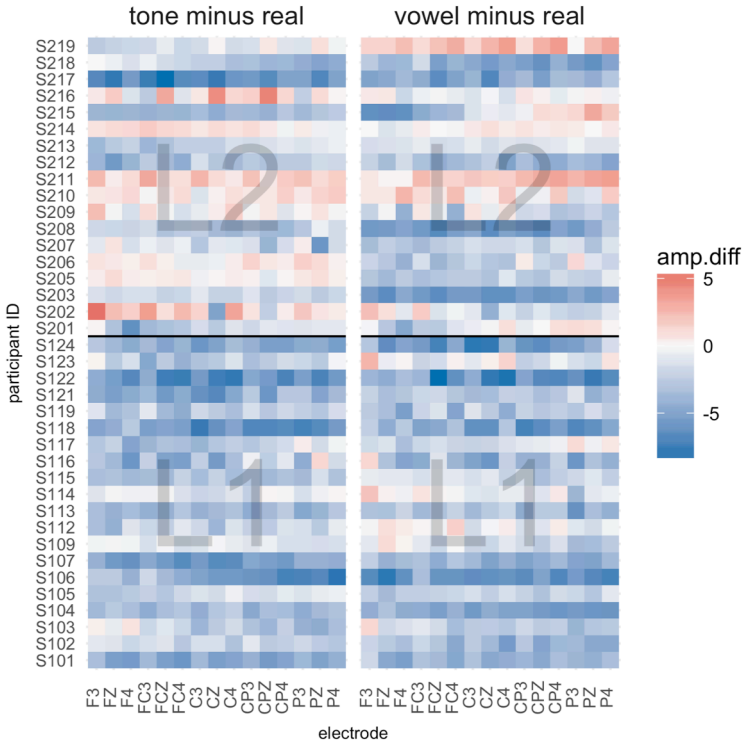


Figure 3.6. Raster plot of average N400 effects for individual participants in tone and vowel nonword conditions (correct trials only). Each row represents a participant and each column represents an electrode. Blue indicates negative amplitude relative to real word trials (i.e., an N400 effect), red indicate positivity relative to real word trials. L2 participants (S201-219) are on the top, L1 participants (S101-S124) are on the bottom.

N400 effects—that is, it results from averaging over the group—this might suggest the actual tone N400 effect for many participants does not differ from the real word response. For such participants, then, successful rejection of tone nonwords often occurred *despite* their not displaying tone sensitivity in the N400. This raises a question of how they might be arriving at their correct decisions, an issue we will return to in Chapter 5.

In sum, ERP results make it clear that correct rejection of vowel nonwords tended to occur as we would expect, namely, when concurrent N400s indicate sensitivity to vowel mismatches in nonwords. Though evidence for tone nonwords is less straightforward, one reasonable interpretation is that only some L2 learners display N400 effects for tonal nonwords, while others find ways to process tones in

an alternative fashion such that, even though they do not display rapid N400 effects, they nevertheless still correctly reject tone nonwords.

3.3.3 L2 Mandarin offline tone knowledge

Stepping back to consider L2 tone knowledge in the offline test, there is evidence of a substantial representational difficulty in L2 tone learning. Specifically, it appears that tones are difficult to encode in *explicit* long-term memory. 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 somewhat more than 10% of the time. The test format did not allow for a comparison with segmental accuracy, but it seems quite likely that tones—as a class of L2 sounds—are harder for L2 learners to remember correctly than consonant or vowel contrasts. We will revisit these issues later, after considering additional data from Experiment 4. However, I note here that the offline vocabulary test format used is likely to *overestimate* L2 tone knowledge and confidence. Because each item on the included Chinese characters as a prompt, it was possible for learners to rely on knowledge of tones for individual characters, even if they did not always know the relevant words in which those characters occurred, or perhaps would have been uncertain about tones without the character prompt. This can happen because some characters occur with extremely high frequency, even though the DS vocabulary they occur in may be relatively less frequent. For example, the extremely common character 和 (*he2* usually meaning 'and') occurs in the two-syllable word 和平 (*he2ping2* 'peace'). Of course, the intention was that *he2ping2* would be familiar to all advanced L2 learners, but if some individual might have otherwise been uncertain

about the tones in *he2ping2*, they might use the character as a cue for guessing. This is unlikely to be a particularly helpful strategy overall (tones must already be known for the characters for this to be useful), but could have lead some learners to indicate higher tone confidence than they actually had.

3.3.4 L1 Mandarin word recognition

Though not the primary focus of the present dissertation, it would be a disservice not to mention L1 results in the present study, as it makes several contributions to our understanding of L1 Mandarin word recognition. First, this study provides additional evidence for ERP responses to pronounceable nonword neighbors of real words. Consistent with previous L1 research (e.g., Friedrich, Eulitz, & Lahiri, 2006; Holcomb & Neville, 1990), including L1 Mandarin research (Y. Liu et al., 2006), L1 participants displayed increased N400s for nonwords relative to real words. As suggested above, this is consistent with the interpretation that the N400 effect indexes difficulty in lexical access, and may capture increasing and/or spreading activation of real word phonological neighbors of the nonword upon failing to access a real word target (cf. Carrasco-Ortiz, Midgley, Grainger, & Holcomb, 2017; Winsler et al., 2018). Second, whereas most previous studies have relied on tone mismatches in monosyllabic stimuli with contextually created expectations (*sentences*: Brown-Schmidt & Canseco-Gonzalez, 2004; X. Li et al., 2008; Schirmer et al., 2005; *prime words*: X. Huang & Yang, 2016; *pictures*: Malins & Joanisse, 2012; J. Zhao et al., 2011), the present research examined disyllabic word recognition in complete isolation. Since the targets were always plausibly words until the arrival of the second syllable, the elicited N400s, while distinctively tied to tone and vowel mismatches,

were ‘purely’ lexical. That is, we were not seeing the online response to the auditory mismatch of a tone or a vowel *per se*, but rather the online response to failure in lexical access as a result of the mismatched cue. In this sense, the present results show rather unequivocally that, all else being equal, tones and vowels are equally important cues in Mandarin word recognition. This is important as some past studies have tended to treat tone and vowel cues differently based on results that included a response to the physical occurrence of the cue itself. In other words, it may be that *noticing* that a cue is deviant may take more or less time, or create larger or smaller neural responses, but this should not be interpreted as evidence of difference in the value of the cues for word recognition *per se*. Present results, then, do not contradict previous studies, but they do help to fill out the picture, indicating how tones and vowels can be equally essential in word recognition when other factors are controlled. Importantly, use of DS words also means that there are different statistical properties guiding lexical access, compared to the MS case where, as noted earlier, homophones and minimal tone neighbors abound. Whether or not there are specific differences in responses for the phonetic cues themselves is an issue that will be taken up in Experiment 4.

Chapter 4: Lexical decision with contextual support

4.1 Overview

The lexical decision experiment reported in Chapter 3 examined L2 responses for words in isolation, with a critical, but indirect test of L2 tone perception. That is, because nonwords were only identifiable *after* the critical mismatching syllable, responses reflected difficulties in lexical access due to tones rather than a direct response to the occurrence of a mismatching tone or vowel. In other words, listeners had no expectations that would drive immediate recognition of the mismatch. The current experiment aims to address L1 and L2 responses to tone and vowel mismatches *as they occur*, by creating strong expectations for specific phonological forms and testing listeners' neural responses to deviations from those expectations.

4.1.1 Background and motivation

While tests of isolated word recognition can be a useful tool for understanding lexical processes, most words do not occur in isolation. Often there are contextual cues that help listeners create expectations about what words they expect to hear. In our previous study (Pelzl et al., 2018), in addition to testing recognition of words in isolation, we also attempted to test L2 learners' ability to use tones and rhymes during lexical recognition in sentential contexts. Unfortunately, this task appears to have been too taxing for most of the L2 participants. While L1 listeners displayed sensitivity in ERP responses to nonwords and mismatching real words, L2 participant showed little neural sensitivity to any of the critical manipulations.

In the current experiment, we will attempt once again to address predictive processes in L2 word recognition through the use of constraining picture cues. While this is of course not a fully ecologically valid approach, it has the potential to address some of the same issues that we had hoped to address previously using constraining sentences. Namely, when contextual cues provide evidence about what words to expect, whether L2 listeners are able to pre-activate those words and then reject nonwords that do not match their expectations.

In Experiment 4, we will once again present listeners with tone and vowel nonwords, but now in the context of a picture-word mismatch task (Desroches, Newman, & Joanisse, 2009), or, more precisely, a picture-*phonology* mismatch task. In this task we will create a strong expectation for a specific word by first presenting participants with an image meant to bring the word to mind. This will allow us to test very early responses that are driven directly by mismatching phonological cues, i.e., the critical tone and vowel cues that distinguish nonwords from real words.

It is important to point out that the lexical demands of this task are quite different than in the LDT. Whereas rejection of a nonword in the LDT required a lexical search on the part of listeners (i.e., to confirm that a nonword does not exist in the lexicon), the Picture-Phonology experiment requires only knowledge of the specific word targeted in a trial. If the listener can successfully bring that word to mind, their task is simply to determine whether the auditory stimulus matches it or not. Thus, while the critical stimuli are still nonwords, the task does not necessarily require the same lexical search processes as in the LDT. This has implications both for the difficulty of the task and the type of ERP responses we expect to observe.

4.1.2 The PMN and LPC responses in ERP research

Just as in the LDT, nonwords in the Picture-Phonology task may evoke N400 effects if listeners have difficulty accessing a word. However, the strongly constraining lexical expectations created by the pictures make two other components equally or even more critical to our analyses.

The first component of interest is the phonological mismatch negativity (PMN) which typically occurs between 200-400ms after stimulus onset and is hypothesized to index neural responses to unexpected/mismatching phonological content in words (Connolly & Phillips, 1994; Desroches et al., 2009; Newman & Connolly, 2009; see also discussion of the "N200" in e.g., Brunellière & Soto-Faraco, 2015; Van Den Brink, Brown, & Hagoort, 2001). The PMN has been consistently observed in previous ERP research of Mandarin spoken words (Malins & Joanisse, 2012; J. Zhao et al., 2011), although it has not always been overtly analyzed or labeled as such (Liu et al., 2006; Pelzl et al., 2018). Of particular relevance is the study by Malins and Joanisse (2012), which used a picture-word paradigm with MS Mandarin words. In their study, all auditory stimuli were real words, and they manipulated the relation to pictures so that either consonants, vowels, tones, or complete syllables matched/mismatched the evoked word. They found significant PMN and N400 effects for all mismatch types.[9]

In the present case, because our nonwords differ from real words only with respect to a tone or a vowel in the first syllable, we expect that PMN responses will

[9] I have chosen not to use their terminology for these conditions as it is somewhat atypical and likely to cause confusion if introduced here.

be evoked as soon as the departure from the target word becomes apparent. Importantly, the PMN should precede any N400 effects, and we might expect that N400 effects will be reduced or non-existent for *nonword* mismatches (cf. Newman & Connolly, 2009).

Along with PMN responses, we also expect to see strong late positive components (LPCs). In sentence processing experiments, late positivities are often classified as P600s and are hypothesized to reflect reanalysis or repair processes when people are confronted by infelicitous syntax (Gouvea, Phillips, Kazanina, & Poeppel, 2010; Kaan & Swaab, 2003; Osterhout & Holcomb, 1992), though similar effects have been observed for lexical violation (e.g., Romero-Rivas, Martin, & Costa, 2015; Schirmer et al., 2005) and phonological mismatches (e.g., Schmidt-Kassow & Kotz, 2009). Importantly, we observed LPCs in our previous sentence processing ERP study when L1 listeners detected tone and rhyme mismatches in nonwords (Pelzl et al., 2018). Although, not a sentence processing study, similar effects—though not analyzed—are also apparent in the later portion of waveforms for vowel and tone mismatches in Malins & Joanisse (2012, Figure 1, p. 2037). Thus, we expect to find LPCs in response to picture-phonology mismatches in the present case. These effects are often described as indexing error detection, repair, reanalysis, or reorientation processes and may be related to more general (i.e. non-linguistic) processing mechanisms (Coulson, King, & Kutas, 1998; Sassenhagen & Bornkessel-Schlesewsky, 2015; Sassenhagen, Schlesewsky, & Bornkessel-Schlesewsky, 2014).

By examining PMN and LPC responses to tone and vowel nonwords, we hope to determine whether L2 listeners are sensitive to tone and vowel mismatches as they

occur (as indexed by the PMN) and whether or not they show the expected late response (LPC).

In summary, the Picture-Phonology experiment aims to create a scenario where L2 listeners are given strong odds of success in recognition of tone mismatches in a lexical context, and, by recording ERPs aims to examine L2 neural responses to the tone and vowel cues as they occur.

4.2 Experiment 4: Picture-Word Mismatch

4.2.1 Experiment 4: research questions and hypotheses

Experiment 4 will utilize a Picture-Phonology mismatch task while recording ERPs in order to address the following research questions regarding advanced L2 learners' behavioral and neural responses to tones and vowels in nonwords.

(1) *Are L2 listeners equally accurate in rejection of nonwords with mismatching tone and vowels cues when strong word expectations are created?*

(2) *Are L2 listeners equally sensitive to vowel and tone mismatches in the phonology of expected disyllabic words (as indexed by the PMN)?*

(3) *Are L2 listeners equally likely to show late positive responses (LPCs) for vowel and tone mismatches in the phonological form of expected disyllabic words?*

Question (1) investigates whether advanced L2 listeners show different levels of behavioral accuracy depending on the nature of phonological mismatches with real words. Based on our previous results (Pelzl et al., 2018), we expected that L2 listeners would be less accurate in rejection of tone nonwords compared to vowel nonwords, demonstrating less ability to use tone cues in online word recognition.

However, as with Experiment 3, the current design is intentionally less difficult than Pelzl et al. (2018), and furthermore also meant to be less difficult than Experiment 3 reviewed above, due to the provision of supporting picture cues that allow successful rejection of nonwords without an exhaustive lexical search. Thus it is possible that we could find highly accurate L2 performance for tone mismatches.

Question (2) asks whether neural responses demonstrate PMN effects to both vowel and tone mismatches. The most straightforward outcome would be that, if behavioral results show less accuracy for tone than vowel mismatches, PMN responses would similarly reflect less robust responses for tones compared to vowels. However, as in Experiment 3, it is possible that we could see different patterns. ERP responses could capture implicit sensitivity to tones despite poor behavioral performance, or might show no indication of PMN sensitivity despite accurate rejection of mismatch trials.

For question (3), although the precise function indexed by LPCs remains unclear, they are expected to align quite tightly with behavioral responses, essentially indexing the attentional processes that lead to decisive rejections. For this reason, it is expected that examination of correct trials should reveal clear LPCs for both L1 and L2 participants.

As in Experiment 3, direct comparison of L1 and L2 groups is not necessarily a major goal of the study. Examination of L1 results will provide support for general interpretations of ERP patterns, and will serve to confirm that stimuli are functioning as intended.

Once again, as in Experiment 3, three additional questions (4-6) consider the impact of L2 learners' word and tone specific knowledge and subjective confidence as a mediator for accuracy and ERP results.

(4) *Does lexical familiarity impact L2 behavioral responses?*

(5) *Do specific tone confusions impact L2 behavioral responses?*

(6) *Does lexical familiarity impact ERP responses?*

Finally, immediately following the critical Picture-Phonology experiment, a secondary Picture-Word experiment will also be conducted. The motivation for including the Picture-Word experiment is to test that our L1 and L2 listeners display standard N400 effects under the picture-word paradigm. Ideally, the word mismatch condition would have been a fourth condition in the Picture-Phonology experiment, but, due to the difficulty of selecting sufficient numbers of appropriate words that L2 listeners would know, we instead made the practical decision to administer it as a separate experiment. Relevant details and results of the Picture-Word will be presented along with the Picture-Phonology experiment below, with additional details available in Appendix A4.

4.2.2 Experiment 4: Participants

Participants were the same as in Experiments 2 and 3 reported above.

4.2.3 Experiment 4: Task and stimulus design

Experiment 4 utilized two types of picture-word trials, delivered as two separate sets of experimental blocks. In the critical Picture-Phonology (Pic-Phono) trials, participants see a picture either followed by a word that matches the picture or

followed by a nonword the mismatches the pronunciation of the word evoked by the picture. In Picture-Word (Pic-Word) trials, participants see a picture, and after the picture disappears, hear a real Chinese word that either matches or mismatches the word evoked by the picture.

Stimuli were based on a set of 96 disyllabic real words. All were highly frequent imageable nouns, chosen so that a corresponding picture could be matched to each one (e.g., *mian4tiao2* 'noodles').

In order to make pictures as easily identifiable as possible, photographic images were used.[10] The majority of images were taken from two freely available picture databases (BOSS: Brodeur, Dionne-Dostie, Montreuil, & Lepage, 2010; Ecological SVLO: Moreno-Martínez & Montoro, 2012), with some images culled from other free photo repositories (e.g., Wikimedia commons[11]). A small number of difficult to find images were purchased from Adobe Stock, and two more images were created specifically for the experiment. An example image is shown in Figure 4.1. All images were placed on a white background. No attempt was made to control colors or luminosity as the neural response to the presentation of the images was not of primary interest. Instead we aimed to make images as recognizable as possible.

To assure that images would evoke the intended vocabulary, two rounds of picture norming were conducted. In each round, ten native Mandarin speakers generated Chinese words for 132 images. Images that were judged to perform

[10] For the words *tian1shi3* 'angel' and *mo2gui3* 'devil', computer generated 3-D cartoon images were used, as no angels or demons were available for photos.

[11] https://commons.wikimedia.org/wiki/Category:Images

Figure 4.1. Example image: *mian4tiao2* 'noodles' (Experiment 4)

inadequately in the first round (less than 70% generation of the target word, or generation of problematic competitor words) were replaced and a second round of norming was conducted with a new group of ten people. The end result was a set of 96 critical images that had an average word generation rate of 86%, though a handful of items (7 total) had rather low naming rates (under 50%). Future work might try to replace either those words or images. Images for filler items were also overall highly identifiable.

Both Pic-Phono and Pic-Word trials drew on the same set of 96 critical picture-word pairs. For the Pic-Phono trials the real words were further manipulated to create two types of nonwords. As in the LDT in Experiment 3, the first syllable of the nonwords mismatched the real word counterpart with respect to either a tone or a vowel. For example, the real word *mian4tiao2* /mian4t^{h}iau2/ became the vowel nonword ***m*ə*n4tiao2*** /mən4t^{h}iau2/ and the tone nonword ***mian3****tiao2* /mian3t^{h}iau2/. As in the LDT, all tone combinations and manipulations were balanced across words and nonwords. There were, however, some differences in the stimuli creation procedures compared to the LDT. Whereas LDT nonwords were constructed so that the first syllables always had plausible second syllable word continuations, because the

images in the Pic-Phono create strong phonological expectations that will either be met or confounded as soon as the rhyme of a first syllable is heard, this restriction was not necessary for nonwords in the Pic-Phono. This allowed for somewhat tighter restriction of vowel mismatches in the Pic-Phono, in most cases being a single phoneme change (i.e., either a single phoneme switch, addition, or subtraction). As much as possible, repetition of first syllables was avoided across stimuli, but this restriction proved impossible to follow with the same rigor as for the LDT stimuli due to the even more limited number of words able to accommodate imageability.

These procedures resulted in a total of 96 critical real word/vowel nonword/tone nonword triplets. An additional 16 real words with accompanying images were selected as fillers. Due to the limitations on words likely to be known by L2 learners, it was not possible to limit selection of fillers to words with a balanced occurrence of tones, and many filler items had neutral tones on the second syllable.

For the Pic-Word trials, only real words were utilized. Half of them were paired with matching images, and half with mismatching images (additional details available in Appendix 4). Thus, in Pic-Word trials, photos were followed by either a matching real word (e.g., a photo of an onion followed by auditory presentation of "*yang2cong1*" [onion]) or a mismatching real word (e.g., a photo of an onion followed by the word "*jian3dao1"* [scissors]).

For both the Pic-Phono and Pic-Word, three lists were constructed to balance images and words across participants. For each list, four unique pseudo-random presentation orders were prepared, with conditions balanced so that no more than three trials in a row would require the same response type (yes or no). Since Pic-

Phono items were repeated in the Pic-Word trials, steps were taken to minimize clues as to which images would be followed by matching or mismatching words (for the brave reader, the tortuous details are available in Appendix A4.1).

4.2.4 Experiment 4: Procedures

Location and equipment were the same as in Experiment 3. Participants first completed the Pic-Phono blocks, then completed the Pic-Word blocks. For the Pic-Phono participants began by completing eight practice items with stimuli not included in the experiment, and then completed 112 Pic-Phono trials. Stimuli were presented in seven blocks of 16 trials, with self-paced breaks between each block. The beginning of each trial was signaled with a 'beep', followed by a fixation cross. After 350 ms, a picture was displayed. Then, after 1.75 seconds the image was replaced by a fixation cross. Still 250 ms later the auditory stimulus was presented, followed by 1.2 sec of silence at which point the fixation cross was replaced by a question prompt: "Did the word match the picture?" After the participant's response, there was a 2 sec pause before the next trial began. The entire Pic-Phono experiment lasted approximately 15 minutes.

The long display time for the images (1.75 sec) was determined after piloting and with the logic that, for this experiment we wanted to maximize the opportunity for L2 learners to recognize images and their associated words. This design allows (but does not compel) participants to utilize explicit knowledge of tones in retrieving target items. The idea was that this design serves as a proof-of-concept for this approach, testing L2 ability to utilize tone cues under near optimal circumstances.

The Pic-Phono blocks were immediately followed by the Pic-Word blocks. After four practice items, participants completed 64 Pic-Word trials in four blocks of 16. Trial structure was the same as for the Pic-Phono blocks. The Pic-Word blocks took about 8 minutes to complete.

4.2.4 Experiment 4: EEG data processing

The same EEG processing procedures were followed as for Experiment 3. Due to equipment failure, data from one L1 participant was excluded.

For the Pic-Phono, data from two additional L1 participants and one L2 participant were excluded due to having greater than 40% artifacts on experimental trials.[12] After excluding these participants, artifact rejection affected 10.55% of experimental trials (L1 8.31%; L2 13.18%). A single average amplitude was obtained for each trial for each electrode for each subject in an early PMN window (200-400 ms) and a later LPC window (400-600 ms). These windows were chosen largely by visual inspection of grand average waveforms. As will be addressed later in more detail, choosing appropriate windows for this data proved a challenge due to significant component overlap.

For the Pic-Word trials, data from two L1 participants and two L2 participants were excluded due to having greater than 40% artifacts on experimental trials. After exclusion of these participants, artifact rejection affected 8.59% of experimental trials (L1 8.95%; L2 8.11%). A single average amplitude was obtained for each trial for

[12] A second L2 participant's data was borderline at 41.67% trials rejected, but was retained due to the difficulty of obtaining advanced L2 data.

each electrode for each subject in an auditory N400 window (200-700ms). This window was chosen on the basis of two criteria. First, unlike the earlier LDT, in this experiment, immediately upon hearing the auditory stimuli listeners could potentially notice a mismatch. This motivated an earlier start for the critical N400 time window. Second, visual inspection of grand average waveforms across all scalp electrodes suggests 200-700 ms would be a reasonable window to capture relevant N400 effects; this duration was also consistent with that of the N400 window in the LDT.

Other processing details were the same as for Experiment 3. After exclusion of incorrect trials, the final Pic-Phono PMN dataset contained 42,613 data points (80.0% out of total possible 53,290 data points: L1=88.1%; L2=70.4%), and the LPC dataset contained 42,610 data points (80.0% out of total possible 53,290 data points. L1= 88.1%; L2=70.4%). The final Pic-Word dataset contained 32,049 data points for the N400 (90.2% out of total possible 35,520 data points: L1=89.9%; L2=90.7%).

4.2.5 Experiment 4: Behavioral results and analysis

In the results and discussion sections below, I will focus on details of interest to critical research questions. For the Pic-Word results, this means some details will be left out. Full results and analysis for the Pic-Word are available in Appendix 4.

Reliability for Pic-Phono trials was high (Pic-Phono: List a: α=.90; List B: α=.92; List C: α=.94).[13] Descriptive results can be seen in Table 4.1 and are depicted

[13] Reliability for Pic-Word lists was not consistently computable due to complete ceiling performance by many participants.

Table 4.1. Descriptive accuracy results for Picture-Phonology Mismatch (Experiment 4)

Group	cond	mean acc. % (sd)
L1 (*n*=24)	real	96 (19)
	vowel	92 (26)
	tone	97 (17)
L2 (*n*=18)	real	87 (33)
	vowel	88 (32)
	tone	66 (47)

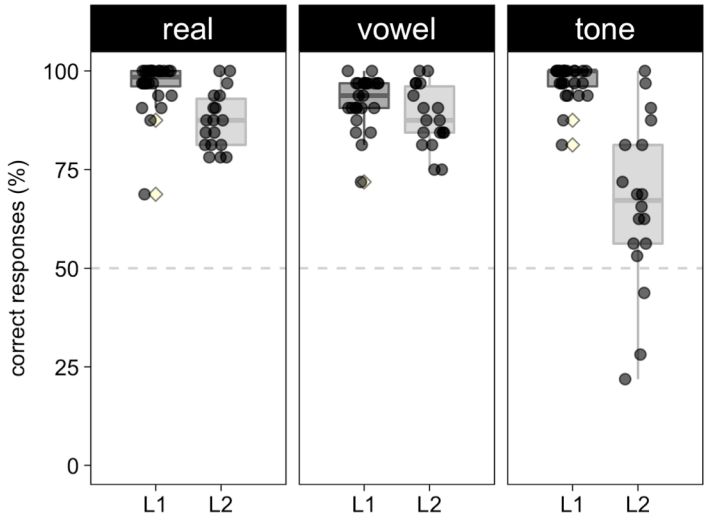


Figure 4.2. Boxplot of accuracy results for Picture-Phonology Mismatch (Experiment 4). Each circle indicates an individual participant's mean score. Diamonds indicate that scores at that level are outliers. The dashed line indicates the level which would be equivalent to chance performance.

visually in boxplots in Figure 4.2. For the Pic-Word, accuracy was near ceiling in all conditions (>97%).

The Pic-Phono model results are summarized in Table 4.2. There were significant main effects for condition (χ^2= 6.13, p=.047) and group (χ^2= 30.23, p<.001), and a significant two-way interaction between condition and group (χ^2= 27.17, p<.001). In the Pic-Word, there were no significant differences in accuracy between groups, and no significant condition-by-group interactions.

Table 4.2. Mixed Model ANOVA Table for accuracy in Picture-Phonology Mismatch (Type 3 tests, LRT-method) (Experiment 4B)

Effect	Df	Chisq.	Chi Df	Pr(>Chisq)	
condition	13	6.13	2	.047	*
group	14	30.23	1	<.001	***
condition × group	13	27.17	2	<.001	***
Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1					

Table 4.3. Planned comparisons for accuracy of Picture-Phonology Mismatch (Experiment 4)

Comparison	Estimate	SE	z value	Pr(>|z|)	
L1 Vowel vs. Tone	-0.79	0.44	-1.79	.073	.
L2 Vowel vs. Tone	2.05	0.40	5.19	<.001	***
L1 V-T vs. L2 V-T	-2.85	0.54	-5.31	<.001	***
Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1					

Planned comparisons for the Pic-Phono are summarized in Table 4.3. There was a marginally significant difference in accuracy between nonword conditions for the L1 group (b=2.05, SE=.40, z=5.19, p<.001), with L1 slightly more accurate for tone than vowel nonwords. In contrast, L2 listeners were significantly more accurate in rejection of vowel nonwords than of tone nonwords (b=2.05, SE=.40, z=5.19, p<.001). They were about three times more likely to incorrectly accept tone nonwords than vowel nonwords (.34/.12=2.83). Compared to L1, the accuracy difference between nonword conditions for L2 was significantly larger (b=-2.85, SE=.54, z=-5.32, p<.001).

In summary, for the Pic-Phono blocks, L2 performed significantly less accurately than L1 and had significant difficulty correctly rejecting tone nonwords. In contrast, accuracy results for the Pic-Word suggest all participants performed equally well in the picture-word blocks regardless of native language.

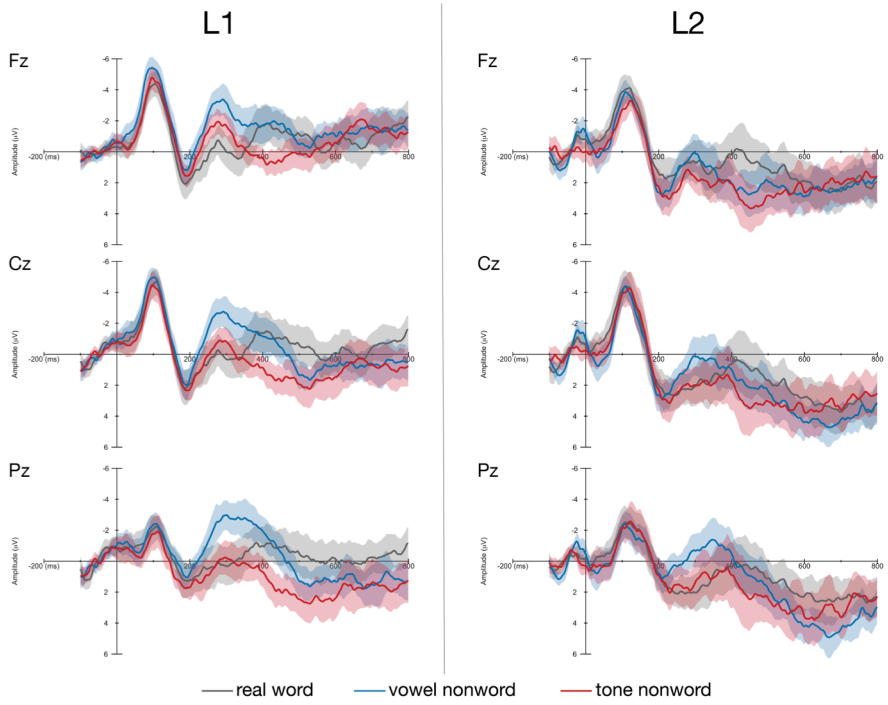


Figure 4.3. Grand average waveforms for L1 and L2 participants in Picture-Phonology Mismatch (Experiment 4). Only correct trials included (40Hz low pass filter). The shaded area around each line represents a 95% within-subjects confidence interval.

4.2.6 Experiment 4: ERP results and analyses

Grand average waveforms for ERP results are displayed visually in Figure 4.3 for the Pic-Phono, and Figure 4.4 for the Pic-Word. For the Pic-Phono there appear to be strong negativities for vowel nonwords in the early PMN window (200-400 ms), though these effects are less distinctive in L2 responses. In the later LPC window (400-600 ms), responses to tone mismatches appear more positive in amplitude than in the other conditions, with the difference once again less distinctive in L2 responses. For the Pic-Word, strong negative deflections in the N400 (200-700 ms) window are apparent for both groups, though these affects appear stronger and earlier in L1 than in L2 responses.

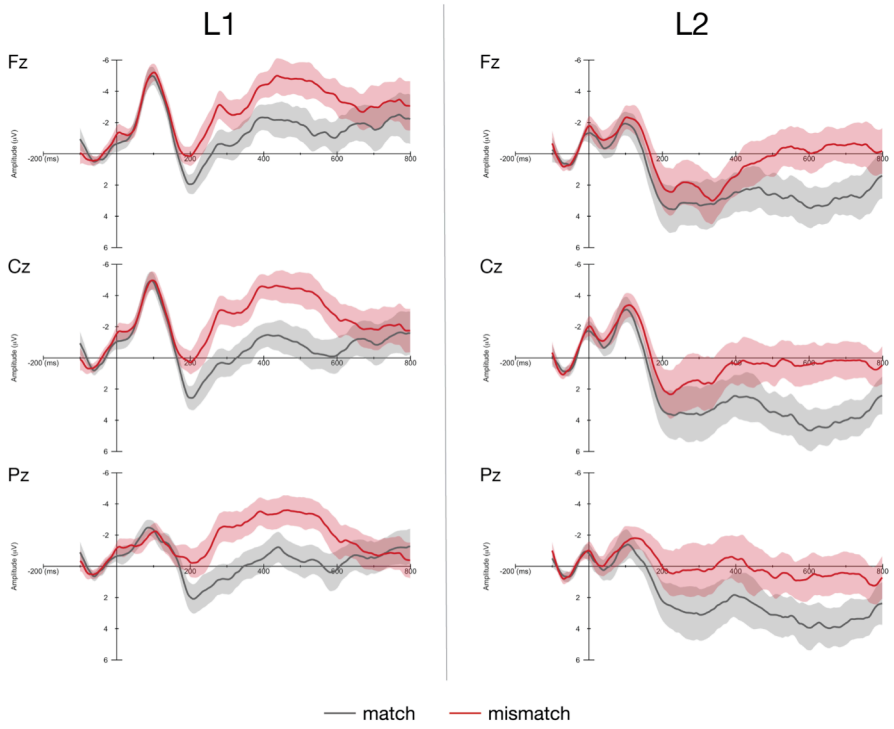


Figure 4.4. Grand average waveforms for L1 and L2 participants in Picture-Word Mismatch. The shaded area around each line represents a 95% within-subjects confidence interval.

Average amplitudes for correct trials in the two windows of the Pic-Phono from 200-400 ms (PMN) and 400-600 ms (LPC), as well as from the 200-700 ms (N400) window in the Pic-Word, were submitted to linear mixed-effects models with fixed effects for *condition* (match, mismatch) and *group* (L1, L2) and their interaction. Model fitting procedures were the same as reported for ERP data in Experiment 3. Convergence difficulties were addressed by specifying uncorrelated random effects. The final maximal models for all data sets were parallel, and included random slopes for subjects and items, with electrodes nested under subjects. The models also included by-subject random intercepts for condition, and by-item random intercepts for condition and group and their interaction. Model results will be reported one at a time.

Table 4.4. Mixed Model ANOVA Table for PMN (200-400ms) amplitude in the Picture-Phonology Mismatch (Type 3 tests, Satterthwaite method) (Experiment 4)

Effect	Df	den Df	F	Pr(>F)
condition	2	95.84	8.07	<.001
group	1	41.60	2.18	.147
condition × group	2	102.09	0.96	.386
Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1				
model formula: amplitude ~ condition * group + (condition || subject / electrode) + (condition * group || item)				

Table 4.5. Planned comparisons for PMN (200-400ms) amplitude in the Picture-Phonology Mismatch (Experiment 4)

Comparison	b	SE	z	p	
L1 real vs. vowel	2.15	0.61	3.53	.001	**
L1 real vs. tone	-0.07	0.96	-0.07	0.941	
L1 vowel vs. tone	-2.22	0.92	-2.42	0.031	*
L2 real vs. vowel	1.04	0.63	1.65	0.299	
L2 real vs. tone	0.035	1.013	0.034	0.972	
L2 vowel vs. tone	-1.01	0.96	-1.05	0.589	
Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1					

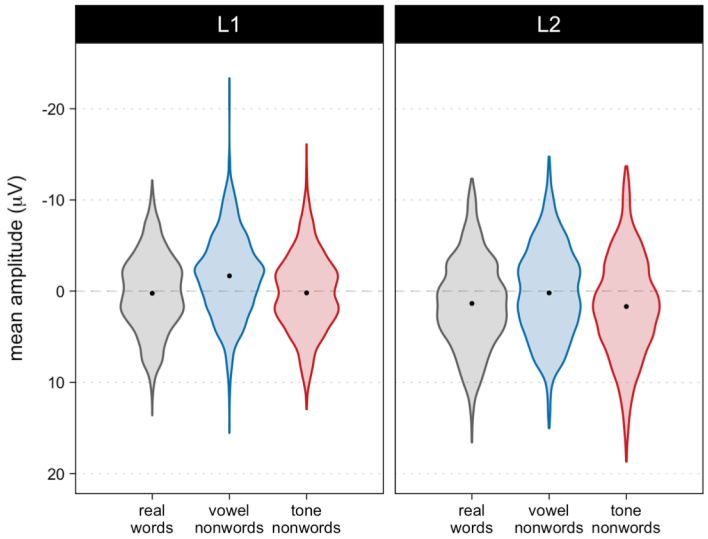


Figure 4.5 Model estimates for PMN (200-400ms) amplitude in the Picture-Phonology Mismatch (Experiment 4B). The black dots represent the model estimated group means for each condition, with shaded areas representing the distribution of estimated responses.

4.2.6.1 Experiment 4: ERP results and analyses for Picture-Phonology PMN

Results for the Pic-Phono are reported in Table 4.4. There was a significant main effect for condition ($F_{2, 95.84}$=8.07, p<.001), but no significant main effect of group, and no significant group-by-condition interaction.

As our research questions are specifically interested in L2 results, we conducted planned comparisons of differences between conditions within each group using the Holm method to correct for multiple comparisons. These are reported in Table 4.5. L1 vowel nonword responses were significantly more negative than real word responses (b=2.15, SE=.61, z=3.53, p=.001) and tone nonword responses (b=-2.22, SE=.92, z=-2.42, p=.031). For L2, there were no statistically significant differences between conditions. Model estimates are depicted visually in violin plots in Figure 4.8.

4.2.6.2 Experiment 4: ERP results and analyses for Picture-Phonology LPC

The Pic-Phono LPC results are reported in Table 4.6. There were no significant main effects or interactions. Once again, in the interest of addressing our research questions, planned comparisons were conducted using the Holm method to

Table 4.6. Mixed Model ANOVA Table for LPC (400-600ms) amplitude in the Picture-Phonology experiment (Type 3 tests, Satterthwaite method) (Experiment 4)

Effect	numer Df	denom Df	F	Pr(>F)
condition	2	95.16	1/89	.157
group	1	40.72	1.54	.221
condition × group	2	93.54	0.95	.390
Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1				
model formula: amplitude ~ condition * group + (condition || subject / electrode) + (condition * group || item)				

Table 4.7. Planned comparisons for LPC (400-600ms) amplitude in the Picture-Phonology experiment (Experiment 4)

Comparison	b	SE	z	p
L1 real vs. vowel	0.10	0.68	0.14	.887
L1 real vs. tone	-1.81	1.07	-1.69	.204
L1 vowel vs. tone	-1.90	1.04	-1,82	.204
L2 real vs. vowel	-0.97	0.71	-1.37	.512
L2 real vs. tone	-1.20	1.13	-1.07	.570
L2 vowel vs. tone	-0.23	1.09	-0.21	.834
Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1				

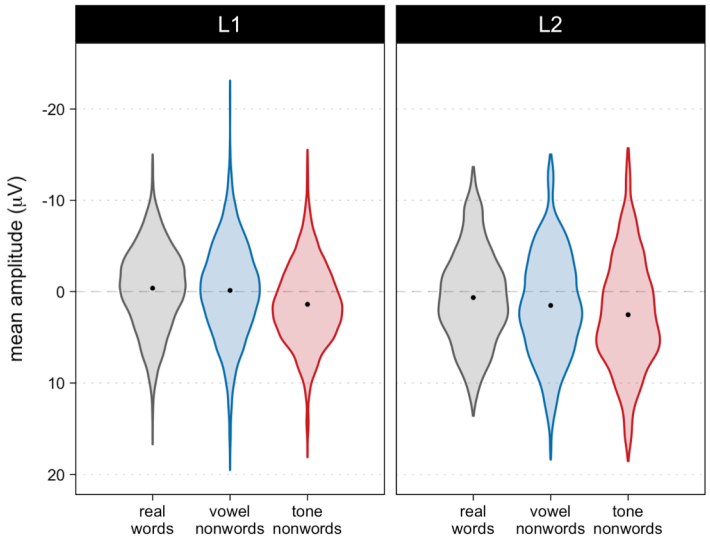


Figure 4.6. Model estimates for LPC (400-600ms) amplitude in the Picture-Phonology Mismatch (Experiment 4). The black dots represent the model estimated group means for each condition, with shaded areas representing the distribution of estimated responses.

correct for multiple comparisons. Results are reported in Table 4.9. No significant differences were found. Model estimates are depicted visually in violin plots in Figure 4.7.

4.2.6.3 Experiment 4: ERP results and analyses for Picture-Word N400

Model results for the Pic-Word N400 are reported in Table 4.8. There were significant main effects of condition ($F_{1,90.62}$=13.31, p<.001) and group ($F_{1,41.82}$=5.71, p=.022). The interaction of condition-by-group was not statistically significant.

Table 4.8. Mixed Model ANOVA Table for N400 (200-700ms) amplitude in the Picture-Word experiment (Type 3 tests, Satterthwaite method) (Experiment 4)

Effect	numer Df	denom Df	F	Pr(>F)	
condition	1	09.62	13.31	<.001	***
group	1	41.82	5.71	.022	*
condition × group	1	82.66	0.31	.581	
Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1					
model formula: amplitude ~ condition * group + (condition || subject / electrode) + (condition * group || item)					

Table 4.9. Planned comparisons for N400 (200-700ms) amplitude in the Picture-Word Mismatch (Experiment 4)

Comparison	b	SE	z	p	
L1 vs. L2	-2.88	1.20	-2.39	.017	*
match vs. mismatch	1.97	0.54	3.65	<.001	***
L1 match vs. mismatch	2.26	0.72	3.13	.002	**
L2 match vs. mismatch	1.68	0.78	2.16	.031	*
Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1					

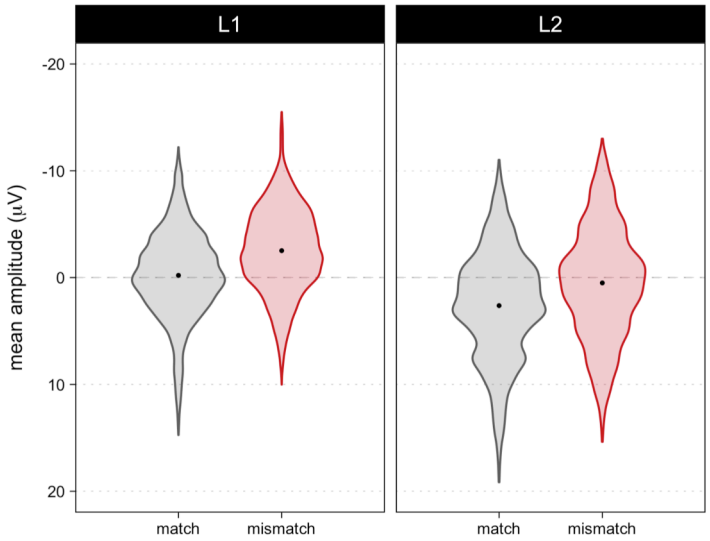


Figure 4.7. Model estimates for N400 (200-700ms) amplitude in the Picture-Word Mismatch (Experiment 4). The black dots represent the model estimated group means for each condition, with shaded areas representing the distribution of estimated responses.

Post-hoc and planned comparisons with Holm corrections are reported in Table 4.12. Match trials were significantly more positive in amplitude than mismatch trials (b=1.97, SE=.54, z=3.65, p<.001). There was also an overall amplitude

difference between L1 and L2 groups, with L2 having more positive amplitude overall than L1 regardless of condition (*b*=-2.88, *SE*=1.20, *z*=*-2.39*, *p*=.017). Although there was no interaction between group and condition, we wanted to confirm that both groups displayed significant N400 effects for mismatching trials. This was indeed the case. There was a significant difference between match and mismatch for both L1 and L2 groups (L1: *b*=2.26, *SE*=.72, *z*=3.13, *p*=.002.; L2: *b*=1.68, *SE*=.78, *z*=2.16, *p*=.031).

4.2.7 Experiment 4: Offline vocabulary test data processing

The steps for processing offline vocabulary test data were the same as for similar data in Experiment 3.

4.2.8 Experiment 4: Offline vocabulary test results

Descriptive results for the offline vocabulary test along with related accuracy for those items in the Pic-Phono task are displayed for tones in Table 4.10, and for vocabulary definitions in Table 4.11. As for Experiment 3, we find that even for high confidence words, explicit tone knowledge is often somewhat inaccurate (overall *mean*=85%), and that overall tone confidence seems to bear a relationship to the

Table 4.10. Results of L2 offline vocabulary test requiring participants to supply tones and tone confidence ratings for nonwords. Tone accuracy indicates whether supplied tones were correct. P-Ph accuracy indicates whether the related nonwords were correctly rejected in the Picture-Phonology matching task.

Confidence ratings and accuracy of L2 supplied tones				
Condition	conf. rating	k (items)	tone acc. %	P-Ph acc. %
Vowel	3 (high)	313	87	92
	2 (mid)	183	51	85
	1 (low)	35	31	86
Tone	3 (high)	318	84	76
	2 (mid)	181	50	53
	1 (low)	40	35	48

Table 4.11. Results of L2 offline vocabulary test requiring participants to supply definitions and definition confidence ratings for nonwords. Tone accuracy indicates whether supplied tones were correct. P-Ph accuracy indicates whether the related nonwords were correctly rejected in the Picture-Phonology matching task.

Confidence ratings and accuracy of L2 supplied definitions				
Condition	conf. rating	k (items)	def. acc. %	P-Ph acc. %
Vowel	3 (high)	496	98	90
	2 (mid)	28	82	79
	1 (low)	7	43	57
Tone	3 (high)	500	99	67
	2 (mid)	25	84	60
	1 (low)	5	40	40

accuracy of the supplied tones. There does seem to be a notable drop in tone confidence for Pic-Phono words compared to LDT words. Whereas there were 762 high confidence items in the LDT, for the Pic-Phono, there are 631, with more items in both the mid and low confidence categories than for the LDT. This suggests less familiarity with the vocabulary in the Pic-Phono overall. Accurate performance in the Pic-Phono for tone nonwords once again appears to be related to offline tone knowledge, whereas for vowel nonwords tone knowledge is not relevant.

For definitions it can be seen that L2 participants' subjective confidence seems to reflect their knowledge quite accurately. High confidence items were correctly defined more than 98% of the time. It appears that high confidence words were also more accurately rejected in Pic-Phono vowel nonword trials, but that the relationship does not directly affect outcomes in the tone nonword condition. This is slightly different than what appeared in the LDT where it seemed that knowledge of definitions did not strongly impact accuracy in rejection of vowel nonwords. If in fact there is a real difference, it seems likely to be related the use of pictures in the Pic-Phono task, which makes semantic aspects of word recognition more relevant than in

the LDT, where recognition of phonological form alone was enough to complete trials.

In sum, as in Experiment 3, we find that L2 offline knowledge suggests some difficulties in accurate encoding of tones in explicit lexical representations, and that this appears to impact accuracy for correct rejection of tone nonwords.

4.2.9 Experiment 4: Exploratory "Best Case Scenario" analysis

4.2.9.1 Overall Accuracy

As in Experiment 3, an exploratory 'Best Case Scenario' analysis was conducted on the accuracy results for the Pic-Phono. In this analysis, we again retain only the subset of trials that targeted nonwords for which an L2 participant had indicated correct and confident knowledge (confidence rating = 3) of both tones and definitions for the real word counterparts. This comprised 263 tone nonword and 265 vowel nonword trials (527 total, 46% of total nonword trial data).

Table 4.12 presents descriptive accuracy results for the two nonword conditions in the 'Best Case Scenario' data for the Pic-Phono. The accuracy results were submitted to a generalized linear mixed effects model following procedures 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.

Table 4.12. Descriptive accuracy results for the 'Best Case Scenario' analysis of the Pic-Phono

group	cond	mean acc. % (sd)
L2 (n=17)	vowel	92 (26)
	tone	79 (41)

Table 4.13. Comparison of conditions for accuracy results in the 'Best Case Scenario' analysis of the Pic-Phono (Type 3 tests, LRT-method) (Experiment 4)

Effect	b	SE	z	p	
Tone vs vowel nonword	-7.99	2.99	-2.67	.008	**
Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1					
model formula: accuracy ~ condition + (condition | subject) + (condition | item)					

Results are displayed in Table 4.13. There was a significant difference in accuracy for tone and vowel nonwords (b=-7.99, SE = 2.99, z = -2.67, p=.008).

In summary, as in Experiment 3, after accounting for offline L2 word knowledge and subjective confidence of that knowledge, L2 still shows a more limited ability to reject tone nonwords than vowel nonwords. In contrast to the LDT however, it does appear that excluding unknown and unconfident trials considerably improved overall performance in both nonword conditions. Accuracy for tone nonwords rose from 66% to 79%, and for vowel nonwords rose from 88% to 92%.

4.2.9.2 Accuracy by tone manipulation

As for the LDT in Experiment 3, we once again consider whether potential tone confusions influenced accuracy. Table 4.18 shows the accuracy of Pic-Phono decisions for tone nonwords according to the tone switch that created the nonword. Once again, we find that tone nonwords that switched T2 and T3 seem related to lower overall accuracy (T2 to T3: *mean*=46%; T3 to T2: *mean*=43%). The 'easy' switches also again appear to trend towards higher accuracy, though not in all cases (T1 to T3: *mean*=71%; T3 to T1: *mean*=94%; T2 to T4: 96%; T4 to T2: 77%). Compared to the LDT, there seems to be more extreme variation, likely due to having somewhat less Best Case Scenario data to work with in the Pic-Phono. Overall, the

Table 4.14 L2 Picture-Phonology accuracy by tone switch in the Best Case Scenario analysis (Experiment 4)

switch type	total data points	accuracy %
T1 to T2	14	100
T1 to T3	21	71
T1 to T4	27	85
T2 to T1	23	87
T2 to T3	13	46
T2 to T4	24	96
T3 to T1	17	94
T3 to T2	23	43
T3 to T4	18	100
T4 to T1	32	78
T4 to T2	22	77
T4 to T3	28	75
	total: 262	*79*

results depict some differential influences for specific tone contrasts, but, as in the LDT, nearly all tones still seem to induce some difficulty.

4.2.9.3 Quality of L2 knowledge for correct trials in ERP data

As in Experiment 3, limited power prevented us from conducting a Best Case Scenario analysis using ERP amplitude data. However, we can once again consider the quality of offline knowledge associated with correct rejections for mismatch trials. For these trials, L2 knowledge of definitions for real word counterparts of nonwords was very accurate (vowel nonwords: *mean*=96%; tone nonwords: *mean*=95%). L2 knowledge of tones was not nearly so high (tone nonwords 77%), and varied rather extremely across participants, with the lowest mean average being 29%, and the highest 96%. The extreme low score was again somewhat atypical of the group. Only two participants scored below 50% (the same two as in the LDT).

In summary, consistent with what was found in the LDT, these results suggest that even for correctly rejected tone nonword trials, for roughly 20% of the trials L2

participants did not necessarily have accurate explicit knowledge of the appropriate tones for target words.

4.3 Experiment 4: Discussion

Results for Experiment 4 can be summarized as follows. Compared to their accuracy for rejection of vowel nonwords, L2 participants were significantly less accurate in rejecting nonwords with mismatching tones. This difference persisted even when we considered only trials where L2 had correct and confident knowledge of tones and definitions for the words associated with picture prompts, though there was an apparent improvement for accuracy overall in such cases. While certain tone confusions (T2 vs. T3 and vice-versa) seemed to impact accuracy more strongly than others, no specific tone or tones can account for the overall degree of L2 tone inaccuracy.

Examination of ERP results suggests that L1 listeners show distinctive early PMN responses to vowel cues that mismatch expectations, but responses to mismatching tone cues were less apparent. There were no other significant effects found in the ERP analysis for the Pic-Phono experiment in either the PMN or LPC time windows, and no evidence that L1 and L2 groups differed significantly in ERP responses for this experiment. For the Pic-Word experiment, while overall amplitude differed between L1 and L2 groups, there were nevertheless clear N400 effects for both groups in responses to mismatching words.

4.3.1 L2 tone accuracy results

As in Pelzl et al. (2018) and the LDT in Experiment 4, results for the Pic-Phono experiment suggest that advanced L2 Mandarin learners have significant difficulties utilizing tone cues to reject nonwords, even when they have correct and confident explicit knowledge for the relevant real words. This was true even though the lexical demands of the Pic-Phono were simpler than those of the LDT. Whereas an LDT requires that listeners have enough confidence in their vocabulary to reject items as nonwords, the Pic-Phono only required that they match a single word to an auditory signal, and consequently the decision to reject a nonword should be easier to make. Considering this, the fact that L2 learners still made errors on roughly 1 of every 5 trials in the Best Case Scenario—when they confidently knew the words and their tones—is rather striking.

The relation of these results to LDT results can be thought of in two ways. First, it may be that Pic-Phono results are straightforwardly consistent with LDT results, indicating the same underlying source or sources of difficulty for tone cues in both cases. Alternatively, it may be that *because* the Pic-Phono trials created lexical expectations, listeners were *more likely* to misperceive tone cues. That is, the pre-activation of the lexical target overrode auditory perception of the acoustic signal so that listeners regularly accepted mismatching tones. This would be something like an L2 version of the familiar Ganong effect (Fox & Unkefer, 1985; Ganong, 1980), where the existence of a real word tone neighbor serves as a kind of magnet for perception of tone nonwords. This would be evidence for the weaker quality of tone categories encoded in some lexical entries. While not inconceivable, this effect would

have to be very strong, as the critical role of tones in the mismatching trials should have been very clear to all participants in the Pic-Phono experiment.[14]

In either case, two conclusions seem merited. First, tone word recognition presents considerable difficulty to advanced L2 learners. Second, it is likely that the difficulty derives from the convergence of multiple sources (i.e., general tone perception weakness, tone knowledge for specific vocabulary, specific tone confusions, lack of confidence).

4.3.2 ERP results

The statistical results from our PMN analysis suggest strong L1 sensitivity to mismatching vowel cues when listeners have lexical expectations. The fact that no other conditions in the PMN or LPC analyses revealed statistically significant differences presents an interpretive challenge. One explanation for these null results is that present analysis is in some way unable to address the relevant differences. Visual inspection of waveforms suggests that considerable component overlap, as well as differences across electrodes (anterior to posterior) may be interfering with our

[14] It had been directly pointed out in the instructions and practice items for the LDT, reinforced through the LDT task itself, and once again illustrated in the feedback on Pic-Phono practice items. So, apart from complete tone deafness (which does not seem to have been the case for any participants), the importance of monitoring tones should have been very clear in the Pic-Phono experiment.

abilities to get clear answers about the true nature of L1 and L2 responses.[15] Alternatively, or additionally, it may be the time windows chosen for the current analyses were not entirely appropriate for capturing the relevant effects. These limitations may justify further exploratory statistical analyses using moving windows and testing for effects of electrode position.

For the moment, I will engage in a more speculative discussion of results, in an attempt to address the fact that, despite a lack of significant statistical differences in ERP analyses, we still need to account the success of L1 listeners in rejection of tone mismatch trials, and L2 listeners in rejection of both vowel and tone mismatch trials.

We can begin by considering L1 listeners. Visual inspection of L1 waveforms suggests that, though it is much weaker than the vowel PMN response, L1 listeners do display PMN sensitivity to tone mismatches. We can posit then, that it is this early sensitivity to tone mismatches that ultimately drives listeners to correctly reject those

[15] One reason why there may be no significant tone PMN is that the response gets washed out by changes in amplitude relative to real word responses across electrodes. While it is more negative in anterior electrodes (e.g., Fz), it is more positive in posterior electrodes (e.g. Pz). Additionally, the PMN appears to be followed by a strong positive shift, which may further distort the PMN for tones. While this positive shift (the LPC) is also observable for vowel responses, because of the strength of the earlier vowel PMN, the vowel LPC ends up essentially overlapping with the real word N400 response during the 400-600 window. Together, these overlaps and shifts undermine our ability to find statistical evidence for differences between conditions.

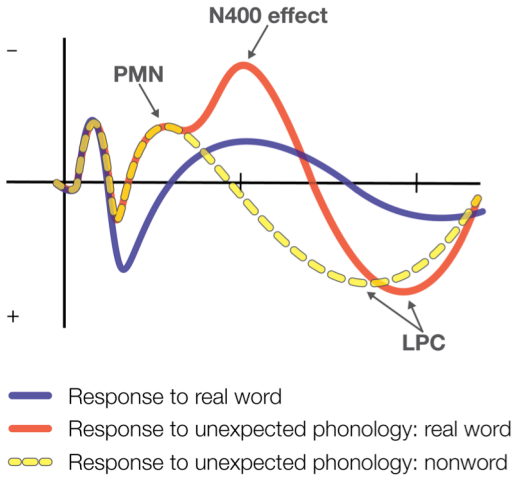


Figure 4.8. Illustration of hypothesized L1 ERP response patterns to Mandarin words and nonwords in phonologically constraining contexts.

trials. In addition to PMNs, we can also observe later positive shifts in both nonword responses—i.e., what we have called the LPC. This positive shift can be understood as re-orientation after detection of the mismatching phonological cue (Sassenhagen & Bornkessel-Schlesewsky, 2015; Sassenhagen et al., 2014). Together then, the PMN and LPC can be interpreted as ERP indices of the processes that lead to correct rejection. It is also worth noting that, because these mismatches involved nonwords in highly predictive contexts, we do not see N400 effects. This null N400 can be contrasted with the strong N400 effect evoked by mismatching real words in the Pic-Word experiment. This suggests that, for nonwords in constraining contexts, it is enough to notice the divergence from phonology to reject the nonword trial. It may be that the LPC then indexes recovery of the intended phonological form—but this is even more speculative than the rest of this discussion. Figure 4.8 attempts to summarize the hypothesized ERP responses discussed in this paragraph.

Shifting our focus to L2 results, the account would be largely the same. PMN and LPC effects, though perhaps somewhat attenuated compared to L1, are indexing

the processes by which L2 learners come to correctly reject nonword trials. The critical question then, is whether L2, like L1, shows a robust PMN for mismatching tone cues. If the answer is affirmative, then we would have evidence of L2 perceptual sensitivity to tone cues. However, another possibility is that, despite a lack of PMN sensitivity to tone cues, L2 nevertheless achieves success in rejecting tone nonwords via an alternative route. In this case, we may still see an LPC effect, suggesting that the nonword has been detected, but this detection is driven by more controlled attentive processes, rather than by automatic and more immediate perception of the mismatching tone. Unfortunately, for the moment we cannot determine which of these is the best approximation of L2 responses.

4.3.4 Conclusion

Chapter 4 has once again provided evidence of weaknesses in tone word recognition by advanced L2 learners. Learners have clear difficulty in encoding tones in explicit long-term memory, and Best Case Scenario results suggest that, even when they do succeed in encoding tones, they do not always succeed at utilizing tones during online Mandarin word recognition. ERP results suggested L1 listeners use early sensitivity to phonological cues to successfully reject mismatching vowels, but there was no clear evidence of other ERP effects in either the L1 or L2 group. In Chapter 5 we will consider the implications of these results, along with those of Experiments 2 and 3, for the three hypotheses that have guided our thinking about L2 tone word recognition.

Chapter 5: Conclusion

5.1 The Tone Perception Hypothesis

The Tone Perception Hypothesis attributes L2 tone word learning difficulties to challenges with low-level perceptual processes, namely, the (in)ability to perceive phonetic distinctions between tones. Results from the experiments reported above suggest that—contrary to my expectations at the outset—continued difficulty with low-level auditory perception of tones is likely to play a persistent role in L2 tone word recognition difficulties.

Experiment 2 showed that, *in DS contexts*, many advanced L2 listeners have persistent low-level perceptual difficulty with tones. These results reinforce similar findings in previous research (Hao, 2012, 2018; Sun, 1998), but with a group of more proficient and more diverse L2 learners (Sun tested college students in intact classrooms, Hao tested learners with 2-3 years of classroom study). Compared to L1 listeners, L2 learners' accuracy dropped somewhat for T1 and T2, and was notably lower for T3. This suggests that DS tone perception is a challenge even for advanced learners, particularly when multiple talkers are involved. Though, as noted above, we should be careful not to exaggerate the magnitude of this difficulty.

In line with previous research (C.-Y. Lee et al., 2009; Pelzl et al., 2018; L. Zhang, 2011), Experiment 2 once again demonstrated that the difficulties L2 learners experience in phonetic tone perception appear limited to contextualized syllables. Experienced L2 learners generally excel at identification of tones in isolated MSs, and even though L2 accuracy dropped for T3 in MS contexts, it was not statistically different from the loss in accuracy shown by L1 listeners. The consistency of these

results across studies should give us increasing confidence that low-level perception of tones *in isolated MS* contexts is not a significant source of difficulty for advanced L2 learners. Unfortunately, this may be of little consolation or practical import, since the relevance of isolated MS tone perception is very limited in real life; even MS words most often occur in context.

5.1.1 The Tone Perception Hypothesis as a simple account of all L2 tone difficulties

Given the observed L2 difficulty in contextualized tone recognition, it is worth considering whether the Tone Perception Hypothesis alone might account for *all* of the present findings. Taken at face value, the level of inaccuracy observed in DS contexts in Experiment 2 (approximately 20% errors) roughly corresponds to the level of inaccuracy observed in rejection of tone nonwords in Experiments 3 and 4 (approximately 20-30% errors in the Best Case Scenario analyses). We could further explain the documented inaccuracy of explicit tone knowledge (from the offline vocabulary test) as the effect of learners' relying purely on explicit memory for those tones without the support of experience-based phonological representations because the tones in the input are inconsistently perceived. That is, if the low-level signal fails to provide information about phonological form, then formation of phonolexical representations will be negatively impacted.

This line of argumentations fits well with research examining naïve or beginning learners, where short-term outcomes seem heavily dependent on individual differences in pitch perception abilities (Bowles et al., 2016; M. Li & DeKeyser, 2017; Perrachione et al., 2011; Wong & Perrachione, 2007). The Perception Hypothesis straightforwardly predicts that such differences will affect the entire

course of L2 learning. Those learners who are able to accurately perceive tones in the input, will be able to successfully encode them in lexical representations, and subsequently use them for word recognition—for example, by rejecting tone nonwords in tasks such as those used in Experiments 3 and 4. This account is appealing as a simple, but powerful explanation for L2 tone word learning and is worth testing further in future research, in particular, by examining individual differences in pitch perception more closely—ideally in a longitudinal context such that early pitch perception aptitude is not confounded with changes in aptitude that take place by learning a tone language. Similar logic applies for considering age of acquisition effects that are known to impact L2 phonological outcomes (e.g., Abrahamsson, 2012).

5.1.2 Some weaknesses in relying only on the Tone Perception Hypothesis

However, there are several ways in which low-level perception seems inadequate as the sole factor in accounting for observed effects. First, despite the superficial similarities in percentages across experiments, the DS difficulty observed in Experiment 2 should not be simplistically equated with effects in the other experiments. Whereas the Tone ID stimuli were produced by multiple talkers, the lexical tasks in Experiment 3 and 4 involved stimuli produced by a single female speaker. This suggests that purely auditory effects observed in the latter experiments ought to be much milder, if indeed they were entirely attributable to low-level perception of tones. In that case, for example, we should have observed near-ceiling performance in our Best Case Scenario analyses for any words that involved T4 switches, as T4 appears to be perceived nearly perfectly even in multi-talker DS

contexts. Instead, we find inconsistent accuracy even when T4 manipulations targeted words for which learners had perfect explicit knowledge.[16] This suggests another layer of difficulty on top of basic auditory perception.

Another puzzle for the Tone Perception Hypothesis is the nature of the context effects found in Experiment 2, which do not necessarily seem to flow directly out of low-level auditory perception. This was discussed earlier, but we can revisit it briefly here. Previously we found (i.e., in Pelzl et al., 2018) that even when tones were clipped from words—and consequently fast and warped by both preceding and following syllables—advanced L2 learners were nevertheless near native in tone perception. It is hard to understand why the mere inclusion of a second syllable—which ought to provide additional helpful cues (as in the case of T3 for L1 listeners)—instead had negative impacts on L2 tone identification. Future research will need to examine additional factors that may be impacting DS perception, such as memory constraints (e.g., *the phonological loop* Baddeley, 1968), L1 prosodic biases that might operate across multiple syllables (Braun et al., 2014; Braun & Johnson, 2011; Schaefer & Darcy, 2014; So & Best, 2010, 2014), and potential ordering effects in the perception of co-articulated tones (Xu, 1994, 1997). These additional factors do not fit comfortably under the current hypothesis as it attempted to limit the space of L2 difficulties to low-level (bottom-up) auditory perceptual issues. The types of

[16] In this respect, results for the Pic-Phono trials are enticing. Where T4 occurs in the nonword stimuli and learners know the correct tones for the real word, they are quite accurate (T1 to T4: 85%; T2 to T4: 96%, T3 to T4: 100%), but this does not hold up for the LDT (T1 to T4: 55%; T2 to T4: 81%, T3 to T4: 73%).

problems raised by the disyllabic results seem to relate to more top-down perceptual processes (i.e., perception of syllable-sized units aligned temporally), and thus may require expanding the hypothesis space, or perhaps revising the framework altogether.

Finally, as we will discuss more below, there are certain aspects of the results that strongly suggest difficulties in tone representation and processing that, while potentially attributable to weakness in auditory perception, also allow for alternative accounts.

5.2 The Tone Representation Hypothesis

If it is correct that not all tone word recognition difficulties are reducible to difficulties in perceiving the phonetic signal, then the next appropriate level to consider is that of tone word representations. As argued above, Experiment 2 provided some evidence for successful toneme abstraction in L2 learners, that is, they successfully learned to identify T3 allotones as belonging to a single tone category. Lexical results in the LDT and Pic-Phono experiments further demonstrate that learners can successfully encode tones (including T3 when it surfaces as T3F) in lexical representations at least some of the time. Although the L2 learners were not as accurate in rejecting tone nonwords as they were for vowel nonwords, they nevertheless performed, as a group, far above chance—in contrast to a similar L2 group in Pelzl et al. (2018).

At the same time, these experiments also provided evidence of substantial weakness in L2 tone word representations. Offline vocabulary tests showed that learners had considerable difficulty remembering tones. Even for words with definitions that they confidently knew, tones were remembered incorrectly around

25% of the time (LDT: 78%; Pic-Phono:72%). Even when learners were highly confident in their tone knowledge, they were still incorrect roughly 15% of the time (LDT: 85%; Pic-Phono: 84%). Assuming these numbers are representative more broadly of learners' vocabularies, and assuming the learners know at least several thousand Mandarin words (Pelzl et al., 2014; H. Shen, 2009), this could mean that their explicit knowledge of Mandarin vocabulary includes hundreds or even thousands of words *with missing or incorrect tones*. It is unclear what such sizeable gap in explicit knowledge means for implicit knowledge of tones. One might believe that, even though explicit knowledge is deficient for a given word, implicit knowledge of that word might be accurate and available for use in automatic processes (DeKeyser, 2003; Suzuki & DeKeyser, 2017). Still, insofar as explicit knowledge has a role to play in L2 word recognition, these results suggest serious challenges for advanced L2 learners in terms of the quality of their lexical representations.

As noted above, one explanation for the extensive weaknesses in explicit L2 tone word representations is that low-level difficulties with tone perception have corrosive cascading effects on all phonolexical representations. However, it is also possible to account for these difficulties through the Tone Representation Hypothesis. In this case we would posit that, although metalinguistic tasks such as tone identification provide evidence of L2 success at forming abstract phonetic tone categories, these L2 tone categories cannot be encoded in (implicit) phonolexical representations. Less extreme positions are also possible, e.g., tones are encoded inconsistently, with poor quality, or only after massive input. In any of these cases,

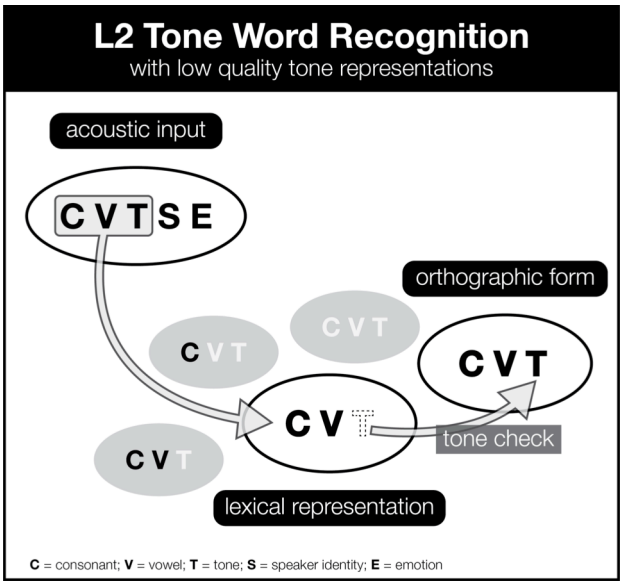


Figure 5.1. L2 tone word representations with poor quality or non-existent phonolexical tone categories might still allow for successful tone word recognition by connections to explicit tone representations, for example, in orthographic word representations.

the difficulty is not to explain L2 tone word recognition failures (how can you recognize what is not encoded?), but to explain L2 tone word recognition success. One possibility is that L2 success in the current tasks depended heavily on explicit knowledge, so that, even if tones are not available in phonolexical representations, they are still accessible via a secondary route such as the orthographic representation of a word (i.e., Pinyin). This alternative path to successful tone word recognition is depicted in Figure 5.1. In this case, successful tone word recognition must also assume successful auditory perception of tones, thus, this account can coexist with the Tone Perception Hypothesis, but adds its own considerable complexity.

5.2.1 Tone representations and L2 Lexical familiarity

As mentioned earlier, Gor and Cook (Cook & Gor, 2015; Gor, 2018; Gor & Cook, 2018) as well as others (Diependaele et al., 2013; Veivo & Järvikivi, 2013) have argued that L2 knowledge for less familiar words (usually lower frequency

items) is characterized by low-quality, or 'fuzzy', phonolexical representations. One key source of evidence for this hypothesis is the apparent lack of inhibition effects for L2 learners in cross-modal priming (see Gor, 2018 for review). Whereas high-frequency/high familiarity words are inhibited by phonological competitors (as occurs for L1 speakers for all words), low-frequency words are either not inhibited, or actually primed, suggesting a uniquely permissive quality to low-familiarity L2 phonolexical representations. In this light, the current results are interesting, in that, with a very coarse measure (accuracy), we find that even mismatches involving highly familiar words lead to incorrect acceptance. This suggests that L2 Mandarin tone difficulties go deeper than just familiarity, and are more akin to problems with difficult L2 speech sounds (Broersma, 2012; Broersma & Cutler, 2008, 2011; Díaz et al., 2012; Sebastián-Gallés & Díaz, 2012). However, as noted in Chapter 1, tones are somewhat unique in this respect as a *class* of sounds, or, more accurately, a *class of suprasegmental sounds*. In other words, tones form a set of phonological contrasts that are orthogonal to segmental contrasts (an idea familiar from Autosegmental Theory, cf. Goldsmith, 1990). This means, unlike other unfamiliar L2 sound categories that might also be considered a class with respect to features (e.g., front-rounded vowels), tones are unique in that, as suprasegmentals, they lie outside the (non-tonal) learner's phonological space, and are contrastive amongst themselves without regard to other segmental features. While lip-rounding, vowel height, or other segmental features may impact F0 to some degree, the contrastive nature of the tone categories is not dependent on these segmental features. Another aspect of tones worth highlighting is that, whereas specific phoneme contrasts will only ever occur in

a subset of vocabulary, tone contrasts are pervasive across every vocabulary item and thus will always play a role in recognition for Mandarin words.

5.3 The Tone Processing Hypothesis

The fact that tones, as a difficult class of sounds, also permeate Mandarin speech leads us to the final hypothesis we wish to discuss. The Tone Processing Hypothesis posits that L2 difficulty with tones derives primarily from the fact that tones are a class of phonetic categories that lie outside the space of specifically *lexical* cues in the learners' L1. Consequently, even if learners have excellent auditory perception of tones, due to their entrenched L1 processing biases, they do not include tones in the normal process of lexical recognition.

Present results allow us to reject the most extreme version of this hypothesis, namely, that L2 learners are *incapable* of ever processing tones lexically. Though attenuated and highly variable, L2 listeners' N400 responses to tone nonwords in Experiment 3 demonstrate that at least some learners are sometimes able to use tone cues in real-time to reject nonwords, just as native listeners do.

While we can reject the most extreme version of the Tone Processing Hypothesis, we cannot easily reject this hypothesis altogether. First, whatever success was apparent in the LDT experiment must be placed in the context of a task that allowed and even encouraged listeners to actively monitor for tones. Thus, while providing some evidence that L2 real-time processing of tones *can* occur, it does not provide evidence that such real-time processing *typically* occurs. To test a stronger role for tone processing, we will need research that places tone word recognition in contexts that do not draw overt attention to tones.

5.3.1 Tone processing as the only L2 problem

I suggested above that the Tone Perception Hypothesis might be a simple and elegant way to account for *all* L2 tone learning difficulties. While it may be less obvious, the same can be said for the Tone Processing Hypothesis.

Under this account, we assume that the Tone Perception Hypothesis is not relevant, and that apparent difficulties in auditorily perceiving tones are not due to low-level perceptual difficulty, but instead due to L1 processing biases. Even though the L2 learner can auditorily perceive the F0 cues in the Mandarin speech signal, those cues are given no weight in the process of lexical access; instead, as depicted earlier, only segmental (C,V) cues are utilized for lexical recognition. This places processing at the head of the word recognition chain, filtering out cues that could have been used to form phonolexical representations. Thus, just as in the case of difficulty in auditory perception, difficulty in tone processing can also account for the observed weaknesses in tone word representations. In the most extreme version, there would be no way for tone categories to ever be encoded, but it is of course not necessary to hold to such an extreme version. In line with present results, we might instead suggest that tones enter into the processing stream *inconsistently*. In this regard, perhaps some of the possible causes of disyllabic tone perception difficulties noted in relation to Experiment 2 (e.g., memory constraints, overshadowing) might be better considered as problems related to processing inconsistency.

Even in the most extreme version of the Tone Processing Hypothesis—that L2 learners can *never* effectively overcome L1 biases—there are still possible alternative routes for successful tone word recognition. Because auditory perception of tones is

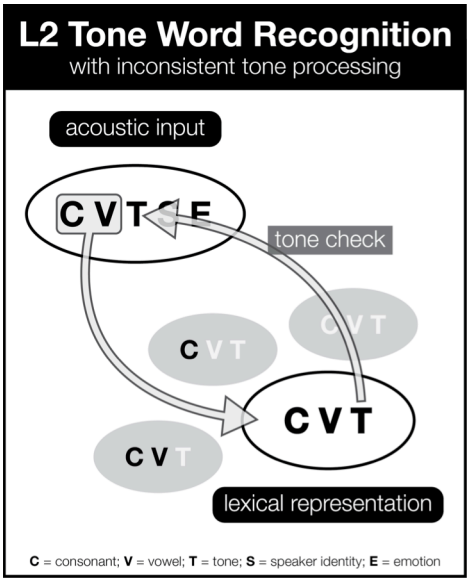


Figure 5.2 L2 tone word processing that fails to utilize tone cues might still allow for successful tone word recognition by checking tones in (implicit or explicit) representations against the acoustic signal held in short-term memory.

possible, the listener might access available representations (or, if necessary, explicit knowledge in orthography or some other form) and then check that against the acoustic signal maintained in memory (Figure 5.2) . This account may be more elaborate than we would like, but is a possible strategy for the L2 listener, especially in tasks like those presented in the present experiments, where there is little time pressure, and use of explicit knowledge is available.

5.3.2 Tone processing and current models of speech recognition

The Tone Processing Hypothesis can be considered in connection to current models of speech processing that posit humans flexibly make use of all available sources of information while attempting to comprehend speech (Noisy Channel models: Gibson et al., 2017; Gibson, Bergen, & Piantadosi, 2013; Ideal adapter models: Kleinschmidt & Jaeger, 2015; Pajak, Fine, Kleinschmidt, & Jaeger, 2016). These models suggest that rational (Bayesian) expectations about the speech signal,

both in terms of the *content* of that signal and the *source* of the signal, will dynamically adapt how different cues in the speech stream are weighted. For L2 learners, this means that L1 expectations (biases) will sometimes stand in the way of properly weighting meaningful speech cues in the L2 (Pajak et al., 2016).

In the case of Mandarin tones, then, the problem would not necessarily be that L2 learners *cannot* process tones, but that they are biased strongly against doing so, in part, because they receive very little lexical information from tones—especially early in the learning process. The question then becomes whether natural Mandarin speech pushes them to rely on tones for comprehension. This is an open question, but, as noted above, at the level of *isolated* lexical items, there is little pressure to pay attention to tones for DS words due to the relatively rare occurrence of tone neighbors. While it might seem that the greater number of tone neighbors for MS words would push learners to attend to tones, as others have pointed out (Wiener & Ito, 2015, 2016; Wiener et al., 2018), because of the large number of homophones that occur for MS words, tones become quite uninformative in many cases for the most frequently occurring syllables. That is, if a word has many homophones, the listener must rely on other cues to word identification besides tone (e.g., frequency, context). Thus, for L2 listeners who have not forged a high sensitivity to tones through massive life-long exposure, and whose smaller vocabularies constrain the number of tone competitors that occur, re-weighting of tone cues for lexical recognition is unlikely to occur without some additional pressure (e.g., personal motivation, demanding teachers). One very relevant and natural pressure that might explain some level of tone processing success across all learners is the need to

produce tones in conversation. This suggests that we should find a tight relationship between L2 ability to accurately produce tone words, and to utilize tones when recognizing those same words.

To wrap up this discussion, as a single account of L2 tone difficulties, the Tone Processing Hypothesis may not be as intuitive as the Tone Perception Hypothesis, but it could be equally as powerful, and has the potential to fit well with current models of human speech comprehension. In order to test it more fairly, future work will need to use tasks that discourage or prevent reliance on overt attention to tones.

5.4 Practical and pedagogical implications

A reader who has come this far might well wonder whether L2 difficulties in tone word recognition have any practical import. On the one hand, it seems highly likely to be the case that tones alone are rarely essential for word recognition in typical speech. Even a tone deaf L2 listener is likely to be able to follow most conversations as long as they know the vocabulary (*sans* tone). However, anecdotally at least, L2 learners who reach more advanced stages often find that tones seem more critical in formal contexts. This might be due to the use of less frequent vocabulary that can be more quickly recognized when tones are accurately perceived. Additionally, as a learner's vocabulary grows closer to the size of an educated native speaker, more tone neighbors will accumulate. Importantly, these tone neighbors will be DS tone neighbors. In yet another contrast to MS words, for DS words, the existence of tone neighbors will make tone cues highly informative, as there are rarely homophones to dilute the information value of tones. Still, practically speaking,

it seems likely that only learners who reach truly high levels of L2 proficiency and use Mandarin in technical and professional contexts will regularly encounter comprehension problems due to tones.

Still, there is one aspect of the tone word recognition issues examined here that seems highly important, namely, incidental learning of vocabulary through listening. Listening is assumed to be a major channel for learning of new vocabulary (cf. Peters & Webb, 2018). Also, in the case of Chinese, unlike many other languages, reading does *not* reinforce the phonological form of a word, especially its tones. Unless you already know the tone for a given character, you will never learn it by looking at the character—and in fact quite a few characters can represent multiple words with different tones. This means that, apart from looking up words in a dictionary or asking an informant about the tones, L2 learners must be able to perceive, process, and encode tones for new vocabulary *through mere exposure*. If they cannot, then they will accumulate ever more toneless phonolexical entries in their mental lexicon.

The present results do not themselves indicate how L2 tone word learning can be improved, but they do indicate certain aspects of tone learning that might be given more attention in L2 pedagogy. Specifically, an emphasis on MS tone perception is clearly not enough, as most L2 difficulties appear to accrue for longer strings of speech. This, of course, includes DS words, but also phrases and sentences. Teachers might explore ways to provide more practice that demands learners utilize tones in such circumstances. Additionally, the pervasive lack of explicit tone knowledge is a

problem that, while severe, can be addressed through carefully designed vocabulary review, and might leverage modern computer-assisted vocabulary review tools.

5.5 Limitations and future directions

One clear limitation of the current study is the small sample size of advanced L2 learners. This does limit the confidence with which we can generalize results to the broader L2 population, and it constrained the types of inferential statistical analyses that could be carried out. Future research will need to find ways to overcome the significant barriers to finding larger samples of advanced L2 learners, particularly if we hope to account for individual differences among them.

Another significant limitation of the present research is its reliance on metalinguistic tasks (Tone ID, LDT) that may not fully capture L2 tone word recognition as it occurs in more ecologically valid circumstances. Future work should seek ways to consider tone word recognition in more complex and meaningful contexts, and might do well to consider factors, such as attention, that might have major impacts on L2 abilities to utilize tone cues.

Discussion above has made it clear that the hypotheses used to frame the present research are difficult to apply to current results without ambiguity. In particular, it is unclear whether the observed difficulties in Experiment 2 for disyllabic tone identification should fall under the scope of tone perception or tone processing. In defining the terms at the outset, I tried to draw a *non-lexical* (auditory perception) vs. *lexical* (processing applied to word recognition) distinction between the hypotheses, but the observed difficulties related to disyllabic tone identification seem to straddle the space between these hypotheses, being neither purely auditory,

nor fully lexical, but rather top-down perceptual processes operating over temporally-aligned abstract units (syllables) but without engaging word recognition. Future work will need to refine (or replace) these hypotheses to address this space.

As reviewed in Chapter 1, disconnects between L2 performance on metalinguistic tasks and lexical tasks are not new in the context of speech learning (Sebastián-Gallés & Díaz, 2012; Strange, 2011). The current study has contributed to such work by going further than most previous tone studies in attempting to account for online performance with respect to the quality of offline knowledge. Future work might attempt to find alternative ways to measure L2 lexical knowledge to better address the source of knowledge used *during* online word recognition, as well as addressing the question of *what tones* were actually perceived. Some eye-tracking paradigms might be able to provide more information along these lines.

5.6 Conclusion

This dissertation has provided new evidence for understanding of L2 tone word recognition. To summarize, the most significant findings are that many advanced L2 learners appear to have persistent low-level difficulty with auditory perception of tones in disyllabic contexts. Most advanced L2 learners also show some level of difficulty in utilizing tone cues during word recognition, as shown by consistently less accurate performance in rejection of tone nonwords compared to vowel nonwords. The behavioral inaccuracy was largely mirrored in ERP responses, with large individual differences across L2 participants. While some showed N400 effects for both tone and vowel nonwords, others showed sensitivity only to vowel nonwords, and some showed little ERP sensitivity to nonwords at all. Finally, offline

measures of L2 tone and definition knowledge showed that, even when L2 learners were quite confident in their explicit knowledge of tones and words, they still had pervasive weaknesses in long-term memory of tones for Mandarin words.

These results were considered in light of three broad hypotheses. While it was not possible to reject any of the hypotheses, it was suggested that the simplest account might be that low-level auditory perception of tones persists for many learners and has cascading effects on higher levels of tone representation and processing. Future research can aim to scrutinize this position more carefully. At the same time, it is quite likely that all three hypotheses have a role to play in accounting for L2 tone learning difficulties, perhaps with different difficulties applying at different stages of learning.

Finally, though no immediate solutions for L2 tone difficulties were presented, I have argued that the practical implications of weaknesses in L2 tone perception are heavily contingent on the needs of individual learners. Those who wish to engage in Chinese at a very high level, such as in formal academic circles, will need to develop highly effective tone word recognition abilities, while learners who have more modest goals may well find that tones are rarely an impediment to successful comprehension of Mandarin, and other issues, such as lack of vocabulary knowledge and background knowledge are more likely to cause communication difficulties. Nevertheless, the present dissertation does provide insights for teachers and learners regarding weaknesses that may not be obvious from simple observation, namely, breakdowns of auditory tone perception in context, and expansive gaps in explicit L2 tone knowledge for words.

Appendices

A1.1 Table summarizing 47 observational studies with experiments targeting non-native perception of Chinese tones

Study	L1	L2 level	Measures	Syllables
Kiriloff (1969)	English	1st semester	Pinyin transcription; 4AFC	MS
Lin (1985)	English	beginner, intermediate, advanced	4AFC	MS
Broselow, Hurtig, & Ringen (1987)	English	naïve	4AFC	MS, DS, TS
Leather (1987)	Dutch, English	naïve	rating task	MS
Repp & Lin (1990)	English	naïve	speeded classification	MS
Lee & Nusbaum (1993)	English	naïve	speeded classification	MS
Lee, Vakoch, & Wurm (1996) *[Exp. 2]*	Canton., English	naïve	AX	MS
Gottfried & Suiter (1997)	English	8 with < 5 years 1 with > 20 years	4AFC	MS
Sun (1998)	English	1st, 2nd, 3rd, & 4th year	4AFC	MS, DS, TS
Klein, Zatorre, & Milner (2001)	English	naïve	AX w/ PET scan	MS
Wang, Jongman, & Sereno (2001)	English	naïve	4AFC (dichotic listening)	MS
Hallé, Chang, & Best (2004)	French	naïve	2AFC; AXB	MS
Alexander, Wong, & Bradlow (2005)	English: music/ non-music	naïve	2AFC	MS
Krishnan, Xu, Gandour, & Cariani (2005)	English	naïve	brainstem frequency following response	MS
Bent, Bradlow, & Wright (2006)	English	naïve	4AFC	MS
Chandrasekaran, Krishnan, & Gandour (2007)	English	naïve	passive discrimination (oddball) w/ ERPs (MMN)	MS
Gottfried (2007)	English: music/ non-music	naïve	4AFC(intact, silent center); AX	MS
Guion & Pederson (2007)	English, Japanese	naïve, ≥4 years	similarity ratings, multi-dimensional scaling	MS
Crinion et al. (2009)	'European'	naïve, 1-4 years experience	structural imaging (MRI)	*NA*
Lee, Tao, & Bond (2009)	English	1st, 2nd, & 3rd year	4AFC (intact, center-only, silent center, onset-only)	MS
Krishnan et al. (2010)	English	naïve	brainstem frequency following response	MS
Krishnan, Gandour, & Bidelman (2010)	English	naïve	brainstem frequency following response	MS
Lee, Tao, & Bond (2010)	English	1st, 2nd, & 3rd year	4AFC (intact, center-only, silent center, onset-only)	MS

Lee, Tao, & Bond (2012)	English	1st, 2nd, 3rd, & 4th year	4AFC (w/ multiple talkers & noise)	MS
Peng et al. (2010)	Canton., German	naïve	AX	MS
So & Best (2010)	Canton., English, Japanese	naïve	4AFC	MS
Yang & Chan (2010)	English	1st, 2nd, & 'advanced'	4AFC; intonation identification	MS, sentences
Huang & Johnson (2011)	English	naïve	difference rating; AX	MS
Braun & Johnson (2011)	Dutch	naïve	ABX	DS
Marie et al. (2011)	French: music/ non-music	naïve	discrimination of syllabic sequences	MS
Zhang (2011)	English	beginner, intermediate	4AFC; pseudoword recognition; multi-dimensional scaling	MS, DS
Hao (2012)	Canton., English	avg. 2.68 years	4AFC	MS, DS
Lee, Tao, & Bond (2013)	English	1st, 2nd, 3rd, & 4th year	4AFC	MS
He & Wayland (2013)	English	3 months, 12 months	4AFC	MS, DS
Liu (2013)	English	naïve	AX	MS
So & Best (2014)	English, French	naïve	AXB; categorization	MS
Lin & Francis (2014)	English	naïve	speeded classification	MS
Chen, Liu, Kager (2015)	Dutch	naïve	AX	DS
Ning, Loucks, & Shih (2015)	English, Korean	naïve, mus., L2 learners	AX, pitch-shift task	MS
Tsukada, Xu, & Rattanasone (2015)	English, Canton. (heritage)	mixed exp., naïve	categorical discrimination test (oddball, 4AFC)	MS
Zou, Chen, Caspers (2016)	Dutch	naïve, beginner, advanced	ABX	MS
Hao & DeJong (2016)	English	intermediate	4AFC	MS
Shen & Froud (2016)	English	naïve, advanced	categorical discrimination, AX	MS
Hao (2017)	English	naïve, 1st year, ≥ 4th year	AXB	MS
Hao (2018)	English	2nd year	4AFC	DS
Pelzl, Lau, Guo, & DeKeyser (2018)	English	advanced	4AFC; LDT; sentence judgment & ERPs	MS; DS; DS in sentences
Shen & Froud (2018)	English	naïve, advanced	categorical discrimination, ERPs	MS

KEY: Canton.=Cantonese; music =musicians; non-music=non-musicians;
2AFC = two alternative forced choice identification;
4AFC = four alternative forced choice identification;
AX = sound discrimination task: *Is the second sound (X) the same as the first (A)?*;
ABX = sound discrimination task: *Is the last sound (X) the same as the first (A) or the second (B)?*;
AXB = sound discrimination task: *Is the second sound (X) the same as the first (A) or the last (B)?*;
oddball = sound discrimination task: *Is the new sound the same as the previous sound?*
MS = monosyllabic; DS = disyllabic; TS = trisyllabic

A1.2 Table summarizing 31 training studies targeting Chinese tone languages

Study	Target Language	L1 (proficiency)	Type	Syllable
Wang et al. (1999)	Mandarin	English (1st year)	pitch	MS
Wang et al. (2003) *fMRI*	Mandarin	English	pitch	MS
Wong & Perrachione (2007)	artificial (Mandarin)	English	lexical	MS
Francis et al. (2008)	Cantonese	English, Mandarin	pitch	MS
Song et al. (2008) *EEG*	artificial (Mandarin)	English	lexical	MS
Chandrasekaran et al. (2010)	artificial (Mandarin)	English	lexical	MS
Liu et al. (2011)	Mandarin	English, Korean	pitch pitch	MS DS
Perrachione et al. (2011)	artificial (Mandarin)	English	lexical	MS
Wang et al. (2011) *behavioral*	Mandarin	English	pitch	MS
EEG	Mandarin	English	pitch	MS
Wong et al. (2011)	artificial (Mandarin)	English	lexical	MS
Cooper & Wang (2012)	artificial (Cantonese)	English (musicians)	lexical	MS
Wang (2013)**	Mandarin	English, Hmong, Japanese	pitch	MS
Cooper & Wang (2013)	artificial (Cantonese)	English	pitch lexical	MS MS
Eng et al. (2013)	Mandarin	English	pitch	MS
Ingvalson et al. (2013)	artificial (Mandarin)	English	pitch lexical	MS MS
Showalter & Hayes-Harb (2013)	artificial (Mandarin)	English	lexical	MS
Braun, Galts, & Kabak (2014)	artificial (Mandarin)	French, German, Japanese	pitch, lexical	DS
Maddox & Chandrasekaran (2014)	Mandarin	English	pitch	MS
Sadakata & McQueen (2014)	artificial (Mandarin)	Dutch	pitch	DS
Saito & Wu (2014)**	Mandarin	Cantonese	lexical	MS
Chang & Bowles (2015)	artificial (Mandarin)	English	lexical	MS, DS
Lu et al. (2015) *EEG*	artificial (Mandarin)	English	pitch	MS
Maddox et al. (2015)	Mandarin	non-tonal L1	pitch	MS
Morett & Chang (2015)	Mandarin	English	lexical	MS
Qi et al. (2015) *MRI*	Mandarin	English	lexical	*NA*
Yang et al. (2015) *fMRI*	artificial (Mandarin)	English	lexical	MS
Zhao & Kuhl (2015)	Mandarin	English	pitch	MS
Bowles et al. (2016)	artificial (Mandarin)	English	lexical	MS, DS
Antoniou & Wong (2016)	artificial (Mandarin-Hindi)	English	lexical	MS
Lee et al. (2017) *MEG*	Mandarin	non-tonal L1 (beginners)	pitch	MS
Wiener et al. (2018)	artificial (Mandarin)	English (2nd year)	lexical	MS

Key: MS = monosyllabic; DS = disyllabic; TS = trisyllabic

pitch = indicates outcomes were measures of phonetic pitch categorization

lexical = indicates outcomes were measures of lexical learning

Most studies in this table used behavioral methods to measure outcomes. Neurolinguistic measures are indicated in *italics* after the study name. Unless indicated in parentheses, participants were naïve, that is, prior to training they had no previous experience learning Mandarin (or another tonal language).

A2.1 Additional statistical reporting for Experiments 1 and 2

Table A2.1.1 Mixed model behavioral accuracy estimates in Experiment 1

	Fixed Effects				*Random Effects* (sd)		
	Estimate	Std.Error	z	Pr(>|z|)	subj	word	talker
(Intercept)	3.283	0.280	11.709	<.001	0.922	0.381	0.178
context1	0.281	0.234	1.203	.229	0.434	0.062	0.391
context2	0.517	0.173	2.987	.003	0.413	0.232	<.001
tone1	0.545	0.298	1.831	.067	0.732	0.091	0.439
tone2	-0.398	0.185	-2.156	.031	0.784	<.001	0.100
tone3	-1.868	0.451	-4.144	<.001	0.708	0.410	0.728
context1:tone1	-0.164	0.194	-0.845	.398	<.001	0.212	<.001
context2:tone1	-0.617	0.205	-3.003	.003	<.001	<.001	0.265
context1:tone2	0.292	0.192	1.523	.128	<.001	<.001	0.258
context2:tone2	-0.634	0.426	-1.490	.136	0.361	<.001	0.794
context1:tone3	0.282	0.553	0.511	.609	0.668	0.635	0.831
context2:tone3	1.467	0.365	4.018	<.001	0.465	0.299	0.566

Table A2.1.1.1 Sum coding applied to model coefficients for behavioral accuracy in Experiment 1

Context	*dummy 1*	*dummy 2*	
MS	1	0	
DS	0	1	
CS	-1	-1	
Tone	***dummy 1***	***dummy 2***	***dummy 3***
Tone 1	1	0	0
Tone 2	0	1	0
Tone 3	0	0	1
Tone 4	-1	-1	-1

Table A2.1.2 Mixed model behavioral accuracy estimates in Experiment 2 (sum coded)

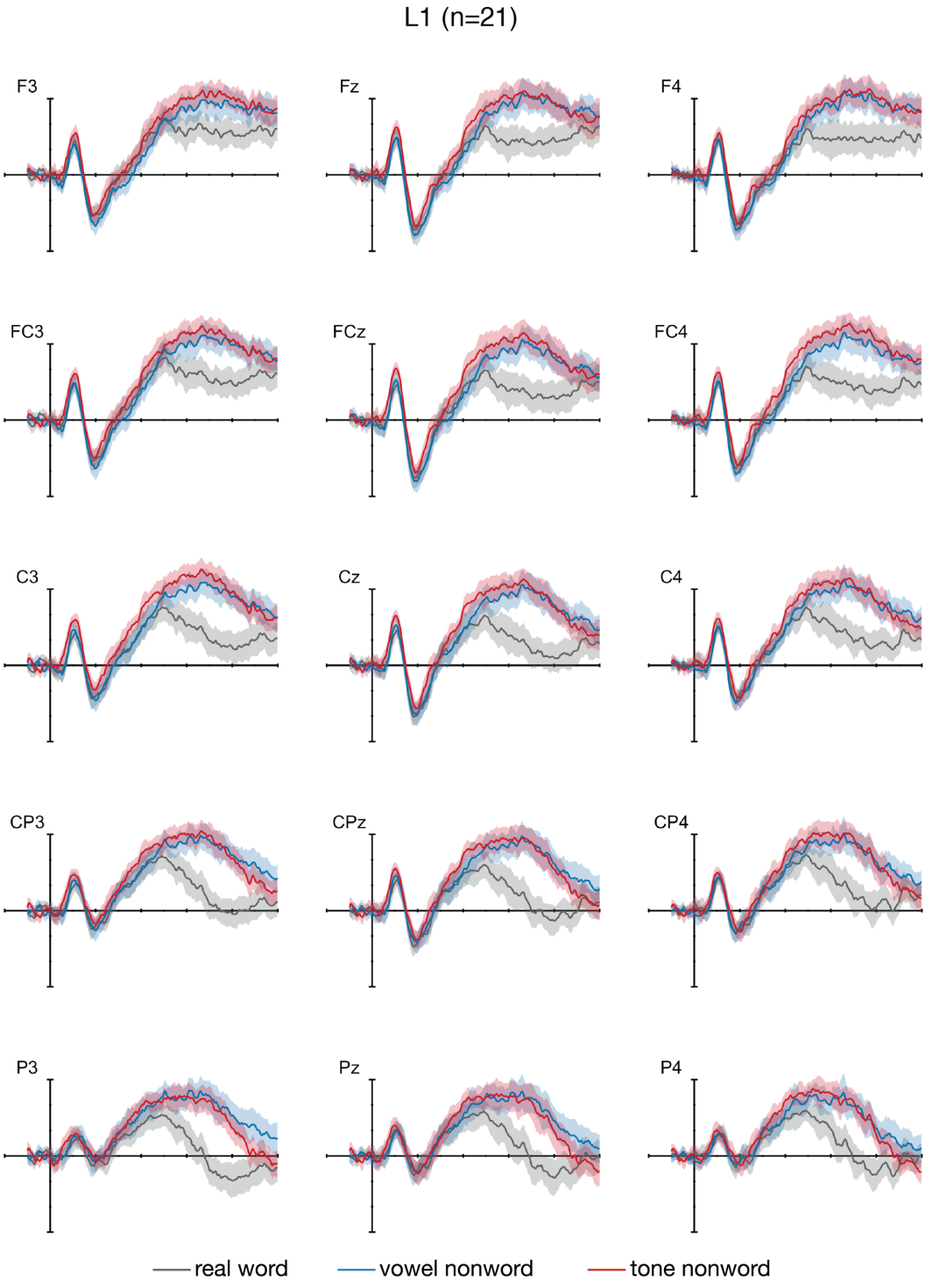
	Fixed Effects				*Random Effects* (sd)		
	Estimate	Std.Error	z	Pr(>|z|)	subj	word	talker
(Intercept)	3.151	0.254	12.404	<.001	1.014	0.242	0.230
context1	0.354	0.093	3.810	<.001	0.085	<.001	0.100
tone1	0.062	0.146	0.422	.673	0.240	<.001	<.001
tone2	-0.561	0.238	-2.358	.018	0.689	0.348	<.001
tone3	-1.158	0.323	-3.590	<.001	0.436	0.012	0.589

group1	0.215	0.192	1.115	.265	—	0.040	0.126
context1:tone1	0.469	0.135	3.486	<.001	<.001	<.001	0.058
context1:tone2	0.007	0.231	0.030	.976	0.022	0.285	0.300
context1:tone3	-0.557	0.375	-1.484	.138	0.531	0.201	0.669
context1:group1	-0.301	0.082	-3.658	<.001	—	<.001	0.072
tone1:group1	-0.267	0.158	-1.692	.091	—	<.001	0.148
tone2:group1	0.137	0.183	0.748	.454	—	0.187	<.001
tone3:group1	0.595	0.154	3.868	<.001	—	0.070	0.160
context1:tone1:group1	-0.375	0.150	-2.508	.012	—	0.143	<.001
context1:tone2:group1	0.214	0.121	1.768	.077	—	<.001	0.139
context1:tone3:group1	-0.031	0.162	-0.189	.850	—	0.186	0.011

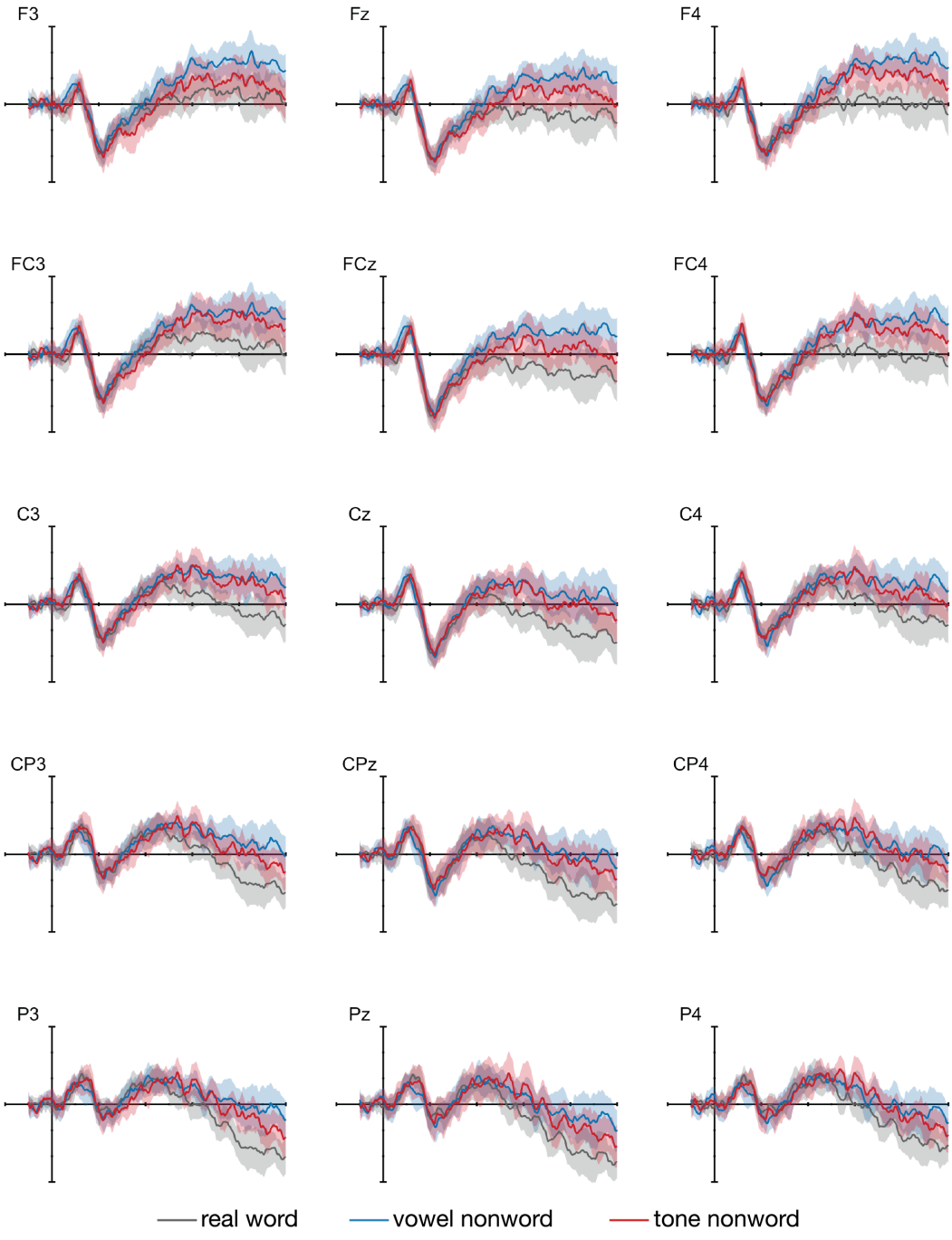
Table A2.1.2.1 Sum coding applied to model coefficients for behavioral accuracy in Experiment 2

Context	*dummy 1*		
MS	1		
DS	-1		
Tone	***dummy 1***	***dummy 2***	***dummy 3***
Tone 1	1	0	0
Tone 2	0	1	0
Tone 3	0	0	1
Tone 4	-1	-1	-1
Group	***dummy 1***		
L1	1		
L2	-1		

A3.1 Grand average waveforms for all electrodes (Experiment 3)



L2 (n=18)


F3
Fz
F4
FC3
FCz
FC4
C3
Cz
C4
CP3
CPz
CP4
P3
Pz
P4
real word
vowel nonword
tone nonword

A3.2 Additional statistical reporting for Experiment 3

Table A3.2.1 Mixed model behavioral accuracy estimates in Lexical Decision Task

	Fixed Effects				*Random Effects* (sd)	
	Estimate	Std.Error	z	Pr(>|z|)	subj	item
(Intercept)	2.990	0.165	18.076	<.001	0.577	0.690
cond1	0.594	0.174	3.405	.001	0.561	0.762
cond2	-0.787	0.151	-5.203	<.001	0.434	0.714
group1	1.144	0.132	8.652	<.001	—	0.455
cond1:group1	0.056	0.157	0.355	.723	—	0.516
cond2:group1	0.407	0.127	3.203	.001	—	0.357

Table A3.2.1.1 Sum coding applied to model coefficients for behavioral accuracy in Lexical Decision Task

Condition	*dummy 1*	*dummy 2*
real	1	0
tone	0	1
vowel	-1	-1
Group	***dummy 1***	
L1	1	
L2	-1	

Table A3.2.2 Mixed model N400 (400-900 ms) amplitude estimates in Lexical Decision Task

	Fixed Effects					*Random Effects* (sd)		
	Estimate	Std.Error	df	t	Pr(>|t|)	elec	subj*	item
(Intercept)	-2.252	0.462	44.26	-4.879	<.001	1.280	2.640	1.495
cond1	1.387	0.269	99.69	5.157	<.001	<.001	1.067	1.905
cond2	-0.781	0.245	104.93	-3.180	.002	<.001	0.905	1.813
group1	-1.662	0.462	44.26	-3.599	.001	—	—	1.500
cond1:group1	0.654	0.266	97.48	2.458	.016	—	—	1.864
cond2:group1	-0.035	0.230	96.65	-0.153	.879	—	—	1.602

*subject random effect nested under electrode

Table A3.2.2.1 Sum coding applied to model coefficients for N400 (400-900 ms) amplitude in Lexical Decision Task

Condition	*dummy 1*	*dummy 2*
real	1	0
vowel	0	1
tone	-1	-1

Group	*dummy 1*
L1	1
L2	-1

Table A3.2.3 Mixed model behavioral accuracy estimates in Best Case Scenario Lexical Decision Task

	Fixed Effects				*Random Effects* (sd)	
	Estimate	Std.Error	z	Pr(>|z|)	subj	item
(Intercept)	1.518	0.249	6.089	<.001	0.707	0.668
cond1	-0.645	0.210	-3.081	.002	0.321	0.944

Table A3.2.3.1 Sum coding applied to model coefficients for behavioral accuracy in Best Case Scenario Lexical Decision Task

Condition	*dummy 1*
tone	1
vowel	-1

A3.3 Table of stimuli for the Lexical Decision Task (Experiment 3)

Real Word	Pinyin	Translation	LogFreq	Citation Tones	Tone Switch	Tone Nonword	Vowel Nonword
春天	chūntiān	*spring*	2.43	11	1/2	chúntiān	chuāntiān
发音	fāyīn	*pronunciation*	2.37	11	1/2	fáyīn	fūyīn
医生	yīshēng	*doctor*	3.49	11	1/3	yǐshēng	yēshēng
咖啡	kāfēi	*coffee*	3.29	11	1/3	kǎfēi	kēfēi
公司	gōngsī	*company*	3.43	11	1/4	gòngsī	guāngsī
飞机	fēijī	*airplane*	3.21	11	1/4	fèijī	fājī
生活	shēnghuó	*life*	3.68	12	1/2	shénghuó	shānghuó
英雄	yīngxióng	*hero*	3.11	12	1/2	yíngxióng	yāngxióng
身材	shēncái	*figure*	2.71	12	1/3	shěncái	shāncái
科学	kēxué	*science*	2.88	12	1/3	kěxué	kāxué
空调	kōngtiáo	*air conditioner*	2.28	12	1/4	kòngtiáo	kāngtiáo
规则	guīzé	*rule*	2.95	12	1/4	guìzé	gūzé
婚礼	hūnlǐ	*wedding*	2.96	13	1/2	húnlǐ	huānlǐ
方法	fāngfǎ	*method*	3.39	13	1/2	fángfǎ	fēngfǎ

思想	sīxiǎng	*thought*	2.82	13	1/3	sǐxiǎng	sāxiǎng
观点	guāndiǎn	*viewpoint*	2.85	13	1/3	guǎndiǎn	gāndiǎn
歌手	gēshǒu	*singer*	2.81	13	1/4	gèshǒu	gūshǒu
机场	jīchǎng	*airport*	2.91	13	1/4	jìchǎng	jūcǎng
兄弟	xiōngdì	*brother*	3.43	14	1/2	xióngdì	xīngdì
书店	shūdiàn	*bookstore*	2.18	14	1/2	shúdiàn	shādiàn
家具	jiājù	*furniture*	2.48	14	1/3	jiǎjù	jiējù
宗教	zōngjiào	*religion*	2.67	14	1/3	zǒngjiào	zēngjiào
车祸	chēhuò	*car accident*	2.75	14	1/4	chèhuò	chāhuò
商店	shāngdiàn	*store*	2.90	14	1/4	shàngdiàn	shēngdiàn
文章	wénzhāng	*article*	2.73	21	2/1	wēnzhāng	wánzhāng
服装	fúzhuāng	*clothing*	2.69	21	2/3	fǔzhuāng	féizhuāng
同屋	tóngwū	*roommate*	1.34	21	2/3	tōngwū	téngwū
白天	báitiān	*daytime*	2.67	21	2/3	bǎitiān	bátiān
阳光	yángguāng	*sunlight*	2.87	21	2/4	yàngguāng	yóngguāng
原因	yuányīn	*reason*	3.58	21	2/4	yuànyīn	yúnyīn
和平	hépíng	*peace*	2.81	22	2/1	hēpíng	hóupíng
银行	yínháng	*bank*	3.01	22	2/1	yīnháng	yánháng
职员	zhíyuán	*office worker*	2.37	22	2/3	zhǐyuán	zhéyuán
邮局	yóujú	*post office*	2.02	22	2/3	yǒujú	yújú
留言	liúyán	*message*	2.91	22	2/4	liùyán	lóuyán
人民	rénmín	*(the) people*	2.85	22	2/4	rènmín	ránmín
门口	ménkǒu	*doorway*	2.83	23	2/1	mēnkǒu	mǐnkǒu
传统	chuántǒng	*tradition*	2.93	23	2/1	chuāntǒng	chúntǒng
存款	cúnkuǎn	*deposit*	2.13	23	2/1	cūnkuǎn	cánkuǎn
情感	qínggǎn	*emotion*	2.74	23	2/4	qìnggǎn	qiánggǎn
结果	jiéguǒ	*result*	3.51	23	2/4	jièguǒ	jiúguǒ
财产	cáichǎn	*property*	2.80	23	2/4	càichǎn	cíchǎn
程度	chéngdù	*degree*	2.98	24	2/1	chēngdù	chóngdù
环境	huánjìng	*environment*	2.96	24	2/1	huānjìng	hánjìng
学校	xuéxiào	*school*	3.39	24	2/3	xuěxiào	xiéxiào
条件	tiáojiàn	*condition*	3.04	24	2/3	tiǎojiàn	táojiàn
模特	mótè	*model*	2.70	24	2/4	mòtè	máotè
毛病	máobìng	*defect*	2.88	24	2/4	màobìng	miáobìng
酒吧	jiǔbā	*bar*	3.15	31	3/1	jiūbā	jiǎbā

傻瓜	shǎguā	*fool*	3.10	31	3/1	shāguā	shǐguā
母亲	mǔqīn	*mother*	3.37	31	3/2	múqīn	mǒuqīn
早餐	zǎocān	*breakfast*	2.92	31	3/4	zàocān	zǒucān
首都	shǒudū	*capital*	2.30	31	3/4	shòudū	shǎodū
果汁	guǒzhī	*fruit juice*	2.48	31	3/4	guózhī	gǔzhī
舞台	wǔtái	*stage*	2.76	32	3/1	wūtái	wǒtái
种族	zhǒngzú	*race*	2.60	32	3/1	zhōngzú	zhěngzú
演员	yǎnyuán	*actor*	3.06	32	3/2	yányuán	yǐnyuán
主题	zhǔtí	*theme*	2.77	32	3/2	zhútí	zhǐtí
导游	dǎoyóu	*tour guide*	1.82	32	3/4	dàoyóu	duǒyóu
小时	xiǎoshí	*hour*	3.63	32	3/4	xiàoshí	xǐshí
选手	xuǎnshǒu	*athlete*	2.81	33	3/1	xuānshǒu	xiǎnshǒu
诊所	zhěnsuǒ	*clinic*	2.55	33	3/1	zhēnsuǒ	zhǎnsuǒ
表姐	biǎojiě	*female cousin*	1.78	33	3/1	biāojiě	bǎojiě
美女	měinǚ	*beautiful girl*	2.97	33	3/4	mèinǚ	mǐnǚ
领导	lǐngdǎo	*leader*	2.79	33	3/4	lìngdǎo	lěngdǎo
水果	shuǐguǒ	*fruit*	2.59	33	3/4	shuìguǒ	shuǎiguǒ
体育	tǐyù	*physical training*	2.60	34	3/1	tīyù	tǔyù
勇气	yǒngqì	*courage*	2.91	34	3/1	yōngqì	yǐngqì
晚饭	wǎnfàn	*dinner*	3.03	34	3/2	wànfàn	wěnfàn
喜剧	xǐjù	*comedy*	2.62	34	3/2	xíjù	xǔjù
比赛	bǐsài	*competition*	3.25	34	3/2	bísài	bǎsài
米饭	mǐfàn	*rice*	1.91	34	3/4	mìfàn	měifàn
辣椒	làjiāo	*chili pepper*	2.15	41	4/1	lājiāo	lùjiāo
帅哥	shuàigē	*handsome guy*	2.20	41	4/1	shuāigē	shuìgē
现金	xiànjīn	*cash*	2.90	41	4/1	xiānjīn	xìnjīn
作家	zuòjiā	*author*	2.72	41	4/2	zuójiā	zàojiā
律师	lǜshī	*lawyer*	3.26	41	4/3	lǚshī	làshī
战争	zhànzhēng	*war*	3.06	41	4/3	zhǎnzhēng	zhènzhēng
话题	huàtí	*topic*	2.92	42	4/1	huātí	huòtí
少年	shàonián	*youth*	2.47	42	4/1	shāonián	shòunián
性格	xìnggé	*disposition*	2.66	42	4/2	xínggé	xiànggé
爱情	àiqíng	*romance*	2.98	42	4/2	áiqíng	àoqíng
大学	dàxué	*university*	3.26	42	4/3	dǎxué	dàixué
距离	jùlí	*distance*	3.00	42	4/3	jǔlí	jìlí

背景	bèijǐng	*background*	2.80	43	4/1	bēijǐng	bàijǐng
办法	bànfǎ	*means*	3.57	43	4/1	bānfǎ	bènfǎ
入口	rùkǒu	*"in" door*	2.58	43	4/2	rúkǒu	rèkǒu
饭馆	fànguǎn	*restaurant*	1.72	43	4/2	fánguǎn	fènguǎn
地铁	dìtiě	*subway*	2.45	43	4/2	dítiě	dàtiě
字典	zìdiǎn	*dictionary*	2.08	43	4/3	zǐdiǎn	zuìdiǎn
报告	bàogào	*report*	3.28	44	4/1	bāogào	bàgào
政治	zhèngzhì	*politics*	2.85	44	4/1	zhēngzhì	zhàngzhì
照片	zhàopiàn	*photograph*	3.39	44	4/2	zháopiàn	zhùpiàn
社会	shèhuì	*society*	3.05	44	4/2	shéhuì	shùhuì
运动	yùndòng	*exercise*	3.04	44	4/3	yǔndòng	yuàndòng
动物	dòngwù	*animal*	3.09	44	4/3	dǒngwù	dàngwù
FILLERS							
将军	jiāngjūn	*general*	2.70	11			
高中	gāozhōng	*high school*	3.04	11			
阿姨	āyí	*aunt*	2.59	12			
新闻	xīnwén	*news*	3.21	12			
餐馆	cānguǎn	*restaurant*	2.79	13			
风景	fēngjǐng	*scenery*	2.50	13			
周末	zhōumò	*weekend*	3.12	14			
黑色	hēisè	*black*	2.83	14			
明星	míngxīng	*celebrity*	3.05	21			
邻居	línjū	*neighbor*	3.04	21			
厨房	chúfáng	*kitchen*	3.02	22			
年级	niánjí	*grade*	2.88	22			
团体	tuántǐ	*organization*	2.52	23			
食品	shípǐn	*foodstuff*	2.67	23			
红色	hóngsè	*red*	2.94	24			
能力	nénglì	*ability*	3.28	24			
海鲜	hǎixiān	*seafood*	1.91	31			
粉丝	fěnsī	*fan*	2.67	31			
语言	yǔyán	*language*	2.87	32			
口红	kǒuhóng	*lipstick*	2.15	32			
想法	xiǎngfǎ	*idea*	3.41	33			
老板	lǎobǎn	*boss*	3.21	33			

广告	guǎnggào	*advertisement*	2.98	34
考试	kǎoshì	*test*	2.69	34
日期	rìqī	*date*	2.62	41
快餐	kuàicān	*fast food*	2.05	41
坏人	huàirén	*bad person*	2.81	42
外婆	wàipó	*wife*	2.21	42
号码	hàomǎ	*number*	3.19	43
路口	lùkǒu	*intersection*	2.20	43
汉字	hànzì	*Chinese character*	1.11	44
教室	jiàoshì	*classroom*	2.36	44

A4.1 Experiment 4: Picture-Word Mismatch additional information

This section reports the further details regarding the Picture-Word Mismatch blocks.

A4.1.1 Creation of lists for Picture-Word blocks

The three Pic-Word lists were constructed so that items were rotated across participants in a manner that minimized clues about which trials would be matching or mismatching, in the perhaps unlikely event that participants would remember which images had occurred in match/mismatch trials in the Pic-Phono experiment. First, all of the real words that had occurred in match trials in a given Pic-Phono list were selected. Next real words corresponding to the vowel nonwords from that same list were selected—these were items that had been mismatch trials in the Pic-Phono. This made for a set of 64 real words. Half of these items (50% originally from of the Pic-Phono real words, 50% originally from Pic-Phono vowel nonwords) were randomly chosen to serve as matches and half were chosen to be mismatches. For the

mismatches, each word was paired with one of 32 pictures that did not match any words in the (i.e., images that had accompanied tone nonwords in the Pic-Phono experiment). Mismatching picture-word pairs were manually checked to be sure the mismatch was obvious, and to avoid strong overlaps on the initial syllables of the picture-evoked word and the auditory stimulus (e.g., the image for *xiong2mao1* 'panda' would not be paired with the word *xiang1jiao1* 'banana' due to the similarity of the onsets of the first syllables).

A4.1.2 Experiment 4: Details of Picture-Word behavioral accuracy results and statistical analysis

Descriptive results are listed in Table A4.1.1 and depicted visually in boxplots in Figure A4.1.1. Performance was near-ceiling across both conditions for both L1 and L2 groups, with two L1 participants performing notably less accurate than all other participants.

Accuracy results were submitted to a generalized linear mixed-effects model, with fixed effects for *condition* (match, mismatch), and *group* (L1, L2) and their interactions. Model fitting procedures were the same as for experiments reported earlier. The final model was the maximal model with by-subject random intercepts and slopes for the effect of condition, and by-item random slopes for the effect of

Table A4.1.1. Descriptive accuracy results for the Picture-Word Mismatch (Experiment 4A)

Group	Condition	Mean Accuracy (sd)
L1 (*n*=22)	Match	.97 (.17)
	Mismatch	.98 (.14)
L2 (*n*=16)	Match	.98 (.14)
	Mismatch	1.00 (.06)

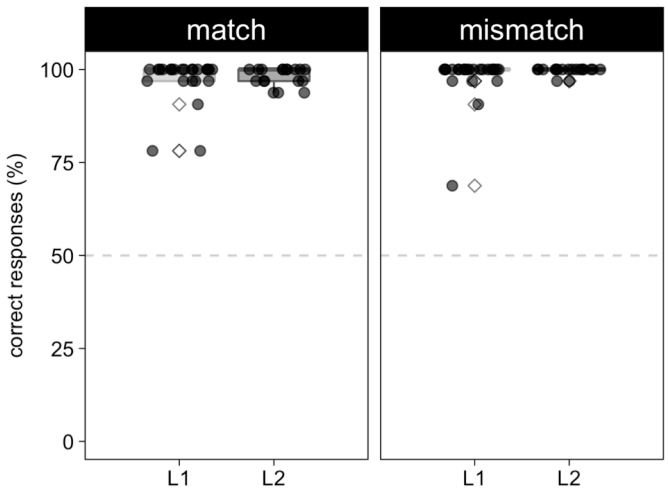


Figure A4.1.1. Boxplot of accuracy results for Picture-Word Mismatch.

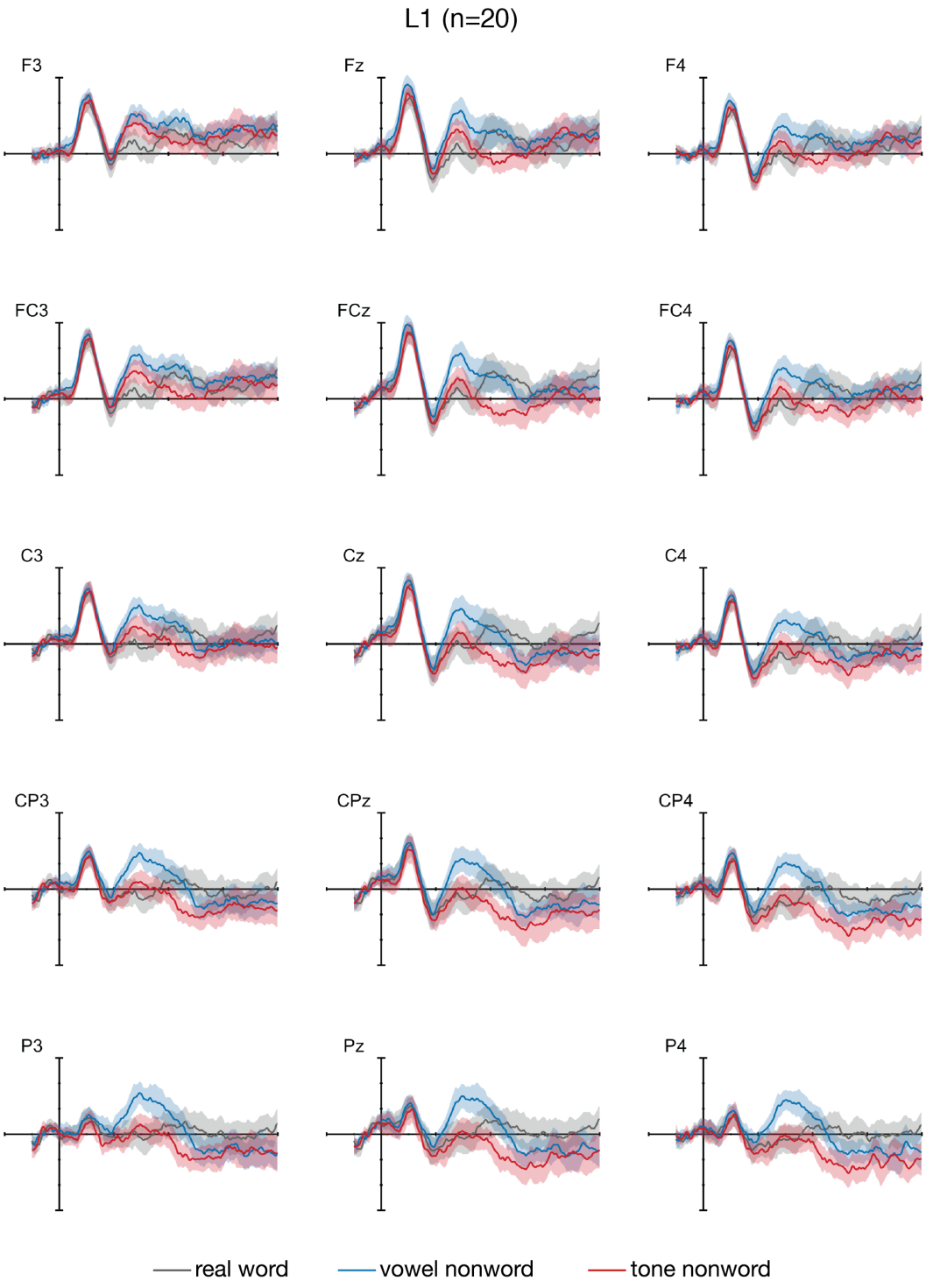
Table A4.1.2. Mixed Model ANOVA Table for accuracy results of Picture-Word Mismatch (Type 3 tests, LRT-method) (Experiment 4A)

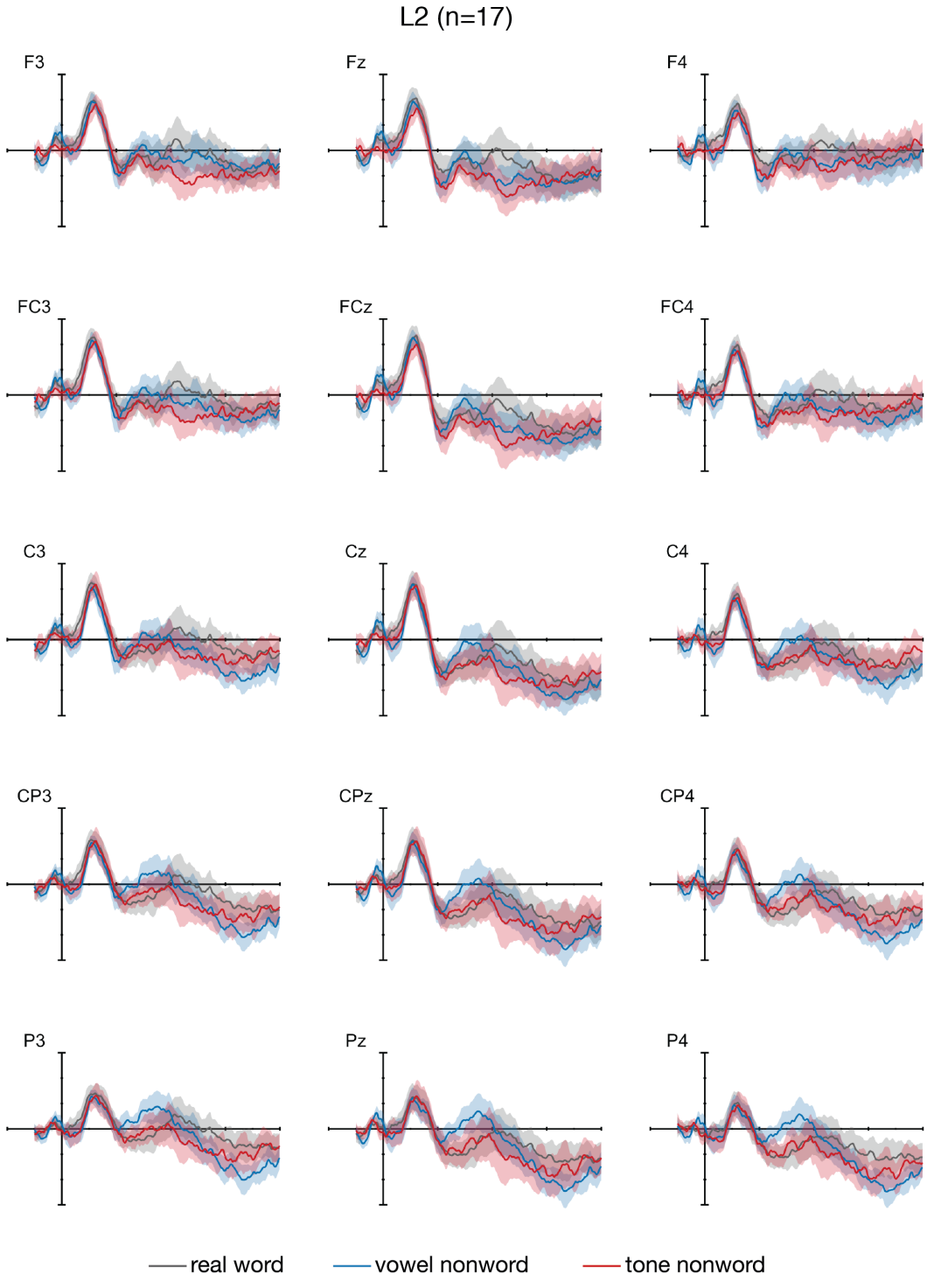
Effect	Df	Chisq.	Chi Df	Pr(>Chisq)	
condition	9	7.67	1	.006	**
group	9	0.32	1	.573	
condition × group	9	1.91	1	.167	
Signif. codes: *** <0.001; **<0.01; *<0.05; . <0.1					
model formula: accuracy ~ condition * group + (condition | subject) + (condition * group | item)					

condition and group and their interaction.

As shown in Table A4.1.2, there was a significant difference in condition—however, this was driven entirely by the relatively poor performance of the two L1 outliers. If these two participants are removed form the data, the difference goes away. There were no significant accuracy differences between groups, and no significant interaction between condition and group. This is not surprising given the near ceiling performance of almost all participants. It appears that, regardless of language group, the image matches and mismatches were as obvious as intended.

A4.2 Grand average waveforms for all electrodes (Experiment 4B: Picture-Phonology)




L2 (n=17)
F3
Fz
F4
FC3
FCz
FC4
C3
Cz
C4
CP3
CPz
CP4
P3
Pz
P4
real word
vowel nonword
tone nonword

A4.3 Additional statistical reporting for Experiment 4

Table A4.3.1 Mixed model behavioral accuracy estimates in Picture-Phonology Mismatch

	Fixed Effects				*Random Effects (sd)*	
	Estimate	Std.Error	z	Pr(>|z|)	subj	item
(Intercept)	3.109	0.192	16.178	<.001	0.801	0.756
cond1	0.207	0.169	1.229	.219	0.623	0.621
cond2	-0.418	0.165	-2.534	.011	0.477	0.736
group1	0.977	0.154	6.351	<.001	—	0.246
cond1:group1	-0.118	0.152	-0.780	.436	—	0.278
cond2:group1	0.770	0.137	5.632	<.001	—	0.255

Table A4.3.1.1 Sum coding applied to model coefficients for behavioral accuracy in Picture-Phonology Mismatch

Condition	*dummy 1*	*dummy 2*
real	1	0
tone	0	1
vowel	-1	-1
Group	***dummy 1***	
L1	1	
L2	-1	

Table A4.3.2 Mixed model PMN (200-400 ms) amplitude estimates in Picture-Phonology Mismatch

	Fixed Effects					*Random Effects* (sd)		
	Estimate	Std.Error	df	t	Pr(>|t|)	elec	subj*	item
(Intercept)	0.255	0.553	40.52	0.461	.647	1.112	3.188	1.490
cond1	0.526	0.314	77.22	1.673	.098	<.001	1.444	1.880
cond2	-1.070	0.290	97.41	-3.685	<.001	<.001	1.171	2.002
group1	-0.822	0.556	41.60	-1.477	.147	—	—	1.616
cond1:group1	0.167	0.326	84.05	0.512	.610	—	—	2.057
cond2:group1	-0.387	0.298	101.66	-1.298	.197	—	—	2.104

**subject random effect nested under electrode*

Table A4.3.2.1 Sum coding applied to model coefficients for PMN (200-400 ms) amplitude in Picture-Phonology Mismatch

Condition	*dummy 1*	*dummy 2*
real	1	0
vowel	0	1

tone	-1	-1
Group	***dummy 1***	
L1	1	
L2	-1	

Table A4.3.3 Mixed model LPC (400-600 ms) amplitude estimates in Picture-Phonology Mismatch

	Fixed Effects					*Random Effects* (sd)		
	Estimate	Std.Error	df	t	Pr(>|t|)	elec	subj*	item
(Intercept)	0.761	0.604	41.44	1.259	.215	1.395	3.469	1.696
cond1	-0.648	0.354	73.57	-1.833	.071	<.001	1.663	2.036
cond2	-0.209	0.339	98.40	-0.616	.539	<.001	1.364	2.353
group1	-0.747	0.602	40.72	-1.242	.221	—	—	1.603
cond1:group1	0.078	0.352	72.85	0.222	.825	—	—	2.012
cond2:group1	-0.458	0.335	96.47	-1.364	.176	—	—	2.302

**subject random effect nested under electrode*

Table A4.3.3.1 Sum coding applied to model coefficients for LPC (400-600 ms) amplitude in Picture-Phonology Mismatch

Condition	*dummy 1*	*dummy 2*
real	1	0
vowel	0	1
tone	-1	-1
Group	***dummy 1***	
L1	1	
L2	-1	

Table A4.3.4 Mixed model behavioral accuracy estimates in Picture-Word Mismatch

	Fixed Effects				*Random Effects (sd)*	
	Estimate	Std.Error	z	Pr(>|z|)	subj	item
(Intercept)	5.599	0.490	11.434	<.001	1.580	<.001
word1	-0.557	0.228	-2.446	.014	0.205	<.001
group1	-0.210	0.368	-0.570	.569	—	<.001
word1:group1	0.318	0.235	1.353	.176	—	0.719

Table A4.3.4.1 Sum coding applied to model coefficients for behavioral accuracy in Picture-Word Mismatch

Condition	*dummy 1*

match	1
mismatch	-1
Group	***dummy 1***
L1	1
L2	-1

Table A4.3.5 Mixed model N400 (200-700 ms) amplitude estimates in Picture-Word Mismatch

	Fixed Effects					*Random Effects* (sd)		
	Estimate	Std.Error	df	t	Pr(>|t|)	elec	subj*	item
(Intercept)	0.040	0.606	42.86	0.065	.948	1.138	3.445	1.614
cond1	0.986	0.270	90.62	3.648	<.001	0.000	1.159	1.529
group1	-1.439	0.602	41.82	-2.389	.022	—	—	1.510
cond1:group1	0.145	0.261	82.66	0.554	.581	—	—	1.405

**subject random effect nested under electrode*

Table A4.3.5.1 Sum coding applied to model coefficients for N400 (200-700 ms) in Picture-Word Mismatch

Condition	*dummy 1*
match	1
mismatch	-1
Group	***dummy 1***
L1	1
L2	-1

Table A4.3.6 Mixed model behavioral accuracy estimates in Best Case Scenario Picture-Phonology Mismatch

	Fixed Effects				*Random Effects* (sd)	
	Estimate	Std.Error	z	Pr(>|z|)	subj	item
(Intercept)	5.972	1.550	3.853	<.001	1.813	4.170
condtone	3.995	1.495	2.673	.008	0.928	3.952

Table A4.3.6.1 Sum coding applied to model coefficients for behavioral accuracy in Best Case Scenario Picture-Phonology Mismatch

Group	*dummy 1*
vowel	1
tone	-1

A4.4 Stimuli for Picture-Phonology & Picture-Word Mismatching tasks (Experiment 4)

Real Word	Pinyin	Translation	LogFreq	Citation Tones	Tone Switch	Tone Nonword	Vowel Nonword
西瓜	xīguā	*watermelon*	1.89	11	1/4	xìguā	xūguā
背包	bēibāo	*backpack*	2.29	11	1/3	běibāo	bībāo
花生	huāshēng	*peanut*	2.27	11	1/2	huáshēng	huīshēng
香蕉	xiāngjiāo	*banana*	2.40	11	1/2	xiángjiāo	xīngjiāo
沙发	shāfā	*sofa*	2.78	11	1/4	shàfā	shīfā
冰箱	bīngxiāng	*refrigerator*	2.74	11	1/3	bǐngxiāng	bēngxiāng
车牌	chēpái	*license plate*	2.38	12	1/4	chèpái	chāpái
钢琴	gāngqín	*piano*	2.50	12	1/3	gǎngqín	gēngqín
公园	gōngyuán	*park*	3.00	12	1/4	gòngyuán	gāngyuán
樱桃	yīngtáo	*cherry*	2.21	12	1/2	yíngtáo	yōngtáo
鲨鱼	shāyú	*shark*	2.31	12	1/2	sháyú	shuāyú
蝙蝠	biānfú	*bat*	2.32	12	1/3	biǎnfú	bānfú
包裹	bāoguǒ	*package*	2.55	13	1/3	bǎoguǒ	bāguǒ
铅笔	qiānbǐ	*pencil*	2.19	13	1/2	qiánbǐ	quānbǐ
天使	tiānshǐ	*angel*	2.88	13	1/2	tiánshǐ	tānshǐ
工厂	gōngchǎng	*factory*	2.67	13	1/4	gòngchǎng	gēngchǎng
黑板	hēibǎn	*chalkboard*	1.77	13	1/4	hèibǎn	hāibǎn
香水	xiāngshuǐ	*perfume*	2.42	13	1/3	xiǎngshuǐ	xīngshuǐ
冰块	bīngkuài	*ice cube*	2.29	14	1/3	bǐngkuài	bāngkuài
雕像	diāoxiàng	*statue*	2.23	14	1/4	diàoxiàng	diūxiàng
鸡蛋	jīdàn	*egg*	2.63	14	1/3	jǐdàn	jūdàn
鞭炮	biānpào	*firecracker*	1.32	14	1/2	biánpào	bīnpào
书架	shūjià	*bookshelf*	1.72	14	1/4	shùjià	shuājià
沙漠	shāmò	*desert*	2.56	14	1/2	shámò	shīmò
钱包	qiánbāo	*wallet*	2.81	21	2/4	qiànbāo	qínbāo
洋葱	yángcōng	*onion*	2.26	21	2/3	yǎngcōng	yíngcōng
黄瓜	huángguā	*cucumber*	1.73	21	2/3	huǎngguā	héngguā
熊猫	xióngmāo	*panda*	1.75	21	2/1	xiōngmāo	xíngmāo
围巾	wéijīn	*scarf*	2.16	21	2/1	wēijīn	wájīn
楼梯	lóutī	*staircase*	2.74	21	2/4	lòutī	láotī
蝴蝶	húdié	*butterfly*	2.25	22	2/1	hūdié	huídié
篮球	lánqiú	*basketball*	2.46	22	2/3	lǎnqiú	lúnqiú
牛排	niúpái	*steak*	2.49	22	2/3	niǔpái	nuópái
长城	chángchéng	*Great Wall*	1.46	22	2/4	chàngchéng	chóngchéng
柠檬	níngméng	*lemon*	2.29	22	2/4	nìngméng	niángméng
足球	zúqiú	*soccer ball*	2.56	22	2/1	zūqiú	zéqiú
苹果	píngguǒ	*apple*	2.65	23	2/1	pīngguǒ	pángguǒ
糖果	tángguǒ	*candy*	2.53	23	2/1	tāngguǒ	tíngguǒ

牙齿	yáchǐ	*teeth*	2.73	23	2/4	yàchǐ	yéchǐ
魔鬼	móguǐ	*devil*	2.79	23	2/4	mòguǐ	múguǐ
牛奶	niúnǎi	*milk*	2.72	23	2/1	niūnǎi	núnǎi
啤酒	píjiǔ	*beer*	3.07	23	2/4	pìjiǔ	pújiǔ
名片	míngpiàn	*business card*	2.51	24	2/4	mìngpiàn	mángpiàn
肥皂	féizào	*soap*	2.38	24	2/1	fēizào	fázào
杂志	zázhì	*magazine*	3.02	24	2/1	zāzhì	zézhì
芹菜	qíncài	*celery*	1.36	24	2/3	qǐncài	qúncài
皮带	pídài	*leather belt*	2.26	24	2/3	pǐdài	púdài
螃蟹	pángxiè	*crab*	1.99	24	2/4	pàngxiè	péngxiè
饼干	bǐnggān	*cracker*	2.77	31	3/4	bìnggān	bǎnggān
火车	huǒchē	*train*	2.80	31	3/2	huóchē	hǒuchē
火鸡	huǒjī	*turkey*	2.33	31	3/1	huōjī	hǔjī
剪刀	jiǎndāo	*scissors*	2.22	31	3/4	jiàndāo	jǐndāo
手机	shǒujī	*cell phone*	3.20	31	3/1	shōujī	shǎojī
纸巾	zhǐjīn	*napkin*	2.18	31	3/2	zhíjīn	zhějīn
彩虹	cǎihóng	*rainbow*	2.10	32	3/1	cāihóng	cǎohóng
草莓	cǎoméi	*strawberry*	2.25	32	3/1	cāoméi	cǎiméi
薯条	shǔtiáo	*French fry*	2.42	32	3/4	shùtiáo	shuǐtiáo
网球	wǎngqiú	*volley ball*	2.27	32	3/2	wángqiú	wěngqiú
警察	jǐngchá	*police officer*	3.44	32	3/2	jíngchá	jiǎngchá
恐龙	kǒnglóng	*dinosaur*	2.09	32	3/4	kònglóng	kǎnglóng
雨伞	yǔsǎn	*umbrella*	1.76	33	3/4	yùsǎn	yísǎn
老虎	lǎohǔ	*tiger*	2.32	33	3/1	lāohǔ	lóuhǔ
老鼠	lǎoshǔ	*mouse*	2.77	33	3/1	lāoshǔ	lóushǔ
手表	shǒubiǎo	*watch*	2.47	33	3/4	shòubiǎo	sháobiǎo
蚂蚁	mǎyǐ	*ant*	2.10	33	3/1	māyǐ	máiyǐ
橄榄	gǎnlǎn	*olive*	2.01	33	3/4	gànlǎn	guánlǎn
眼镜	yǎnjìng	*eyeglasses*	2.70	34	3/4	yànjìng	yǐnjìng
短裤	duǎnkù	*shorts*	2.37	34	3/1	duānkù	dǎnkù
口袋	kǒudài	*pocket*	2.83	34	3/1	kōudài	kǎodài
礼物	lǐwù	*present*	3.26	34	3/2	líwù	lv3wù
土豆	tǔdòu	*potato*	2.39	34	3/2	túdòu	tuǒdòu
领带	lǐngdài	*necktie*	2.51	34	3/4	lìngdài	lǎngdài
信封	xìnfēng	*envelope*	2.35	41	4/2	xínfēng	xiànfēng
衬衫	chènshān	*shirt*	2.75	41	4/2	chénshān	chànshān
蛋糕	dàngāo	*cake*	2.92	41	4/1	dāngāo	dùngāo
面包	miànbāo	*bread*	2.86	41	4/3	miǎnbāo	mànbāo
汽车	qìchē	*car*	3.12	41	4/3	qǐchē	qùchē
电梯	diàntī	*elevator*	2.72	41	4/1	diāntī	dàntī
棒球	bàngqiú	*baseball*	2.71	42	4/1	bāngqiú	bèngqiú
地图	dìtú	*map*	2.76	42	4/1	dītú	dàitú

电池	diànchí	*battery*	2.51	42	4/3	diǎnchí	dànchí
面条	miàntiáo	*noodle*	2.23	42	4/3	miǎntiáo	mèntiáo
教堂	jiàotáng	*church*	2.96	42	4/2	jiáotáng	jiàtáng
太阳	tàiyáng	*sun*	2.89	42	4/2	táiyáng	tìyáng
报纸	bàozhǐ	*newspaper*	2.99	43	4/1	bāozhǐ	bàizhǐ
厕所	cèsuǒ	*toilet*	3.02	43	4/1	cēsuǒ	cìsuǒ
电脑	diànnǎo	*computer*	3.08	43	4/3	diǎnnǎo	dànnǎo
热狗	règǒu	*hotdog*	2.45	43	4/3	rěgǒu	rùgǒu
玉米	yùmǐ	*corn*	2.58	43	4/2	yúmǐ	yòumǐ
汉堡	hànbǎo	*hamburger*	2.59	43	4/2	hánbǎo	hènbǎo
大象	dàxiàng	*elephant*	2.27	44	4/1	dāxiàng	dùxiàng
护照	hùzhào	*passport*	2.44	44	4/2	húzhào	huìzhào
面具	miànjù	*mask*	2.48	44	4/3	miǎnjù	mànjù
项链	xiàngliàn	*necklace*	2.48	44	4/2	xiángliàn	xìngliàn
瀑布	pùbù	*waterfall*	2.00	44	4/3	pǔbù	pàbù
电视	diànshì	*television*	3.33	44	4/1	diānshì	dènshì
FILLERS							
狮子	shīzi	*lion*	2.32	10			
鹦鹉	yīngwǔ	*parrot*	1.96	13			
蜂蜜	fēngmì	*honey*	2.00	14			
吉他	jíta	*guitar*	2.47	20			
猴子	hóuzi	*monkey*	2.62	20			
葡萄	pútao	*grape*	2.19	20			
龙虾	lóngxiā	*lobster*	2.19	21			
轮胎	lúntāi	*tire*	2.42	21			
蘑菇	mógū	*mushroom*	2.12	21			
骨头	gǔtou	*bone*	2.78	30			
椅子	yǐzi	*chair*	2.88	30			
月亮	yuèliang	*moon*	2.40	40			
袜子	wàzi	*sock*	2.53	40			
钥匙	yàoshi	*key*	3.19	40			
气球	qìqiú	*balloon*	2.27	42			
键盘	jiànpán	*keyboard*	1.97	42			

Bibliography

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

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

Amengual, M. (2016). The perception of language-specific phonetic categories does not guarantee accurate phonological representations in the lexicon of early bilinguals. *Applied Psycholinguistics*, *37*(05), 1221–1251. https://doi.org/10.1017/S0142716415000557

Baddeley, A. D. (1968). How does acoustic similarity influence short-term memory. *Quarterly Journal of Experimental Psychology*, *20*(3), 249–264.

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., Jiang, N., & Idsardi, W. J. (2016). Similarity in L2 Phonology: Evidence from L1 Spanish late-learners' perception and lexical representation of English vowel contrasts. *Second Language Research*, *32*(3), 367–395. https://doi.org/10.1177/0267658316630784

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

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious Mixed Models. *ArXiv:1506.04967 [Stat]*. Retrieved from 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

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J. Goodman & H. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167–224). Cambridge, MA: MIT Press. Retrieved from http://www.haskins.yale.edu/SR/SR107/SR107_01.pdf

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). Amsterdam: John Benjamins.

Blicher, D. L., Diehl, R. L., & Cohen, L. B. (1990). Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: Evidence of auditory enhancement. *Journal of Phonetics*, *18*(1), 37–49.

Boersma, P., & Weenink, D. (2010). Praat: Doing phonetics by computer (Version 5.3.51). Retrieved from www.praat.org

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

Braun, B., Galts, T., & Kabak, B. (2014). Lexical encoding of L2 tones: The role of L1 stress, pitch accent and intonation. *Second Language Research*, *30*(3), 323–350.

Braun, B., & Johnson, E. K. (2011). Question or tone 2? How language experience and linguistic function guide pitch processing. *Journal of Phonetics*, *39*, 585–594.

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The Bank of Standardized Stimuli (BOSS), a New Set of 480 Normative Photos of Objects to Be Used as Visual Stimuli in Cognitive Research. *PLoS ONE*, *5*(5), e10773. https://doi.org/10.1371/journal.pone.0010773

Broersma, M. (2012). Increased lexical activation and reduced competition in second language listening. *Language and Cognitive Processes*, *27*(7–8), 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*(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). New York: 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.

Brunellière, A., & Soto-Faraco, S. (2015). The interplay between semantic and phonological constraints during spoken-word comprehension: Semantic and phonological constraints. *Psychophysiology*, *52*(1), 46–58. https://doi.org/10.1111/psyp.12285

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS One*, *5*(6), e10729.

Caldwell-Harris, C. L., Lancaster, A., Ladd, D. R., Dediu, D., & Christiansen, M. H. (2015). Factors influencing sensitivity to lexical tone in an artificial language. *Studies in Second Language Acquisition*, *37*(02), 335–357. https://doi.org/10.1017/S0272263114000849

Carrasco-Ortiz, H., Midgley, K. J., Grainger, J., & Holcomb, P. J. (2017). Interactions in the neighborhood: Effects of orthographic and phonological neighbors on N400 amplitude. *Journal of Neurolinguistics*, *41*, 1–10. https://doi.org/10.1016/j.jneuroling.2016.06.007

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

Chao, Y. R. (1968). *Grammar of Spoken Chinese*. Berkeley, CA: Unversity of California Press.

Chen, Y., & Xu, Y. (2006). Production of weak elements in speech: evidence from F0 patterns of neutral tone in standard Chinese. *Phonetica*, *63*(1), 47–75. https://doi.org/10.1159/000091406

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

Connolly, J. F., & Phillips, N. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Cognitive Neuroscience, Journal Of*, *6*(3), 256–266.

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

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.

Coulson, S., King, J. W., & Kutas, M. (1998). ERPs and domain specificity: beating a straw horse. *Language and Cognitive Processes*, *13*(6), 653–672.

Cutler, A. (2012). *Native Listening: Language Experience and the Recognition of Spoken Words*. Cambridge, MA: The MIT Press.

Cutler, A., Weber, A., & Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics*, *34*(2), 269–284. https://doi.org/10.1016/j.wocn.2005.06.002

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America*, *116*(6), 3668. https://doi.org/10.1121/1.1810292

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

Darcy, I., Dekydtspotter, L., Sprouse, R. A., Glover, J., Kaden, C., McGuire, M., & Scott, J. H. (2012). Direct mapping of acoustics to phonology: On the lexical encoding of front rounded vowels in L1 English- L2 French acquisition. *Second Language Research*, *28*(1), 5–40. https://doi.org/10.1177/0267658311423455

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). Malden, MA: 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

Desroches, A. S., Newman, R. L., & Joanisse, M. F. (2009). Investigating the time course of spoken word recognition: Electrophysiological evidence for the influences of phonological similarity. *Journal of Cognitive Neuroscience*, *21*(10), 1893–1906.

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

Duanmu, S. (2007). *The Phonology of Standard Chinese* (2nd edition). New York, New York: Oxford University Press.

Escudero, P., Hayes-Harb, R., & Mitterer, H. (2008). Novel second-language words and asymmetric lexical access. *Journal of Phonetics*, *36*(2), 345–360. https://doi.org/10.1016/j.wocn.2007.11.002

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research* (pp. 233–277). Timonium, MD: York.

Fox, R. A., & Qi, Y.-Y. (1990). CONTEXT EFFECTS IN THE PERCEPTION OF LEXICAL TONE / 语境对词调感知的影响. *Journal of Chinese Linguistics*, *18*(2), 261–284.

Fox, R. A., & Unkefer, J. (1985). THE EFFECT OF LEXICAL STATUS ON THE PERCEPTION OF TONE / 成词状况对声调感知的影 响. *Journal of Chinese Linguistics*, *13*(1), 69–90.

Friedrich, C. K., Eulitz, C., & Lahiri, A. (2006). Not every pseudoword disrupts word recognition: an ERP study. *Behavioral and Brain Functions*, *2*(1), 1.

Gandour, J. T. (1983). Tone perception in Far Eastern languages. *Journal of Phonetics*, *11*, 149–175.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110.

Gårding, E., Kratochvil, P., Svantesson, J.-O., & Zhang, J. (1986). Tone 4 and Tone 3 Discrimination in Modern Standard Chinese. *Language and Speech*, *29*(3), 281–293. https://doi.org/10.1177/002383098602900307

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*(20), 8051–8056. https://doi.org/10.1073/pnas.1216438110

Gibson, E., Tan, C., Futrell, R., Mahowald, K., Konieczny, L., Hemforth, B., & Fedorenko, E. (2017). Don't Underestimate the Benefits of Being Misunderstood. *Psychological Science*, *28*(6), 703–712. https://doi.org/10.1177/0956797617690277

Goldsmith, J. A. (1990). *Autosegmental and metrical phonology*. Blackwell.

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). Amsterdam: John Benjamins.

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and Cognitive Processes*, *25*(2), 149–188. https://doi.org/10.1080/01690960902965951

Hallé, P. A., Chang, Y.-C., & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of Phonetics*, *32*, 395–421.

Hao, Y.-C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, *40*(2), 269–279. https://doi.org/10.1016/j.wocn.2011.11.001

Hao, Y.-C. (2018). Contextual effect in second language perception and production of Mandarin tones. *Speech Communication*, *97*, 32–42. https://doi.org/10.1016/j.specom.2017.12.015

Holcomb, P. J., & Neville, H. J. (1990). Auditory and visual semantic priming in lexical decision: A comparison using event-related brain potentials. *Language and Cognitive Processes*, *5*(4), 281–312.

Howie, J. (1976). *Acoustical studies of Mandarin vowels and tones*. Cambridge, UK: Cambridge University Press.

Huang, J., & Holt, L. L. (2009). General perceptual contributions to lexical tone normalization. *The Journal of the Acoustical Society of America*, *125*(6), 3983–3994. https://doi.org/10.1121/1.3125342

Huang, T., & Johnson, K. (2010). Language specificity in speech perception: perception of Mandarin tones by native and nonnative listeners. *Phonetica*, *67*(4), 243–267. https://doi.org/10.1159/000327392

Huang, X., & Yang, J.-C. (2016). The time course of lexical competition during spoken word recognition in Mandarin Chinese: an event-related potential study. *NeuroReport*, *27*(2), 67–72. https://doi.org/10.1097/WNR.0000000000000492

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)

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

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

Kaan, E., & Swaab, T. (2003). Repair, revision, and complexity in syntactic analysis: An electrophysiological differentiation. *Cognitive Neuroscience, Journal Of*, *15*(1), 98–110.

Kazanina, N., Bowers, J. S., & Idsardi, W. (2017). Phonemes: Lexical access and beyond. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-017-1362-0

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203. https://doi.org/10.1037/a0038695

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.

Ladd, D. R., & Morton, R. (1997). The perception of intonational emphasis: continuous or categorical? *Journal of Phonetics*, *25*(3), 313–342. https://doi.org/10.1006/jpho.1997.0046

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

Leather, J. (1987). F0 pattern inference in the perceptual acquisition of second language tone. In J. Leather & A. James (Eds.), *Sound Patterns in Second Langauge Acquisition* (pp. 59–80). Providence: Foris.

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, C.-Y., Tao, L., & Bond, Z. S. (2010). Identification of acoustically modified Mandarin tones by non-native listeners. *Language and Speech*, *53*(2), 217–243. https://doi.org/10.1177/0023830909357160

Lee, C.-Y., Tao, L., & Bond, Z. S. (2013). Effects of speaker variability and noise on Mandarin tone identification by native and non-native listeners. *Speech, Language and Hearing*, *16*(1), 46–54. https://doi.org/10.1179/2050571X12Z.0000000003

Lee, L., & Nusbaum, H. C. (1993). Processing interactions between segmental and suprasegmental information in native speakers of English and Mandarin Chinese. *Perception & Psychophysics*, *53*(2), 157–165.

Lee, W.-S., & Zee, E. (2008). Prosodic characteristics of the neutral tone in Beijing Mandarin/北京话轻声的韵律特征. *Journal of Chinese Linguistics*, *36*(1), 1–29.

Lee, W.-S., & Zee, E. (2014). Chinese phonetics. In C.-T. J. Huang, Y.-H. A. Li, & A. Simpson (Eds.), *The handbook of Chinese linguistics* (pp. 369–399). Malden, MA: Wiley Blackwell.

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, C., Wang, M., & Idsardi, W. (2015). The effect of orthographic form-cuing on the phonological preparation unit in spoken word production. *Memory & Cognition*, *43*(4), 563–578. https://doi.org/10.3758/s13421-014-0484-0

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.

Lin, W. C. J. (1985). Teaching Mandarin Tones to Adult English Speakers: Analysis of Difficulties with Suggested Remedies. *RELC Journal*, *16*(2), 31–47.

Ling, W., Schafer, A., & Grüter, T. (2016, September). *Identification and discrimination of tone by L2 learners of Mandarin*. Presentation presented at the The 35th annual Second Language Research Forum (SLRF), Columbia Teacher College, New York.

Lisker, L., & Abramson, A. S. (1964). A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements. WORD, *20*(3), 384–422. https://doi.org/10.1080/00437956.1964.11659830

Liu, L., & Kager, R. (2014). Perception of tones by infants learning a non-tone language. *Cognition*, *133*(2), 385–394. https://doi.org/10.1016/j.cognition.2014.06.004

Liu, Yanni, Shu, H., & Wei, J. (2006). Spoken word recognition in context: Evidence from Chinese ERP analyses. *Brain and Language*, *96*(1), 37–48. https://doi.org/10.1016/j.bandl.2005.08.007

Liu, Yuehua, Yao, T., Bi, N.-P., Ge, L., & Shi, Y. (2008). *Integrated Chinese, Level 1, Part 1* (Third edition). Boston, MA: Cheng & Tsui.

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

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Lukianchenko, A. (2014). *From Sound to Meaning: Quantifying Contextual Effects in Resolution of L2 Phonolexical Ambiguity*. University of Maryland, College Park.

MacWhinney, B., & Bates, E. (Eds.). (1989). *The Cross-linguistic Study of Sentence Processing*. New York, New York: 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

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1–2), 71–102. https://doi.org/10.1016/0010-0277(87)90005-9

McClelland, J. L., & Elman, J. L. (1986). The TRACE Model of Speech Perception. *Cognitive Psychology*, *18*, 1–86.

McLaughlin, J., Osterhout, L., & Kim, A. (2004). Neural correlates of second-language word learning: minimal instruction produces rapid change. *Nature Neuroscience*, *7*(7), 703–704.

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *The Journal of the Acoustical Society of America*, *102*(3), 1864–1877. https://doi.org/10.1121/1.420092

Moreno-Martínez, F. J., & Montoro, P. R. (2012). An Ecological Alternative to Snodgrass & Vanderwart: 360 High Quality Colour Images with Norms for Seven Psycholinguistic Variables. *PLoS ONE*, *7*(5), e37527. https://doi.org/10.1371/journal.pone.0037527

Newman, R. L., & Connolly, J. F. (2009). Electrophysiological markers of pre-lexical speech processing: Evidence for bottom–up and top–down effects on spoken

word processing. *Biological Psychology*, *80*(1), 114–121. https://doi.org/10.1016/j.biopsycho.2008.04.008

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*, 189–234.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395. https://doi.org/10.1037/0033-295X.115.2.357

Ong, J. H., Burnham, D., & Escudero, P. (2015). Distributional learning of lexical tones: A comparison of attended vs. unattended listening. *PloS One*, *10*(7). Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4516233/

O'Seaghdha, P. G., Chen, J.-Y., & Chen, T.-M. (2010). Proximate units in word production: Phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition*, *115*(2), 282–302. https://doi.org/10.1016/j.cognition.2010.01.001

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elcited by syntactic anomaly. *Journal of Memory and Language*, *31*, 785–806.

Ota, M., Hartsuiker, R. J., & Haywood, S. L. (2009). The KEY to the ROCK: Near-homophony in nonnative visual word recognition. *Cognition*, *111*(2), 263–269. https://doi.org/10.1016/j.cognition.2008.12.007

Pajak, B., Fine, A. B., Kleinschmidt, D. F., & Jaeger, T. F. (2016). Learning Additional Languages as Hierarchical Probabilistic Inference: Insights From First Language Processing: Learning Languages as Hierarchical Inference. *Language Learning*, *66*(4), 900–944. https://doi.org/10.1111/lang.12168

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*(6), 445–449. https://doi.org/10.1111/1467-9280.00383

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Pelzl, E., Lau, E. F., Guo, T., & DeKeyser, R. (2018). Advanced second language learners' perception of lexical tone contrasts. *Studies in Second Language Acquisition*, 1–28. https://doi.org/10.1017/S0272263117000444

Pelzl, E., Vafaee, P., Chrabaszcz, A., Cook, S., Gor, K., Jackson, S. R., … Zhou, Q. (2014). Linguistic Correlates of Proficiency. Presented at the East Coast Organization of Language Testers (ECOLT), Columbia Teacher College, New York.

Perfetti, C. (2007). Reading Ability: Lexical Quality to Comprehension. *Scientific Studies of Reading*, *11*(4), 357–383. https://doi.org/10.1080/10888430701530730

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.

Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition*, 1–27. https://doi.org/10.1017/S0272263117000407

Pierrehumbert, J. B. (2016). Phonological Representation: Beyond Abstract Versus Episodic. *Annual Review of Linguistics*, *2*(1). https://doi.org/10.1146/annurev-linguist-030514-125050

Pleco Chinese Dictionary for iOS. (2018). (Version 3.2.30).

Potter, C. E., Wang, T., & Saffran, J. R. (2016). Second Language Experience Facilitates Statistical Learning of Novel Linguistic Materials. *Cognitive Science*. https://doi.org/10.1111/cogs.12473

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/

Repp, B. H., & Lin, H.-B. (1990). Integration of segmental and tonal information in speech perception: a cross-linguistic study. *Journal of Phonetics*, *18*, 481–495.

Romero-Rivas, C., Martin, C. D., & Costa, A. (2015). Processing changes when listening to foreign-accented speech. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00167

Sassenhagen, J., & Bornkessel-Schlesewsky, I. (2015). The P600 as a correlate of ventral attention network reorientation. *Cortex*, *66*, A3–A20. https://doi.org/10.1016/j.cortex.2014.12.019

Sassenhagen, J., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2014). The P600-as-P3 hypothesis revisited: Single-trial analyses reveal that the late EEG positivity following linguistically deviant material is reaction time aligned. *Brain and Language*, *137*, 29–39. https://doi.org/10.1016/j.bandl.2014.07.010

Schaefer, V., & Darcy, I. (2014). Lexical function of pitch in the first language shapes cross-linguistic perception of Thai tones. *Laboratory Phonology*, *5*(4). https://doi.org/10.1515/lp-2014-0016

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.

Schmidt-Kassow, M., & Kotz, S. A. (2009). Event-related brain potentials suggest a late interaction of meter and syntax in the P600. *Journal of Cognitive Neuroscience*, *21*(9), 1693–1708.

Sebastián-Gallés, N., & Díaz, B. (2012). First and second language speech perception: Graded learning. *Language Learning*, *62*(s2), 131–147.

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

Shen, G., & Froud, K. (2018). Electrophysiological correlates of categorical perception of lexical tones by English learners of Mandarin Chinese: an ERP study. *Bilingualism: Language and Cognition*, 1–13. https://doi.org/10.1017/S136672891800038X

Shen, H. H. (2009). Size and Strength: Written Vocabulary Acquisition among Advanced Learners. *Shijie Hanyu Jiaoxue (Chinese Teaching in the World)*, *23*(1), 74–85.

Shen, J., Deutsch, D., & Rayner, K. (2013). On-line perception of Mandarin Tones 2 and 3: Evidence from eye movements. *The Journal of the Acoustical Society of America*, *133*(5), 3016–3029. https://doi.org/10.1121/1.4795775

Shen, X. S., & Lin, M. (1991). A perceptual study of Mandarin tones 2 and 3. *Language and Speech*, *34*(2), 145–156.

Shi, F. (2009). *Shiyan Yinxixue Tansuo (Exploration of Experimental Phonology)*. Beijing, China: Peking University Press.

Shi, J. (2007). On teaching tone three in Mandarin. *Journal of the Chinese Language Teachers Association*, *42*(2), 1–10.

Shi, R., Santos, E., Gao, J., & Li, A. (2017). Perception of Similar and Dissimilar Lexical Tones by Non-Tone-Learning Infants. *Infancy*, *22*(6), 790–800. https://doi.org/10.1111/infa.12191

Shuai, L., & Malins, J. G. (2017). Encoding lexical tones in jTRACE: a simulation of monosyllabic spoken word recognition in Mandarin Chinese. *Behavior Research Methods*, *49*(1), 230–241. https://doi.org/10.3758/s13428-015-0690-0

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2017). Afex: analysis of factorial experiments (Version R package version 0.17-8). Retrieved from http://cran.r-project.org/package=afex

Sjolander, K. (2000). WaveSurfer (Version 1.8.8p5-1701261420). Retrieved from https://sourceforge.net/projects/wavesurfer/

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.

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*(02), 195–221. https://doi.org/10.1017/S0272263114000047

Sparvoli, C. (2017). From phonological studies to teaching Mandarin tone: some perspectives on the revision of the tonal inventory. In I. Kecskes & Sun, Chaofen (Eds.), *Key Issues in Chinese as a Second Language Research* (pp. 81–100). New York, NY: Routledge.

Speer, S. R., Shih, C.-L., & Slowiaczek, M. L. (1989). Prosodic structure in language understanding: evidence from tone sandhi in Mandarin. *Language and Speech*, *32*(4), 337–354.

Stagray, J. R., & Downs, D. (1993). Differential sensitivity for frequency among speakers of a tone and nontone language. *Journal of Chinese Linguistics*, *21*, 143–163.

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

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.

Suzuki, Y., & DeKeyser, R. (2017). The Interface of Explicit and Implicit Knowledge in a Second Language: Insights From Individual Differences in Cognitive Aptitudes: Interface of Explicit and Implicit Knowledge. *Language Learning*, *67*(4), 747–790. https://doi.org/10.1111/lang.12241

Van Den Brink, D., Brown, C., & Hagoort, P. (2001). Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 versus N400 effects. *Cognitive Neuroscience, Journal Of*, *13*(7), 967–985.

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

Wagner, M., & Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, *25*(7–9), 905–945. https://doi.org/10.1080/01690961003589492

Wang, T., & Saffran, J. R. (2014). Statistical learning of a tonal language: the influence of bilingualism and previous linguistic experience. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00953

Wang, Y., Jongman, A., & Sereno, J. A. (2001). Dichotic Perception of Mandarin Tones by Chinese and American Listeners. *Brain and Language*, *78*(3), 332–348. https://doi.org/10.1006/brln.2001.2474

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.

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, *50*(1), 1–25. https://doi.org/10.1016/S0749-596X(03)00105-0

Wiener, S., & Ito, K. (2015). Do syllable-specific tonal probabilities guide lexical access? Evidence from Mandarin, Shanghai and Cantonese speakers. *Language, Cognition and Neuroscience*, *30*(9), 1048–1060. https://doi.org/10.1080/23273798.2014.946934

Wiener, S., & Ito, K. (2016). Impoverished acoustic input triggers probability-based tone processing in mono-dialectal Mandarin listeners. *Journal of Phonetics*, *56*, 38–51. https://doi.org/10.1016/j.wocn.2016.02.001

Wiener, S., Ito, K., & Speer, S. R. (2016a). Perception and production of newly learned words in an L2: A distributional learning account. Presented at the Linguistic Society of America, Washingtong, D.C.

Wiener, S., Ito, K., & Speer, S. R. (2016b, September). *Speaker variability and explicit awareness of novel speech cues modulate word recognition in an L2*. Presentation presented at the The 35th annual Second Language Research Forum (SLRF), Columbia Teacher College, New York.

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

Winsler, K., Midgley, K. J., Grainger, J., & Holcomb, P. J. (2018). An electrophysiological megastudy of spoken word recognition. *Language, Cognition and Neuroscience*, 1–20. https://doi.org/10.1080/23273798.2018.1455985

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.

Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience*, *10*, 420–422. https://doi.org/10.1038/nn1872

Xu, Y. (1994). Production and perception of coarticulated tones. *The Journal of the Acoustical Society of America*, *95*(4), 2240–2253.

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

Yang, B. (2012). The gap between the perception and production of tones by American learners of Mandarin – An intralingual perspective. *Chinese as a Second Language Research*, *1*(1), 33–53. https://doi.org/10.1515/caslar-2012-0003

Yang, R. (2015). The role of phonation cues in Mandarin tonal perception. *Journal of Chinese Linguistics*, *43*(1B), 453–472. https://doi.org/10.1353/jcl.2015.0035

Zhang, H. (2014). The Third Tone: Allophones, Sandhi Rules and Pedagogy. *Journal of the Chinese Language Teachers Association*, *49*(1), 117–145.

Zhang, J., & Lai, Y. (2010). Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology*, *27*(01), 153–201. https://doi.org/10.1017/S0952675710000060

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

Zhao, T. C., & Kuhl, P. K. (2015). Effect of musical experience on learning lexical tone categories. *The Journal of the Acoustical Society of America*, *137*(3), 1452–1463.

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