**A1. Details of screening measures**

In order to ensure that the label of ‘advanced’ was reasonably applied to L2 participants, all potential L2 participants first completed a set of three online screening tests. The first test was a *Can-do* self-assessment of functional proficiency. Participants read 25 sentences describing a range of tasks, from simple (“I can use Chinese to bargain for items in a tourist shop.”) to more complex (“I can use Chinese to explain the plot of a movie in detail with appropriate and specific vocabulary.”). Sentences described tasks that learners were likely to be familiar with in order to maximize the reliability of their self-assessments. For each sentence they gave a rating from 1 to 5, with the following explanation of the rating scale (modified from Brown, Dewey, & Cox, 2014):

*1= I could not do this, even with extensive preparation.*

*2= I am unsure as to whether I could or could not do this.*

*3= I could do this with extensive preparation (for example, with up to an hour of preparation or extensive use of a dictionary).*

*4= I could do this with minimal preparation (for example, with a few minutes preparation or looking up a few words in a dictionary).*

*5= I could do this without any preparation.*

The total of the rating scale answers for the 25 items was used as a *Can-do* score for each test taker. A very small pilot of five L2 learners known to the author was carried out prior to the current study. Pilot participants were known to be either below the ideal proficiency for experimental materials, or of a very high proficiency such that the materials would be appropriate. Based on their scores, a cut-off of approximately 100 seemed reasonable for the *Can-do* test in the current study, but to avoid being too restrictive, a cut-score of 90 was used for determining eligibility. In total, 46 individuals completed the online Can-do self-assessment; 33 of them achieved a score of 90 or higher (*m*= 97.4). The Can-do assessment (n=46) had a Rasch item reliability of .97.

A second self-assessment of written vocabulary knowledge (VSA) was also completed online. In this test, participants saw a Chinese word presented on the screen with the question, “Do you know this word?” and responded by pressing ‘Y’ for *yes* or ‘N’ for *no*. The test consisted of 105 Chinese words drawn from the SUBTLEX-CH corpus (Cai & Brysbaert, 2010). To ensure a reasonable spread of difficulty, 15 words were sampled from each 1000-word group between 1-1000, 1001-2000, and so on, up to 7000. This measure was based on an earlier pilot study the first author carried out with 48 L2 participants with widely varying levels of experience studying Chinese or living in a Mandarin-speaking country. Based on the earlier results, a cut-score of 70% or above was used for establishing eligibility. 27 of 46 participants completed this task with scores above 70% (*m* =87.8%). The VSA (n= 94) had a Rasch item reliability of .95.

Originally, an auditory listening comprehension test was intended to serve as a third screening measure, but due to technical difficulties encountered in online delivery in China, only about half of the participants were able to complete the test. For a few early participants, eligibility was established on the basis of meeting the criterion score on two of three tests, but due to the difficulties faced in completing the listening comprehension test, after the first few participants, results of this test were no longer considered for the purposes of establishing eligibility.

Participants who met criterion scores in both the *Can-do* and vocabulary self-assessments were invited to come to the laboratory to participate in the full study. This was a total of 24 individuals, 14 of whom eventually completed the full study. Additionally, one of the earliest participants was accepted despite having a *Can-do* score of 84. In this case eligibility was determined based on the strength of her listening comprehension score. An additional early participant was accepted with a low vocabulary score of 60.1. This participant was one of the first to encounter technical difficulties with the listening comprehension test. Because character recognition ability was a minor focus of the study, and based on the participant’s strong language background, with work and study both requiring immersion in Chinese language, the participant was deemed eligible.

**A2. Details of stimuli creation**

To construct the sentence stimuli, a list of 270 high frequency disyllabic words likely to be familiar to all advanced learners of Mandarin was created. Words were selected due to frequency and/or occurrence in commonly used beginning textbooks for Chinese (e.g., *Integrated Chinese,* Liu, Yao, Bi, Ge, & Shi, 2008). Words with homophones were avoided (this was another reason to use disyllabic words). For each word, two nonwords were also created. As in the LDT, nonwords differed from real words in either the tone or rhyme of the first syllable*.* For each real word, a constraining sentence was created such that the context prior to the target word would create a high expectation for seeing that word. Target words occurred at variable locations across sentences such that, apart from the fact that all target words are nouns, there are no reliable cues as to when the target word will occur in a given sentence. In total 270 sentences were created by the author, a non-native speaker of Mandarin, using vocabulary and structures considered likely to be comprehensible for intermediate and advanced L2 speakers. All sentences were examined by a native speaker for naturalness and grammaticality, and final versions were agreed upon over several rounds of editing and adaptation.

To ensure that the sentences were reasonably constraining, sentence stems consisting of the sentences up to, but not containing, the target word were presented to native speakers in three sets to elicit cloze probabilities for the target words (1 item was accidentally omitted). Of the original 270 sentences, 30 sentences were removed either because they were judged to be potentially difficult for L2 listeners, or had target words with very low cloze probabilities. This resulted in a set of 240 sentences with an average cloze probably of 60%.

Next, the semantic mismatch condition was created by randomly assigning each expected condition word to an additional mismatched sentence.[[1]](#footnote-1) Each item was examined to verify that the sentence-word pairings were semantically odd.

Using Praat (Boersma & Weenink, 2010), all 240 sentences were recorded by a native speaker of Mandarin from Beijing with a standard Chinese accent. The speaker was seated in a sound-attenuated booth and recordings were occasionally monitored to make sure that the speaker was not speaking either too fast for L2 listeners or unnaturally slowly.

When selecting target words, those with liquid, nasal, or vowel sounds as the onsets were avoided to minimize coarticulation effects in audio recording of sentences. To further avoid coarticulation effects, each sentence was recorded first with a blank (pause) in the position of the target word, then again with each of the four conditions’ target words, resulting in a sentence frame and four sentence conditions. Sentence frames were each recorded three times, and sentence conditions were recorded once unless the speaker felt the target had been mispronounced, in which case a second (or third) take was also recorded. When editing, target words for each condition were extracted from their sentence contexts and pasted into the sentence frames. This allowed for control of the auditory context across all four conditions leading up to and following critical words.

The decision to record sentences with blanks resulted in lengthening effects prior to the target words. Overall, the lengthening effects were quite subtle, but in the case of some highly frequent function words (e.g., the particle *de*), lengthening effects were felt to be highly noticeable. In these cases, the function words were replaced with tokens of those words from other unrelated sentences, thus removing the lengthening effects and creating very natural sounding sentences. Though the lengthening effects are certainly less than ideal, some type of coarticulation artifact in this type of stimuli is perhaps unavoidable.[[2]](#footnote-2)

To avoid effects of T3 sandhi in the LDT and SJT, we took the following steps. First, no critical word stimuli had two T3 syllables (i.e., T3-T3). Second, for tonal nonwords, tone manipulations avoided T3 if it would result in sandhi (e.g., diányǐng, not diǎnyǐng). Finally, for sentences, T3 was also avoided immediately before critical words/nonwords that began with T3.

For the current ERP experiment, 120 individual sentences were used, each one occurring with four target word conditions. In other words, there were a total of 480 individual sentences, of which only 120 were heard by each participant, with target word conditions balanced across sentences. The target words from the sentences were also used in isolation for the LDT in Experiment 2. Additionally, stimuli for the Tone ID were created by extracting syllables from 60 of the target words. Finally, filler sentences were also created using the expected word versions of 60 other sentences. The filler items were not manipulated in any fashion, that is, they were single takes of complete and normal sentences without any splicing of critical words.

A unique design aspect of the present study is that target words from the ERP stimuli were also used as stimuli in the LDT, and served as the raw material for the Tone ID stimuli. One potential concern with this approach is that participants could be exposed to the same items repeatedly across experiments. To control for this in the LDT, target items used in the ERP experiment were balanced across four presentation lists such that, if a participant heard expected condition words for items 1-30 in the ERP experiment, that participant would hear the semantic condition words for items 1-30 in the LDT. Other conditions were similarly switched across experiments so that no individual heard target words in the same condition in both the ERP and LDT experiments.

**A3. Explicit vocabulary knowledge and correlational analysis**

Table A3.1 summarizes individual scores for all participants for the EVKT and the three experiments. In general, it can be seen that participants with strong scores for tones and definitions on the EVKT seem to be generally strong on all tests—with notable inconsistency for tonal nonwords and mismatches. Participants A and B (and to a lesser extent M) were the strongest for tonal nonwords and mismatches, and also showed consistently strong performance across all tasks, suggesting, unsurprisingly, that increasing proficiency leads to increasingly strong outcomes for tones. At the same time, the weakest participants (O and P) were also consistently weak.

In order to constrain interpretation of the individual results, we conducted an exploratory correlational analysis (Table A3.2) between all test outcomes, using the false discovery rate to adjust p-values (Benjamini & Hochberg, 1995). The strongest tone-related result was for the relationship between accuracy for tone knowledge in the EVKT and overall performance on the Tone ID (*r*=.81, *p*<.01), suggesting a strong relationship between explicit knowledge of tones in words and the accurate perception of tones. Additionally, accuracy in rejection of tonal nonwords in the LDT and SJT was significantly correlated (*r*=.68, *p*=.03). Perhaps the most interesting correlation was between accuracy in rejection of tonal nonwords in the LDT and of semantically mismatching words in the SJT (*r*=.64, *p*=.04). This relationship might suggest that general comprehension abilities are tied to phonological tone perception. Though the direction of this relationship is unclear, it potentially lends a bit of urgency to the need for L2 learners to master lexical tones. Beyond these correlations, results are mainly informative in what they do not find, namely relationships between explicit knowledge of tones, performance on the Tone ID, and performance in the tone conditions of the LDT and SJT. In other words, while Table 10 suggests that there are broad relationships between overall proficiency and outcomes across most tests, the results for tonal nonwords and mismatches do not seem to tie strongly to overall proficiency. We acknowledge that, due to the small sample in the current study, these trends must be interpreted with caution. Further research along these lines will be needed to provide more confidence in interpretations. A larger L2 sample might allow for power to find smaller effects that went undetected here.

|  |
| --- |
| **Table A3.1. Individual scores for L2 participants across tests and experimental tasks** |
|  | **EVKT** | **Tone ID** | **LDT** | **SJT** |
| Participant ID |  | overall tone accuracy | definitions | composite score | real words | segmental nonwords | tonal nonwords | expected |  semantic mismatch | segmental mismatch | tonal mismatch |
| A |  | **0.98** | **1.00** | **0.97** | **1.00** | **0.93** | **0.80** | **0.87** | **0.93** | **0.97** | **0.80** |
| B |  | **0.95** | **1.00** | **0.88** | **0.97** | **0.93** | **0.70** | 0.70 | **0.90** | **0.97** | **0.50** |
| C |  | **0.93** | **1.00** | **0.92** | **0.98** | **0.87** | 0.27 | **0.77** | **0.87** | **0.87** | **0.43** |
| D |  | **0.90** | **1.00** | **0.88** | **0.92** | **0.97** | **0.43** | **0.73** | **0.83** | **0.90** | 0.33 |
| E |  | **0.88** | **1.00** | **0.82** | 0.88 | **1.00** | **0.37** | **0.77** | **0.97** | **0.90** | **0.47** |
| F |  | **0.88** | **0.98** | 0.73 | 0.85 | **0.93** | 0.33 | **0.73** | 0.67 | **0.80** | 0.07 |
| G |  | 0.82 | **0.98** | **0.82** | 0.88 | **0.90** | **0.57** | 0.47 | **0.83** | **0.90** | **0.53** |
| H |  | 0.81 | **0.98** | **0.83** | 0.85 | **0.87** | **0.37** | 0.50 | 0.70 | 0.70 | 0.33 |
| I |  | **0.92** | **0.97** | **0.85** | **1.00** | **0.90** | 0.07 | 0.53 | **0.83** | **0.93** | **0.43** |
| J |  | 0.81 | **0.97** | 0.80 | **0.95** | 0.77 | **0.37** | **0.93** | 0.53 | 0.57 | 0.37 |
| K |  | 0.54 | **0.97** | 0.55 | **0.92** | 0.60 | 0.10 | **0.80** | 0.67 | 0.70 | 0.27 |
| L |  | **0.96** | **0.95** | **0.83** | **0.92** | **0.93** | 0.23 | **0.87** | 0.50 | 0.43 | 0.20 |
| M |  | **0.89** | **0.95** | **0.83** | 0.83 | **0.93** | **0.60** | 0.70 | **0.83** | 0.67 | **0.63** |
| N |  | **0.87** | 0.88 | **0.87** | **0.92** | 0.57 | 0.10 | **0.80** | 0.43 | 0.53 | 0.20 |
| O |  | 0.73 | 0.82 | **0.87** | 0.85 | 0.70 | 0.30 | 0.57 | 0.60 | 0.63 | **0.57** |
| P |  | 0.72 | 0.78 | 0.72 | 0.85 | 0.70 | 0.03 | 0.70 | 0.40 | 0.67 | 0.30 |
|  *L2 AVG* | *0.85* | *0.95* | *0.82* | *0.91* | *0.84* | *0.35* | *0.71* | *0.72* | *0.76* | *0.40* |
| *L1 AVG* |  |  | *0.87* | *0.94* | *0.96* | *0.91* | *0.92* | *0.91* | *0.95* | *0.84* |
| Individual L2 participant results across the study, including scores on the EVKT, Tone ID (composite), LDT, SJT. L2 and L1 average scores are listed at the bottom. Scores that are at or above the L2 average are bolded. Scores that are at or above the L1 average are highlighted in gray. |

|  |
| --- |
| **Table A3.2. Correlations between EVKT, Tone ID, LDT, and SJT scores** |
|  | EVKT:def | Tone ID | LDT:Real | LDT:Seg | LDT:Tone | SJT:Exp | SJT:Sem | SJT:Seg | SJT:Tone |
| EVKT: tone | 0.44 | **0.81** | 0.41 | **0.71** | 0.47 | 0.14 | 0.42 | 0.31 | 0.23 |
| EVKT: def |  | 0.19 | 0.44 | **0.64** | 0.51 | 0.13 | **0.73** | 0.54 | 0.13 |
| Tone ID |  |  | 0.38 | 0.44 | 0.50 | -0.04 | 0.38 | 0.29 | 0.53 |
| LDT: Real |  |  |  | 0.10 | 0.11 | 0.35 | 0.31 | 0.38 | 0.22 |
| LDT: Seg |  |  |  |  | 0.59 | -0.11 | **0.70** | 0.55 | 0.29 |
| LDT: Tone |  |  |  |  |  | 0.02 | **0.64** | 0.48 | **0.68** |
| SJT: Exp |  |  |  |  |  |  | -0.20 | -0.30 | -0.15 |
| SJT: Sem |  |  |  |  |  |  |  | **0.85** | 0.60 |
| SJT: Seg |  |  |  |  |  |  |  |  | 0.46 |
| Pearson correlations between all tests. **Bold values** indicate significant *p-values* (>.05). All p-values adjusted using *false discovery rate*. |

**A4. D-prime scores for LDT and SJT**

Table A4.1 reports d’ scores for the LDT and SJT. In calculating d’, scores for individuals who performed with either perfect accuracy (*d’*=1) or complete inaccuracy (*d’*=0) were replaced using Laplace smoothing (cf. Barrios, Namyst, Lau, Feldman, & Idsardi, 2016).

|  |
| --- |
| **Table A4.1. D’ scores for L1 and L2 groups in the LDT and SJT** |
| *Experiment* | *Group* | *Condition* | *d’ avg.* | *d’ std. dev* |
| LDT | L1 | Segmental Nonword | 3.35 | 0.57 |
|  |  | Tonal Nonword | 3.03 | 0.55 |
|  | L2 | Segmental Nonword | 2.64 | 0.83 |
|  |  | Tonal Nonword | 1.19 | 1.11 |
| SJT | L1 | Semantic Mismatch | 2.95 | 0.58 |
|  |  | Rhyme Mismatch | 3.18 | 0.59 |
|  |  | Tone Mismatch | 2.58 | 0.56 |
|  | L2 | Semantic Mismatch | 1.47 | 0.67 |
|  |  | Rhyme Mismatch | 1.31 | 0.70 |
|  |  | Tone Mismatch | 0.34 | 0.68 |

**A5. Analysis for ERP results incorporating correct/incorrect behavioral response as a fixed effect**

In this supplementary analysis we re-examine the ERP data presented in Experiment 3 by including the accuracy of behavioral responses as a fixed effect in our models. Though the new models do not substantively change results or our interpretations, they may still be of some interest for readers looking to dig deeper into our data.

This alternative analysis comes with a major caveat. Due to the high inaccuracy of L2 behavioral responses overall (about 30% incorrect) and especially in the critical tone mismatch condition (around 60% incorrect), consideration of only correct L2 trials, whether separately or as a factor in the model, leaves the models very underpowered.

We also wish to stress that the original intent of the SJT was only to encourage participants to pay attention. It was not meant to measure responses to critical words specifically. In the case of sentence rejections, we can have no certainty that the critical manipulated word drove the observed behavioral response. For sentence acceptance, particularly in the case of L2 listeners, a listener may well have noticed the mismatching critical word, but nevertheless been reluctant to reject the sentence due to some type of uncertainty (e.g., doubting their vocabulary knowledge, unsure what they heard). For these reasons, the behavioral responses relate to ERP responses in a rather loose fashion. This is a further reason we did not originally pursue this more complex analysis.

*Analysis and Results*

To account for a possible relationship between behavioral and neural responses, we examined additional mixed-effects models along the same lines as reported in the main study. In this case, along with *group* and *condition*, *behavioral* *response* (correct/incorrect) was included as a fixed effect along with all three-way interactions. The rest of the models (for N400 and LPC data) was kept the same as what was reported in the main study. Due to the inclusion of three-way interactions, model output is much more complex. However, as our main interest is in results for correct responses, we will not discuss results for incorrect trials in much detail, though the output from *lme4* is reported in Tables A5.1 and A5.2.

Results for the N400 time window (Table A5.1) are largely comparable to our original analysis, but some interesting differences emerge after accounting for a potential relationship with behavioral responses. For L1, the strong N400 responses to mismatch conditions persist, though the response to tone mismatches is marginally significant, whereas it was originally significant (*b*=-1.67, *SE*=0.85, *t*=-1.97). A bit counter-intuitively, incorrect acceptance of sentences with mismatches on critical words is related to strong increases in amplitude of N400s for both semantic and tone mismatches (semantic: *b*= -1.62, *SE*=0.78, *t*=-2.08; tone: *b*=-3.32, *SE*=0.78, *t*=-4.26), whereas responses for rhyme mismatches moved in the expected direction, becoming marginally more positive (*b*=1.81, *SE*=0.93, *t*=1.93).

L2 N400s are once again weak, as in the original analysis, with a similar decreasing order of magnitude from semantic, to rhyme, to tone. However, the response to semantic mismatches is now marginally significant (*b*=1.26, *SE*=0.75, *t*=1.68). This might lend support to the view that, for those trials where sentences were comprehended and mismatches noticed, L2 listeners displayed an N400, though the effect is weak, perhaps due to the sparseness of data for correct trials.

For the LPC window (Table A5.2), L1 results for correct trials were again comparable to our original analysis, with strong LPCs for rhyme and tone mismatches relative to the expected word condition (*rhyme:* *b*=2.28, *SE*=1.02, *t*=2.25; *tone:* *b*=4.11, *SE*=0.90, *t*=4.56). The LPC for tone mismatches is now significantly stronger (more positive) than that for rhyme mismatches (*b*=-1.83, *SE*=0.89, *t*=-2.06).

For L2 LPCs, results for correct trials are consistent with the N400 data, suggesting stronger negativity for semantic responses relative to the expected word condition, though the effect is marginally significant (*b*=1.87, *SE*=0.98, *t*=1.92). There is a corresponding increase in the difference between the semantic and tone mismatch conditions (*b*=-2.42, *SE*=1.06, *t*=-2.27), now significant, whereas it was marginal in the original model. The LPCs for the expected word condition and the rhyme mismatch are more similar than in the previous model (*b*=0.97, *SE*=1.01, *t*=0.96), and, as the tone condition is largely unchanged, the difference between the rhyme and tone mismatch conditions is now marginal (*b*=-1.51, *SE*=0.90, *t*=-1.69), whereas it was significant in the original model. As for the N400, these small changes in the pattern of results suggest some L2 sensitivity to semantic mismatches in correct trials, otherwise results are consistent with the original model in showing no significant late positivities for the rhyme and tone mismatch conditions.

In summary, results from this alternative analysis do not diverge strongly from the original analysis, especially as pertains to the critical L2 results. There is some weak evidence of L2 sensitivity to semantic mismatches from increased negativity in both the N400 and LPC windows, however, the effect is marginal, and the loss of power due to evaluating trials according to the correctness of behavioral responses casts further doubt on the effect. Results for rhyme and tone mismatches are still weak and statistically insignificant.

|  |
| --- |
| **Table A5.1. Results of linear mixed-effects model for ERP experiment: 300-500ms** ***BEHAVIORAL RESPONSE as fixed effect*** |
| *model formula (lmer):* mean.amp ~ 1 + word \* group \* behav.resp + ( 1 + word | subject ) + ( 1 + word \* group | item ) |
|  | **Fixed Effects** | **Random Effects** |
|  |  |  |  |  | *subjects* | *items* |
|  | *b* | *Std. Error* | *t-value* |  | *SD* | *SD* |
| Intercept (L1/expected word/correct) | 1.64 | 0.65 | 2.53 |  | 1.92 | 4.80 |
| semantic mismatch | -2.34 | 0.79 | -2.94 | \* | 1.90 | 6.82 |
| rhyme mismatch | -2.25 | 0.87 | -2.59 | \* | 2.59 | 6.35 |
| tone mismatch | -1.67 | 0.85 | -1.97 | . | 2.69 | 5.67 |
| incorrect | 0.87 | 0.57 | 1.53 |  | — | — |
| semantic mismatch × incorrect | -1.62 | 0.78 | -2.08 | \* | — | — |
| rhyme mismatch × incorrect | 1.81 | 0.93 | 1.93 | . | — | — |
| tone mismatch × incorrect | -3.32 | 0.78 | -4.26 | \* | — | — |
| L2 (expected word/correct) | -0.22 | 0.95 | -0.23 |  | — | 6.88 |
| L2 × semantic mismatch | 1.07 | 1.13 | 0.95 |  | — | 9.40 |
| L2 × rhyme mismatch | 1.22 | 1.24 | 0.99 |  | — | 8.53 |
| L2 × tone mismatch | 1.25 | 1.29 | 0.97 |  | — | 8.72 |
| L2 (expected word) × incorrect | -1.77 | 0.68 | -2.61 | \* | — | — |
| L2 × semantic mismtch × incorrect | 3.65 | 0.97 | 3.76 | \* | — | — |
| L2 × rhyme mismatch × incorrect | -0.40 | 1.08 | -0.37 |  | — | — |
| L2 × tone mismatch × incorrect | 4.41 | 0.94 | 4.71 | \* | — | — |
| **Additional comparisons** (correct trials) |  |  |  |  |  |  |
| L1 rhyme – L1 tone | -0.58 | 0.75 | -0.78 |  |  |  |
| L2 expected – L2 semantic | 1.26 | 0.75 | 1.68 | . |  |  |
| L2 expected – L2 rhyme | 1.02 | 0.87 | 1.17 |  |  |  |
| L2 expected – L2 tone | 0.42 | 0.98 | 0.43 |  |  |  |
| L2 rhyme – L2 tone | -0.60 | 0.82 | -0.73 |  |  |  |
| “\*” = statistically significant (|t| > 2.00); “.” = marginal (|t| > 1.65) |

|  |
| --- |
| **Table A5.2. Results of linear mixed-effects model for ERP experiment: 550-800ms** ***BEHAVIORAL RESPONSE as fixed effect*** |
| *model formula (lmer):* mean.amp ~ 1 + word \* group \* behav.resp + ( 1 + word | subject ) + ( 1 + word \* group | item ) |
|  | **Fixed Effects** | **Random Effects** |
|  |  |  |  |  | *subjects* | *items* |
|  | *b* | *Std. Error* | *t-value* |  | *SD* | *SD* |
| Intercept (L1/expected word/correct) | 2.18 | 0.66 | 3.29 |  | 1.98 | 4.84 |
| semantic mismatch | -1.10 | 0.99 | -1.10 |  | 2.87 | 7.53 |
| rhyme mismatch | 2.28 | 1.02 | 2.25 | \* | 3.06 | 7.38 |
| tone mismatch | 4.11 | 0.90 | 4.56 | \* | 2.94 | 5.71 |
| incorrect | 1.48 | 0.64 | 2.30 | \* | — | — |
| semantic mismatch × incorrect | -3.67 | 0.88 | -4.15 | \* | — | — |
| rhyme mismatch × incorrect | 0.79 | 1.06 | 0.75 |  | — | — |
| tone mismatch × incorrect | -4.76 | 0.88 | -5.38 | \* | — | — |
| L2 (expected word/correct) | 0.20 | 0.99 | 0.20 |  | — | 7.20 |
| L2 × semantic mismatch | -0.77 | 1.41 | -0.55 |  | — | 10.01 |
| L2 × rhyme mismatch | -3.25 | 1.46 | -2.23 | \* | — | 10.03 |
| L2 × tone mismatch | -3.56 | 1.36 | -2.61 | \* | — | 8.53 |
| L2 (expected word) × incorrect | -2.54 | 0.77 | -3.30 | \* | — | — |
| L2 × semantic mismatch × incorrect | 5.81 | 1.10 | 5.29 | \* | — | — |
| L2 × rhyme mismatch × incorrect | -1.98 | 1.22 | -1.62 |  | — | — |
| L2 × tone mismatch × incorrect | 5.48 | 1.06 | 5.17 | \* | — | — |
| **Additional comparisons** (correct trials) |  |  |  |  |  |  |
| L1 rhyme – L1 tone | -1.83 | 0.89 | -2.06 | \* |  |  |
| L2 expected – L2 semantic | 1.87 | 0.98 | 1.92 | . |  |  |
| L2 expected – L2 rhyme | 0.97 | 1.01 | 0.96 |  |  |  |
| L2 expected – L2 tone | -0.55 | 1.04 | -0.53 |  |  |  |
| L2 rhyme – L2 tone | -1.51 | 0.90 | -1.69 | . |  |  |
| L2 semantic – L2 tone | -2.42 | 1.06 | -2.27 | \* |  |  |
| “\*” = statistically significant (|t| > 2.00); “.” = marginal (|t| > 1.65) |

**REFERENCES**

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

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Socieity*, *57*(1), 289–300.

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

Brown, N. A., Dewey, D. P., & Cox, T. L. (2014). Assessing the validity of can-do statements in retrospective (then-now) self-assessment. *Foreign Language Annals*, *47*(2), 261–285. https://doi.org/10.1111/flan.12082

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

1. The full set of 240 sentences was intended for use in a separate experiment. As the present experiment used only a subset of 120 sentences, this means that the frequency information of the semantic condition was not as tightly controlled as it would be for the full set of items. [↑](#footnote-ref-1)
2. The decision to use blanks was motivated in part by a concern that the native Chinese speaker would epenthesize vowels at the end of the target words thus creating coarticulation effects between the offset of the target and the onset of the following word resulting in notable differences between the L1 and L2 speaker stimuli for use in a separate experiment. [↑](#footnote-ref-2)