N400 Evidence That the Early Stages of Lexical Access Ignore Knowledge About Phonological Alternations

Stephen Politzer-Ahles [ID]

Jueyao Lin [ID]
Hong Kong Polytechnic University, Hong Kong

Lei Pan
Hong Kong Polytechnic University, Hong Kong
Nanyang Technological University, Singapore

Ka Keung Lee
Hong Kong Polytechnic University, Hong Kong

Abstract

Hearing a word that was already expected often facilitates comprehension, attenuating the amplitude of the N400 event-related brain potential component. On the other hand, hearing a word that was not expected elicits a larger N400. In the present study, we examined whether the N400 would be attenuated when a person hears something that is not exactly what they expected but is a viable alternative pronunciation of the morpheme they expected. This was done using Mandarin syllables, some of which can be pronounced with different lexical tones depending on the context. In two large-sample experiments (total $n = 160$) testing syllables in isolation and in phonologically viable contexts, we found little evidence that hearing an alternative pronunciation of the expected word attenuates the N400. These results suggest that comprehenders do not take advantage of their knowledge about systematic phonological alternations during the early stages of prediction or discrimination.

Keywords

Corresponding author:
Stephen Politzer-Ahles, Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University, Hong Kong.
Email: stephen.politzerahles@polyu.edu.hk

Introduction

To understand language, comprehenders must have mechanisms for tolerating variation. Linguistic input that comprehenders hear or see may vary from the expected or citation forms because of differences across speakers (e.g., some speakers' overall speech rate, pitch, or voice quality may be different from others), across speaking contexts (e.g., speech in emotional or noisy contexts may sound different from speech in neutral or quiet contexts), and random variation. Input may also vary because of systematic phonological changes; for example, the /m/ in *some* may be instead pronounced as [ɱ] (with the lower lip touching the upper teeth, instead of touching the upper lip) in a sentence such as *Can we get some forks?* The present study uses event-related brain potentials (ERPs) to examine whether participants use phonological knowledge about these systematic alternations to account for potential alternative forms of words they are expecting. Our study focuses on alternations of lexical tone in Mandarin Chinese.

This is not the first study to examine the role of systematic phonological alternation in lexical access. The well-known study by Gaskell and Marslen-Wilson (1996) provided evidence that listeners can use altered surface forms to access underlying lexical representations in English. Their experiment took advantage of the fact that a word such as *lean* is sometimes pronounced as "*leam*" (in, e.g., *lean bacon*) and showed, using cross-modal priming, that participants hearing *leam* are able to activate the lexical entry for *lean* (i.e., participants seeing "LEAN" were able to respond to it more quickly after having heard *lean*, pronounced as either [lin] or [lim], than after having heard an unrelated word such as *brown*). While that study did not directly examine responses to the phonologically altered word itself, the downstream responses provide evidence that the phonologically altered word allowed listeners to access the intended citation form. In the present study, we directly probe processing of phonologically altered words by examining the brain response when those words are heard.

Instead of the kind of segmental alternation tested by, for example, Gaskell and Marslen-Wilson (1996), we test tone alternations in Mandarin Chinese. (We had no a priori reason to expect that these alternations would work differently than segmental ones in this respect; the choice to use Mandarin was a matter of convenience.) Mandarin has four citation tones (labeled in this paper as H[igh], R[ising], L[ow], and F[alling]). The Low tone is sometimes pronounced as (or very similar to—see Endnote 2) a Rising tone instead. Specifically, when two Low tones occur next to each other within the same intonational unit, the first one becomes Rising—for example, /ni^L/[1] (你, "you") is pronounced with a Low tone, but /ni^L hao^L/ (你好, "hello" [literally "you good"]) is actually pronounced as [ni^R hao^L], with a Rising tone on the first syllable. (The above is a simplified overview of what happens in this particular tonal alternation pattern, known as "third tone sandhi." For further details about the complex phonetic and phonological aspects of this phenomenon and its interactions with things such as morphology and syntax, see, for example, Chen, 2000; Duanmu, 2000; Kuo et al., 2007; Zhang & Lai, 2010).

Do listeners take advantage of their knowledge about this systematic alternation when they hear Mandarin sounds? If some preceding context makes a listener expect to hear ni^L, will the listener also anticipate its potential variant form ni^R?[2] Or, to put it another way: if a listener hears ni^R, will they recognize that it might be a variant form of ni^L?

Of course, the answer to the latter question must, eventually, be affirmative, as there is simple phonological evidence that listeners hearing a Rising tone can interpret it as a variant of a Low tone. For example, the character 祈 ("pray") corresponds to a morpheme pronounced /qi^R/, which has Rising tone underlyingly. However, some native speakers, when shown this character, have reported to us that they think its pronunciation is /qi^L/. This is likely a backformation that occurs

because this morpheme is most commonly used in the compound /qi^R dao^L/ (祈祷, "pray"). In this word, since qi^R is occurring in a context that could license a syllable's tone to change from Low to Rising, some speakers may believe that qi^R is not the underlying form, but the result of tone alternation. This demonstrates that some speakers are able to hear a Rising-tone syllable and interpret it as a variant of a Low-tone underlying form.

Another piece of evidence is how people complete a sentence such as the following (based on Speer & Xu, 2008):

1) $jin^H tian^H$ $hai^L bian^H$ yu^R hen^L . . .
    today   by the sea   ???   very . . .

Because yu^R occurs before a Low-tone syllable hen^L (很, "very"), it could be interpreted either as yu^R with an underlyingly Rising tone, or as yu^L with a Low tone which has been pronounced as Rising because of tone alternation. Both of these correspond to real words which fit the context: /yu^R/ is the phonological form of 鱼 ("fish"); and /yu^L/ the phonological form of 雨 ("rain"). Thus, if people complete the sentence with a word such as duo^H (多, "many"), it suggests that they interpreted yu^R as underlyingly Rising (as it yields a sentence meaning "today at the seaside there are many fish"). Sometimes, however, people complete this sentence with a word such as da^F (大, "big", or in this case "heavy"), yielding a sentence meaning "today at the seaside the rain is very heavy." A completion such as this demonstrates that a person hearing yu^R can indeed interpret it as a variant form of yu^L.

These forms of phonological evidence, however, are based on end-state responses. It is not yet known whether knowledge about this tone alternation impacts the earliest stages of lexical activation (or, earlier yet, forward predictions generated before the input is even encountered). To examine the brain-level processing of these kinds of phonologically altered words at the moment they are encountered, we adopt ERPs, focusing specifically on the N400. The N400 is a component of brain activity which tends to be larger (more negative) in proportion to how difficult the lexical entry or semantic knowledge about some meaningful stimulus is to access (Lau et al., 2008; see also Bornkessel-Schlesewsky & Schlesewsky, 2019), for a mechanistic account of the N400 as reflecting prediction error signals)—for example, words that are unexpected in their context, have not been primed, or are just rare, all tend to elicit larger N400s than expected, primed, and/or common words. Since our aim is to examine whether listeners' recognition of phonologically altered forms is aided by phonological knowledge, the N400 can serve as a useful measure of how easily listeners were able to recognize, and activate the associated lexical entry of, the auditory input they hear.

Many previous studies have shown that lexical tone matters in Mandarin lexical access and/or prediction. Malins and Joanisse (2012), for example, had participants hear Mandarin monosyllables (such as $tang^H$, "soup"), immediately after having seen a picture that would let them expect that syllable (a picture of a bowl of soup) or having seen a picture that would have led them to expect a syllable with the same segments but a different tone (e.g., a picture of a candy, the Mandarin word for which is $tang^R$). The auditory syllables elicited a greater N400 when their tone mismatched the expected syllable, compared to when they matched the expected syllable. Similar patterns have also been found for stimuli embedded in sentence contexts in Mandarin (Ho et al., 2019; Li et al., 2008) and Cantonese (Schirmer et al., 2005).[3]

These studies show that hearing a syllable with an unexpected tone elicits a greater N400 than hearing a syllable with an expected tone. They did not, however, examine systematic phonological relationships between different tones. The present study examines how participants react when hearing a tone that is different than, but phonologically related to, the expected tone. For example,

we know from the previous studies that hearing $tang^{R}$ should yield a smaller N400 when $tang^{R}$ was expected, compared to when $tang^{H}$ was expected. But what about when $tang^{L}$ was expected? Since $tang^{R}$ is a possible pronunciation of $tang^{L}$, will this relationship attenuate the magnitude of the N400 effect?[4]

We test this question in two experiments. The first experiment uses monosyllabic stimuli. Since this tone alternation only occurs when licensed by a following syllable, the second experiment uses two-syllable stimuli to create an environment where the phonologically altered syllable is a viable variant of the expected syllable.

2 Experiment 1: monosyllables

Experiment 1 tested whether the effect of phonetic mismatch between expected and heard syllables was attenuated when the unexpected heard syllable shared a morpho-phonological relationship with the expected one. In each trial, participants heard a Mandarin monosyllable (e.g., shi^{L}), preceded by a Chinese character. The preceding Chinese character either represented a morpheme with the same sound as the critical auditory syllable (e.g., 使, pronounced shi^{L}), a morpheme with the same segments but a different tone (e.g., 实, pronounced shi^{R}, or 是, pronounced shi^{F}), or a morpheme pronounced with different segments and tone (e.g., 御, pronounced yu^{F}). This yields four conditions: Match (when the critical syllable is the same as the expected syllable); Total mismatch[5] (when the critical syllable is unrelated to the expected syllable); Phonologically-related mismatch (when the critical syllable has a different tone than the expected syllable, but these tones are related through tone alternation); and Phonologically-unrelated mismatch (when the critical syllable has a different tone than the expected syllable, and these tones are not related).

While it is known that tone mismatch yields a more negative N400 effect relative to complete match (Malins & Joanisse, 2012), we predicted that the phonologically-related mismatch would attenuate this effect. That is to say, since shi^{R} (for example) is a possible variant of shi^{L} in some contexts, whereas it is not a possible variant of shi^{F} in any context, hearing shi^{R} when shi^{L} was expected (the *Phonologically-related mismatch* condition) would yield a smaller N400 effect than hearing shi^{R} when shi^{F} was expected (the *Phonologically-unrelated mismatch* condition).

A potential confound to this prediction, however, is that Low and Rising tones are more phonetically similar than any other pair of tones. Previous research has shown that they are confused with one another more often than other tones are (Chuang & Hiki, 1972; Sharma et al., 2015), they are less perceptually distant from one another as determined by mismatch negativity (Chandrasekaran et al., 2007), and they prime one another in implicit production priming (Chen et al., 2012). Thus, if what we call the phonologically-related mismatch condition elicits a smaller N400 effect than the phonologically-unrelated mismatch condition, that might occur simply because the phonologically-related mismatch consists of two phonetically similar tones, whereas the phonologically-unrelated condition does not; it might have nothing to do with knowledge about phonological alternations. Therefore, the experiment needed to be designed in a way to rule out this alternative explanation.

To rule out this confound, we tested both possible directions of presentation, that is, hearing shi^{L} when shi^{R} or shi^{F} is expected, and hearing shi^{R} when shi^{L} or shi^{F} is expected. It is possible that the N400 would be somewhat attenuated for phonologically-related mismatch in both of these conditions because of the abovementioned phonetic similarity issue. But it should be even more attenuated when hearing shi^{R}, as this is a possible surface realization of an underlyingly shi^{L} syllable. On the other hand, the attenuation should be less substantial when hearing shi^{L}, as this is not a possible surface realization of an underlyingly shi^{R} syllable (in standard Mandarin, there is no phonological process that changes Rising tones to Low tones). Thus, we would only take the results as indicating
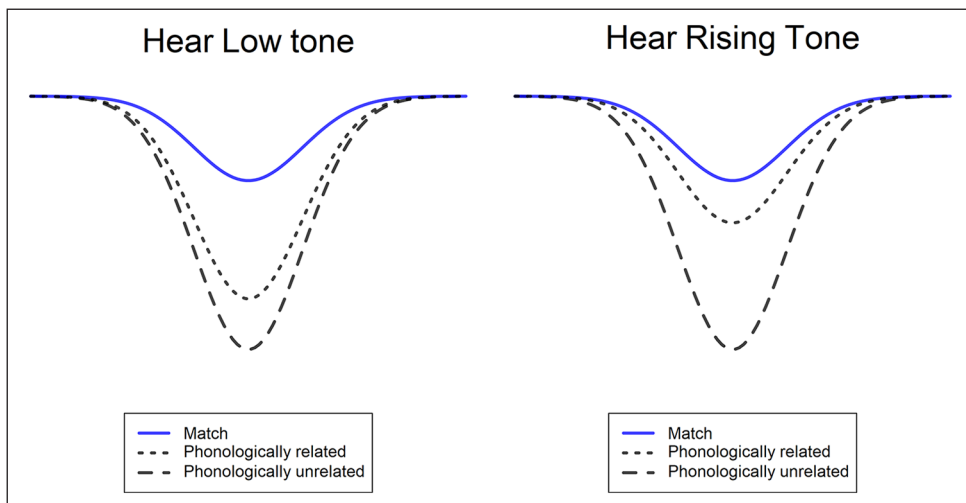
Figure 1. The predicted pattern of results; see text for details.

that phonological knowledge influences the N400 process if such an interaction is observed, with greater N400 attenuation when hearing Rising tone than when hearing Low tone.

This prediction is illustrated in Figure 1. When hearing Low tone, the phonologically-related mismatch condition may elicit a slightly smaller N400 than the phonologically-unrelated mismatch condition because the Low and Rising tones are phonetically similar. Hence, the left-hand panel of Figure 1 shows a dashed black line representing the ERP elicited by the phonologically-unrelated mismatch condition, which is substantially more negative than the blue line representing the ERP elicited by the match condition—this is the N400 effect—and a dotted black line representing the ERP elicited by the phonologically-related mismatch condition, which is more negative than the match condition but not as negative as the unrelated mismatch condition—this is the attenuation of the N400 effect. When hearing Rising tone, on the other hand, the phonologically-related mismatch condition should elicit a greatly reduced N400 compared to the phonologically-unrelated mismatch condition, because in addition to being phonetically similar to the expected tone, the heard tone is also a possible variant of the expected tone. Hence, the right-hand panel of Figure 1 shows a similar situation as the left-hand panel, except with much more attenuation—the ERP for the phonologically-related mismatch condition is much closer to the match condition, and much farther from the mismatch condition.

All methods for the experiment were pre-registered at https://osf.io/a5beh/.

2.1 Methods

2.1.1 Participants. Ninety-two right-handed native speakers of Mandarin (average age 22.2 years, range 18–34, 35 men and 56 women and one who did not list a gender; see https://osf.io/fhsn4 for detailed demographic information). Twelve participants were removed during data preprocessing because their electroencephalogram (EEG) data had substantial artifacts; this left 80 participants in the final dataset. This target sample size was chosen based on a simulation-based power analysis (https://politzerahles.shinyapps.io/ERPpowersim/) suggesting that a large sample would be needed to detect the potentially subtle effects we predicted. Procedures for the experiment were approved
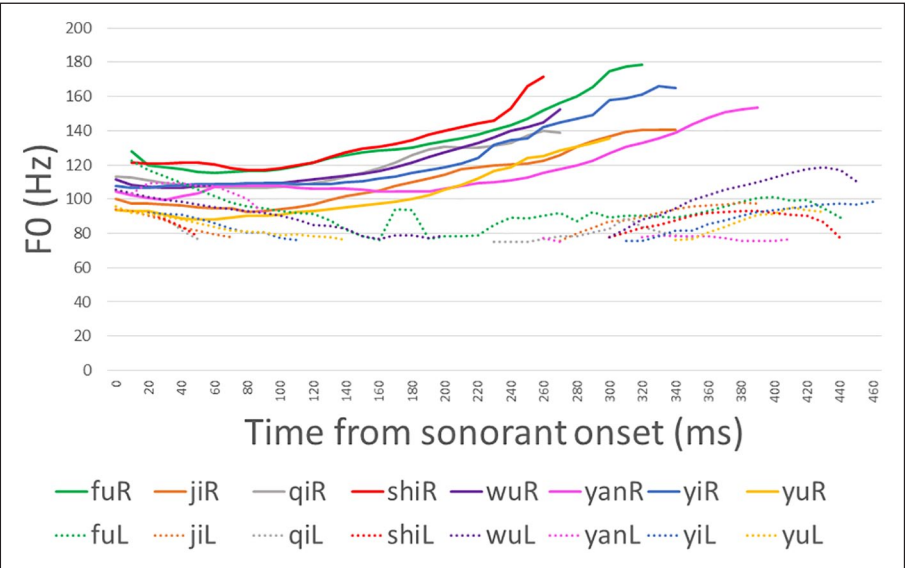
Figure 2. Fundamental frequency contours of the 16 critical stimulus tokens.

Table 1. Sample stimulus set: pronunciation of each character is indicated in parentheses.

	Low-tone target: *shi*^{L}	Rising-tone target: *shi*^{R}
Match	屎 (*shi*^{L})	食 (*shi*^{R})
Phonologically-related mismatch	食 (*shi*^{R})	屎 (*shi*^{L})
Phonologically-unrelated mismatch	士 (*shi*^{F})	士 (*shi*^{F})
Total mismatch	厌 (*yan*^{F})	厌 (*yan*^{F})

by the Human Subjects Ethics Sub-committee at the Hong Kong Polytechnic University. All participants provided written informed consent and were paid for their participation.

2.1.2 Materials. The critical audio stimuli comprised eight segmental syllables (*fu, ji, qi, shi, wu, yan, yi*, and *yu*), each recorded in all four tones, yielding 32 recordings. Each syllable formed an existing morpheme in all four tones. Several repetitions of each stimulus were produced, in a random order, in isolation, by a male native speaker of Mandarin from Changchun. We selected the clearest token of each stimulus and saved it into its own sound file. Fundamental frequency (F0) curves of all the critical stimuli are shown in Figure 2. Note that the Low-tone syllables used the full (falling–rising) contour of the Low tone, which generally only occurs when the syllable is pronounced in isolation or at the end of an intonational phrase, rather than the arguably more common low-falling contour which results from "half-third" sandhi when a Low-tone syllable occurs before any non-Low-tone syllable in the same intonational phrase. There are also noticeable gaps in most of the Low-tone F0 contours because the speaker pronounced these with a creaky voice, which is common.

Each audio stimulus could be preceded by a variety of characters, depending on the condition. The critical conditions are illustrated in Table 1. Each segmental syllable was paired with several possible characters for each condition to allow for more variety across trials. For example, in some

trials, the match character for Low-tone target shi^L was 屍, and in other trials it was 使, 史, or 始; all of these characters are pronounced shi^L. We did not use any characters which had multiple pronunciations, and when possible we tried to avoid having the same radical appear in multiple characters in a set. A full list of stimuli and the corresponding characters for each condition is available at https://osf.io/fhsn4/, as are all the audio recordings used in the experiment.

Several types of fillers were also included. To ensure that characters pronounced with Low or Rising tones were sometimes used in unrelated mismatch trials as well as match and phonologically-related trials, characters from one set were also used as stimuli in other sets. For example, while 鱼 (pronounced yu^R) would serve as the prime character in match or phonologically-related mismatch conditions in the *yu* set where it would be paired with yu^R or yu^L auditory stimuli, it was also used as a filler in the *shi* set where it would serve as a segmentally-unrelated mismatch when followed by a shi^L auditory stimulus. Thus, each Low or Rising tone auditory stimulus was associated with five kinds of preceding characters: the four critical conditions described above and in Table 1, and this filler condition. Furthermore, auditory stimuli with High or Falling tone were also used as fillers, to ensure that auditory targets were not always Low or Rising tone. Each High or Falling auditory target could be preceded by a character whose pronunciation completely matched the target, one which had the right segments but had Low or Rising tone (this ensured that when participants saw a character pronounced with Low or Rising tone they cannot be sure that they will hear a Low or Rising auditory target), or one with the wrong segments. The full list of fillers is available at https://osf.io/fhsn4/.

Across the critical conditions and fillers, a given Low-tone or Rising-tone character would be followed by a matching auditory stimulus 40% of the time and a mismatching stimulus 60% of the time; for example, 石 (shi^R) is followed by shi^R twice, shi^L once, shi^F once, and yan^L once. A given auditory syllable used in critical conditions would be preceded by a matching character 33% of the time and a mismatching character 66% of the time. For example, shi^L occurs eight times with matching characters (two each of 使, 屍, 史, and 始), and 16 times with mismatching characters (characters pronounced shi^R: {实, 石, 时, 十}; characters pronounced shi^H or shi^F: {湿, 师, 市, 试}; or characters with totally unrelated pronunciations that share neither segments nor tone with shi^L: {乌, 淤, 御, 厌, 鱼, 余, 伏, 虞}).

2.1.3 Procedure. The character-audio pairings described above resulted in a total of 544 trials (the full list can be seen on the stimulus spreadsheet or the Presentation experiment control file available at https://osf.io/fhsn4/). These trials were presented in a fully random order. Stimulus presentation, response logging, and output of event markers to the EEG acquisition system was handled by Presentation v20.0 (Neurobehavioral Systems, Inc.).

Throughout the experiment, the screen background was black and visual stimuli were shown in white. Each trial began with a fixation point "*" presented in 88pt font at the center of the screen for 250 milliseconds. Next, the character was presented in 88pt FangSong font, and remained on the screen until the auditory stimulus ended. The auditory stimulus began 800–1200 milliseconds after the appearance of the character (the onset time was determined randomly for each trial at runtime). After the sound finished playing, the character disappeared from the screen and was replaced by a "???" prompt for participants to make a judgment; just below this prompt, on the left and right sides of the screen were shown the prompts "不一致" ("inconsistent") and "一致" ("consistent"), respectively, to help participants remember which button was which. Their task was to press the left or right arrow key (using the index and middle fingers of their dominant hand) to indicate whether the syllable they heard matched the character they saw (they were not given any instructions to respond quickly, and were in fact encouraged to use this time if they needed to take a break to blink, stretch, etc.). After they indicated their judgment, an inter-trial interval with a

blank screen was presented for 1500 milliseconds, after which time the next trial commenced. The experiment was divided into eight blocks with equal numbers of trials, and participants were given an opportunity to rest between blocks. Before the main experiment, 20 practice trials (using syllables *cai* and *an* as the auditory stimuli, recorded by the same speaker as the experiment proper) were presented, after which the program seamlessly moved into the real experimental trials. The Presentation script and stimulus files are available at https://osf.io/fhsn4/.

Before the recording, participants provided informed consent, filled out a demographic information sheet and read the instructions, and then the EEG cap was put on them. The entire session (including this setup, the experiment proper, and clean-up and debriefing afterwards) lasted about 1.5 hours.

2.1.4 Acquisition and preprocessing. The continuous EEG was recorded using a SynAmps 2 amplifier (NeuroScan, Charlotte, NC, U.S.A.) connected to a 64-channel Ag/AgCl electrode cap. The channels followed the 10–20 system. Polygraphic electrodes were placed above and below the left eye (forming a bipolar channel to monitor vertical electrooculogram (EOG)), at the left and right outer canthi (forming a bipolar channel for horizontal EOG), and at the right and left mastoids (to be digitally averaged offline for re-referencing). A channel located halfway between Cz and CPz served as the reference during the recording and AFz as the ground. Impedances were kept below 5 kΩ. The EEG was digitized at a rate of 1000 Hz with an analog bandpass filter of 0.03–100 Hz. A Cedrus Stimtracker interfaced between the experiment control software and the EEG acquisition software; in addition to recording event markers sent from Presentation, it also recorded the onset and offset of the audio stimuli via an auditory channel, but these markers were not used in the present analysis.

As many datasets had substantial blink artifact, perhaps because of the relatively long time participants needed to keep their eyes open per trial (1050–1450 milliseconds before the onset of the auditory stimulus, plus the duration of the stimulus itself), we cleaned artifacts using independent component analysis (ICA) decomposition (Makeig et al., 1996). While this step was not included in our pre-registration plan, we never looked at the data pattern without ICA, and thus the choice to use ICA was not motivated by how it would or would not support our hypothesis.

The specific preprocessing steps were as follows. For each participant, data were imported to EEGLAB version 13.16.5b (Delorme & Makeig, 2004) and, if necessary, up to three bad channels were interpolated. All scalp channels were re-referenced to the average of both mastoids. The 320 critical trials (64 trials for each of the Match conditions in Table 1, and 32 trials for each of the Mismatch conditions) were segmented into epochs from 500 milliseconds before to 1500 milliseconds after each time-locking event (this epoch is longer than the pre-registered epoch to allow more data to be used for ICA decomposition, but the ERP analysis only focused on a shorter epoch as pre-registered). For the obstruent-initial stimuli (*fu, ji, qi*, and *shi*) the time-locking event was the manually measured vowel onset, since the initial consonant carries little to no information about the tone of the syllable. For the other stimuli (*wu, yan, yi, yu*) it was the onset of the sound file. The data were then composed into as many independent components as there were channels (minus the two mastoids and any bad channels). For each participant, up to six components associated with blinks or saccades were identified based on manual inspection and removed, and then the epochs were manually inspected to identify any with a remaining artifact. The number of trials remaining per participant per condition is shown on the participant information spreadsheet at https://osf.io/fhsn4/. Finally, the data were low-pass filtered with a 30 Hz high cutoff (using the default filter in EEGLAB 13.16.5b), baseline-corrected using the mean of the 200-milliseconds pre-stimulus interval, and the remaining trials for each condition were averaged.

Table 2. Each condition's average reaction time (in milliseconds) and accuracy (in percentage correct). Note that for the phonologically-related mismatch condition, responses of "mismatch" were coded as correct and responses of "match" as incorrect, but participants were not given any instructions or feedback regarding which responses we would consider correct.

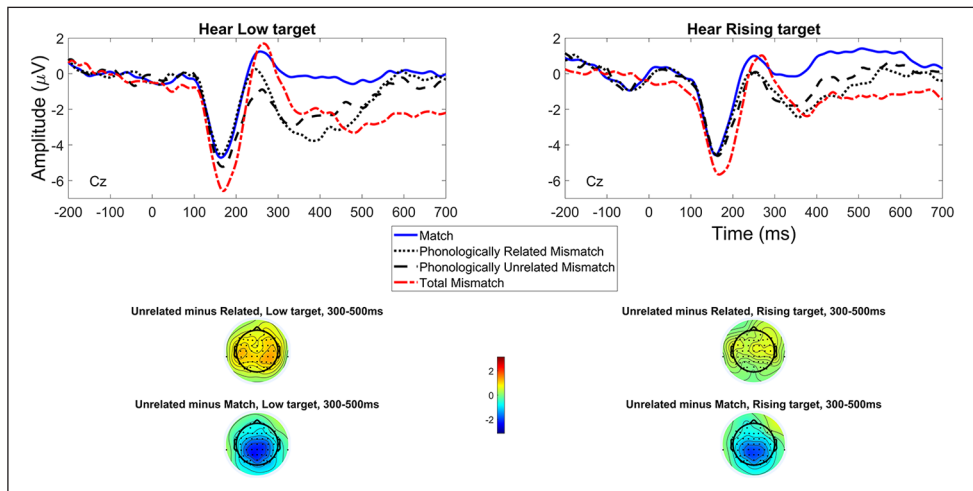|	Hear Low target	Hear Rising target
Match	841 (94.5)	821 (95.5)
Phonologically-related mismatch	1018 (92.3)	863 (96.4)
Phonologically-unrelated mismatch	861 (97.8)	841 (99.0)
Total mismatch	840 (99.5)	912 (99.6)

Figure 3. Event-related brain potentials (ERPs) at electrode Cz (top portion) and topographic difference plots for the 300–500 milliseconds time window. The upper left panel shows the ERPs in four conditions when hearing Low targets; the upper right, ERPs in four conditions when hearing Rising targets. The lower half of the figure shows the topographic distribution of these effects (in the 300–500 milliseconds time window). Note that the "Total mismatch" condition is presented only for information but is not part of the statistical analysis, and any direct comparisons between this condition and other conditions are problematic since the point at which participants can detect a mismatch is earlier in this condition than it is in the other conditions.

2.2 Results

The data are available at https://osf.io/fhsn4/. Behavioral accuracy and reaction times are shown in Table 2. ERPs from electrode Cz, which are fairly representative of the pattern over the rest of the scalp, are shown in Figure 3. For both Low-tone and Rising-tone targets, all types of mismatch elicited more negative ERP amplitudes compared to the match condition, in the time window typically associated with the N400 (starting from about 200 to 300 milliseconds). This pattern helps confirm that our manipulation was effective. The crucial question, however, is whether the N400 effect was attenuated in the phonologically-related mismatch condition, relative to the phonologically-unrelated mismatch condition; and whether this attenuation was greater when hearing the Rising target than when hearing the Low target. It is clear that this is not the case. In the typical N400 time window (300–500 milliseconds after the onset of the sonorant part of the syllable),
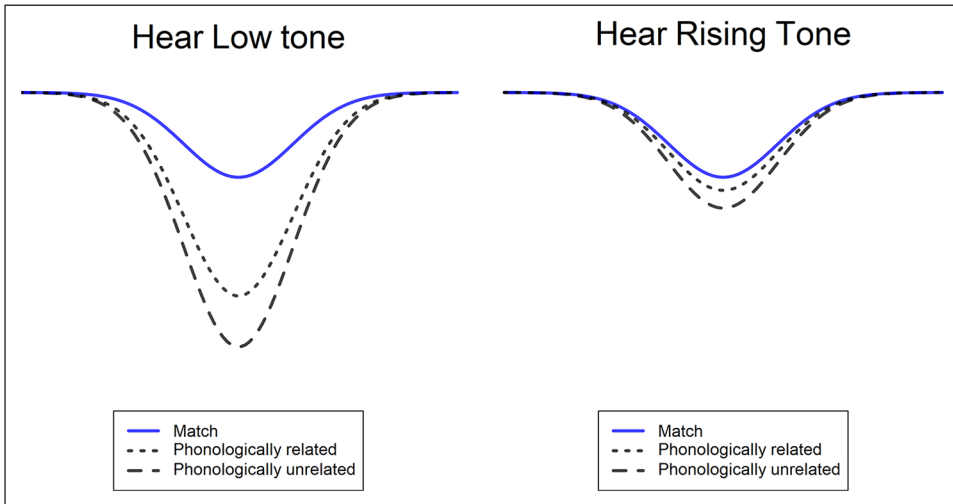
Figure 4. Another possible pattern of results consistent with our prediction; see text for details.

ERPs elicited by the phonologically-related mismatch were *more* negative, rather than less negative, than ERPs elicited by the phonologically-unrelated mismatch. In an earlier time window (about 150–250 milliseconds) the phonologically-related mismatch does appear to be less negative, but this effect is numerically larger in Low targets than Rising targets, which is opposite our prediction.

We supplemented visual inspection of the ERP data with statistical analysis, as pre-registered at https://osf.io/a5beh/. The goal of the analysis was to select the spatiotemporal window of analysis in a data-driven but unbiased (i.e., independent of our main hypothesis) way, and to directly quantify the extent of N400 attenuation. We quantified N400 attenuation in terms of the ERP amplitude for the phonologically-related condition relative to both the phonologically-unrelated and match conditions. Recall that our prediction was that the N400 for phonologically-related mismatch would be more attenuated (relative to phonologically-unrelated mismatch) when hearing Rising tone than when hearing Low tone. This was schematically illustrated in Figure 1, where the ERP for the phonologically-related condition is substantially more positive than that for the phonologically-unrelated, and this difference is larger when hearing Rising tone than when hearing Low tone. However, simply comparing phonologically-related and phonologically-unrelated mismatch may not be sufficient to capture this pattern, because the overall N400 magnitude may be intrinsically different when hearing Low versus Rising tones. This possibility is illustrated in Figure 4. In Figure 4, the raw difference between the ERPs for phonologically-related and phonologically-unrelated mismatch is smaller when hearing Rising tone than when hearing Low tone. The amount of N400 *attenuation*, however, is larger. Specifically, when hearing Low tone, the ERP for phonologically-related mismatch is still closer to that for phonologically-unrelated than it is to that for match, whereas when hearing Rising tone the ERP for phonologically-related mismatch is closer to that for match, indicating that the N400 effect of mismatch is almost fully attenuated. Therefore, rather than directly comparing the raw ERPs for phonologically-related and phonologically-unrelated mismatch, we instead quantified how much of the effect was attenuated, as a proportion of the original effect magnitude.

Specifically, we first identified the spatiotemporal time window of the N400 effect using cluster-based permutation tests (Maris & Oostenveld, 2007). This test compared the ERP for match trials

(averaged across Low and Rising targets) to that for phonologically-unrelated mismatch trials (again averaged across Low and Rising targets) to identify the spatiotemporal cluster underlying the difference between these, that is, to identify the mismatch effect. The test was implemented in FieldTrip (2020-03-20 release; Oostenveld et al., 2011), from -200 to 00 milliseconds relative to the time-locking point, including all EEG channels, with clusters identified using *t*-tests with a threshold of one-tailed α = 0.05 and the *minnbchan* parameter set to 2. A robust mismatch effect consistent with an N400 effect was identified (*p* = 0.002), driven by a cluster encompassing most of the post-stimulus epoch and all channels (Figure 5). For each participant and each condition, the amplitudes of all channel-sample pairs that were part of this cluster were averaged to yield a mean amplitude in the N400 spatiotemporal window. Finally, separately for Low and Rising targets, the extent of attenuation

was calculated by the following formula: $\dfrac{(Phonologically\ related) - (Phonologically\ unrelated)}{(Match) - (Phonologically\ unrelated)}$.

For example, imagine that for Rising-tone targets in one participant, the amplitude in the N400 window for the match condition was -5 µV, in the phonologically-related mismatch it was -8 µV, and in phonologically-unrelated mismatch it was -10 µV. Thus, the difference between match and phonologically-unrelated mismatch (i.e., the N400 effect) is 5 µV. The difference between phonologically-related and phonologically-unrelated mismatch is 2 µV. 2 is 40% of 5, or in other words, the N400 effect would have been attenuated by 40%. Thus, this number quantified the extent to which the N400 was attenuated, such that a higher value meant more attenuation. This value could be above 100% or below 0%, in cases where the ERP for phonologically-related mismatch did not fall between the two other conditions. Finally, after this value was calculated for all participants for Low and Rising targets separately, the attenuation values for Low and Rising targets were compared via a paired-sample *t*-test with a one-tailed *p* value. This test was not significant, *t*(79) = 0.51. In other words, just as Figure 3 suggests, the results did not provide evidence that the N400 effect was attenuated by phonological knowledge in the way we predicted it would be.

2.3 Discussion

We did not find evidence that the N400 amplitude would be modulated by phonological knowledge. Contra our expectation, the mismatch effect elicited by tonal alternations (legal variants of the expected tone) was no smaller than that elicited by outright violations (tones from a completely different category not related to the expected tone). While tone mismatch between the target and the expected sound elicited a robust difference that was consistent with an N400 effect, this effect was not attenuated when the mismatching tone was phonologically related to the expected one.

It is worth noting that the way we quantified N400 attenuation for the statistical analysis may not have accurately isolated the N400 component. As can be seen in Figure 5, the cluster identified by the test extended well beyond the typical time window of the N400. In fact it looks more like two distinct effects: an earlier one consistent with an N400 (more long-lived on posterior channels than anterior channels); and a later-emerging frontal negativity. As an exploratory analysis, we re-ran the cluster test with slightly different settings (a clustering alpha threshold of 0.01 rather than 0.05, and *minnbchan* of 4 instead of 2, which would make it less likely for two distinct effects to be connected into one cluster by a narrow "bridge" of channels). This analysis did successfully reveal a more N400-like cluster (extending only from about 200 milliseconds to about 550 milliseconds), but the difference in N400 attenuation was still nonsignificant and was in the opposite direction relative to what we predicted, *t*(79) = -0.29. Furthermore, visual inspection of the ERP waveforms (Figure 3; figures for a broader selection of channels are available at https://osf.io/fhsn4/) suggests no trend in the predicted direction. Thus, we conclude that the data do not support the original hypothesis,
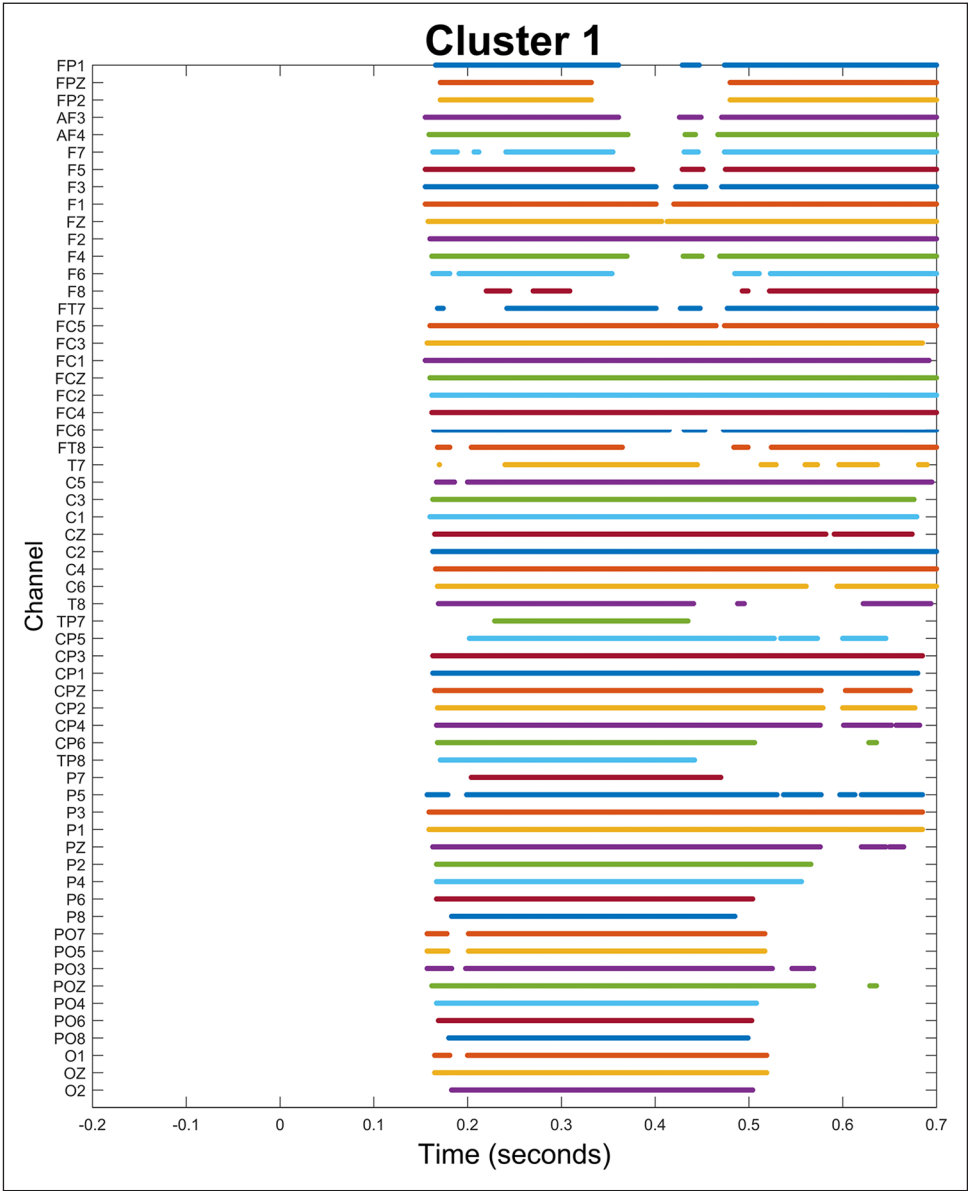
Figure 5. A raster plot (Groppe et al., 2011) illustrating the spatiotemporal extent of the cluster we used to quantify the N400 effect. The x-axis represents time, and on the y-axis the 60 scalp channels are organized roughly from anterior to posterior. In the plot, each row represents one channel, and each coloured dot in that row represents one timepoint at which that channel was part of the cluster. When many subsequent timepoints on the same channel are included in the cluster, the dots placed closely together look instead like a horizontal line.

regardless of how they are statistically evaluated. Nonetheless, the issue of how to effectively isolate modulation of the N400 from other overlapping components in this paradigm deserves further attention. It is important to keep in mind that even effects within the canonical "N400 time window" may

not be modulations of the N400 component itself, but modulations of other components which spatiotemporally overlap with the N400 (Luck, 2005).

We then speculated that the failure to observe an effect of phonological knowledge may have been due to our use of monosyllables in isolation. The phonological alternation between Low and Rising tones only occurs with tones in a phrasal context (when one Low tone is followed by another). Therefore, participants may have avoided activating allophonic variants of Low tone because the tones were not in a position that licensed that alternation. Gaskell and Marslen-Wilson's (1996) study on English coronal alternation supports this idea; auditory presentation of *leam* primed the visual stimulus LEAN better when *leam* occurred in a context that licenses the change of [n] to [m] (*leam bacon*) than when it occurred in a context that does not license it (*leam gammon*). Thus, the results of Experiment 1 are consistent with two quite opposite conclusions. On the one hand, the results might be taken as evidence that Mandarin-speaking participants do not use phonological knowledge about tone alternations to inform their predictions or lexical access. On the other hand, they could be taken as evidence that these participants use that phonological knowledge so well that they do not activate or predict alternative pronunciations of a word if the word does not appear in a phonological context that licenses them. To disentangle these possibilities, we ran a new version of the experiment in which the critical targets were all the first syllable of a disyllabic pseudoword.

3   Experiment 2: disyllables

Experiment 2 had roughly the same design as experiment one, but the critical syllables were placed in a context that licenses the tone alternation. Participants still saw a single character before hearing the audio target, but then they heard a two-syllable pseudoword, and were instructed to judge whether the first syllable they heard matched the character they had just seen.

The predictions for Experiment 2 were similar to those for Experiment 1: we expect that phonologically-related mismatch should attenuate the N400 effect (relative to phonologically-unrelated mismatch) more when hearing Rising tone than when hearing Low tone. An additional complication, however, is that Rising targets in a disyllable context are temporarily ambiguous: until the listener hears the second syllable, they do not know whether the Rising syllable will be followed by a Low syllable (in which it could underlyingly have been either Rising or Low) or followed by another syllable (in which case it must be underlyingly Rising). This ambiguity may further complicate the processing of Rising targets; for example, they may elicit slightly more negative N400s even in the match condition, compared to the unambiguous condition Low-tone match would elicit, given that they induce temporary ambiguity. Thus, the comparison across Low and Rising targets is less straightforward than in Experiment 1. However, within each of these targets, we still predict that the N400 should be attenuated for phonologically-related mismatch, compared to that for phonologically-unrelated mismatch. As the phonetic similarity of Low and Rising tones apparently did not trigger attenuation of the N400 in Experiment 1, we do not consider it as serious a confound as we did during Experiment 1, and thus we consider the interaction between condition and target tone less crucial in this experiment.

Methods for Experiment 2 were pre-registered at https://osf.io/jtr95/.

3.1 Methods

3.1.1 Participants. Eighty-nine right-handed native speakers of Mandarin (average age 23.9 years, range 18–36, 24 men and 65 women; see https://osf.io/c8qv9/ for detailed demographic information) participated. Nine participants were removed during data preprocessing because their EEG data had substantial artifacts, leaving 80 participants in the final dataset. The characteristics of the

participant sample were otherwise the same as in Experiment 1. None of the participants in Experiment 2 had previously participated in Experiment 1.

3.1.2 Materials. The stimuli followed the same design as Experiment 1, with a few key changes. First of all, each of the target monosyllables was combined with a second syllable to form a two-syllable pseudoword. The second syllable was always an accidental gap—that is, a phonotactically legal syllable that does not exist in Mandarin. The accidental gaps used were *bou, kei, pe, pong, ten*, and *tiu*, each in all four tones. We used pseudowords in order to minimize effects of semantics, lexical frequency, and morphological structure, so we could focus instead on phonology. Furthermore, if we used two-syllable real words, it would not have been possible to use the same second syllable in every item, since a syllable that forms a real compound word when combined with, for example, shi^L does not necessarily form a real compound word when combined with, for example, yu^L. The 32 targets (eight syllables times four tones) were combined with these 24 pseudo-syllables (six syllables times four tones) to yield 720 disyllables (Low–Low combinations were excluded, as these sequences are disallowed within an intonational unit because of the tone alternation that changes Low tones to Rising tones). Thus, for example, while the critical auditory stimuli for Experiment 1, shown in Table 1, were monosyllables such as shi^L or shi^R, in Experiment 2 they were meaningless two-syllable words such as $shi^L tiu^H$ or $shi^R tiu^H$. These were recorded by a 21-year-old female native Mandarin speaker from Dalian. After the stimuli were segmented into separate .wav files, they were normalized to have the same average intensity using Praat (Boersma & Weenink, 2018).

In the critical conditions, the critical target (the Low or Rising initial syllable) was always followed by a non-Low tone, to avoid eliciting brain activity that could be related to the processing of the tone alternation itself. In fillers (where the target syllable is High or Falling), the second syllable could be any tone. To make participants still willing to treat initial syllables as ones that *could* be followed by a Low tone and to make sure apparent tone alternation does occur in the context of the experiment, we added 128 filler trials in which the first syllable was Rising and the second syllable was Low tone. These fillers comprised 16 trials for each segmental syllable item (e.g., 16 *ji* trials with Low tone in the second syllable, 16 *qi*, etc.). In these 16 trials, eight were ambiguous match trials (e.g., hearing ji^R in the first syllable, when either ji^L or ji^R was expected; since tone alternation would change ji^L to ji^R when it precedes another Low tone, the ji^R that the participant hears could match either a Low or Rising expected syllable), four were phonologically-unrelated mismatches (e.g., hearing ji^R when expecting ji^H or ji^F), and four were total mismatches.

The practice trials were also disyllables, with the same pseudo-syllables used in the second syllable position. All the recordings are available at https://osf.io/c8qv9/.

3.1.3 Procedure. The procedure was identical to Experiment 1, except that participants were instructed to judge whether the *first syllable they heard* on a given trial matched the character they had just seen. On each trial, the pseudo-syllable to be used as the second syllable was randomly selected at runtime (within the design constraints, for example, for critical Low and Rising targets a Low second syllable was never selected, and for the fillers described in 3.1.2 above a Low second syllable was always selected). The Presentation script and audio stimuli are available at https://osf.io/c8qv9/.

3.1.4 Analysis. Data preprocessing was carried out as in Experiment 1.

3.2 Results

The data are available at https://osf.io/c8qv9/. Behavioral accuracy and reaction times are shown in Table 3. ERP results at Cz are shown in Figure 6. As in Experiment 1, all mismatches elicited

Table 3. Each condition's average reaction time (in milliseconds) and accuracy (in percentage correct).

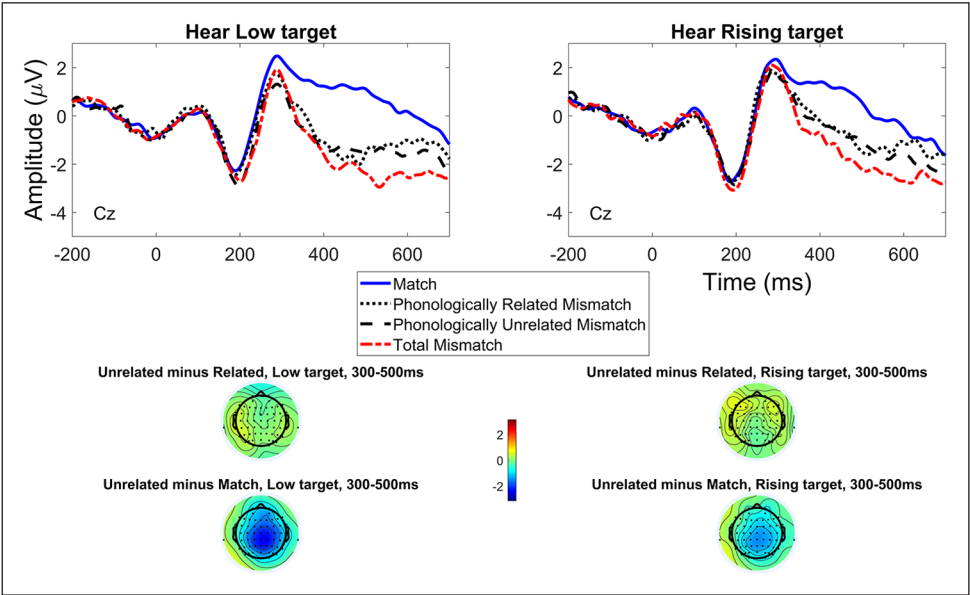|	Hear Low target	Hear Rising target
Match	742 (93.9)	754 (94.3)
Phonologically-related mismatch	843 (88.7)	734 (97.0)
Phonologically-unrelated mismatch	720 (95.9)	714 (98.6)
Total mismatch	675 (99.5)	723 (99.6)

Figure 6. Event-related brain potential results at electrode Cz for Experiment 2, shown in the same format as the results for Experiment 1 (see Figure 3).

more negative ERPs in the time window associated with the N400. As in Experiment 1, however, there was no clear indication that response for the phonologically-related mismatch was attenuated relative to that for the phonologically-unrelated mismatch. As the pattern of results is not even in the direction predicted, we did not conduct any statistical analysis of it. Figures showing a broader selection of electrodes are available at https://osf.io/c8qv9/.

3.3 Discussion

The results from Experiment 2 replicated the results of Experiment 1 with monosyllables: again, we did not find any evidence that unexpected tones with a phonological relationship to the expected tone are treated as less serious mismatches than unexpected tones with no phonological relationship to the expected tone. In each case, the syllable mismatching the expected syllable appeared to elicit a larger N400 than the matching syllable did, regardless of the presence or absence of a systematic phonological relationship between them.

The fact that we observed this pattern of results twice, each time with a very large sample of participants, makes it unlikely that our failure to observe the predicted attenuation is just a Type II

error. It is more likely that there is some systematic reason that phonologically-related mismatches do not attenuate the N400 (relative to phonologically-unrelated mismatches) in the present study.

Above we suggested that a possible explanation for that observation in Experiment 1 was that the stimuli were not in a context that licenses the tone change, and thus participants may have had good reason not to consider a Rising-tone syllable as a viable variant of a Low-tone syllable. In Experiment 2, on the other hand, the critical syllables that participants heard were in a two-syllable context, which has the potential to license the tone change. This does not guarantee, however, that participants will consider the critical tone to be in a context that really licenses the tone change. In fact, the relevant tone change is only licensed when the syllable following the critical syllable has Low tone. In this study, the critical syllables were never in this context; they were always followed by a non-Low tone, and only filler stimuli had a Low tone in the following context. However, when participants are hearing this critical syllable, they do not yet know what the following syllable will be. We therefore assumed that, if they really use knowledge about phonological alternations at the earliest stage of prediction and lexical access, they would process the initial syllable as if it *could* be in a viable context for the tone change. This does not appear to have been the case, however. The median onset time of the F0 contour of the second syllable (i.e., the earliest time participants could possibly determine whether the second syllable is Low or not) was 371 milliseconds after the ERP onset (minimum: 200 milliseconds; maximum: 584 milliseconds; inter-quartile range: 97 milliseconds), whereas the onset of the N400 effect shown in Figure 6 is near 300 milliseconds. Thus, it seems that participants tended to categorize the critical syllables as having the wrong tone even before the tone of the following syllable became available, and certainly before they had time to process it (presumably if a determination of whether the tone was a "match" or "mismatch" were delayed until hearing the second syllable, then an N400 effect elicited by this information would emerge a couple hundred milliseconds after that).

Thus, a potential explanation for our results could be that participants wait to hear the following tone before they are willing to consider that a surface Rising tone could be a variant pronunciation of an underlying Low tone, and thus their initial processing of the critical syllable has not yet taken the phonological context into account because they have not heard it yet. If they do that, however, that would suggest that participants do not actively predict possible different surface realizations of words they are expecting, and at the initial stages of lexical access they do not entertain the idea that the input they heard could have derived from a different underlying form. One could interpret our results as showing that the process underlying the N400 is not sensitive to knowledge about alternations, or that participants simply hold off on using that knowledge until a later processing stage; in either case, the results appear to strongly demonstrate that knowledge about alternations does not appear to play a role at the initial processing stage examined in this paradigm.

Note, however, that the results from these two experiments could be looked at in a different way than what we have done here. We compared ERP responses elicited by the same target audio stimuli as a function of what character preceded them. Alternatively, given that the participants' task involves comparing the incoming auditory signal to a previously-generated expectation, we could instead have compared ERP responses to different auditory stimuli following the same initial character, which might be a more direct way of examining whether responses to violations of Low-tone predictions are different than responses to violations of Rising-tone predictions. While the present experiment was not designed to make this comparison and thus several aspects of the experiment design are not conducive to this comparison, we examined it as an exploratory analysis and found only very weak evidence that violated Low-tone predictions (i.e., hearing a Rising-tone syllable after having predicted a Low-tone one) elicit less surprise than violated Rising-tone predictions. This analysis is reported in more detail in the Online Supplementary File 1.

4   General discussion

Across two experiments with large sample sizes, the present study did not find evidence that listeners take advantage of their knowledge about systematic phonological alternations at the initial stage of word recognition or checking input against predictions. In other words, we did not find evidence that listeners attending to the pronunciation shi^L also make forward predictions that they might hear its allophonic variant shi^R; likewise, we did not find evidence that listeners hearing shi^R entertain the possibility that it could be an allophonic variant of shi^L. How can these findings be reconciled with what is already known about the processing of phonological variability?

4.1 Comparison with previous studies

These results are in fact consistent with some recent results calling into question the notion of prediction of words' surface forms. It was once widely accepted that the parser makes graded forward predictions of phonetic forms. The strongest evidence for this was a study by DeLong et al. (2005), which showed that the amplitude of ERPs elicited by the English articles *a* or *an* in the time-window associated with N400 was correlated with the likelihood that that particular article will appear in that position in that sentence. Since *a* and *an* are not content words, and the distinction between them is driven purely by the form of the following content word (e.g., *a kite* vs. *an airplane*), evidence that readers are sensitive to the likelihood of *a* or *an* is also evidence that they had made predictions about the form of the following word (specifically, whether the first segment is a vowel or a consonant). Recently, however, this account has been challenged. In particular, Nieuwland et al. (2018) attempted a large-scale replication of the study by DeLong et al. (2005), but were unable to convincingly replicate the abovementioned finding.[6] Since the publication of that study there have been many more studies attempting to investigate this form of gradient prediction of forms, with mixed results. Overall, then, while the notion of gradient prediction of phonological form is not dead, it has gone from being accepted orthodoxy (after DeLong et al., 2005) to hotly contested (after Nieuwland et al., 2018). It is in this sense that our study, which failed to find evidence that people predict alternative pronunciations of Mandarin morphemes, seems broadly consistent with this literature which has struggled to find evidence that people predict the alternative pronunciation of an article such as English *a* (albeit in a sentence context, which is different than the paradigm used in the present study). And from the perspective of bottom-up activation as opposed to top-down prediction, a companion study of ours, using auditory long-lag priming instead of character–audio paired priming as in the present study, also failed to find evidence that previous activation of one variant of a Mandarin morpheme facilitates later activation of another variant of the same morpheme (Politzer-Ahles et al., 2019b). Overall, then, the results of the present study fit with similar recent studies.

On the other hand, however, it is clear that knowledge about phonological alternation must play some role, at some point during comprehension. As discussed in the Introduction, we already know, based on basic phonological evidence, that Mandarin speakers must have a mechanism for interpreting a Rising tone as a variant of Low tone. However, this may take place at a different stage of processing than what was probed in the present study; the present study examines initial processing of a syllable at the moment it is heard, whereas the sorts of metalinguistic judgments described above may be occurring at a later, post-lexical stage.

There is also ample evidence from other paradigms that knowledge about alternations should influence processing. The study by Gaskell and Marslen-Wilson (1996), described above, suggests that phonologically altered forms activate their corresponding underlying representations better when

they occur in a context that licenses the alternation, compared to when they do not. This, however, may also be tapping in to a later stage of processing than what we observed in the present study, given that the present study measures activation by examining ERPs elicited by the altered word itself, whereas Gaskell and Marslen-Wilson (1996) measure it by examining reaction times to a later-presented word. Some ERP evidence for sensitivity to phonological knowledge, albeit from a very different paradigm, comes from a study by Lawyer and Corina (2017). They had participants listen to English words in which the prefix *in-* may have been appropriately assimilated to the following context (e.g., *i[m]possible*) or inappropriately altered to a wrong place of articulation (e.g., *i[m]conceivable*). They found a less robust violation response to these than they did for the prefix *un-*, which they argue to be consistent with the claim that the place of articulation of the consonant in *in-*, but not *un-*, is underspecified, and thus is more tolerant of variability (i.e., the parser is more likely to allow a wider range of inputs to activate the underlying representation of *in-*). Another piece of ERP evidence for the role of phonological knowledge, this time in Mandarin tone specifically, comes from Politzer-Ahles et al. (2016). Their study showed, using the mismatch negativity ERP component elicited in a passive oddball paradigm, that Mandarin Low tone has a qualitatively different mental representation than other tones do, and that this effect seems to be due to language-specific knowledge (as Low tone only elicited different brain responses than other tones among fluent Mandarin speakers, not among non-Mandarin-speaking controls). This language-specific knowledge is likely to be knowledge about this tone's involvement in a phonological alternation. All of these studies provide evidence that phonological knowledge matters at some stage of online processing.

The counterintuitive situation we are left with, then, is that there appears to be psycholinguistic and neurolinguistic evidence for both the importance and the unimportance of phonological knowledge. There is evidence that abstract phonological knowledge plays a role at a very early stage (the mismatch negativity, for instance, occurs earlier than the N400 that was examined in the present study) and at a very late stage (offline, end-state responses). On the other hand, there is also evidence that abstract phonological knowledge does *not* play a role in at least some N400 studies. How can these both be true?

4.2 How to reconcile these seemingly contradictory findings?

We would like to suggest that these results highlight the importance of viewing event-related potential components in terms of the cognitive *processes* that underlie them, instead of in terms of linguistic *phenomena*. Much neurolinguistics research, particularly in the early days of the field but still today as well, has focused on trying to find electrophysiological correlates of linguistic concepts, such as syntax, phonology, pragmatics, etc. When it became clear that the brain does not work that way, there began to be much research looking into more specific linguistic primitives or phenomena: rather than looking for neural correlates of, for example, syntax, phonology, and pragmatics, we would instead look for correlates of things such as subject–verb agreement, sonority sequencing violations, or scalar implicatures (e.g., Politzer-Ahles et al., 2013). These, however, are not the primitives that the brain uses, and thus we will never find a brain response that is uniquely sensitive to one of these linguistic concepts. In other words, the processes and representations of interest in linguistics are *ontologically incommensurable* with the processes and representations going on in the brain (Poeppel & Embick, 2005).

Rather than viewing brain activations as direct reflection of certain linguistic operations, we can view them as instruments to probe when or how efficiently certain cognitive operations occur (van Berkum, 2010); this requires detailed understanding of the *mapping* (Poeppel, 2012) between basic neurocognitive operations and the linguistic phenomena that emerge from them. (For a recent proposal for what this mapping may look like in phenomena that elicit negative-going components in language processing, see Bornkessel-Schlesewsky & Schlesewsky, 2019.) What this means, then, is that ERP

components cannot be treated as an insight into phonological alternation itself, but only as insight into the steps that a person takes to process that alternation *in the context of the task they are performing*.

Take, for instance, the observation that abstract knowledge about alternation appears to play a role in the study by Politzer-Ahles et al. (2016) and not the present study. This might occur because that study used a task that examines how people take an incoming signal and eventually encode it into a more abstract representation, whereas the present study examines how people make forward predictions for upcoming forms or how they take incoming signals and try to match these signals to a specific predicted representation, with no need for any abstraction. From this perspective, it makes sense that knowledge about alternation may play a role in one of these operations and not another.

4.3 Alternative explanations

Another potential explanation for the present results could come from frequency of a syllable's allomorphs. While the Rising-tone form of an underlyingly Low-tone syllable is indeed a legal variant, it might not be a very frequent variant; thus, participants may not have expected it. This may vary across words; there are some Mandarin Low-tone syllables that occur in sandhi-licensing more frequently than other contexts, and others that rarely occur in sandhi-licensing contexts. While the present study did not involve a wide enough selection of critical syllables to test this systematically, a purely frequency-based account might predict the attenuation of the mismatch response to be related to how frequently a given character occurs in sandhi-licensing contexts.

Yet another possible explanation involves the role of conscious, attentional processes. The above discussions of participants' processing of the stimuli in this study assume that participants are making active forward predictions about what form they expect to hear. This, however, may not be an accurate model of how participants complete the task involved in this study. Rather than expecting to hear a certain word, participants may instead be generating a model of a certain sound and then comparing the input to that model (e.g., if they see a character whose pronunciation is shi^L, they hold the pronunciation shi^L in their mind and, regardless of whether they actually expect to encounter that pronunciation, they compare the incoming stimulus to that mental model in order to decide which button to press). Under this account, the fact that Low-tone syllables are sometimes realized as Rising-tone may not matter if participants are not conscious enough of that fact to actively think of the Rising-tone variant and hold that pronunciation, along with the Low-tone pronunciation, in memory while preparing to compare the incoming stimulus to it.

Finally, it should also be noted that there are multiple accounts of the mechanisms through which prediction occurs. While much of the preceding discussion assumes that the objects of interest are predictions themselves, there are other well-articulated accounts that focus mainly on prediction errors (differences between predicted input and actually encountered input), such as the predictive coding framework (see, e.g., Bornkessel-Schlesewky & Schlesewky, 2019, on how prediction errors can account for negative-going ERP components) and the discriminative learning framework (see, e.g., Nixon, 2020; Ramscar et al., 2010). These frameworks may offer additional valuable insight into both how variation in Mandarin Low tone is dealt with, and specifically how it impacts ERPs. Examining Mandarin tone sandhi through these frameworks would be a valuable direction for future study.

5 Conclusion

The present study examined whether people hearing a morpheme that does not exactly match their prediction, but is a viable alternative pronunciation of the predicted morpheme, treat this as a less serious mismatch than a morpheme that has no systematic phonological relationship with the one they predicted. We found little evidence for this. Rather, hearing a different pronunciation of the

expected word elicited just as much of a mismatch response as hearing a wrong word. These results suggest that participants do not use their knowledge about phonological alternation at the very early stages of prediction or lexical activation.

ORCID iDs

Stephen Politzer-Ahles iD https://orcid.org/0000-0002-5474-7930

Jueyao Lin iD https://orcid.org/0000-0003-1530-1597

Supplemental material

Supplemental material for this article is available online.

Notes

1. Transcriptions are given in Hanyu Pinyin, the Romanization system used for Mandarin within the People's Republic of China, but with tones indicated using superscript letters rather than diacritics or superscript numbers.
2. Whether the tone that a person would expect would be ni^R, or some very similar (but still phonologically different) tone, is a point of controversy. There are slight acoustic differences between the Rising tone that is derived via tone sandhi and the Rising tone that is realized from phonologically (underlyingly) specified Rising tone (Peng, 2000; Zhang & Lai, 2010), that is, it is a case of incomplete neutralization in production. Most perception research into this topic has concluded that listeners cannot reliably perceive this difference (e.g., Peng, 2000), but there are some results suggesting that they sometimes can (Politzer-Ahles et al., 2019a), in which case the neutralization may be incomplete in perception as well. Phonologically, some analyses claim that that sandhi-derived Rising tone is a different phonological category than true Rising tone (Yuan & Chen, 2014), whereas others do not (Durvasula & Du, 2020). Given that the differences between sandhi-derived and non-sandhi-derived Rising tones are very subtle and listeners have difficulty discriminating them (even in Politzer-Ahles et al. (2019a), participants were barely above chance), we treat them as the same here, although we acknowledge that this decision is controversial. Nevertheless, given that participants often cannot tell the difference between these tones, our predictions for the present study would remain the same even if we treat sandhi-derived Rising and non-sandhi-derived Rising tones as belonging to different, but perceptually almost indistinguishable, phonological categories.
3. In some studies, these results are discussed in light of a putative "Phonological Mapping Negativity" component instead of or in addition to the N400. We do not distinguish these here, as we are not convinced that these are qualitatively different components of brain activity (see e.g., Lewendon et al., 2020), and because the way they are typically distinguished in most recent studies (by analyzing a slightly earlier and slightly later time window and referring to one as "PMN" and another as "N400") is not a reliable way to distinguish different components, because of issues such as component overlap (Luck, 2005).
4. Note that word recognition, and thus potential effects of tone, may occur much earlier than the N400 (see, e.g., MacGregor et al., 2012) . However, for the purposes of the present study, we are only interested in examining the mismatch response that occurs when people hear a syllable with an unexpected tone, and particularly looking into whether or not that mismatch response is attenuated when the unexpected tone is phonologically-related to the expected one. Therefore, we chose to focus on the N400 because previous literature on Mandarin suggests that the N400, rather than earlier potential indices of lexical processing, is where that "mismatch response" occurs in the present paradigm.
5. Calling this a "total mismatch" condition is a slight oversimplification because in some cases the expected and heard syllables may have overlapped in more subtle aspects such as syllable structure, or

the presence or absence of distinctive features such as voicing, vowel height, etc.—these were not carefully controlled. However, we had no a priori expectation that these would elicit substantial priming on the N400, and this condition is only a manipulation check that is not crucial to the main comparison in the experiment (see Analysis below); we were confident that this mismatch condition would still trigger large N400 effects even in the face of some potential overlaps of the sort described here.

6. While the study by Nieuwland et al. (2018) is often cited as failing to find evidence for prediction, that is not strictly true; Nieuwland and colleagues found suggestive evidence for a very small prediction effect, but this was significantly smaller than the effect observed by DeLong et al. (2005), and in fact was so small that it could not be detected to a reasonable standard of statistical significance even with a sample of over 300 participants. Thus, the main take-home message from their study is that the effect of gradient form prediction, if it exists at all, may be so small that it cannot be meaningfully studied without truly massive sample sizes. The effect may be larger than Nieuwland and colleagues described, however, if analyzed in a different way (Yan et al., 2017).

References

Boersma, P., & Weenink, D (2018). Praat: doing phonetics by computer [Computer program]. http://www.praat.org/

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2019). Toward a neurobiologically plausible model of language-related, negative event-related potentials. *Frontiers in Psychology*, *10*, 298. https://doi.org/10.3389/fpsyg.2019.00298

Chandrasekaran, B., Krishnan, A., & Gandour, J. (2007). Mismatch negativity to pitch contours is influenced by language experience. *Brain Research*, *1128*, 148–156. https://doi.org/10.1016/j.brainres.2006.10.064

Chen, J., Chen, T., & Dai, Y. (2012). Cognitive mechanisms and locus of tone sandhi during Chinese spoken word production. Talk presented at Tonal Aspects of Languages-2012, paper S2-01.

Chen, M. (2000). *Tone sandhi: patterns across Chinese dialects*. Cambridge University Press.

Chuang, C., & Hiki, S. (1972). Acoustical features and perceptual cues of the four tones of standard colloquial Chinese. *Journal of the Acoustical Society of America*, *52*(1), 146. https://doi.org/10.1121/1.1981919

DeLong, K., Urbach, T., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121. https://doi.org/10.1038/nn1504

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Duanmu, S (2000). *The phonology of Standard Chinese*. Oxford University Press.

Durvasula, K., & Du, N. (2020). Phonologically complete, but phonetically incomplete. *Proceedings of the 18th Conference on Laboratory Phonology*.

Gaskell, G., & Marslen-Wilson, W. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(1), 144–158. https://doi.org/10.1037//0096-1523.22.1.144

Groppe, D., Urbach, T., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology*, *48*(12), 1711–1725. https://doi.org/10.1111/j.1469-8986.2011.01273.x

Ho, A., Boshra, R., Schmidtke, D., Oralova, G., Moro, A., Service, E., & Connolly, J. (2019). Electrophyisiological evidence for the integral nature of tone in Mandarin spoken word recognition. *Neuropsychologia*, *131*, 325–332. https://doi.org/10.1016/j.neuropsychologia.2019.05.031

Kuo, Y., Xu, Y., & Yip, M. (2007). The phonetics and phonology of apparent cases of iterative tone change in Standard Chinese. In Aditi Lahiri (Ed.), *Phonology and phonetics: Tones and tunes*, Volume 2 (pp. 212–237). Mouton de Gruyter.

Lau, E., 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

Lawyer, L., & Corina, D. (2017). Putting underspecification in context: ERP evidence for sparse representations in morphophonological alternations. *Language, Cognition and Neuroscience*, *33*(1), 50–64. https://doi.org/10.1080/23273798.2017.1359635

Lewendon, J., Mortimore, L, & Egan, C. (2020). The Phonological Mapping (Mismatch) Negativity: history, inconsistency, and future direction. *Frontiers in Psychology*, *11*, 1967. https://doi.org/10.3389/fpsyg.2020.01967

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. https://doi.org/10.1016/j.brainres.2008.05.031

Luck, S. (2005). *An introduction to the event-related potentials technique*. MIT Press.

MacGregor, L., Pulvermüller, F., van Casteren, M., & Shtyrov, Y. (2012). Ultra-rapid access to words in the brain. *Nature Communications*, *3*, 711. https://doi.org/10.1038/ncomms1715

Makeig, S., Bell, A., Jung, T., & Sejnowski, T. (1996). Independent component analysis of electroencephalographic data. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in neural information processing systems*, Volume 8 (pp. 145–151). MIT Press.

Malins, J., & Joannisse, M. (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

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Nieuwland, M., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D., Rousselet, G., Ferguson, H., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Matthew Husband, E., Donaldson, D., Kohút, Z., Rueschemeyer, S.-A., & Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468. https://doi.org/10.7554/eLife.33468

Nixon, J. (2020). Of mice and men: Speech sound acquisition as discriminative learning from prediction error, not just statistical tracking. *Cognition*, *197*, 104081. https://doi.org/10.1016/j.cognition.2019.104081

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 156869. https://doi.org/10.1155/2011/156869

Peng, S. (2000). Lexical versus 'phonological' representations of Mandarin sandhi tones. In Broe, M. B. & Pierrehumbert, J. (Eds.), *Language acquisition and the lexicon: Papers in laboratory phonology V*, (pp. 152–167). Cambridge University Press.

Poeppel, D. (2012). The maps problem and the mapping problem: Two challenges for a cognitive neuroscience of speech and language. *Cognitive Neuropsychology*, *29*(1–2), 34–55. https://doi.org/10.1080/02643294.2012.710600

Poeppel, D., & Embick, D. (2005). Defining the relation between linguistics and neuroscience. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 103–120). Lawrence Erlbaum.

Politzer-Ahles, S., Connell, K., Pan, L., & Hsu, Y. (2019a). Mandarin third tone sandhi may be incompletely neutralizing in perception as well as production. *19th International Congress of Phonetic Sciences*, Melbourne, Australia. https://icphs2019.org/icphs2019-fullpapers/pdf/full-paper_290.pdf

Politzer-Ahles, S., Fiorentino, R., Jiang, X., & Zhou, X. (2013). Distinct neural correlates for pragmatic and semantic meaning processing: An event-related potential investigation of scalar implicature processing using picture-sentence verification. *Brain Research*, *1490*, 134–152. https://doi.org/10.1016/j.brainres.2012.10.042

Politzer-Ahles, S., Pan, L., Lin, J., & Lee, L. (2019b). Cross-linguistic differences in long-lag priming. Poster presented at 2nd Hanyang International Symposium on Phonetics and Cognitive Sciences of Language, Seoul, South Korea, May 24–25, 2019. https://site.hanyang.ac.kr/documents/379191/572223/HISPhonCog2019_P52_Politzer-Ahles_etal.pdf/996c18d6-4559-4a35-bd10-a3dbd156afff

Politzer-Ahles, S., Schluter, K., Wu, K., & Almeida, D. (2016). Asymmetries in the perception of Mandarin tones: Evidence from mismatch negativity. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(10), 1547–1570. https://doi.org/10.1037/xhp0000242

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*(6), 909–957. https://doi.org/10.1111/j.1551-6709.2009.01092.x

Schirmer, A., Tang, S., Penney, T., Gunter, T., & Chen, H. (2005). Brain responses to segmentally and tonally induced semantic violations in Cantonese. *Journal of Cognitive Neuroscience*, *17*(1), 1–12. https://doi.org/10.1162/0898929052880057

Sharma, B., Liu, C., & Yao, Y. (2015). Perceptual confusability of Mandarin sounds, tones and syllables. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. University of Glasgow. Paper number 0890. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0890.pdf.

Speer, S., & Xu, L. (2008). Processing lexical tone in third-tone sandhi. *Talk presented at 11th Laboratory Phonology conference*, Wellington, New Zealand.

van Berkum, J. (2010). The brain is a prediction machine that cares about good and bad. *Rivista di Linguistica*, *22*(1), 181–208. http://hdl.handle.net/11858/00-001M-0000-0012-C6B0-9

Yan, S., Kuperberg, G., & Jaeger, T. (2017). Prediction (or not) during language processing. A commentary on Nieuwland et al. (2017) and DeLong et al. (2005). *bioRXiv* preprint. https://www.biorxiv.org/content/10.1101/143750v1.full

Yuan, J., & Chen, Y. (2014). 3rd tone sandhi in Standard Chinese: A corpus approach. *Journal of Chinese Linguistics*, *42*(1), 218–237.

Zhang, J., & Lai, Y. (2010). Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology*, *27*(1), 153–201. https://doi.org/10.1017/S0952675710000060