Chapter 5

Tone sandhi cues modulate comprehenders' lexical expectations, but to a limited extent

5.1 Introduction

Chapter 3 provided some evidence that listeners of Mandarin Chinese can use informative tone sandhi cues to generate lexical predictions in real-time. However, the size of prediction effects was small, and Bayesian analysis ultimately could not provide substantial evidence supporting such predictions. Therefore, the role of tone sandhi cues as an input to real-time lexical predictions appears rather small.

One possible explanation for the limited effect of tone sandhi cues on *real-time* lexical prediction is that comprehenders may simply lack sufficient time or cognitive resources to quickly compute such lexical predictions on the fly, as speech input unfolds at a fast pace. In the current chapter, therefore, I present experiments using variants of the cloze task (a.k.a. the sentence completion task) to investigate the influence of tone sandhi cues on comprehenders' expectations about upcoming linguistic material. By using

the sentence completion task, we are able to investigate listeners' predictions in a setting where they are not constrained by the speed of real-time speech input or demands of real-time comprehension. This approach enables us to determine the extent to which tone sandhi cues can affect comprehenders' expectations in an "offline" manner, when they are given unrestricted time and cognitive resources.

5.1.1 Cloze probability as a measurement of prediction

The sentence completion task, also known as the cloze task, is a well-established and widely used tool in psycholinguistic research for estimating the predictability of a word within a given context (Taylor, 1953). In this task, participants are typically presented with an incomplete sentence fragment and asked to provide the word or phrase they believe is most likely to complete the sentence. Lexical predictability in the sentence completion task is operationalized as *cloze probability*—defined as the proportion of participants who produce a specific target word as their continuation. Higher cloze probabilities indicate greater predictability of a word in a given linguistic context. The sentence completion task has been particularly valuable because it provides an intuitive and quantifiable way of assessing how well language comprehenders can use contextual cues to anticipate upcoming linguistic material.

According to the processing model of cloze responses proposed by Staub et al. (2015), cloze probability reflects the likelihood that a given candidate word accumulates the most activation and reaches the response threshold first. In this model, words that are more predictable in a particular context receive greater activation from the context. As a result, these words accumulate activation more quickly and are more likely to reach the response threshold earlier. Consequently, words with higher predictability exhibit higher cloze probabilities.

The sentence completion task can provide valuable insights into predic-

tive processes in language processing beyond providing a measure of lexical predictability. According to Staub et al. (2015), the underlying processes involved in generating a cloze response may closely parallel those involved in generating a lexical prediction. In both cases, contextual information preactivates likely words, facilitating either their selection as a response (in the sentence completion task) or their processing during real-time comprehension. As such, the sentence completion task can be understood as an "offline" prediction measure: participants are not constrained by the continuous real-time language input where a target can be presented before the comprehender has time to generate a prediction. In the sentence completion task, participants are able to consider the context without the time constraints, freely activating likely candidates and ultimately producing the word that reaches the activation threshold first—that is, the most predictable word given the context, regardless of how much time this word needs to reach the activation threshold.

5.1.2 Phonological cues in the sentence completion task

Phonological information can change the predictability of upcoming lexical items. For example, the English indefinite article has two distinct phonological forms (*a* and *an*), with their selection determined by the onset of the following word: *a* precedes a consonant, while *an* precedes a vowel. Consequently, the phonological form of the article constrains the phonological properties of the upcoming word, narrowing the set of possible lexical items that can follow. Similar phonological constraints are observed in other languages, such as the French singular definite article (*le*/*la* before a consonant vs. *l'* before a vowel) and its Italian counterpart (*il*/*la* vs. *l'*). Other examples include Mandarin Chinese tone sandhi and Kansai Japanese pitch accent sandhi (see Chapter 1 Sections 1.2.1.2 and 1.4.2).

When such informative phonological cues are present in a sentence fragment, one may ask whether they assert any influence on cloze responses.

One simple possibility, the null hypothesis, is that informative phonological information simply does not have any influence on cloze responses, although this is not highly likely, given some initial evidence that comprehenders may be able to use phonological cues to generate lexical predictions even in real-time tasks (A. Ito & Hirose, 2025; Liu et al., 2025) (also see Chapter 3), suggesting that phonological cues should at least have some effect on how comprehenders expect upcoming language.

If phonological cues *can* affect cloze probability, its impact can still take different forms. One possibility, what we here call the Strong Influence Hypothesis, is that comprehenders only generate continuations compatible with the phonological cue. For example, in English, a consonant-initial word cannot follow *an*, and a vowel-initial word cannot follow *a*. In this sense, the phonological cue imposes a constraint on the upcoming word: lexical items incompatible with the cue are not phonologically "grammatical" continuations. Accordingly, when presented with a fragment ending in *an*, for instance, comprehenders may only respond with words that begin with a vowel such as *umbrella* or *apple*.

A second possibility, what we call the Weak Influence Hypothesis, is that phonological cues increase the activation of compatible words, thereby raising their cloze probabilities, but do not necessarily exclude or inhibit incompatible words. This hypothesis would predict that following *an*, for instance, while words like *umbrella* and *apple* would become more likely responses, some participants might still respond with incompatible words such as *raincoat* or *banana*, albeit at lower rates.

Although sentence completion tasks have not been widely used in the study of phonological information in lexical prediction, prior research using sentence completion tasks to assess stimulus predictability provides valuable insights. For instance, Martin et al. (2013) reported two cloze probability tests with Spanish-English bilinguals as part of their stimulus development process for an ERP experiment. In the first test, participants completed

constraining sentence fragments truncated before a target noun phrase (e.g., "*She has a nice voice and always wanted to be…*"). Sentences with a high-cloze response (e.g., "*…a singer*", mean cloze probability = 69%) were selected for a second test.

In the second test, participants read the same sentences, this time the sentence fragment also included an English article mismatched with the highest-cloze noun (e.g., "*She has a nice voice and always wanted to be* ***an****…*"). The cloze probability of the previously expected noun ("*singer*") dropped drastically (mean cloze probability = 3.5%), while the cloze probability of some unexpected but phonologically compatible words (e.g. "*artist*") increased, to 37.4% from 0.8% in the first test. The drastic change in cloze probability following the introduction of an indefinite article suggest that phonological cues can have substential influence on comprehenders' cloze responses, that is to say on their expectations about upcoming words.

However, the tests by Martin et al. (2013) leaves some questions unanswered. Firstly, it remains unclear whether the observed modulation in cloze responses depends on the modality in which the stimuli were presented. In the written sentence completion task, the critical phonological cue (*a*/*an*) was encoded in their distinct orthographic forms, which may have heightened the cue's salience. Additionally, written tasks typically allow participants to read and reread the sentence fragments freely, potentially reinforcing the effect of the informative cue by providing multiple opportunities for processing. As a result, it is uncertain whether the effect of informative phonological cues observed in the written modality would extend to spoken language processing, where phonological cues are presented transiently and cannot be revisited.

Secondly, Although the drastic decrease of the most expected word's probability (*singer* from 69% to 3.5%) seem to align with the Strong Influence Hypothesis, Martin et al. (2013) only reported data for a limited set of responses in the second test—the prior most-expected word ("*singer*", 3.5%)

and the specific unexpected word used in their ERP study ("*artist*", 37.4%). The authors did not report what other responses were produced by participants (the rest 59.1% of responses), or whether those began with vowels or consonants. As a result, it remains whether these data more closely align with the strong or the Weak Influence Hypothesis.

To address these limitations, the present study employs auditory versions of the sentence completion task, in which participants respond to spoken sentence fragments. Furthermore, we analyze the full range of participant responses to determine precisely how phonological cues impact cloze responses. Across two experiments, participants listen to incomplete sentences ending with a phonologically informative syllable in an otherwise non-constraining context. According to the Strong Influence Hypothesis, nearly all cloze responses should be phonologically compatible with the cue, producing a ceiling effect. By contrast, under the Weak Influence Hypothesis, compatible responses should increase in probability, but incompatible responses may still occur.

5.1.3 Mandarin Chinese tone sandhi in the sentence completion task

5.1.3.1 From real-time predictions to offline predictions

In Chapter 3 Section 3.6, I summarized the findings regarding the subtle role of tone sandhi cues as input to *real-time* lexical prediction. Given these findings, an important remaining question is whether this subtle effect is primarily due to the rapid pace of natural speech input, or whether it reflects a more fundamental limitation in how tone sandhi cues effect lexical predictability. In other words, if it is difficult to find evidence for tone-sandhi-based real-time lexical predictions, is this because listeners cannot compute such predictions rapidly enough in real time, or because tone sandhi cues do not strongly influence which words are likely, regardless of the time or cognitive resources available to the listener?

To further investigate the role of Mandarin Chinese tone sandhi in lexical prediction and to determine whether the apparent inability to use them to predict stems from the time constraints of real-time language input, in the current experiment, we use the sentence completion task to test lexical prediction in an "offline" manner, in which participants produce continuations of sentence frames without time pressure, allowing ample time to adjust their expectations.

5.1.3.2 Base tone vs. sandhi tone

Chapters 2 and 3 revealed that for the T3 sandhi, prediction effects are notably stronger when listeners hear the sandhi tone T2, compared to the base tone T3. In two out of three experiments (Experiment 1 in Chapter 2, and Chapter 3), separate analyses revealed that prediction effects emerged only when participants heard the sandhi tone T2; while no such effects were observed when they heard the base tone T3[1]. Interestingly, this pattern does not hold for the *yi* sandhi: across all three eye-tracking experiments in Chapters 2 and 3, analyses comparing trials with *yi4* and those with *yi2* revealed no significant prediction effects for either type of trial[2], see Appendix B.

These findings suggest that, at least for the T3 sandhi, the base tone (T3) and the sandhi tone (T2) may play distinct roles in prediction. One explanation is that the sandhi tone T2 poses a stronger constraint on upcoming language input than the base tone T3: hearing T2 signals that the next syllable must be T3, whereas hearing the base tone T3 only indicates that the next syllable is not T3—leaving much greater uncertainty about the tone of the upcoming syllable. Accordingly, a sandhi tone T2 may be more readily used to generate lexical predictions.

A similar asymmetry in the strength of constraint is present in the *yi* sandhi: *yi2* constrains the following syllable to T4, while *yi4* indicates any

[1] In Experiment 3 of Chapter 2, no prediction effects were found for either type of trial.

[2] Although in Chapter 3, combining both trial types yielded a significant overall prediction effect. The absence of effects upon separate analysis could stem from reduced statistical power.

tone other than T4. Nevertheless, our experiments did not find differences in predictive eye movements between hearing *yi2* and hearing *yi4*.

In Chapter 2, Section 2.2.3, I discussed a hypothesis that using a *yi* sandhi cue to generate real-time lexical predictions may be more difficult than using a T3 sandhi cue, as a result of the difference in frequency between the two sandhi patterns. Additionally, the less constraining base tones may generally be less readily used for prediction than their more constraining sandhi counterparts. These two factors could jointly influence how readily used each type of cue is, yielding possible hierarchies such as: T2 < T3 < *yi2* < *yi4*, or T2 < *yi2* < T3 < *yi4*. Either scenario is compatible with our eye-tracking results: under the naturalistic speech rate used in Chapters 2 and 3 (about four syllables per second), prediction effects based on the sandhi tone T2 was the most notable, while effects based on the other three cues were much subtler.

However, results from the acceptability judgment tasks in Chapters 2 and 3 indicate that the role of base versus sandhi tones might be more fundamental than their effectiveness in real-time. In the acceptability rating experiments presented in Chapters 2 and 3, we found reduced sensitivity to T3 sandhi violations in sandhi-inducing contexts. Specifically, listeners found consecutive T3-T3 sequences more acceptable than other types of tone sandhi violations (e.g., producing a sandhi T2 when not followed by a T3). For the *yi* sandhi, a much smaller asymmetry was found: while listeners were somewhat more sensitive to violations in the sandhi-inducing context (*yi4*-T4) versus the non-sandhi context (*yi2*-T1/2/3), this difference was considerably less pronounced than for T3 sandhi.

The reduced sensitivity to T3-T3 sequences suggests listeners may be less certain about what can or cannot follow a T3 syllable. When a T3 is heard, listeners might continue to accept the possibility of another T3, making the base tone T3 a poor predictive cue regardless of how much time and resources is available for processing. Again, this is compatible with our eye-

tracking results: listeners predicted based on a T2 as its presence significantly alters the probability of the upcoming syllable's tone, while they did not predict based on a T3 possibly because all tones (including another T3) remain likely for the upcoming syllable. Nonetheless, another finding from Chapters 2 and 3 suggests that this possibility may be less likely: there was no correlation between the extent to which listeners were sensitive to T3-T3 sequences and the extent to which they used T3 sandhi to predict.

In sum, the observed differences between base tones and sandhi tones in real-time lexical prediction could either stem from a difference in how readily used these cues are in real-time prediction, or a more fundamental issue that in the case of the T3 sandhi, only the sandhi tone T2 alters lexical predictability, while the base tone T3 does not, or at least to a much smaller extent. In the current study, therefore, we also aim to reconcile the different roles played by base tones and sandhi tones in prediction. By including Tone (base vs. sandhi) as a factor during data analysis, we aim to determine whether base tones and sandhi tones differently affect lexical predictability.

5.1.4 The present study

In the present study, we employ the sentence completion task to examine how T3 sandhi and *yi* sandhi cues influence comprehenders' expectations about upcoming words.

In Experiment 1, we conducted both an auditory sentence completion task (experimental condition) and a written sentence completion task (control condition). In the auditory task, participants listened to unconstraining sentence frames ending either with the numeral *yi* (undergoing the *yi* sandhi) or with *liang* (undergoing the T3 sandhi). The critical syllable—*yi* or *liang*—was presented either in the base tone (*yi4* or *liang3*) or in the sandhi tone (*yi2* or *liang2*). In the written task, we capitalised on the fact that tone sandhi is not encoded in Chinese characters, and presented the same sentence frames as written stimuli to a separate group of participants, providing no tone sandhi cues to constrain the set of possible upcoming words. Partic-

ipants were asked to produce the most likely continuation of each sentence—verbally in the auditory task and in writing in the written task. The dependent variable was whether participants' responses were compatible with the tone of the critical syllable.

Experiment 2 improved upon the design of Experiment 1 in several key ways. First, we introduced a more comparable control condition in which participants heard sentence frames without the critical syllable and thus without informative tone sandhi cues. Second, we replaced the free production task with a forced-choice task: participants listened to the sentence frame and then selected the better continuation from two visually presented options. This design minimized the efforts in generating candidates that were compatible with the sandhi cue. Third, we expanded the context in which the T3 sandhi was tested: in addition to the numeral *liang*, we included T3 adjectives such as *xiao3* ("small") or *lao3* ("old"), as well as the first syllable in T3-initial compound nouns such as *lü* in *lü3dian4* ("inn") versus *lü2dian4* ("hotel").

If tone sandhi cues have no influence on comprehenders' expectations, we should find no significant effects of the experimental manipulation (i.e. cue availability) on cloze responses. If tone sandhi cues influence comprehenders' expectations, we expect to find a significant effect of cue availability on the proportion of cue-consistent responses: the proportion of responses consistent with the critical syllable's tone should be significantly higher in the experimental condition than in the control condition. However, the Strong and Weak Influence Hypotheses predict different sizes of this effect: according to the Strong Influence Hypothesis, cue-consistent responses in the experimental condition should approach ceiling levels. In contrast, the Weak Influence Hypothesis predicts a higher proportion of cue-consistent responses in the experimental versus control conditions, but not necessarily at ceiling.

Given these complex picture of base versus sandhi tones in prediction,

the present study analyzes cloze response data with Tone (base vs. sandhi) included as a factor. If the distinction between base and sandhi tones reflects a fundamental difference in how each cue influences the probability of upcoming tones, we expect an interaction between cue availability and Tone for T3 sandhi: specifically, that an informative sandhi tone T2 will have a larger impact on cloze responses than the base tone T3. For the *yi* sandhi, by contrast, we expect any such interaction to be absent or considerably weaker, in line with the smaller difference in acceptability ratings.

Alternatively, if the difference between base and sandhi tones is primarily a matter of how readily used each cue is in real-time lexical prediction, no interaction between cue availability and Tone should be observed for either the T3 sandhi or the *yi* sandhi in the sentence completion tasks, due to the non-real-time nature of the task.

5.2 Experiment 1

In the current experiment, using a between-subjects design, we employed versions of the sentence completion task (a.k.a the cloze task) to investigate the effect of tone sandhi cues on lexical predictability. Participants in the experimental group listened to incomplete sentence frames and completed the sentences verbally. Meanwhile, participants in the control group *read* the same sentence frames and completed the sentences in writing. Crucially, in the experimental condition, the sentence frames ended with a numeral (*liang* or *yi*) that was informative about the upcoming syllable's tone, as a result of tone sandhi. In the control condition, in contrast, the end-of-sentence numeral was uninformative, because tone sandhi was not encoded in Chinese characters.

The main experimental items consisted of 40 non-constraining sentence frames. In addition, we included 40 filler items with constraining sentences where a specific word can be highly predictable, while the tone of the critical syllable (i.e. the numeral) either matched or mismatched with the ex-

pected noun (e.g. "*At Starbucks, Anne bought yi4/yi2…*", expected: *yi4 bei1 ka1fei1* "a cup of coffee"). The filler items provided an additional opportunity to investigate whether a tone sandhi cue that mismatched with an expected word can significantly decrease the cloze probability of that word, similar to Martin et al. (2013). The relevant analysis and results are presented in Appendix D.

In addition, we included a proportion of trials where the weakly or non-constraining sentence frame ended with the first syllable of a multisyllabic noun. Crucially, switching the lexical tone of this critical syllable (i.e. switching to another word with the same segments but different tones) results in a different set of possible target nouns that are equally plausible in the sentence context (e.g. "*Ming bought some snacks from the store, they were…* ***bing1****qi2lin2*/冰淇淋 'ice cream' vs. ***bing3****gan1*/饼干 'cookies' "). We call these trails lexical tone trials, as they were designed to verify that listeners were sensitive to *lexical tones* themselves, and would respond according to the tone of the critical syllables. As such, if no effects were found in tone sandhi trials, we can attribute the null effects to an insensitivity to tone sandhi, rather than a fundamental insensitivity to lexical tones. Notably, the critical syllables in these trials were equally informative for the experimental and the control group, as different versions of the critical syllable were in fact different words/morphemes and had different written forms. Therefore, for these trials, we expect that participants respond according to the tone of the critical syllable in the vast majority of trials, regardless of modality.

5.2.1 Methods

5.2.1.1 Participants

Participants were 39 native Mandarin Chinese speakers, in which 20 were allocated to the experimental group (19 female, age $M = 24.4$, $SD = 3.4$), while 19 were allocated to the control group (16 female, age $M = 21.1$, $SD = 3.3$).

5.2.1.2 Stimuli

A female native Mandarin Chinese speaker recorded the auditory stimuli. All recordings' volume was normalised. The complete set of experimental and filler items can be found in Appendix H.

Tone sandhi trials The stimuli for the tone sandhi trials comprised of 40 sentence fragments, each containing an unconstraining context and ending with a numeral (the critical syllable). We manipulated the stimuli in a 2 × 2 × 2 mixed design, crossing within-subject factors sandhi (T3 sandhi, *liang* vs. *yi* sandhi, *yi*), tone (base tone vs. sandhi tone), and a between-subject factor modality (auditory/experimental group vs. written/control group).

Since the critical syllable was a numeral, grammatical continuations should be a classifier and a noun, according to Chinese noun phrase structure. Crucially, the tone of the numeral poses restrictions on what classifiers might follow. For example, in the *liang*-sandhi (*liang2*) condition, the tone of the numeral *liang2* indicates that a T3 sandhi must have been triggered by a tone-3 (T3) classifier, making T3 classifiers the only type of phonologically grammatical continuations. Table 5.1 presents an example set of stimuli.

Lexical tone trials Stimuli for the lexical tone trials were 40 pairs of sentence fragments containing weakly or non-constraining context and ending with the first syllable of a target noun (the critical syllable). Two versions of each sentence fragment were created, differing only in the critical syllable: critical syllables of the two versions were words or morphemes that shared the same segments but differed in tone. As such, the two versions had different set of possible targets, although they were equally plausible in the sentential context. This set of stimuli was also manipulated in modality (auditory vs. written task), however, different from tone sandhi stimuli, critical syllables in the lexical tone trials were equally informative about the set of possible target words regardless of modality. Table 5.2 presents an example stimuli set.

Sandhi	Tone	Modality	Sentence fragment	Critical syllable	Compatible continuation(s)	Incompatible continuation(s)
T3 sandhi	base	auditory	*Lao3-Zhang1 jia1 ke4ting1 zhuo1zi shang4 bai3-zhe* English: "On the table in Zhang's living room, there is/are"	*liang3* "two"	T1/T2/T4	T3
		written	老张家客厅桌子上摆着	两	T1/T2/T3/T4	N/A
	sandhi	auditory	*Lao3-Zhang1 jia1 ke4ting1 zhuo1zi shang4 bai3-zhe*	*liang2*	T3	T1/T2/T4
		written	老张家客厅桌子上摆着	两	T1/T2/T3/T4	N/A
yi sandhi	base	auditory	*Lao3-Zhang1 jia1 ke4ting1 zhuo1zi shang4 bai3-zhe*	*yi4* "one"	T1/T2/T3	T4
		written	老张家客厅桌子上摆着	一	T1/T2/T3/T4	N/A
	sandhi	auditory	*Lao3-Zhang1 jia1 ke4ting1 zhuo1zi shang4 bai3-zhe*	*yi2*	T4	T1/T2/T3
		written	老张家客厅桌子上摆着	一	T1/T2/T3/T4	N/A

Table 5.1: An example set of stimuli used in Experiment 1, tone sandhi trials. See Appendix H for the complete set of stimuli.

Modality	Sentence frame	Critical syllable	Compatible continuation(s)
auditory	*Xiao3ming2 fang4xue2 hou4 zai4 xiao3 chao1shi4 mai3 ling2shi2, mai3de shi4* English: "After school, Ming bought some snacks from the shop, they were"	*bing1* (version a)/ *bing3* (version b)	*bing1qi2lin2* "ice cream", ... / *bing3gan1* "cookies", ...
written	小明放学后在小超市买零食，买的是	冰/ 饼	冰淇淋 "ice cream", ... / 饼干 "cookies", ...

Table 5.2: An example set of stimuli used in Experiment 1, lexical tone trials. See Appendix H for the complete set of stimuli.

Fillers and presentation lists Filler items were 40 highly constraining sentence frames ended with a numeral whose tone was either compatible with the highly expected noun or incompatible. A more detailed discription of the filler items and relevant data analysis are presented in Appendix D.

Four presentation lists were created, such that each participant only saw one version of each experimental item. Within each presentation list, participants were presented with 20 experimental items with the T3 sandhi, in which 10 contained a critical syllable in the base tone (*liang3*), while the other 10 contained a critical syllable in the sandhi tone (*liang2*). Similarly, each participant was presented with 20 experimental trials with the *yi* sandhi, in which 10 contained *yi4*, and 10 contained *yi2*. In addition, each participant was presented with 40 lexical tone trials, as well as 20 filler trials with the T3 sandhi (10 having *liang* in the expected tone and 10 having *liang* in the unexpected tone), in addition to 20 filler trials with the *yi* sandhi, 10 in the expected tone and 10 in the unexpected tone.

5.2.1.3 Procedure

The experiment is implemented using Psychopy 2022.2.5 for the experimental group (auditory cloze task) (Peirce et al., 2019). Participants sat in front of a 24inch monitor and listened to the auditory stimuli through a pair of Beyerdynamic DT 297 headphones. Participants' responses were recorded using the condenser microphone attached to the headphones. As illustrated in Figure 5.1, during each trial, a black fixation cross appeared on the screen with a white background. After 500ms, an audio recording of the sentence fragment began to play. The fixation cross stayed on the screen during stimulus presentation. As the auditory stimulus finished playing, a which question mark appeared on the screen while the screen background turned from white to grey, which prompted the participants to respond. At the same time, the microphone began to record participants' responses for 4000ms. Afterwards, the screen background turned back to white, and a written instruction appeared on the screen to ask the participant to press the space key at their

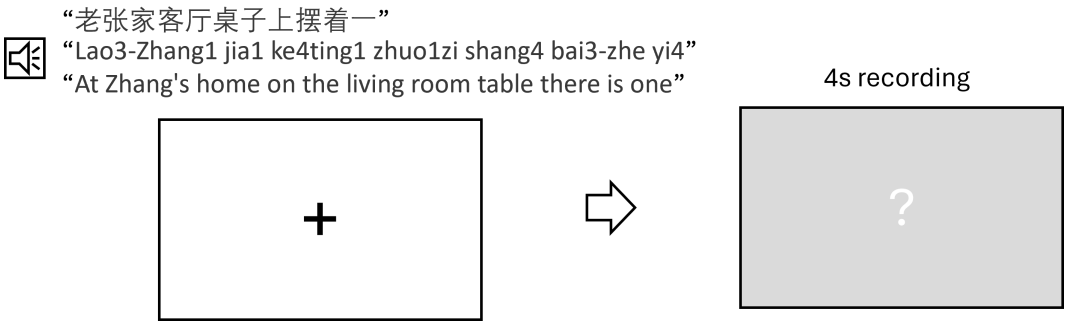

Figure 5.1: Experiment procedure as experienced by the experimental group (auditory cloze task), Experiment 1. Participants listened to incomplete sentence frames and were prompted to verbally complete the sentence by the presentation of a question mark as well as a change in the screen's background colour.

own pace to start the next trial.

For the control group (written cloze task), the experiment was hosted online on Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). In each trial, the participants were presented with the sentence frame written in Chinese characters, while they were asked to complete the sentence by typing Chinese characters in a text entry box below the sentence frame. After completing the sentence, the participant pressed a button or pressed the space key to start the next trial.

Each participant first completed a block of 40 lexical tone trials, followed by a block of 40 tone sandhi trials mixed with 40 filler trials. Participants were given a break every 20 trials. For the experimental group, at the beginning of each block, the participant was informed that they will hear sentence frames that ended with the onset of a noun (the lexical tone trials), or a numeral (the tone sandhi trials), repectively. This was designed to help participants correctly access the lexical item for the critical syllable, as some critical syllables have a large set of homophones (e.g. *yi2*/疑 “to doubt”; *yi4*/意 “meaning”, etc.). This instruction was not given to the control group (written task), as the Chinese characters unambiguously indicated the identity of the lexical item.

Following the main task, participants completed a questionnaire,

hosted on Gorilla Experiment Builder, about their experiences with Mandarin Chinese tone sandhi and their strategies during the auditory cloze task. The questionnaire first briefly introduced the concept of the T3 sandhi and the *yi* sandhi, then asked the participant to respond whether they were formally instructed about the tone sandhi in school, and whether they knew about the tone sandhi before reading the questionnaire. For the experimental group, the questionnaire additionally included questions asking whether the participant noticed that the critical syllable were sometimes pronounced in the base tone and sometimes in the sandhi tone, as well as whether the participant would agree that the tone of the critical syllable affected how they responded. A complete questionnaire can be found in Appendix I.

5.2.1.4 Coding of responses and data analysis

For the tone sandhi experience questionnaire, responses was categorized, and chi-squared tests were used to determine whether participants' experience differed significantly between the T3 sandhi and the *yi* sandhi.

For the experimental group (auditory task), two native speakers of Mandarin Chinese coded participants' responses in the experimental task. Coding was done on Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). Recordings were presented in a randomised order to each coder. For each response, the coders played the recording as many times as needed, and transcribed the speech they heard into pinyin. The coders were also asked to indicate whether the participants made a self-correction in each case. Only trials in which the two coders agreed on a transcription were included in data analyses. For the control group, participants' typed responses were converted into pinyin (phonological forms) using Python and the pypinyin package.

For experimental trials, transcribed responses were then coded for whether the response was compatible with the tone of the critical syllable or not (compatible = 1, incompatible = 0). For instance, if the participant heard *liang2* and responded with *ben3 shu1*, this response would be coded as compatible, whereas if they responded with *zhi1 bi3*, this would be incom-

patible. For the control group (written cloze task), the response was coded according to whether it was compatible or incompatible with the tone of the critical syllable as presented in the same item in the experimental group (auditory cloze task).

Coded responses were analysed in R 4.4.1 (R Core Team, 2019) using the lme4 package (Bates et al., 2015) and the emmeans package (Lenth, 2024). Logistic regression models were fitted to the coded responses to experimental items. The model included three sum-coded fixed factors: Sandhi (T3 vs. *yi*), Tone (base tone vs. sandhi tone), and cue availability or Modality (auditory vs. written), and their two-way and three-way interactions. Where an interaction was found, it was followed up by conducting an estimated marginal means analysis using the emmeans package.

For lexical tone trials, the responsed were coded by a native Mandarin Chinese speaker on two dimensions: (1) whether the response constituted a real word when combined with the heard critical syllable (e.g. *qi2lin2* following *bing1* makes up the word *bing1qi2lin2* "ice cream"), and (2) whether it constituted a real word when combined with the critical syllable of the other version (e.g. *qi2lin2* following *bing3* does not correspond to any real word). A response is considered a target response if it constitutes a real word when combined with the syllable they heard. Meanwhile, a response is considered an infelicitous response if it constitutes a real word when combined with the syllable in the unheared tone. Proportions of target and infelicitous responses were calculated for the experimental as well as the control group.

5.2.2 Results

5.2.2.1 Questionnaire results

Chi-squared tests suggested a significant difference between pariticipants' experiences with the *yi* sandhi and the T3 sandhi, such that compared to the T3 sandhi, more participants received formal instructions about the *yi*

	yi sandhi			T3 sandhi		
	yes	no	do not remember	yes	no	do not remember
explicit knowledge	13 (61.9%)	8 (38.1%)		14 (66.7%)	7 (33.3%)	
classroom experience	12 (57.1%)	5 (23.8)%	4 (19%)	14 (66.7%)	4 (19%)	3 (14.3%)

Table 5.3: Participants self-reported tone sandhi experience, Experiment 1. See Appendix I for the complete questionnaire.

sandhi ($X^2 = 20.31$, $p < 0.001$), and were aware of the *yi* sandhi's patterns ($X^2 = 7.29$, $p = 0.007$).

5.2.2.2 Sentence completion results

Logistic regression models revealed significant effects of Modality ($\beta = 0.24$, $SE = 0.07$, $z = 3.56$, $p = 0.0004$) and Tone (base vs. sandhi) ($\beta = 0.91$, $SE = 0.07$, $z = 13.74$, $p < 0.0001$). Further, there was a significant interaction between Sandhi and Tone ($\beta = -0.98$, $SE = 0.07$, $z = -14.86$, $p =< 0.0001 >$), as well as a three-way interaction among the factors ($\beta = -0.30$, $SE = 0.07$, $z = -4.57$, $p < 0.0001$).

Estimated marginal means (EMMs) using the emmeans package revealed that the effect of Modality was significant only in the sandhi tone for *yi* (*yi2*) ($\beta = 1.09$, $SE = 0.36$, $z = 30.3$, $p = 0.002$), and the base tone for *liang* (*liang3*) ($\beta = 1.08$, $SE = 0.21$, $z = 5.14$, $p < 0.0001$). No significant effect of Modality was found for base tone *yi* (*yi4*) ($\beta = 0.10$, $SE = 0.20$, $z = 0.50$, $p = 0.62$) and sandhi tone *liang* (*liang2*) ($\beta = -0.37$, $SE = 0.26$, $z = -1.43$, $p = 0.15$).

These results suggest that the proportion of phonologically grammatical responses were significantly higher in the experimental task than the control task, but only for trials involving *yi2* and *liang3*, not *yi4* or *liang2*. This indicates that listeners were sensitive to the constraints that tone sandhi poses on upcoming lexical items, but only in some cases.

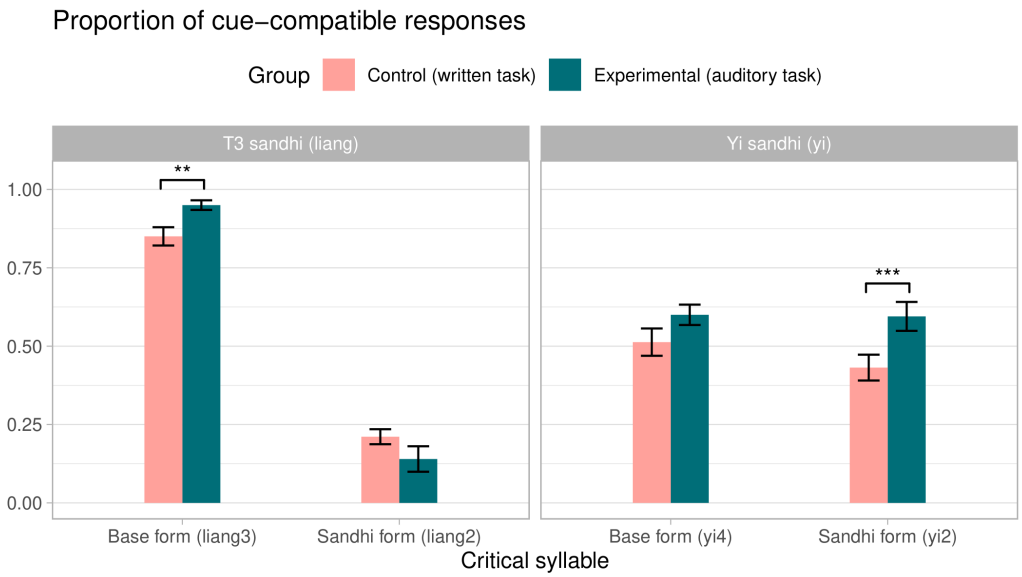

Figure 5.2: Proportions of cue-compatible responses. Error bars show standard errors of the means.

5.2.2.3 Excluding *ge4* responses

Since the critical syllable was always a numeral, participants almost always responded with a classifier and a noun. As the first analysis revealed an effect of Modality only for *yi2* trials and *liang3* trials, one potential explanation for this effect was that the general classifier *ge4*, which is compatible with *yi2* and *liang3*, can be more frequently used in spoken language than written language.

We therefore run an additional analysis excluding responses that began with the general classifier *ge4* from both the experimental group (auditory task, n = 334) and the control group (written task, n = 218) (34% of all trials) (Figure 5.3). Again, significant main effects of Modality ($\beta = 0.30$, $SE = 0.08$, $z = 3.96$, $p < 0.0001$) and Tone ($\beta = 1.17$, $SE = 0.08$, $z = 15.25$, $p < 0.0001$) were found, in addition to a significant interaction between Sandhi and Tone ($\beta = -0.21$, $SE = 0.08$, $z = -2.71$, $p = 0.007$), as well as a three-way interaction ($\beta = -0.21$, $SE = 0.08$, $z = -2.75$, $p = 0.006$).

Estimated marginal means (EMMs) revealed significant effects of Modality for base tone *yi* (*yi4*) ($\beta = 0.67$, $SE = 0.30$, $z = 2.26$, $p = 0.02$),

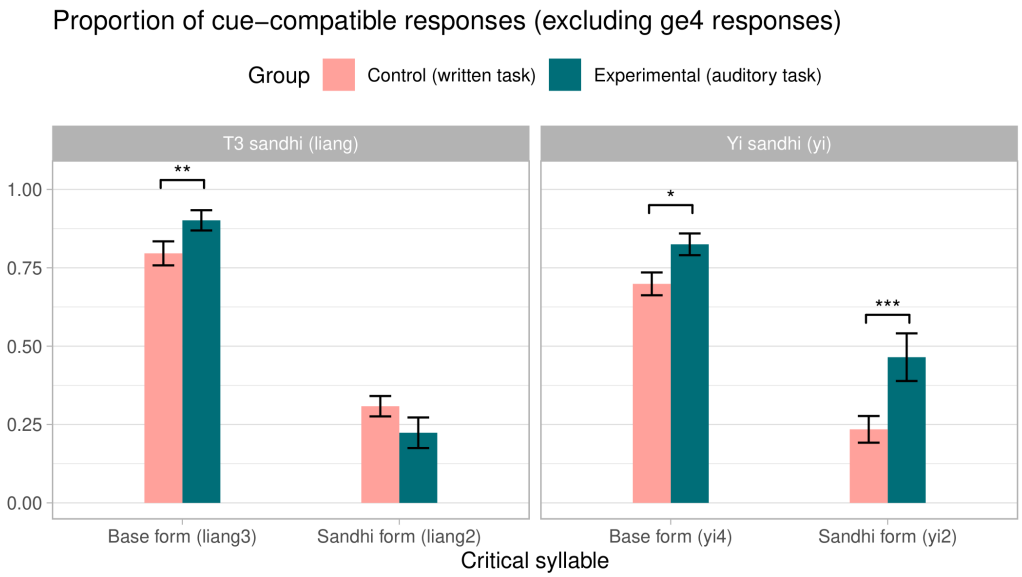

Figure 5.3: Proportions of cue-compatible responses, excluding responses containing the general classifier *ge4*. Error bars show standard errors of the means.

sandhi tone *yi* (*yi2*) ($\beta = 0.96$, $SE = 0.37$, $z = 2.60$, $p = 0.009$), and base tone *liang* (*liang3*) ($\beta = 1.09$, $SE = 0.27$, $z = 4.07$, $p < 0.0001$), but not for sandhi tone *liang* (*liang2*) ($\beta = -0.30$, $SE = 0.27$, $z = -1.09$, $p = 0.28$). In addition, an interaction between Modality and Tone was found for the T3 sandhi ($\beta = -0.32$, $SE = 0.12$, $z = -2.73$, $p = 0.006$), but not for the *yi* sandhi ($\beta = -0.1$, $SE = 0.10$, $z = -1.04$, $p = 0.29$). However, the interaction in T3 sandhi trials suggested that the effect of Modality was *larger* for the base tone than the sandhi tone, opposite from what we predicted if the base tone T3 was fundamentally a weaker predictive cue than the sandhi tone T2.

These results suggest that the proportion of phonologically grammatical responses were significantly higher in the experimental task than the control task for trials involving *yi4*, *yi2*, and *liang3*, but not *liang2*. This suggests that when the difference in classifier use between spoken and written language is controlled, listeners were generally more likely to produce phonologically grammatical (or sandhi-consistent) responses in the experimental task than the control task, except for trials containing *liang2*.

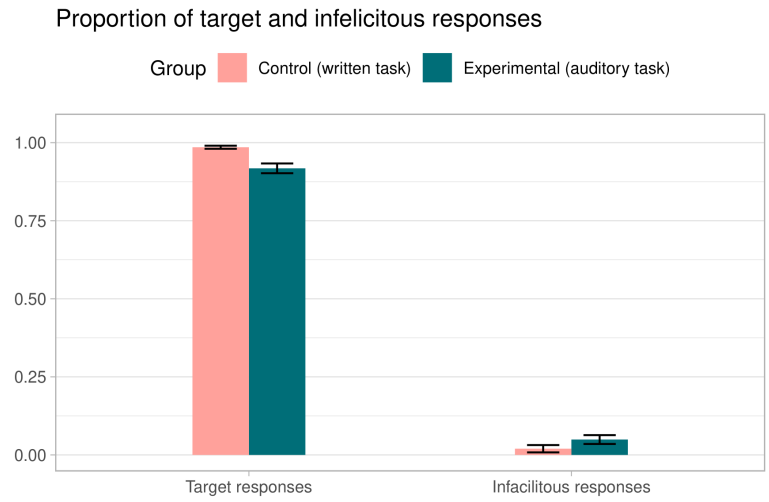

Figure 5.4: Proportions of target and infelicitous responses in lexical tone trials, Experiment 1. Error bars show standard errors of the means.

5.2.2.4 Lexical tone trials

For the experimental group (auditory task), the proportion of target responses (i.e. responses forming a real word when combined with the heard tone) was 92% ($se = 0.02$), while that in the control task was 99%s ($se = 0.01$). The proportion of infelicitous responses was minimal for both groups (experimental: $M = 0.05$, $se = 0.01$; control: $M = 0.02$, $sd = 0.01$) (Figure 5.4).

5.2.3 Discussion

5.2.3.1 Summary of results

The current experiment used the sentence completion task to investigate whether informative tone sandhi cues modulate comprehenders' cloze responses. The results indicate that participants in the experimental group, who had informative tone sandhi cues, produced a higher proportion of cue-compatible responses than the control group, who did not receive informative cues. However, this effect was specific to certain cues: a significant effect of cue availability (Modality) on the proportion of cue-compatible responses was observed only in trials involving *liang3* and *yi2*, but not in those involving *liang2* or *yi4*.

However, the observed differences between cue types may be explained

by the use of the general classifier *ge4*, which can be more common in spoken language than in written language. Supporting this, we found that responses beginning with *ge4* were more frequent in the experimental group (n = 334) than in the control group (n = 218). After excluding trials in which responses began with the general classifier *ge4*, a significant effect of Modality was observed for *liang3*, *yi2*, and *yi4*, but not for *liang2*.

Overall, these findings suggest that cue availability does affect cloze responses, but not uniformly across all cases or cue types. To further disentangle the effects of cue availability from those of modality, Experiment 2 will manipulate cue availability without changing the modality of stimuli.

5.2.3.2 Tone sandhi versus simple lexical tones

Notably, in the *lexical tone trials* of the current experiment, where comprehenders listened to or read sentence frames ending with the first syllable of a multisyllabic noun, participants' responses only formed a real word when combined with the critical syllable in the presented tone, in the vast majority (>90%) of trials. This indicates that participants responded according to the lexical tone of the critical syllable, confirming that cloze responses are highly sensitive to the lexical tone of the critical syllable.

Despite this fundamental high sensitivity to lexical tones, an interesting pattern emerged in the experimental trials involving tone sandhi: although the presence of informative cues increased the proportion of cue-compatible responses for some conditions, there remained a substantial proportion of cue-incompatible responses (>40%) in the experimental group, except in trials involving *liang3*. For *liang3*, the baseline proportion of cue-compatible responses in the control group was already very high, likely because the vast majority of classifiers (87.7%) in Mandarin Chinese are not in T3 and thus compatible with *liang3* (Cai & Brysbaert, 2010).

Overall, the observation that Modality effects were only present for a subset of tone sandhi cues, along with the persistence of a large proportion of cue-incompatible responses when an informative tone sandhi cue was

present, supports the Weak Influence Hypothesis. That is, while informative T3 and *yi* sandhi cues increase the cloze probability of compatible words, participants still produced incompatible responses in a notable number of trials. This suggests that tone sandhi cues increase the activation level of compatible words but do not entirely inhibit the activation of incompatible alternatives.

5.2.3.3 Limitations and alternative explanations

Nonetheless, alternative explanations should be considered. In the current experiment, participants might pre-activate contextually plausible words before hearing the critical syllable, especially when the current experiment included constraining filler items where a particular word is highly predictable. Upon hearing the critical syllable, participants must then inhibit any pre-activated but phonologically incompatible candidates and search for a phonologically compatible alternative. Failures to successfully inhibit the initial candidate or to retrieve a suitable alternative could result in participants defaulting to the contextually plausible but phonologically incompatible response, despite recognizing its incompatibility. This could explain the substantial proportion of cue-incompatible responses without assuming the Weak Influence Hypothesis.

An interaction between Tone and Modality was found for the T3 sandhi, but not for the *yi* sandhi, after excluding responses beginning with *ge4*. However, the direction of the interaction was opposite to what we predicted: there was a larger effect of Modality for the base tone T3 than for the sandhi tone T2. We suspect that this interaction does not reflect differential informativeness between the base tone T3 and the sandhi tone T2. Instead, it may be attributed to a difficulty in retrieving appropriate candidates, as discussed in the previous paragraph, alongside with the low frequency of T3 classifiers in Mandarin Chinese. According to the SUBTLEX-CH word frequency corpus (Cai & Brysbaert, 2010), T3 classifiers constitute only 12.3% of all clas-

sifier occurrences[3]. The relative infrequency of T3 classifiers could make it difficult for participants to retrieve an appropriate classifier when the cue was *liang2*, as only T3 classifiers are compatible. Therefore, the absence of a Modality effect for *liang2* could reflect this retrieval difficulty, rather than a lack of informativeness of the sandhi tone T2.

To minimize the impact of lexical retrieval difficulty, Experiment 2 will replace the free sentence completion task in favour of a forced-choice task, reducing the cognitive load associated with generating sentence completions. Instead of requiring participants to retrieve words from memory, they will be presented with two preselected options following the auditory sentence frame. This design ensures that participants have immediate access to cue-compatible alternatives, allowing for a more precise examination of how tone sandhi influences lexical expectations.

Furthermore, Experiment 2 will adopt a fully within-subjects design that manipulates cue availability without varying the modality of stimulus presentation. This approach offers a more direct comparison between cue-present and cue-absent conditions.

Lastly, Experiment 2 will expand the scope of tone sandhi phenomena under investigation. While the current experiment focused on numerals (*yi* and *liang*), Experiment 2 will also examine T3 sandhi in adjectives and nouns, extending the syntactic environments in which comprehenders' sensitivity to tone sandhi cues is tested beyond a specific kind of noun phrases. This expansion will provide further insight into whether tone sandhi effects are specific to numeral-classifier-noun phrases or if it can be generalized to different syntactic environments.

[3]Meanwhile, the most frequent classifier tone is T4, comprising 61.3% of all classifier occurrences. Since T4 is compatible with the sandhi tone *yi2*, participants would likely find it easier to retrieve a cue-compatible classifier when hearing *yi2* than when hearing *liang2*, even though both are sandhi tones and constrain the upcoming classifier to a single tone.

5.3 Experiment 2

In the current experiment, we employ a forced-choice sentence-completion task in which participants listen to incomplete sentence frames and select one of two written phrases as the more likely continuation. All sentence frames provided a context in which both options were semantically plausible. Crucially, only one of the two written phrases triggered a tone sandhi (T3 sandhi or *yi* sandhi) in the first syllable.

In Experiment 2, we expanded the syntactic environment in which the T3 sandhi was tested. In addition to numerals that undergo the T3 sandhi (*liang*), we included trials where the critical syllable was a monosyllabic adjective that undergoes the T3 sandhi (*xiao3* "small" or *lao3* "old"), or the first syllable in T3-initial compound nouns (*lü* in *lü3dian4* "inn" versus *lü2guan3* "hostel").

Using a within-subjects design, we manipulated availability of an informative tone sandhi cue by presenting the first syllable of the target (the critical syllable) in the experimental condition, but not in the control condition. We hypothesized that the proportion of target responses should be higher in the experimental condition than the control condition. More specifically, according to the Strong Influence Hypothesis, the proportion of target responses in the experimental condition should approach the ceiling; while according to the Weak Influence Hypothesis, although the proportion of target responses should be higher in the experimental condition than the control condition, we may still observe a notable proportion of non-target responses in the experimental condition.

A few key changes were made in the current experiment compared to Experiment 1. First, we no longer manipulated the modality of stimuli, varying tone sandhi cue availability in the same auditory modality. Second, instead of asking participant to freely complete the sentence frames, we presented two possible options in writings on the screen while asking participants to select the more likely option. This is to minimise the cognitive

load needed to retrieve plausible and phonologically compatible candidates. Third, we did not include any "lexical tone trials", as a fundamental sensitivity to lexical tones was already established in Experiment 1. Four, we used a different set of filler items. A proportion fillers served to ensure that the target phrase did not always begin with a *yi* or a T3 syllable, while the other filler items contained an implausible option, serving to ensure participants' attention on the task. Finally, we added trials in which an informative T3 sandhi cue appeared in an adjective or the first syllable of a noun compound, expanding the syntactic environment in which T3 sandhi cues were tested.

5.3.1 Methods

5.3.1.1 Participants

Participants were 48 native Mandarin Chinese speakers aged between 18 and 35 years (23 female, age $M = 21.33$ years, $SD = 2.10$). All participants gave informed consent and was paid for their time.

5.3.1.2 Stimuli

Experimental items Experimental items consisted of 124 sets of sentence frames, paired with a target phrase and a competitor phrase; both phrases were plausible in the sentential context. Crucially, the two phrases shared the same initial syllable, however, this syllable was realised in different tones because it underwent tone sandhi in one of the two phrases. The last syllable of the sentence frame, namely the critical syllable, took on either its base tone or its sandhi tone, and was always only compatible with the target phrase and not the competitor phrase.

All items were manipulated for cue availability, creating an experimental condition where the critical syllable was presented in the spoken sentence frame, and a control condition where the sentence frame was truncated before the critical syllable. Among the 124 experimental items, 40 involved a numeral as the critical syllable (*liang* "two", the T3 sandhi, or *yi* "one", the *yi* sandhi), 24 involved a monosyllabic T3 adjective as the critical syllable

(e.g. *xiao3* "small"), and 60 involved a T3 syllable that was the onset of a multisyllabic noun (e.g. *lü3* in *lü3dian4* "inn"). For items involving a numeral, we manipulated the critical syllable in a 2 × 2 × 2 factorial design, crossing factors Sandhi (T3 sandhi, *liang* vs. *yi* sandhi, *yi*), Tone (base tone vs. sandhi tone), and Cue Availability (cue present vs. cue absent). For items involving a T3 adjective or noun, we manipulated the critical syllable in a 2 × 2 Cue Availability (cue present vs. cue absent) as well as the critical syllable's tone (base tone vs. sandhi tone). A set of sample stimuli is given in Table 5.4. The full set of experimental and filler items can be found in Appendix H.

A female native Mandarin Chinese speaker recorded the auditory stimuli. All recordings' volume was normalised. To make sure the sentence frame's intonation did not contain an end-of-sentence pitch drop or increased duration, a "dummy" syllable was selected for each experimental item prior to recording. The recorder read the sentence frame followed by the dummy syllable. The dummy syllable was chosen from three tokens, /di/, /da/, /du/, representing a plosive consonant followed by a front closed, front open, and a back closed vowel respectively. The dummy syllable was chosen such that its vowel was as far away as possible on the vowel chart from both the target word's and the competitor word's first vowel (excluding the critical syllable), in order to avoid any possible coarticulation bias. For instance, for the item where *xiao3 mao1* was the target and *xiao2 gou3* was the competitor, the recorder read the sentence frame followed by *xiao3 di1*.

Figure 5.5 shows average pitch countours of the critical syllables. The average durations were as follows: *yi*: 393 ms; *liang*: 435 ms; T3 adjectives: 475 ms; and the first syllable of T3-initial nouns: 449 ms.

After recording, the onset of the dummy syllable was identified, and the recording was truncated to remove the dummy syllable. This resulted in 248 audio recordings for the experimental trials. In order to make recordings for the control trials, the onset of the critical syllable was identified, and the experimental trials' recordings were truncated to remove the critical syllable.

Sandhi	Type	Tone	Sentence frame	Target phrase	Competitor phrase
yi sandhi	Numeral (*yi*)	base	餐桌上摆放着 (一) *Can1zhuo1 shang4 bai3fang4zhe (**yi4**)* On the dinning table, there is (a)	一根黄瓜 *yi4 gen1 huang2gua1* A cucumber	一串葡萄 *yi2 chuan4 pu2tao2* A bunch of grapes
		sandhi	餐桌上摆放着 (一) *Can1zhuo1 shang4 bai3fang4zhe (**yi2**)* On the dinning table, there is (a)	一串葡萄 *yi2 chuan4 pu2tao2* A bunch of grapes	一根黄瓜 *yi4 gen1 huang2gua1* A cucumber
T3 sandhi	Numeral (*liang*)	base	餐桌上摆放着 (两) *Can1zhuo1 shang4 bai3fang4zhe (**liang3**)* On the dinning table, there are (two)	两根黄瓜 *liang3 gen1 huang2gua1* Two cucumbers	两把勺子 *liang2 ba3 shao2zi* Two spoons
		sandhi	餐桌上摆放着 (两) *Can1zhuo1 shang4 bai3fang4zhe (**liang2**)* On the dinning table, there are (two)	两把勺子 *liang2 ba3 shao2zi* Two spoons	两根黄瓜 *liang3 gen1 huang2gua1* Two cucumbers
	T3 Adjectives	base	老王特地准备了 (好) *Lao3wang2 te4di4 zhun3bei4le (**hao3**)* Wang specially prepared some (good)	好茶 *hao3 cha2* good tea	好酒 *hao2 jiu3* good wine
		sandhi	老王特地准备了 (好) *Lao3wang2 te4di4 zhun3bei4le (**hao2**)* Wang specially prepared some (good)	好酒 *hao2 jiu3* good wine	好茶 *hao3 cha2* good tea
	T3-initial Nouns	base	游客们在晚餐后回到了 (旅) *You2ke4men zai4 wan3can1 hou4 hui2dao4le (**lü3**)* After dinner, the tourists returned to the (lü)	旅店 *lü3dian4* inn	旅馆 *lü2guan3* hostel
		sandhi	游客们在晚餐后回到了 (旅) *You2ke4men zai4 wan3can1 hou4 hui2dao4le (**lü2**)* After dinner, the tourists returned to the (lü)	旅馆 *lü2guan3* hostel	旅店 *lü3dian4* inn

Table 5.4: An example set of stimuli, Experiment 2. The participants listened to spoken sentence frames while seeing the two phrases in Chinese characters. The critical syllable is presented in parentheses, which was presented in the spoken sentence frame only in the experimental condition (cue present), but not in the control condition (cue absent). See Appendix H for the complete set of stimuli.

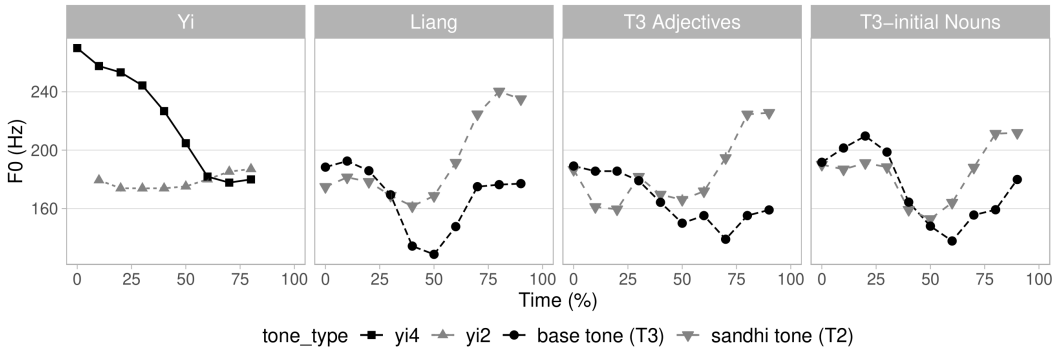

Figure 5.5: Average pitch (F0) contour of the critical syllables, Experiment 2, within a normalized time window. Error bars represent ± 1 standard error of the mean. The average durations of the syllables were as follows: *yi*: 393 ms; *liang*: 435 ms; T3 adjectives: 475 ms; the first syllable of T3-initial nouns: 449 ms.

Filler items and presentation lists Eight presentation lists were created, such that participants receiving each list will see only one version of each experimental item. Each presentation list also contained 60 filler items. Among these, 20 fillers contained a numeral (*yi* or *liang*) as its critical syllable and had an implausible competitor. Another 40 filler items had a non-T3 critical syllable that was either an adjective or the first syllable of a noun. Among these, half had an implausible competitor and the other half a plausible competitor. In both types of fillers, the critical syllable was pronounced in the audio recording in half of the items and not pronounced in the other half. Filler sentences were recorded directly without dummy syllables.

5.3.1.3 Procedure

The experiment was hosted online on Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). At the beginning of the experiment, participants read the information about the experiment and gave consent, followed by a sample audio with the same volume as the auditory stimuli with which participants could make sure they could hear the recordings clearly. Subsequently, participants read the instructions and completed 4 practice trials before starting the experiment.

In each experimental trial, participants saw a visual display of two writ-

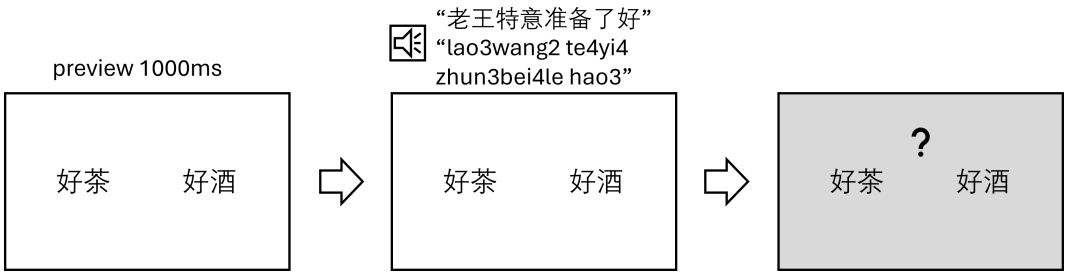

Figure 5.6: Experiment trial procedure, Experiment 2. Participants viewed the written options for 1s before the spoken sentence frame began to play. Following the offset of the sentence frame, the participant was prompted to select which option was the better sentence continuation.

ten phrases for 1 second before the audio recording of the sentence frame began to play. The visual display stayed unchanged as the audio stimuli were presented. Immediately after the audio stimuli finished playing, the screen changed to a grey background and a question mark appeared at the centre of the screen to prompt participants to respond. Participants were asked to select among the two phrases the more suitable continuation for the sentence frame they heard by pressing one of two keys on the keyboard. Here, they were also given the option to report an error if the spoken sentence frame did not successfully play, a known occasional issue of Gorilla Experiment Builder. The next trial began once the participant has made a selection or reported an error. Experimental trials and filler trials were mixed and pseudo-randomised. Figure 5.6 illustrates the trial procedure. Following the sentence completion task, participants completed the same tone sandhi experience questionnaire as that in Experiment 1.

5.3.1.4 Data analysis

Data was analysed using the brms package (Bürkner, 2017) in R (R Core Team, 2019), version 4.5.1. Participants' response accuracy in filler items with implausible competitors was used to check their attention to the experiment. All participants had very high accuracy in these trials (all > 0.9, average accuracy = 0.98), thus none was excluded from further data analysis.

Answers to the questions in the questionnaire was categorized, and chi-

squared tests were used to determine whether participants' experience differed significantly between the T3 sandhi and the *yi* sandhi.

Participants' responses in the forced-choice sentence completion task were separately analysed for the four trial types (trials containing the numeral *yi*, the numeral *liang*, T3 adjectives, and T3 nouns). Target responses were coded as 1 (success), and competitor responses were coded as 0 (failure). Trials in which the auditory stimuli did not play successfully was excluded from analysis (n=2).

Bayesian logistic regressions were fitted to the coded data, with sum-coded factors Cue Availability (experimental/cue-present vs. control/cue-absent), and Tone (base tone vs. the sandhi tone), and interactions between the two factors. The prior distribution chosen for the intercept of the model was $\mathcal{N}(0, 1.5)$, while the prior distribution chosen for the predictors was $\mathcal{N}(0, 1.5)$. These priors represent a principle that before observing the data, we are uncertain whether the experimental manipulations will introduce a null effect or a relatively large effect (Nicenboim, Schad, & Vasishth, 2025).

5.3.2 Results

5.3.2.1 Accuracy in filler items

All participants showed ceiling-level accuracy in fillers containing implausible competitors (mean $= 0.98$, $SD = 0.02$, min $= 0.95$, max $= 1$). Therefore, no participant was excluded from further data analysis.

5.3.2.2 Tone sandhi experience

Chi-squared tests revealed a significant difference between participants' experiences with the T3 sandhi and the *yi* sandhi. Specifically, there were significantly more participants who reported receiving formal instructions regarding the *yi* sandhi ($X^2 = 23.43$, $p < 0.001$), as well as being aware of the the patterns of the *yi* sandhi ($X^2 = 7.24$, $p = 0.007$), than the T3 sandhi.

	yi sandhi			T3 sandhi		
	yes	no	do not remember	yes	no	do not remember
explicit knowledge	35 (72.92%)	13 (27.08%)		28 (58.33%)	20 (41.67%)	
classroom experience	32 (66.67%)	6 (12.5)%	10 (20.83%)	27 (56.25%)	9 (18.75%)	12 (25%)

Table 5.5: Participants' self-reported tone sandhi experience, Experiment 2. See Appendix I for the complete questionnaire.

5.3.2.3 Forced choice responses

Bayesian logistic regression revealed a significant and strong main effect of Cue Availability (proportion of target responses when cue present > cue absent, BF10's > 6) in all types of materials except for *liang* trials (*yi* trials: $\beta = -0.18$, $SE = 0.05$, 95%CI $= [-0.28, -0.07]$, $BF10 = 100$; *liang* trials: $\beta = -0.1$, $SE = 0.05$, 95%CI $= [-0.20, -0.01]$, $BF10 = 2.56$; Adj T3 trials: $\beta = -0.15$, $SE = 0.05$, 95%CI $= [-0.25, -0.05]$, $BF10 = 20$; Noun T3 trials: $\beta = -0.18$, $SE = 0.04$, 95%CI $= [-0.24, -0.10]$, $BF10 = +\infty$). For *liang* trials, $BF10 = 2.56$ suggests weak effect of Cue Availability, although note that the data set for *liang* (20 trials per participant) was the smallest among all T3 sandhi cues (T3 adjectives: 24 trials per participant; T3-initial nouns: 60 per participant).

Notably, as Figure 5.7 suggests, the proportion of target responses was far from ceiling-level in the experimental condition, when an informative tone sandhi cue was present.

An unexpected main effect of Tone was found in all models, suggesting an overall difference in the proportion of target responses for trials containing the base tone vs. the sandhi tone of the critical syllable, regardless of Cue Availability. These main effects of Tone may suggest some overall bias towards one option in some of the experimental items. Nevertheless, the focus of the current experiment was the effect of Cue Availability (ex-

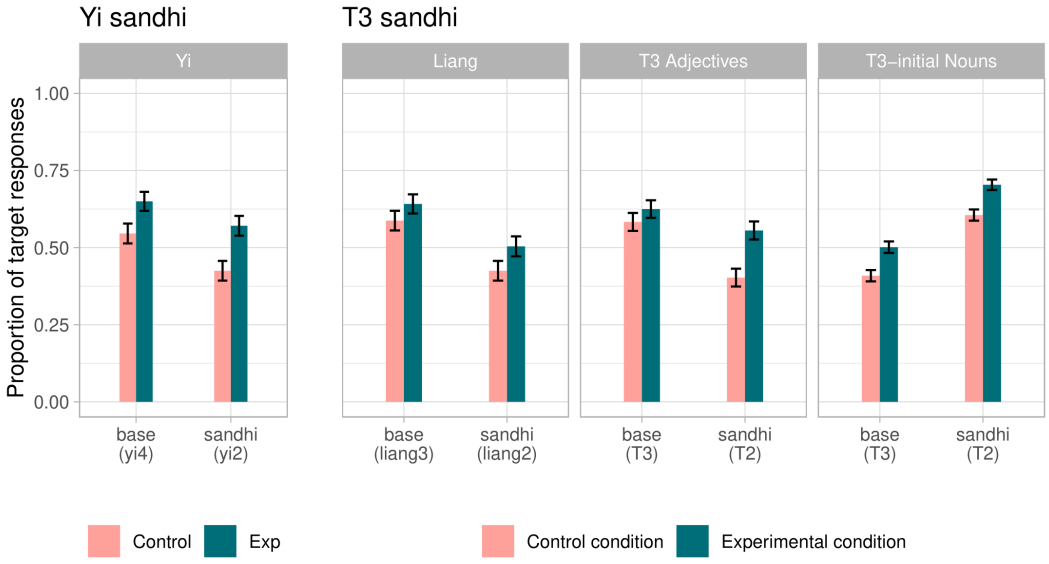

Figure 5.7: Proportion of target responses in each condition, Experiment 2. Error bars show standard errors of the means.

perimental - control). Since the proportion of target responses never surpassed 60% in the control versions, we saw no reason to believe the potential bias towards one option over the other caused any ceiling effect in the control condition which might reduce statistical power. For *yi* trials ($\beta = 0.14$, $SE = 0.05$, 95%CI $= [0.04, 0.25]$, $BF10 = 16.67$), *liang* trials ($\beta = 0.21$, $SE = 0.05$, 95%CI $= [0.10, 0.32]$, $BF10 = +\infty$), as well as Adj T3 trials ($\beta = 0.19$, $SE = 0.05$, 95%CI $= [0.09, 0.29]$, $BF10 = +\infty$), there was significantly more target responses when the sentence frame contained the base tone than the sandhi tone. This pattern is reversed in Noun T3 trials ($\beta = -0.36$, $SE = 0.04$, 95%CI $= [-0.43, -0.29]$, $BF10 = +\infty$).

More importantly, we found no interaction between Tone and Cue Availability in all of the models, even though visual inspection of the plot (Figure 5.7) suggested a potential smaller effect of Cue Availability for the base tone T3. BE10's was between 1/3 and 1 for all models except for Adjective T3 trials, where it was above 1 but below 3 ($BF10 = 1.92$), providing no substantial evidence for an interaction and suggesting that the effect of Cue Availability was present regardless of whether the cue was in the base tone or sandhi tone.

5.3.3 Discussion

The present experiment employed a forced-choice sentence completion task to investigate the effect of tone sandhi cues on lexical predictability (i.e., cloze responses). Participants listened to incomplete sentence frames while viewing two written phrases on a screen and were prompted to select the phrase they considered to be the more likely continuation.

The results revealed a significantly higher proportion of target responses in the experimental condition (cue present) compared to the control condition (cue absent), indicating that informative tone sandhi cues modulate cloze responses by increasing the probability of compatible words. Nevertheless, despite the significant effect of Cue Availability, we still observed a substantial proportion of competitor responses in the experimental condition (over 25%). This finding aligns with the Weak Influence Hypothesis, according to which tone sandhi cues affect predictability simply by increasing the predictability of compatible words, while not completely inhibiting incompatible continuations.

Additionally, we found no significant interaction between Cue Availability and Tone across all trial types (both the T3 sandhi and the *yi* sandhi). This pattern indicates that informative tone sandhi cues modulate lexical predictability regardless of base tone or sandhi tone. Therefore, the difference between how *liang3* and *liang2* were used in real-time lexical predictions (see Chapter 2 and 3) likely does not stem from an inherent difference in how the base tone and the sandhi tone affect predictability. Rather, it could stem from a difference in how readily the two types of cues can be used to compute lexical predictions in real-time.

Compared to Experiment 1, the present study successfully improved the experimental design to better isolate the effect of tone sandhi on sentence completion. By presenting all stimuli auditorily in both conditions and providing preselected response options to minimize the cognitive loads associated with lexical retrieval, we controlled for the effect of modality and the

effect of differences in retrieval difficulty arising from the uneven frequency of classifiers in each tone. With these improvements, the current results demonstrate significant effects of Cue Availability on cloze probability regardless of base or sandhi tone, providing clear evidence for the role of tone sandhi cues in modulating lexical predictability.

5.4 General discussion

In the present study, we investigated the the role of informative tone sandhi cues in comprehenders' expectations about upcoming lexical items. Using versions of the sentence completion task (a.k.a. the cloze task), we showed that informative T3 sandhi cues and *yi* sandhi cues can modulate cloze responses. More specifically, when an informative tone sandhi cue is present, the cloze probabilities of cue-compatible words increase, but cue-incompatible words are not entirely inhibited (appearing in more than 25% of trials).

This pattern is consistent with our Weak Influence Hypothesis about tone sandhi's role in lexical predictability, such that hearing an informative tone sandhi cue boosts the activation level of compatible words without necessarily inhibiting the activation of incompatible ones.

It is worth highlighting that the weak influence on cloze probablity is a property of tone sandhi, rather than lexical tones themselves. In Experiment 1, we included a proportion of "lexical tone trials" where the critical syllable is the onset of a multisyllabic noun (e.g. *Ming bought some snacks from the store, they were bing1 ...* $\rightarrow$ *bing1qi2lin2* "ice cream" vs. **bing3gan1* "cookies"), and found that participants' responses were overwhelmingly consistent with the heard tone, showing a clear ceiling effect. This indicates that lexical tones, by themselves, strongly shape comprehenders' predictions—listeners can immediately rule out words with incompatible tones as possible sentence continuations.

This contrasts sharply with our findings on tone sandhi cues. Across

both experiments, when sentence frames ended with a syllable carrying an informative tone sandhi cue—whether across word boundaries (e.g., *On the dining table, there are liang3... → liang3 gen1 huang3gua1* "two cucumbers" vs. **liang2 ba3 shao2zi* "two spoons") or within a word's boundary (e.g., *After dinner, the tourists returned to the lü2... → lü2guan3* "hostel" vs. **lü3dian4* "inn")—listeners did not strongly inhibit cue-incompatible continuations.

In sum, the present study demonstrates that listeners adjust their expectations based on informative tone sandhi cues, but do not fully suppress cue-incompatible candidates, suggesting that tone sandhi cues only assert a weak influence on lexical predictability.

5.4.1 Indications for real-time lexical predictions

The present study found that although listeners' expectations about upcoming lexical items do change according to tone sandhi cues, this change may be rather subtle. This may explain why real-time lexical prediction using tone sandhi cues is difficult to observe. In Chapter 2, we tested this by presenting listeners with non-constraining sentences that identified one of the two objects in the visual display in a noun phrase containing a numeral, a classifier, and the head noun. The spoken sentence either contained a tone sandhi cue in the numeral that was informative about the identity of the target (Different Tones condition) or an uninformative numeral (Same Tones condition). Eye-tracking data revealed that listeners did not fixate on the target object any earlier when a tone sandhi cue was available compared to when it was absent, suggesting that listeners did not use the informative tone sandhi cue to anticipate the upcoming classifier and noun.

In Chapter 3, we improved on Chapter 2's experiment design and explicitly presented classifiers and nouns in writing in the visual display. This paradigm ensured that the classifier—positioned between the numeral carrying the tone sandhi cue and the target noun—was explicitly presented on the display, maximizing the opportunity to detect a prediction effect. With

this improved design, we found a statistically significant effect of prediction, however Bayes Factors was not able to provide strong support for the alternative hypothesis, despite a large sample size.

The present study's findings offer insights into why lexical prediction based on tone sandhi may be difficult to detect in real-time. Consider the requirements for successful lexical prediction based on tone sandhi cues in the visual world eye-tracking paradigm. Upon hearing an informative tone sandhi cue, listeners need to (uniquely) identify the target phrase or object. For example, when presented with two options (*liang2 ben3 shu1* "two books" vs. *liang3 zhang1 chuang2* "two beds"), hearing *liang2* must enable listeners to infer that *liang3 zhang1 chuang2* is highly unlikely to be the target. This inference depends on listeners adjusting their expectations in response to the tone sandhi cue, effectively ruling out incompatible continuations (*liang3 zhang1 chuang2*) and subsequently shifting their gaze toward the most likely target.

However, based on the results of the present study, we argue that listeners may not fully rule out incompatible continuations upon hearing a tone sandhi. In Chapter 2 and 3's eye-tracking experiments, this means that upon hearing the informative numeral, listeners may still consider both options as viable continuations. In other words, there may not be a significant change in the likelihood of either option, even after hearing the informative tone sandhi cues at the numeral. While listeners may slightly adjust their expectations in favour of the target, the small effects of tone sandhi on predictability observed in this study suggest that this modulation may be rather weak. If the bias towards the target object is weak, then anticipatory eye movements may remain subtle and difficult to observe in the visual world paradigm.

Notably, in the present study, listeners had ample time to adjust their expectations about the upcoming word following the tone sandhi cues. In Experiment 1, listeners were recorded for a total of 4 seconds following sentence frame offset. In Experiment 2, participants had no time limit to choose

their preferred continuations. Yet, results suggest that, even when given ample time to adjust their expectations, comprehenders still did not fully consider the phonological constraints tone sandhi poses on the upcoming words. The present study can be considered a study of prediction in an offline manner: what comprehenders would predict when they had all the cognitive resources and all the time to compute such predictions. What the present study's results show is that, essentially, tone sandhi cues do not result in significant changes in comprehenders' lexical expectations, even when they are not constraint by the speed of natural speech input.

5.4.2 Offline expectations versus real-time predictions

The current chapter demonstrates that tone sandhi cues weakly modulates lexical expectations, even when participants are given unlimited processing time. However, a key question remains: is this weak effect of tone sandhi cues on lexical predictability fully evident in real time, or is there still a discrepancy between the extent to which tone sandhi cues can modulate lexical predictability and the degree to which listeners actually use them to generate lexical predictions on the fly?

A careful comparison of the current chapter's results with those from Chapters 2 and 3 suggests that such a discrepancy between offline expectations and real-time predictions may indeed exist. Specifically, the Bayesian analyses reported in Chapters 2 and 3 never provided substantial support for real-time lexical predictions based on tone sandhi cues (all BF10's between 1 and 3), indicating that the data were never much more likely under the alternative hypothesis than under the null. In contrast, the present (offline) study found strong Bayesian evidence (BF10's > 20 in three out of four types of cue), clearly demonstrating that tone sandhi cues do shape lexical predictions when time pressure is removed, as in the sentence completion task.

Taken together, this pattern reveals a contrast: the influence of tone

sandhi on lexical expectation is robust and readily observable in offline (untimed) tasks, but remains difficult to detect under real-time conditions.

This discrepancy aligns well with what the prediction-as-memory-retrieval framework says about the time course of lexical predictions (Chow et al., 2016), which emphasizes that the effectiveness of predictive cues in real-time prediction depends not only on the nature of information retrieved from memory, but also on the time course in which it is accessed. If lexical retrieval is too slow, the information may not be accessible early enough to impact immediate, real-time predictions—even if it can eventually refine predictions with more processing time.

Evidence supporting this view comes from studies on verb prediction and argument structure. The amplitude of the N400 ERP component is known to be inversely proportional to word predictability. However, Chow et al. (2018) found that the N400 is not immediately sensitive to a verb's change in predictability when the preverbal arguments were reversed. For example, the verb "*served*" is likely in a context such as "*can you tell me which customer the waitress …*", but much less so when arguments are role-reversed, "*… which waitress the customer …*". However, despite a clear difference in the verb's cloze probabilities depending on the arguments' roles, the N400 elicited by the verb "*served*" was identical between the canonical and role-reversed conditions (Chow et al., 2018, 2016; Hoeks, Stowe, & Doedens, 2004; Kuperberg, Sitnikova, Caplan, & Holcomb, 2003). An N400 effect emerged only when the verb was moved to appear later in the sentence to allow comprehenders more time to process the arguments, by inserting adjuncts such as "*yesterday*" between the arguments and the verb. This suggests that time is required for comprehenders to adjust their predictions based on argument structure.

This short-lived insensitivity to role-reversal suggests that, at the earliest stage of prediction, comprehenders may not incorporate specific argument role information, thus continuing to predict verbs like "*serve*" even

in role-reversed contexts. In other words, although comprehenders may be able to immediately recognise that the action of serving is highly frequent between a waitress and a customer, they may need additional time to consider that this is in fact unlikely if the customer is the agent and the waitress is the patient, and subsequently inhibit "*serve*" form their verb predictions.

A similar phenomenon may be at play in the processing of tone sandhi cues in Mandarin Chinese. When generating predictions based on tone sandhi, listeners must first infer the set of possible tones for the upcoming syllable before retrieving corresponding lexical candidates—this is a multi-step process that may be more demanding and time-consuming than making predictions based on semantic or syntactic cues. Critically, this processing must occur rapidly, before the onset of the upcoming syllable, which is the target of prediction. This possibility is supported by findings from Liu et al. (2025), who observed prediction effects in real-time when the presentation rate of stimuli was slowed, providing comprehenders with additional time to complete the prediction process.

The discrepancy between the present study's robust offline effects and the subtle real-time effects in Chapters 2 and 3 likely reflects differences in the amount of time or cognitive resources available for prediction in real time. In the sentence completion (offline) task, unlimited time allows comprehenders to use tone sandhi cues to inform their predictions. In real-time tasks, however, lexical prediction based on tonal cues may require more than is available.

5.4.3 Base versus sandhi tones

In the present study, we did not find a statistically significant difference in the extent to which base tone cues and sandhi tone cues modulate lexical predictability. In Experiment 1, we observed an interaction between Cue Availability and Tone for the T3 sandhi condition (*liang*); however, this likely reflects a general difficulty in retrieving a T3 classifier following *liang2*, as T3 classifiers are the least frequent among the four Mandarin tones. In contrast,

Experiment 2 required participants to choose from a preselected pair of options, thereby reducing the cognitive load associated with lexical retrieval. Under these conditions, no statistically significant interaction between Cue Availability and Tone was observed for either T3 sandhi or *yi* sandhi trials.

These results suggest that base tone and sandhi tone cues do not significantly differ in how they influence lexical predictability. Nevertheless, findings from Chapters 2 and 3 indicated a different pattern in real-time processing: specifically for the T3 sandhi, sandhi tone T2 cues appeared to support lexical prediction to a greater extent than base tone T3 cues.

If both the base tone T3 and the sandhi tone T2 modulate the predictability of upcoming words, why might they differ in their impact on real-time prediction? Once again, we find one possible answer in the time course of lexical prediction.

As discussed in the introduction, base and sandhi tones differ in the strength of their constraints. A sandhi tone poses a stronger phonological constraint by limiting the following syllable to a specific tone (T2 $\rightarrow$ T3, as in *yi2* $\rightarrow$ T4). In contrast, a base tone merely excludes one possibility, thereby allowing a set of alternatives (e.g., T3 $\rightarrow$ T1/2/4, *yi4* $\rightarrow$ T1/2/3). This difference could affect how quickly listeners can use base versus sandhi tone cues to generate lexical predictions during real-time processing—base tone cues may require more time to yield specific lexical predictions compared to sandhi tone cues.

The lack of interaction in the present study thus supports the view that although both base tones and sandhi tones modulate probabilities of the upcoming word, the modulation from sandhi tones is more readily used during real-time comprehension to generate predictions than base tones, due to a stronger phonological constraint on the upcoming syllable's tone.

5.4.4 Summary

The present study investigated how tone sandhi cues influence comprehenders' expectations about upcoming lexical items. Using offline sentence

completion tasks, we found that informative tone sandhi cues weakly increased the likelihood of cue-compatible responses. This is consistent with the Weak Influence Hypothesis where informative tone sandhi cues increase the predictability of compatible words without completely inhibiting incompatible ones. In addition, we found that base tones and sandhi tones equally affect lexical predictability.

Taken together, the present study demonstrates a role of informative tone sandhi cues, regardless of base or sandhi tone, in lexical expectations when comprehenders have ample time to compute such predictions. These findings suggest that the uncertain effects of prediction as well as the differential effects of base versus sandhi tones observed in the previous chapters could be attributed to a lack of processing time for computing lexical predictions in real-time.