LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

EMPIRICAL STUDY

The Influence of Native Phonology, Allophony, and Phonotactics on Nonnative Lexical Encoding: A Vocabulary Training Study

Qi Zheng (she/her) and Kira Gor (she/her)
University of Maryland

Abstract: Second language (L2) speakers often experience difficulties in learning words with L2-specific phonemes due to the unfaithful lexical encoding predicted by the fuzzy lexical representations hypothesis. Currently, there is limited understanding of how allophonic variation in the first language (L1) influences L2 phonological and lexical encoding. We report how the Mandarin Chinese L1 phonemic inventory and allophonic variation subject to phonotactic constraints predict phonological encoding problems for novel L2 English words with the /v/–/w/ contrast. L1 English and L1 Chinese participants speaking two varieties of Mandarin Chinese differing as to the

CRediT author statement – **Qi Zheng**: conceptualization (lead); formal analysis; investigation; methodology (lead); software; writing – original draft preparation (lead); writing – review & editing (equal). **Kira Gor**: conceptualization (supporting); methodology (supporting); writing – original draft preparation (supporting); writing – review & editing (equal).

A one-page Accessible Summary of this article in non-technical language is freely available in the Supporting Information online and at https://oasis-database.org

We would like to thank the anonymous reviewers and the Associate Editor, Professor Theres Grüter, for their insightful suggestions regarding this article throughout the review process. We are also very grateful to Robert DeKeyser, Nan Jiang, Steven J. Ross, Gregory R. Hancock, Tracy Sweet, Yang Liu, Maria Polinsky, Marek Majer, Catherine Pulupa, Caitlin Burke, Chae Mamayek, Jason Struck, Matt Coss, Wei Yi, Xu Li, Hongyang Zhao, Yun Lü, Yuting Sun, Yaxin Zhang, Zhiyuan Deng, Liuan Yang, and Rachel Chang for their support with various aspects of the study.

Correspondence concerning this article should be addressed to Qi Zheng, University of Maryland, College Park, 3215 Jimenez Hall, University of Maryland, College Park, MD 20742, United States. Email: qiz@terpmail.umd.edu

The handling editor for this manuscript was Theres Grüter.

DOI: 10.1111/lang.12581

presence of [ʋ]–[w] allophonic variation for the /w/ phoneme participated in a vocabulary learning task. The novel L2 words with the /v/–/w/ contrast were systematically less robustly encoded than the control words on the day of training and 24 hours later. The degree of fuzziness in lexical representations was jointly predicted by L1 allophonic variation subject to phonotactic constraints and L2 phonological categorization.

Keywords lexical encoding; nonnative listeners; phonological categorization; allophonic variation; memory consolidation

Introduction

Spoken word recognition in a second language (L2) is not always successful and effortless, with recent years witnessing a growing interest in understanding how L2 spoken words are perceived, encoded, and accessed (e.g., Barrios & Hayes-Harb, 2021; Cutler et al., 2006; Darcy et al., 2012, 2013; Escudero et al., 2008, 2014; Llompart, 2021). Adult L2 learners are sensitive to the phonological contrasts in their first language (L1), whereas they experience difficulties in discriminating novel L2 contrasts (Best, 1995; Best & Tyler, 2007; Flege, 1995; Flege & MacKay, 2004). For example, Japanese speakers show great difficulties in the perception and production of English words with the /ɹ/–/l/ contrast (Jamieson & Morosan, 1986; Logan et al., 1991), and Finnish speakers can hardly discriminate the English /i/–/ɪ/ contrast (Ylinen et al., 2010). Although numerous findings about reduced sensitivity to L2 phonological contrasts have been reported, a remaining critical issue is how problems with phonological categorization of L2 sounds impact L2 word perception, storage, and access.

According to the fuzzy lexical representations hypothesis, difficulties with the phonological categorization of L2 contrasts may lead to imprecise or low-resolution, or fuzzy, phonological encoding of L2 lexical representations (referred to as phonolexical encoding) in the learners' mental lexicon and contribute to problems with L2 spoken word recognition (Bordag et al., 2022; Cook & Gor, 2015; Cook et al., 2016; Gor & Cook, 2020; Gor et al., 2021). Fuzziness in phonolexical encoding can lead to fuzzy form–meaning mappings and lexical confusions: For example, *light* (/l/) and *right* (/ɹ/) may be poorly differentiated by Japanese speakers. While the fuzzy lexical representations hypothesis claims that L2 phonolexical encoding is generally effortful and unfaithful because L2 phonetic cues do not match L1 phonetic cues, it also predicts that the words with difficult L2 contrasts will have even less precise (fuzzier) encoding that will be more resistant to improvement with additional input (Gor et al., 2021).

The empirical evidence supporting this hypothesis comes from several sources: lexical decision tasks or phonological and repetition priming experiments (Cook & Gor, 2015; Darcy et al., 2012, 2013; Gor & Cook, 2020; Pallier et al., 2001), visual-world eye-tracking experiments (Cutler et al., 2006; Weber & Cutler, 2004), and (pseudo-)semantic priming (Cook et al., 2016). The most relevant findings for the present study come from vocabulary and lexical training experiments, which explore the learning of novel L2 words with problematic phonological contrasts (Escudero et al., 2008, 2014).

Whereas there is considerable experimental support for the fuzziness at different levels of lexical representation in novel L2 word learning, little is known about how L1 allophonic variation may influence L2 phonological and, particularly, lexical encoding (Barrios et al., 2016; Eckman et al., 2003; López Velarde & Simonet, 2020; Shea, 2014). Against this background, the current study explores how L1 Mandarin Chinese speakers with English as their L2 learn novel English-like /v/- and /w/-initial words after a short training period and again after 24-hour consolidation. Crucially, these L1 Mandarin Chinese (hereafter Chinese) speakers differ in whether they produce /w/ with allophonic variation ([ʋ]–[w]) in their L1, therefore enabling a more comprehensive investigation of the contributions of the L1 phonemic inventory to L2 phonolexical encoding. The study tests the fuzzy lexical representations hypothesis (Gor et al., 2021) by comparing the learning outcomes for the novel L2 words with easy (/k/–/g/) and difficult (/v/–/w/) phonological contrasts immediately after lexicalization and after 24-hour consolidation. It extends the claims of the fuzzy lexical representations hypothesis to complex L1–L2 phonological mapping involving free allophonic variation in the L1.

Background Literature

One of the widely explored reasons for L2 learners' less efficient L2 word recognition is their perceptual problems: L2 learners have been reported to experience difficulties in discriminating confusable L2-specific phonemic contrasts (Darcy et al., 2012, 2013; Eckman et al., 2003). The absence of perceptual sensitivity leads to unfaithful phonolexical encoding of the words stored in the L2 mental lexicon and fuzzy L2 lexical representations (Gor et al., 2021). For instance, Pallier et al. (2001) reported that Spanish-dominant but not Catalan-dominant Spanish–Catalan bilinguals showed repetition priming effects for minimal pairs containing Catalan-specific contrasts (/ɛ/–/e/, /o/–/ɔ/, /s/–/z/). The results suggested that the Spanish-dominant bilinguals activated both words in the minimal pair when given the L2 input, whereas the Catalan-dominant speakers activated only the words presented in the input. Using the eye-tracking visual-world paradigm, Weber and Cutler (2004) explored real-time lexical competition in L1-Dutch L2-English learners with the English-specific contrast /ɛ/–/æ/. The authors found that when given the auditory instruction to click on the picture of *panda* (phoneme /æ/), L1 Dutch speakers fixated longer on the picture of *pencil* (phoneme /ɛ/) than on less confusable distracters. Not only did English speakers show no such fixation differences, but the L2 confusability was also unidirectional: When given *pencil* as the target, *panda* was not more confusable than other distracters. Cutler et al. (2006) also reported similar patterns of asymmetry with L1 Japanese speakers who encountered perceptual and articulatory difficulties with the English /ɹ/–/l/ contrast. These observed asymmetries are often associated with situations where two L2 phonemes are mapped onto one L1 category with differing degrees of similarity (cf. the perceptual assimilation model; Best, 1995; Best & Tyler, 2007). The more familiar L2 category (i.e., the dominant category) is typically thought to be more robustly encoded than the other (i.e., the nondominant category).

Concerning the locus of learners' difficulty, Darcy et al. (2012, 2013) reported learners' successful phonetic categorization (AXB discrimination) of L2 contrasts together with their nonnative-like repetition priming effects in a lexical decision task and their nonnative-like asymmetric patterns of lexical access in an auditory lexical decision task. Based on the observed disassociation between phonetic categorization and lexical encoding, the authors suggested that the problems experienced by L2 speakers in auditory word recognition stemmed from phonolexical encoding in addition to phonetic discrimination and categorization. Using an auditory lexical decision task to locate learners' difficulty, Barrios and Hayes-Harb (2021) observed contrastive accuracy patterns for the English /l/–/ɹ/ contrast (e.g., "[l]ecture," "[ɹ]ecture"). Native English speakers showed highly accurate and symmetric classification for words and pseudowords (interpreted as indicative of precise perceptual and lexical encoding), Chinese L2 learners exhibited higher accuracy for [l] words and [l] pseudowords (interpreted as perceptual coding difficulty), and the Korean L2 learners exhibited more accurate performance for [l] for words and [ɹ] for pseudowords (interpreted as lexical encoding difficulty). In fact, recent years have seen increased attention to how various patterns should be interpreted with respect to the sources of observed asymmetries, as this may well vary depending on the task used, the stimuli chosen, and the relationships between participants' languages (Barrios & Hayes-Harb, 2021; Cutler et al., 2006; Llompart, 2021).

L2 vocabulary training experiments offer opportunities to control the properties of the materials and the learning conditions (e.g., audio only, combining pictures and sounds) in order to investigate different aspects of phonolexical encoding of new words (Bürki et al., 2019; Escudero et al., 2008, 2014). For example, Escudero et al. (2008) attempted to replicate the previous eye-tracking studies (e.g., Weber & Cutler, 2004) with novel L2 (pseudo)words. Dutch learners of English were instructed to learn pseudowords containing the perceptually difficult English /ɛ/–/æ/ contrast that were accompanied by images of novel objects. Subsequently, the learner group that received only auditory input confused /ɛ/- and /æ/-pseudowords symmetrically, whereas the group receiving the auditory input together with the spelled forms (e.g., [tɛnzə], "tenzer") displayed a similar asymmetric pattern to that seen in earlier studies (e.g., Cutler et al., 2006; Weber & Cutler, 2004). The observed asymmetry was viewed as evidence that learners were building separate lexical representations for the words that included the difficult L2 contrast, and explicit information (orthographic input) was suggested to be necessary, with the auditory-only input being insufficient. Later studies on the role of orthographic input (Bürki et al., 2019; Escudero et al., 2014) further demonstrated that orthography supported lexical encoding when it was systematically consistent with the sound shapes of words. Hence, congruent orthography for the dominant phonological category in the L1 may support the development of a corresponding dominant L2 phonological category.

Vocabulary training studies in the L1 and L2 share a set of methodological challenges regarding the choice of properties for the words in the training set. Indeed, in addition to their form, auditory or visual, real words have a meaning, and researchers in each study must decide whether the meaning will be engaged (cf. Gaskell & Dumay, 2003, who trained only the orthographic form of novel L1 words), whether the meaning will be novel, and how it will be introduced (e.g., via pictures or by using definitions or translations). Each methodological choice has its advantages and disadvantages; criticism has been raised about the use of pseudowords associated with nonobjects (e.g., as in Escudero et al., 2008, 2014) on the grounds that when a novel lexical representation includes a concept that has not been previously encountered, and especially if the nonobject cannot be easily categorized (i.e., as an animal, a tool, etc.), its lexical encoding becomes more effortful and less reliable (Havas et al., 2018). In the present study, we opted for the use of pseudowords matched with pictures of low-frequency real objects in order to avoid the potential noise generated by the difficulties associated with encoding the meanings of pseudowords.

Additionally, delayed posttests are less common than immediate ones in the literature (but see Lindsay & Gaskell, 2013; Liu & Wiener, 2020), although there is a clear need to explore the consolidation of initially formed phonolexical representations in vocabulary training studies. Indeed, according to the complementary learning systems account (Gaskell & Dumay, 2003; Lindsay & Gaskell, 2013; McClelland et al., 1995), newly acquired words are processed in two learning systems. The first is responsible for the fast storage of episodic representations and their encoding into lexical representations intended for long-term storage. The other, slower system receives the encoded lexical representations and integrates them within the existing lexical networks. The process of integration of the newly encoded lexical representations into the mental lexicon (i.e., lexical consolidation) takes place over an extended period of time (Dumay & Gaskell, 2007; Gaskell & Dumay, 2003; Lindsay & Gaskell, 2013; Tamminen et al., 2017), and in lexical training studies, lexical consolidation is usually observed after 24 hours (Dumay & Gaskell, 2007; Tamminen et al., 2017). The present study therefore makes use of a delayed test to examine learners' phonological encoding of the contrasts of interest after 24-hour consolidation, in order to further test the fuzzy lexical representations hypothesis: Initial fuzziness may resist further encoding and make it more difficult for consolidation to take place.

To summarize, previous studies exploring phonolexical encoding in L2 word learning and processing have demonstrated that difficulty with the precise perceptual coding of L2 phonological contrasts contributes to the problems with the phonological encoding and subsequent retrieval of new lexical items, that is, fuzzy lexical representations. They have demonstrated asymmetries in performance that may reflect the precision with which dominant and nondominant members of a contrast are lexically encoded. Furthermore, the relationships between learners' languages, the use of words or pseudowords, and the consistency of orthographic encoding in either the L1 or L2 may be important variables influencing successful lexicalization.

Whereas previous findings support the role played by the L1 phonological category that is phonetically similar to its L2 counterpart in establishing a L2 dominant category, not much has been done on a learning scenario with a particularly high degree of difficulty, that involving an allophonic split, where the learners' L1 has both phones but lacks the phonological contrast, that is, separate phonemes in the L2 are allophones of the same phoneme in the L1 (Eckman et al., 2003; López Velarde & Simonet, 2020). Moreover, little is known about whether L1 phonotactic rules can constrain the learning patterns for problematic L2 contrasts (see, however, Weber & Cutler, 2006, a study demonstrating the role of L1 phonotactics in the segmenting of continuous speech in the L2); this study aims to explore this novel issue under the allophonic split learning scenario in order to investigate the role of L1 phonology, allophony, phonotactics, and memory consolidation in the phonolexical encoding of novel words with problematic L2 contrasts. It tests the predictions of the fuzzy lexical representations hypothesis that novel word learning will be less successful for the words with a difficult L2 phonemic contrast, as evidenced in the novel word recognition task, even when the participants show sensitivity to this contrast in the phonetic categorization task. Furthermore, the study evaluates whether free L1 allophonic variation and phonotactic constraints on the variation create additional problems with the encoding of the already difficult L2 contrast.

The Present Study

The present vocabulary training study focuses on the English phonemic contrast /v/–/w/ and explores the perceptual discrimination and lexical encoding patterns pertaining to this contrast demonstrated by L1-Chinese L2-English learners (hereafter Chinese L2 learners) speaking two varieties of Chinese. Specifically, the existence of a historically region-based difference in pronunciation in Chinese makes it possible to compare the speakers of a one-allophone ([w]-only) versus a two-allophone ([ʋ]–[w]-mixed) variety of Chinese rather than comparing speakers of two different L1s, who may differ on multiple dimensions.

English /v/–/w/ is considered a novel L2 contrast for Chinese native speakers: There is only the phoneme /w/ and a corresponding written "w" in pinyin, the pronunciation-based orthography for Chinese people in Mainland China. Meanwhile, in one variety of Chinese, /w/ is consistently pronounced as [w], whereas in another, /w/ can be pronounced as a labiodental approximant [ʋ] and a labiovelar approximant [w] interchangeably, that is, [ʋ] and [w] are in free allophonic variation and both represent the phoneme /w/ in Chinese (Wiener & Shih, 2013). Due to the mobility of the population and the role of nationwide media (Wang, 2007), strict regional criteria historically available for the categorization of Chinese [ʋ]–[w] use are now less accurate; accordingly, we describe our learners in this study based on their actual [ʋ]–[w] production. Phonetically, the Chinese [ʋ] is close to the English /v/, and the [w] in Chinese is almost the same as the English /w/ (see Wiener & Shih, 2013). This constitutes a scenario of an allophonic split, in which difficulties in perceiving the L2 contrast have been reported mostly for allophonic variation in complementary distribution (Barrios et al., 2016; Eckman et al., 2003; Shea, 2014) and, rarely, in free variation (López Velarde & Simonet, 2020).

More crucially, the interchangeable use of [w] and [ʋ] is disallowed before high back vowels /ɔ/ and /u/: No native speaker of Chinese will accept the pronunciation of [ʋɔ]/[ʋu] instead of [wɔ]/[wu]; [wɔ] and [wu] are in fact not in free variation within the [ʋ]–[w] two-allophone L1 variety (Wang, 2007; Wiener & Shih, 2013). On this basis, the current study is novel in testing how free allophonic variation in the L1 modulates L2 phonolexical encoding under different scenarios: when it is unconstrained by the vowel context in the L1 and when it is disallowed before particular L1 vowels (i.e., [wɔ] and [wu]). Although no clear predictions about the role of L1 phonotactics can be derived from the literature, we hypothesize that unconstrained free variation in the L1 may further impede L2 phonolexical encoding, and any L1 phonotactic constraints will modulate the transferability of the pattern of free allophonic variation from the L1 to the L2.

Additionally, orthographic support is believed to promote phonolexical encoding only when consistent phoneme–grapheme mappings are available (Escudero et al., 2008, 2014). While this study does not explore the role of orthography in the L2, the pinyin orthography "w" for the phoneme /w/ supports [w] as the dominant allophone of the phoneme /w/ in the L1, and this may contribute to the dominant status of the corresponding L2 phoneme /w/ in the /v/–/w/ contrast. Furthermore, the [ʋ]–/w/–"w" phoneme–grapheme mismatch at the subphonemic level may make the L2 contrast more difficult to perceive and lexicalize for the Chinese speakers of the [ʋ]–[w] variety.

The current study uses an auditory-only novel vocabulary training paradigm including a posttest and a delayed posttest to measure the accuracy and the processing speed of L2 lexicalization. The research also includes the results of an AX discrimination task (accuracy and processing speed) to demonstrate L2 phonological categorization, as well as a read-aloud task in Chinese to determine the use of [w] or [ʋ]/[w] in learners' L1 Chinese production (i.e., L1 [ʋ]–[w] preference). By teasing apart the effects of phonological categorization and lexicalization, the study will establish whether the locus of fuzziness in phonolexical representations is at the perception or lexicalization level (or both). By accounting for the phonotactic and allophonic patterns in the L1, the study will single out the role of the allophonic split in shaping the patterns of phonological categorization and lexicalization in the L2. The following research questions are addressed:

1. To what extent is the English /v/–/w/ contrast auditorily confusable for Chinese L2 learners of English?
2. In learning novel English "words" containing English /v/ and /w/ phonemes, to what extent do Chinese L2 learners experience difficulty in the encoding and lexicalization of the /v/–/w/ contrast, leading to fuzzy lexical representations?

3. To what extent does the presence of L1 [ʋ]–[w] free allophonic variation influence Chinese L2 learners' ability to perceive and lexically encode the English /v/–/w/ contrast, and what role is played by L1 phonotactic constraints on this variation?
4. During a 24-hour consolidation period, does consolidation take place for fuzzy L2 lexical representations to the same extent as for the novel words with robust phonolexical encoding?

Specifically, L2 learners' lower accuracy rates and higher reaction times (RTs) in the AX discrimination task for the /v/–/w/ contrast compared to the control contrast will demonstrate L2 /v/–/w/ discrimination difficulties (Research Question 1). Learners' lower accuracy and higher RTs in recognizing the words involving the /v/–/w/ contrast on the posttest will support the increased level of L2 lexical encoding difficulties (Research Question 2). Regarding Research Question 3, lower accuracy and higher RTs are expected for the /v/–/w/ contrast in the AX discrimination task and the posttest and delayed posttest for Chinese speakers of the [ʋ]–[w] two-allophone variety compared to the [w]-only one-allophone variety; it is further predicted that test stimuli involving the L1–L2 congruent phonotactic constraint (i.e., unconstrained free variation) will be more difficult to lexicalize than the others. Lastly, novel L2 words with the /v/–/w/ contrast are expected to be less robustly encoded than the control words for Chinese L2 learners after 24-hour consolidation (Research Question 4).

Method

Participants

The participants were undergraduate or graduate students recruited from a public university in the United States. They included 25 native speakers of English ($M_{\text{age}} = 19.80$ years, 95% CI [19.38, 20.22]) and 51 Chinese L2 learners of English ($M_{\text{age}} = 23.72$ years, 95% CI [22.80, 24.64]). The Chinese L2 learners' self-reported English proficiency corresponded to the university's international student admission requirement (minimum TOEFL scores: speaking, 22; listening, 24; reading, 26; writing, 24), indicating that these learners were advanced-level L2 learners. The 51 Chinese L2 learners came from different regions in China, showing variability in their [ʋ]–[w] preference in the read-aloud task (26 [w]-only, 25 [ʋ]–[w]-mixed). Experience related to language learning was carefully screened (see Appendix S1 and S2 in the Supporting Information online for the background questionnaire and the inclusion criteria, respectively). All the participants were right-handed, which ensured that the potential preference for the right or left response key was controlled. All participants signed an informed consent form and were paid for participating in the study.

Materials

The materials used in this study are publicly available via OSF at https://osf.io/un9m7. The target contrast for this study is English /v/–/w/, with /k/–/g/ chosen as the control contrast, that is, the baseline reference for the target–control comparison. Standard Mandarin has no underlying voiced stops (e.g., /g/), and the L1 phonemic contrast is actually /k^{h}/–/k/. However, the unaspirated /k/ is treated as “g” in pinyin romanization (and /k^{h}/ as “k”), thereby codifying /k^{h}/ and /k/ as separate phonemes (Duanmu, 2007).

Auditory Training and Posttest Materials

Two male and two female native speakers of the same American English variety each recorded a list of 24 disyllabic English pseudowords containing L2 target and control contrasts conforming to English phonotactic constraints (see Table 1, List A). We matched all pseudowords on phonological neighborhood density using the *Irvine Phonotactic Online Dictionary* (IPhOD) calculator (Vaden et al., 2009; see Appendix S3 in the Supporting Information online for details).

The contrasts /v/–/w/ and /k/–/g/ occurred in the word-initial position to match the phonotactic rules of Chinese that allow these consonants only in the word-initial position. Six familiar vowels that represent separate phonemes in Chinese were used in the second position in order to create three L1–L2 phonotactic conditions in relation to the preceding consonants /v/ and /w/: *congruent*, *incongruent*, and *new* conditions (see Table 1, Target). Specifically, in the *congruent* condition, the consonant–vowel combinations follow the phonotactic rules in L1 Chinese, whereas in the *incongruent* condition, two of the four combinations violate the L1 rules (in Chinese, /wɔ/ and /wu/ are allowed, whereas [ʋɔ] and [ʋu] violate the rules). The target *new* condition does not occur in L1 Chinese for either [w] or [ʋ]. Each of the six pairs of target pseudowords was paired with control pseudowords using the same vowel context. Since the incongruent condition is relevant only for the /v/–/w/ contrast, the /k/–/g/ contrast occurred only in two conditions: /kaɪ/–/gaɪ/, /keɪ/–/geɪ/, and /ku/–/gu/ (congruent); /kɔ/–/gɔ/, /ke/–/ge/, and /ki/–/gi/ (new).

Auditory Posttest-Only Materials: The Lures

We created another list of pseudowords, or lures, for use in the posttest and delayed posttest (see Table 1, List B). These were recorded by the same native speakers of American English. The lures were identical to the pseudowords

Table 1 IPA transcriptions of the target and control pseudowords and lures

Condition		Contrast	List A: pseudowords[a]		List B: paired lures[b]	
Target /v/–/w/	L1–L2 congruent	/waɪ/–/vaɪ/	/waɪrən/	/vaɪdəm/	/**v**aɪrən/	/**w**aɪdəm/
		/weɪ/–/veɪ/	/weɪmər/	/veɪzən/	/**v**eɪmər/	/**w**eɪzən/
	L1–L2 incongruent	/wu/–/vu/	/wukərn/	/vugən/	/**v**ukərn/	/**w**ugən/
		/wɔ/–/vɔ/	/wɔltaɪl/	/vɔsli/	/**v**ɔltaɪl/	/**w**ɔsli/
	L1–L2 new	/we/–/ve/	/wetʃər/	/vepənt/	/**v**etʃər/	/**w**epənt/
		/wi/–/vi/	/witlər/	/vinən/	/**v**itlər/	/**w**inən/
Control /k/–/g/	L1–L2 congruent	/kaɪ/–/gaɪ/	/kaɪtərn/	/gaɪmən/	/**g**aɪtərn/	/**k**aɪmən/
		/keɪ/–/geɪ/	/keɪdən/	/geɪsər/	/**g**eɪdən/	/**k**eɪsər/
		/ku/–/gu/	/kuləm/	/gulənt/	/**g**uləm/	/**k**ulənt/
	L1–L2 new	/kɔ/–/gɔ/	/kɔkərt/	/gɔrmɪl/	/**g**ɔkərt/	/**k**ɔrmɪl/
		/ke/–/ge/	/keʃər/	/geklɪn/	/**g**eʃər/	/**k**eklɪn/
		/ki/–/gi/	/kiθəm/	/giklər/	/**g**iθəm/	/**k**iklər/

Note. The switched word-initial consonants in List B are in boldface. L1 = first language; L2 = second language. [a]Used both in training and in posttests; [b]used only in posttests.

created for the training (Table 1, List A), with only one exception: The word-initial consonants were switched with their pairs in the contrast (e.g., for the pseudowords /waɪrən/ and /vaɪdəm/ in training, the corresponding lures are /vaɪrən/ and /waɪdəm/). These lures worked as potential "competitors" to the pseudowords that participants were required to learn (Table 1, List A), our aim being to establish whether (a) participants could discriminate the lures from the pseudowords they had learned; and (b) there was an asymmetric competition effect from one stimuli category to another, but not in the opposite direction.

All the recordings were examined by native speakers of English to ensure that production of consonants and vowels was accurate, and that the stimuli in lists A and B sounded like real English words. In order to minimize the individual speaker effect, including gender, in the training and exit test sessions, we exposed participants to the pseudowords produced by one male and one female native speaker randomly selected for each participant from the four participating native speakers. During the posttest and delayed posttest, participants were presented with the stimuli produced by the other male and the other female speaker (i.e., those whose stimuli had not been used in the training and exit test sessions). The purpose of introducing new speakers was to eliminate reliance on idiosyncratic pronunciation features and their later use as cues in the posttests.

Visual Materials

Twenty-four line drawings of real objects were taken from the public domain and were adjusted for visual salience by manipulating their brightness and contrast. These line drawings were associated with the 24 pseudowords in List A (Table 1). To remove potential drawing-specific pairing bias (e.g., one auditory stimulus is found to be easier or harder to associate with one specific picture due to undetected or unknown reasons), all drawings were randomly paired with the list of auditory stimuli for each participant. Each drawing was of a real object named by an English noun of low frequency (see Appendix S4 and S5 in the Supporting Information online for these line drawings and the related noun frequency information, respectively). These choices made it possible to rely on both form and meaning in learning novel words and avoid the unnecessary cognitive load associated with mapping the word form to nonexistent objects or concepts.

Procedure

The study followed a posttest–delayed posttest design, with approximately 24 hours (min. = 22 hours, max. = 25 hours) between the posttest and delayed

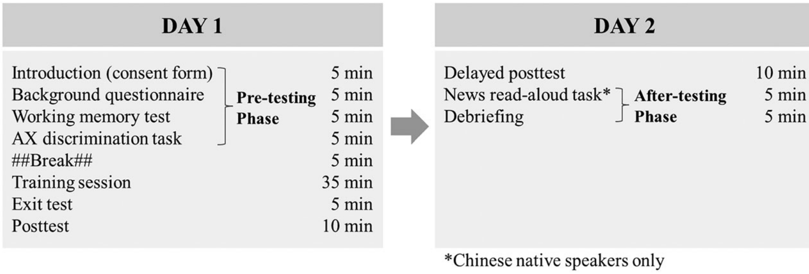

Figure 1 Summary of the training and testing procedure.

posttest sessions. A summary of the training and testing protocol is provided in Figure 1. Participants worked individually on a Dell computer. Line drawings were presented on Dell P2217Hb/P2419H monitors, and the auditory stimuli were presented over EarForce Z11 headphones. Participants were told that they would learn 24 words of an English dialect that shared the pronunciation system of English and that was passed down orally from generation to generation with no specific spelling associated with the oral words. The "English dialect" scenario was designed to reduce the learning load, since learning words from a language that the participants already know is less challenging (e.g., Escudero et al., 2008; Kroll & Sunderman, 2003). Participants were debriefed about the made-up nature of the "English dialect" immediately after the completion of the experiment so as to remove any misconceptions.

Pretest Working Memory Test

Working memory was used as a covariate in this study because previous research has suggested that it can mitigate the outcomes of word learning as a critical cognitive component (e.g., Masoura & Gathercole, 2005). A 5-min digit-span task (Stone & Towse, 2015) with 81 trials was administered as a pretest to assess the memory span of the participants. Participants' accuracy score ($M_{\text{L1 English}}$ = 60.89%, 95% CI [57.39, 64.40]; $M_{\text{L1 Chinese}}$ = 75.11%, 95% CI [72.46, 77.75]) was treated as a covariate in the later analyses (see Appendix S6 in the Supporting Information online for details and related covariate discussions). The split-half test reliability was .76.

Pretest AX Discrimination Task

The second pretest was an AX discrimination task that tested whether the target and control contrasts were auditorily confusable for Chinese L2 learners. The auditory stimuli were disyllabic "words" that conformed to English phonotactic constraints and that either were identical or differed as to the word-initial consonant: /v/ versus /w/, or /k/ versus /g/. As in the training and testing materials, the vowels following these consonants were /aɪ/, /eɪ/, /ɔ/, /u/, /e/, and /i/ (see Appendix S7 in the Supporting Information online for the AX list). The materials were recorded by the same native speakers as those who recorded the materials for the training and exit test sessions, and two different voices were used for the two stimuli in each trial to elicit phonological categorization (cf. Darcy et al., 2013). The test materials were presented via DMDX (Forster & Forster, 2003): Participants heard two auditory stimuli in one trial and were instructed to decide whether they were congruent by pressing the right ("Yes") or the left ("No") shift key on the keyboard. A total of 48 trials (24 matches, 24 mismatches) were presented in random orders for each participant, and the presentation order for the AX pairs was counterbalanced. Eight practice trials (with feedback) were given before the start of the task (with no feedback). Items in two of the 48 trials turned out to be real words, and the final analysis excluded these cases. Accuracy and RT data were collected by DMDX. The test reliability was .82 (Cronbach's alpha).

Training Session

The training session was composed of two subsessions: the initial training session and the self-testing–feedback session. During the initial training session, participants were presented with two sound icons together with the meaning-associated line drawing in a self-running PowerPoint presentation where the slides appeared at a predetermined speed. They were instructed to click on both sound icons to listen to the pronunciation of the English "word" produced by a male and a female native speaker. To facilitate memorization, the 24 items appeared in different combinations: 24 one-by-one trials, 6 four-by-four trials, 6 eight-by-eight trials, and 1 all-in-one trial were presented (see Figure 2 for examples).

The second session was a self-testing–feedback session to help participants prepare for the following exit test, which has a pass rate accuracy requirement of 90%. The 2 × 2 testing slides included two sound-playing icons for each stimulus with four line drawings (one correct answer and three distracters). Participants were instructed to click on the sound icons and select the picture matching the "word" they had just heard (Figure 3, left), and feedback would be given (Figure 3, right). A total of 24 trials were presented via PowerPoint.

Exit Test

After the training session, an exit test was conducted via DMDX (Forster & Forster, 2003). Figure 4 illustrates the presentation sequence. Participants were

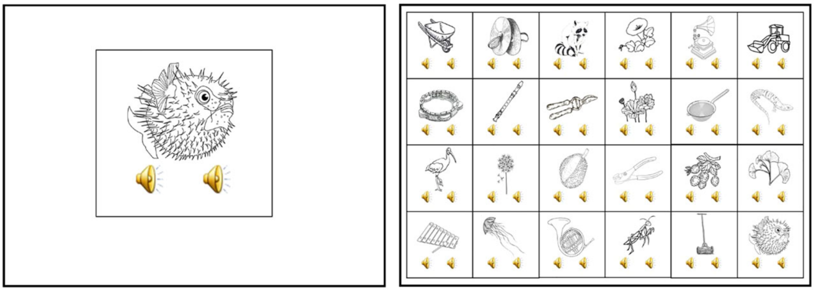

Figure 2 Sample initial training slides: one-by-one trial (left), 24 all-in-one trial (right).

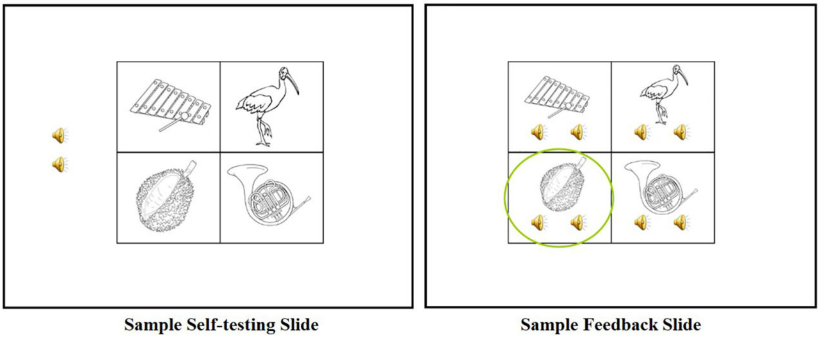

Figure 3 Sample self-testing-feedback session slides: self-testing (left), feedback (right).

instructed to complete a match–mismatch task so we could verify their successful learning of the pseudowords. For example, learners were instructed to press “Yes” (right shift key) when they heard /waɪrən/ and saw its corresponding line drawing in the training session, and to press “No” (left shift key) when they saw an irrelevant line drawing (randomly selected from the remaining 23 out of the 24 line drawings). A total of 48 trials (24 matches, 24 mismatches) including 4 practice trials were presented (see Figure 4). Specifically, all 48 audio stimuli (24 male, 24 female) used in the training session were randomly assigned as either correct answers or distracters in the exit test, and each stimulus appeared only once. Participants had to meet the 90% accuracy requirement in the exit test before proceeding to the posttest. Participants who failed to meet this requirement (3 L1 Chinese [w]-only users; 3 L1 Chinese [ʋ]–[w]-mixed users; 2 L1 English participants) had an additional chance to review the training slides and a second chance to complete a differently randomized exit

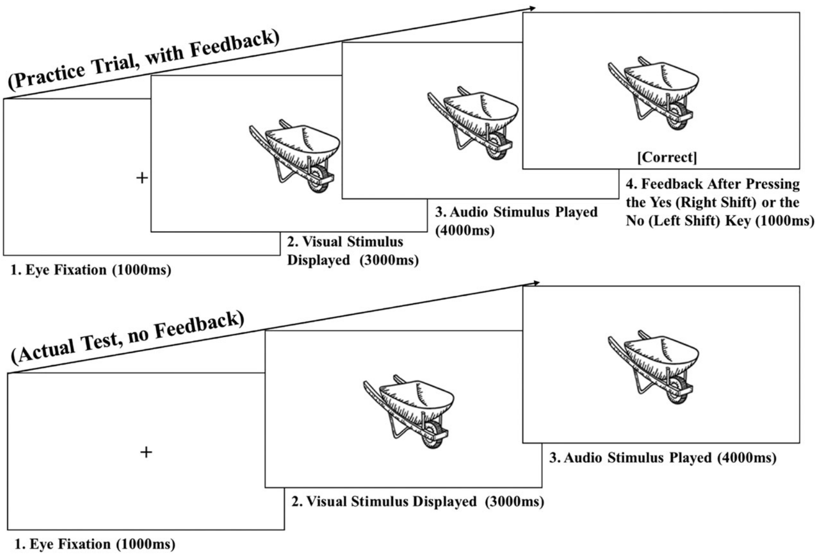

Figure 4 Presentation sequence for visual and audio stimuli: practice trial and the actual test trial in the exit test and posttests.

test; if they were successful in scoring over 90% on this second attempt, they proceeded to the posttests.

Posttest and Delayed Posttest

The purpose of the posttests was to examine if participants had accurately lexicalized the pseudowords they had learned by testing whether they could reject the lures with switched word-initial consonants. The posttest and the delayed posttest (a differently randomized version of the posttest) were also conducted via DMDX (Forster & Forster, 2003). Between the two posttests, participants had no additional training. The reliability of the posttest was .91 (Cronbach's alpha).

In the posttests, participants performed a similar match–mismatch task (see Figure 4 for the presentation sequence): Participants would hear a "word" and press "Yes" (right shift key) on the keyboard if the word they heard matched the picture on the screen, and press "No" (left shift key) if it did not. Unlike in the exit test, in the "Yes" condition, participants would hear the learned pseudowords (Table 1, List A) produced by a "new" native speaker, whereas for the "No" condition, they would hear the corresponding lures (Table 1, List

B). Eight practice trials with feedback (see also Figure 4, top panel) preceded the posttests with 96 trials (Figure 4, bottom panel). Importantly, participants were warned that they would hear words that would sound very similar to the words they had just learned, and that only the word-initial consonants would be slightly different. This instruction was intended to focus participants' attention on the critical word segments, reduce their uncertainty about the criteria to use in their decisions, and thereby increase their chances of a correct response. If participants failed to give a correct response, this could be attributed to encoding problems rather than incorrect response strategies. Data on accuracy at rejecting the lures and RTs were collected by DMDX.

After-Testing Phase: News Read-Aloud Task

All Chinese participants were asked to read five Chinese passages, edited versions of real news items published online, containing Chinese words with "w" occurring in different phonotactic positions, that is, consonant-vowel combinations (see Appendix S8 in the Supporting Information online for these passages). Recordings were coded by two native speakers of Chinese with prior linguistic knowledge to document L1 speakers' actual [ʋ] or [w] production. Participants were categorized as either [w]-only (i.e., 26 consistently used only [w]) or [ʋ]–[w]-mixed users (i.e., 25 interchangeably used [ʋ] and [w]; see Appendix S9 in the Supporting Information online for detailed [ʋ]–[w] distribution among participants). The two coders first coded two sample recordings together to reach a consensus, and then each of them independently coded the experiment recordings. The overall interrater reliability was 91.19%, and in case of discrepancies, the raters discussed them and reached a final agreement.

Data Analysis

To address the research questions and to optimize model fit, we computed multilevel modeling with crossed random effects of participant and item (Baayen et al., 2008) in the open-source statistical programming environment R (R Core Team, 2018), using the lme4 package (Bates et al., 2015) with the BOBYQA and the Nelder Mead optimization. We plotted predicted performance and interaction effects using the sjPlot package (Lüdecke, 2020). Multilevel logistic mixed-effects models were computed for accuracy analyses (1 = correct responses, 0 = incorrect responses), and multilevel mixed-effects models were applied to RTs. Table 2 lists all independent variables and covariates examined in the model-building procedure. Dummy coding was used for all categorical variables; accordingly, model effects are simple effects. Centering within clusters and centering at the grand mean were applied for different continuous

Table 2 Summary of variables

Variable	Coding	Model level	Centering	Related RQs
Independent variables				
Target	1 = Target, 0 = Control	1		1, 2, 4
Group	1 = L1 Chinese speakers, 0 = L1 English speakers	2		1, 2, 4
Consolidation	1 = Day 2, 0 = Day 1	1		2, 3, 4
Perception	(Continuous) AX discrimination accuracy or RT (aggregated by vowel context)	1	CWC	2, 3, 4
Congruent	1 = L1–L2 phonotactically congruent condition, 0 = others	1		3, 4
Incongruent	1 = L1–L2 phonotactically incongruent condition, 0 = others	1		3, 4
New	1 = L1–L2 phonotactically new condition, 0 = others	1		3, 4
Production	Chinese [ʋ]–[w] production: 1 = [w]-only, 0 = [ʋ]–[w]-mixed	2		3, 4
Lure	1 = /w/-lure, 0 = /v/-lure	1		3, 4
Covariate				
Working memory	(Continuous) Digit span accuracy score	2	CGM	1, 2, 3, 4

Note. RQ = research question; RT = reaction time; CWC = centering within clusters; CGM = centering at the grand mean.

variables (see Table 2), as appropriate (see Appendix S10 in the Supporting Information online for further discussion of centering). We entered all variables and interactions into the models through a stepwise procedure from Level 1 to Level 2 in order to examine potential fixed and random effects. The model building stopped once adding additional fixed and/or random effects no longer improved model fit significantly. The more parsimonious model at the last stage of comparison was chosen as the final model. The statistical significance level was set at alpha .05. For simple effects, statistical comparisons of interest were performed by changing the reference level of each categorical variable (i.e., releveling) and refitting the models (see Appendix S11 in the Supporting Information online for additional model outputs). We compared logarithm, inverse, and square-root transformations of the RTs in order to adjust distributional skewness (Kliegl et al., 2010; see Appendix S12 in the Supporting Information online for further discussion of each model).

Results

Research Question 1: Phonological Categorization

In order to address the effect of L1 (i.e., L1 English vs. L1 Chinese) on the phonological categorization of English /v/–/w/, we analyzed accuracy and RTs from both speaker groups. Only accurate trials were entered into the RT analysis, resulting in the exclusion of 7.52% of the data. Overall, L1 English speakers showed a high discrimination accuracy for both the target contrast (M = 96.83%, 95% CI [95.32, 98.35]) and the control contrast (M = 96.73%, 95% CI [94.74, 98.72]), whereas Chinese L2 learners showed a nativelike accuracy only for the control contrast (M = 96.52%, 95% CI [95.34, 97.71]), with a lower accuracy for discriminating the target contrast (M = 84.72%, 95% CI [81.01, 88.43]). Table 3 summarizes the final models for phonological categorization. The intercepts represent the logit accuracy (Model A) and the inverse-transformed RTs (Model B) of L1 English speakers in successfully discriminating the control contrast.

No significant accuracy differences were found between the L1 English and Chinese groups on the control contrast, group (reference level = L1 English speakers) $b = -0.33$, $SE = 0.43$, $p = .45$, or between the target and the control contrast within the L1 English speakers, target $b = 0.01$, $SE = 0.42$, $p = .99$ (see Table 3, Model A). The L1 English speakers significantly outperformed the L1 Chinese speakers only on the target contrast, group × target $b = -1.81$, $SE = 0.39$, $p < .001$. Similarly, in RTs (see Table 3, Model B), the two groups differed significantly only in discriminating the target contrast, group × target $b = -0.000112$, $SE = 0.0000279$, $p < .001$. Figure 5 presents the

Table 3 AX discrimination task: accuracy and reaction time (RT)

	Model A: perception accuracy					Model B: perception RT[a]				
Fixed effects	Estimate	95% CI	*SE*	*z*	*p*	Estimate	95% CI	*SE*	*t*	*p*
Intercept	4.17	[3.39, 4.95]	0.40	10.51	< .001	0.00123000	[0.00110766, 0.00134351]	0.00006060	20.24	< .001
Working memory	1.99	[−0.76, 4.75]	1.40	1.42	.16	0.00022800	[−0.00037907, 0.00083579]	0.00031100	0.73	.46
Target	0.01	[−0.82, 0.83]	0.42	0.01	.99	−0.00000651	[−0.00006930, 0.00005627]	0.00003210	−0.20	.84
Group	−0.33	[−1.16, 0.51]	0.43	−0.76	.45	−0.00006270	[−0.00020977, 0.00008492]	0.00007560	−0.83	.41
Group × Target	−1.81	[−2.57, −1.04]	0.39	−4.61	< .001	−0.00011200	[−0.00016675, −0.00005691]	0.00002790	−4.01	< .001
Random effects	Variance		*SD*			Variance		*SD*		
Item-Intercept	0.59		0.77			0.00000001		0.00007667		
Participant-Intercept	0.76		0.87			0.00000006		0.00024470		
Participant-Target						0.00000001		0.00008233		
Residuals						0.00000007		0.00026010		

Note. [a]Inverse-transformed RTs.

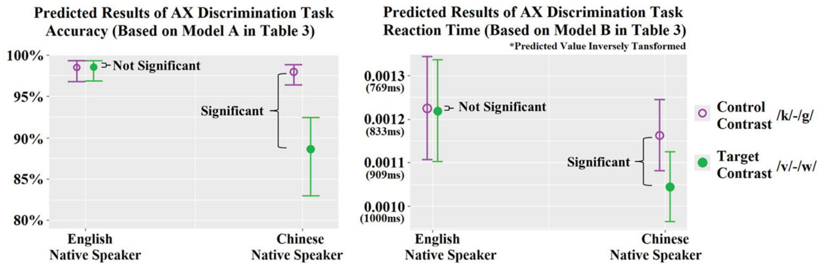

Figure 5 AX discrimination task illustrating a two-way interaction between first language background and contrast type: accuracy, based on Model A in Table 3 (left), reaction time, based on Model B in Table 3 (right). The error bars represent standard errors. The significance labels indicate the simple effect for target within each speaker group after releveling the reference level (group) and rerunning Model A and Model B (see Appendix S11 for the releveled models).

interaction effect for both models. Our results supported the conclusion that L1 English and L1 Chinese speakers differed significantly in discriminating only the English-specific /v/–/w/ contrast. Working memory was not a statistically significant covariate in either model.

Research Question 2: Lexicalization

In order to examine the effect of L1 (L1 English vs. L1 Chinese) on the lexicalization of English pseudowords, we analyzed accuracy and RTs from the posttest and delayed posttest. Only accurate trials were entered into the RT analysis, resulting in the exclusion of 27.82% of the overall data. Compared to native speakers, Chinese L2 learners encountered great difficulty when lexicalizing the English /v/–/w/ contrast (posttest accuracy M = 44.69%, 95% CI [37.63, 51.75]; delayed posttest accuracy M = 49.59%, 95% CI [42.28, 56.90]), with a substantial amount of variability within speakers. In particular, the low accuracy rates also resulted in the exclusion of 52.86% of the Chinese native RT responses on the /v/–/w/ contrast, compared to the exclusion of 17.65% of the /k/–/g/ RT responses. With chance-level accuracy, the variance in /v/–/w/ RT performance might be driven by guessing as opposed to lexical retrieval; however, in the current analysis we chose to retain only the accurate trials for an overall L1–L2 target–control RT comparison based on the statistically significant L2 RT differences observed between the accurate and inaccurate /v/–/w/ trials (posttest $M = 112$ ms, $t = -2.99$, $p = .002$; delayed posttest $M = 157$ ms, $t = -4.65$, $p < .001$), as well as the nonsignificant L2 RT differences observed between the inaccurate /k/–/g/ and inaccurate /v/–/w/

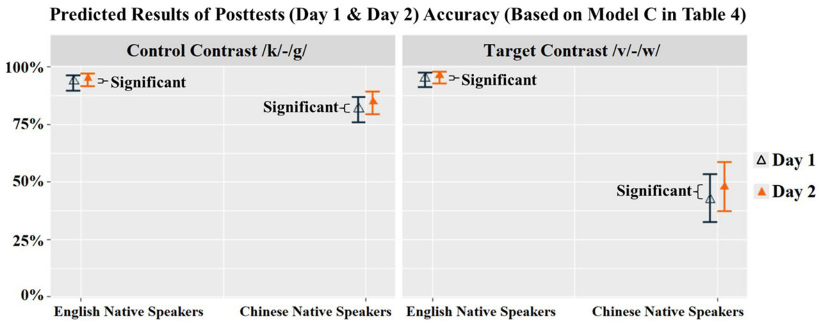

Figure 6 Posttest and delayed posttest: consolidation effect and two-way interaction between first language background and contrast type (based on Model C in Table 4). The error bars represent standard errors. The significance labels indicate the simple effect for consolidation within each speaker group after releveling the reference level (group and target) and rerunning Model C (see Appendix S11 for the releveled models).

trials (posttest $M = 57$ ms, $t = 1.03$, $p = .30$; delayed posttest $M = 79$ ms, $t = 1.48$, $p = .14$). Importantly, the results must be treated with caution given that a large proportion of L2 speakers' /v/–/w/ RT responses might reflect noise rather than successful lexicalization. Table 4 summarizes the best fitting models for L1 English and Chinese speakers' lexicalization, in which the intercepts represent the logit probability (Model C) and the log-transformed RTs (Model D) of L1 English speakers correctly rejecting the control lures in the posttest after controlling for phonological categorization and working memory. Unlike RTs in the AX discrimination task (inverse-transformed), posttest RTs were log-transformed to best correct for skewness (see Appendix S12 for further discussion).

The lexicalization accuracy model (Table 4, Model C) detected a significant native–nonnative group difference, group $b = -1.22$, $SE = 0.35$, $p < .001$: In both the posttest and the delayed posttest, English L1 speakers significantly outperformed Chinese L2 learners in accurately rejecting lures, after phonological perception (i.e., AX discrimination accuracy) was controlled for. Additionally, English L1 speakers performed equally well in lexicalizing the control and the target contrast, target $b = 0.27$, $SE = 0.30$, $p = .36$, whereas Chinese L2 learners performed significantly better on the control than on the target contrast, target $\times$ group $b = -2.08$, $SE = 0.34$, $p < .001$. Figure 6 presents the Target $\times$ Group interaction and an overall significant consolidation effect across both groups, which is discussed in the Research Question 4 section.

Table 4 Posttest and delayed posttest: accuracy and reaction time (RT)

	Model C: lexicalization accuracy					Model D: lexicalization RT[a]				
Fixed effects	Estimate	95% CI	*SE*	*z*	*p*	Estimate	95% CI	*SE*	*t*	*p*
Intercept	2.72	[2.16, 3.27]	0.28	9.61	< .001	6.80	[6.68, 6.93]	0.06	105.09	< .001
Target	0.27	[−0.31, 0.85]	0.30	0.91	.36	0.17	[0.10, 0.24]	0.04	4.81	< .001
Group	−1.22	[−1.91, −0.53]	0.35	−3.46	< .001	0.26	[0.10, 0.42]	0.08	3.23	.001
Consolidation	0.22	[0.10, 0.34]	0.06	3.49	< .001	−0.04	[−0.11, 0.04]	0.04	−0.94	.35
Perception	0.97	[0.48, 1.46]	0.25	3.87	< .001					
Working memory	3.88	[1.28, 6.48]	1.33	2.93	.003	−0.58	[−1.22, 0.06]	0.32	−1.79	.07
Group × Target	−2.08	[−2.76, −1.40]	0.34	−6.03	< .001	−0.01	[−0.08, 0.05]	0.03	−0.41	.68
Group × Consolidation						−0.03	[−0.13, 0.06]	0.05	−0.69	.49
Target × Consolidation						−0.09	[−0.15, −0.04]	0.03	−3.17	.002
Group × Consolidation × Target						0.13	[0.06, 0.21]	0.04	3.41	<.001
Random effects	Variance		*SD*			Variance		*SD*		
Participants-Intercept	0.93		0.96			0.07		0.26		
Participants-Target	0.94		0.97			0.01		0.08		
Participants-Consolidation						0.03		0.17		
Item-Intercept	0.23		0.48			0.01		0.08		
Item-Group	0.33		0.58							
Residuals						0.11		0.33		

Note. [a]Log-transformed RT.

Table 5 AX discrimination task: second language (L2) accuracy

	Model E: L2 perception accuracy				
Fixed effects	Estimate	95% CI	*SE*	*z*	*p*
Intercept	1.86	[1.27, 2.45]	0.30	6.20	< .001
Working memory	2.45	[−1.06, 5.96]	1.79	1.37	.17
Production	0.83	[0.16, 1.49]	0.34	2.44	.02
Random effects	Variance		*SD*		
Participants-Intercept	0.99		1.00		
Item-Intercept	0.77		0.88		

Additional differences were observed among the recorded RTs for the lures that participants successfully rejected (Table 4, Model D). In both the posttest and delayed posttest, Chinese L2 learners had significantly longer RTs in rejecting the lures in comparison to English L1 speakers, group $b = 0.26$, $SE = 0.08$, $p = .001$. Before consolidation, significantly longer RTs were found for the target compared to the control condition for both English L1 speakers, target $b = 0.17$, $SE = 0.04$, $p < .001$, and Chinese L2 learners, $b = 0.16$, $SE = 0.03$, $p < .001$ (see Appendix S11 for the releveled model), and this effect changed after consolidation for the nonnative group, target × group × consolidation $b = 0.13$, $SE = 0.04$, $p < .001$. Detailed results regarding consolidation are illustrated in the Research Question 4 section. Interestingly, the covariate working memory was statistically significant in the accuracy model (Model C) but not significant when considering RT performances (Model D).

Research Question 3: L1 Influence and Direction of Confusion

We then focused on the performance of L1 Chinese speakers by adding L1 [ʋ]–[w] preference and L1 phonotactic constraints into the model-building procedure to analyze the L1 influence on L2 phonological perception, L2 lexicalization, and the direction of confusion. Table 5 (Model E) summarizes the final model of L1 influence on L2 perception accuracy. The intercept represents the logit accuracy (Model E) of Chinese L1 [ʋ]–[w]-mixed users with sample mean working memory in discriminating the /v/–/w/ contrast. The L2 perception RT model fails to detect any significant fixed effects (see, however, Appendix S13 in the Supporting Information online for an extended discussion on age of acquisition and its influence on L2 perception and lexicalization).

Table 6 Posttest and delayed posttest: accuracy

	Model G: L2 lexicalization accuracy				
Fixed effects	Estimate	95% CI	*SE*	*z*	*p*
Intercept	−0.28	[−0.82, 0.26]	0.28	−1.01	.31
Consolidation	0.08	[−0.18, 0.34]	0.13	0.61	.54
Production	0.54	[−0.14, 1.22]	0.35	1.56	.12
Congruent	−0.68	[−1.28, −0.07]	0.31	−2.19	.03
Lure	−0.62	[−1.10, −0.15]	0.24	−2.57	.01
Perception	1.02	[0.34, 1.71]	0.35	2.92	.003
Working memory	2.15	[−1.31, 5.60]	1.76	1.22	.22
Production × Consolidation	0.39	[0.02, 0.76]	0.19	2.05	.04
Production × Congruent	0.70	[0.09, 1.31]	0.31	2.25	.03
Production × Lure	0.59	[0.14, 1.05]	0.23	2.56	.01
Congruent × Lure	1.24	[0.41, 2.06]	0.42	2.94	.003
Production × Congruent × Lure	−1.18	[−1.97, −0.39]	0.40	−2.91	.004
Random effects	Variance		*SD*		
Participants-Intercept	1.20		1.09		
Participants-Congruent	0.36		0.60		
Participants-New	0.85		0.92		
Item-Intercept	0.13		0.36		

In the L2 perception accuracy model (Table 5, Model E), the L1 [w]-only production group significantly outperformed the [ʋ]–[w]-mixed group in accurately discriminating the English /v/–/w/, production (reference level = L1 [ʋ]–[w]-mixed group) $b = 0.33$, $SE = 0.34$, $p = .02$: The [w]-only group had an averaged AX discrimination task accuracy of 88.94% (95% CI [84.61, 93.28]), whereas the [ʋ]–[w]-mixed group had an averaged accuracy of 80.33% (95% CI [74.45, 86.22]). Working memory was not a significant covariate in this model.

Table 6 (Model G) summarizes the best fitting model of L1 influence on L2 lexicalization. The intercept of this model is the logit probability of a randomly selected L1 [ʋ]–[w]-mixed user correctly rejecting a L1–L2 incongruent /v/-lure in the posttest, controlled for phonological categorization. Similarly, given that L2 speakers' /v/–/w/ lexicalization accuracy was at chance (overall accuracy 47.14%), one cannot assume that the accurate /v/–/w/ RTs reflected true lexical retrieval. Moreover, due to the complete absence of data in some key phonotactic conditions, conclusions based solely on the L2 /v/–/w/

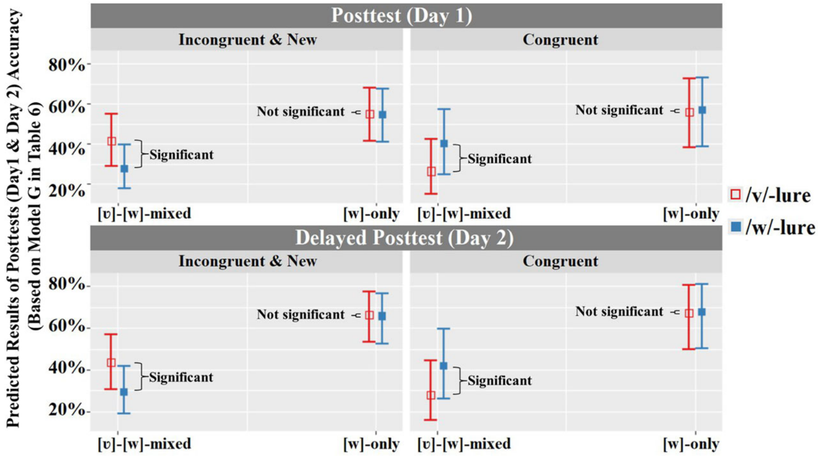

Figure 7 Three-way interaction between first language [ʋ]–[w] preference, phonotactic constraints, and lure type in predicting second language lexicalization accuracy (based on Model G in Table 6). The error bars represent standard errors. The significance labels indicate the simple effect for lure within each Chinese native speaker subgroup after releveling the reference level (production and consolidation) and rerunning Model G (see Appendix S11 for the releveled models).

lexicalization RTs are only tentative and will not be discussed here (see Appendix S14 in the Supporting Information online for further discussion).

Regarding L1 influence on L2 lexicalization accuracy (Table 6, Model G), the L1 [w]-only production group significantly outperformed the L1 [ʋ]–[w]-mixed group in discriminating all contrasts under all phonotactic constraints, with the two exceptions of the incongruent and new /v/-lures, production $b = 0.54$, $SE = 0.35$, $p = .12$). A significant three-way interaction between L1 [ʋ]–[w] production preference (reference level = L1 [ʋ]–[w]-mixed group), L1–L2 phonotactic constraints (reference level = L1-L2 incongruent condition), and lure type (reference level = /v/-lure) was observed, production × congruent × lure $b = -1.18$, $SE = 0.40$, $p = .004$, and their influence on L2 lexicalization is represented in Figure 7. Asymmetric lure interference reflected a joint effect of L1 [ʋ]–[w] preference and L1 phonotactic constraints: No significant directional bias in the competition effects was observed in the L1 [w]-only group, $b = -0.03$, $SE = 0.24$, $p = .90$ (see Appendix S11 for the releveled models). Meanwhile, for the L1 [ʋ]–[w]-mixed users, there was a significant competition bias effect with nondominant /v/-lures rejected less often than dominant /w/-lures in the congruent condition, congruent × lure

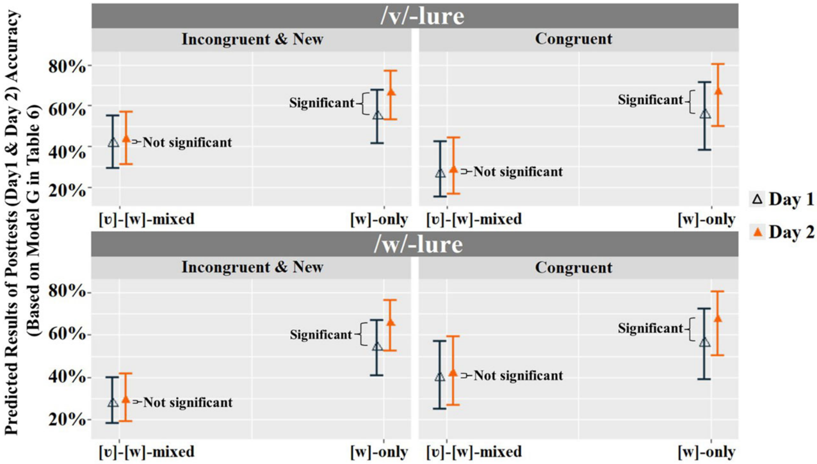

Figure 8 Second language lexicalization: accuracy improvement by first language [ʋ]–[w] preference groups (based on Model G in Table 6). The error bars represent standard errors. The significance labels indicate the simple effect for consolidation within each Chinese native speaker subgroup after releveling the reference level (production and lure) and rerunning Model G (see Appendix S11 for the releveled models).

$b = 1.24$, $SE = 0.42$, $p = .003$, and a significantly stronger /w/ to /v/ (i.e., dominant-to-nondominant) competition bias effect in the incongruent and new conditions, production × lure $b = 0.59$, $SE = 0.23$, $p = .01$. The covariate working memory was not significant in this model.

Research Question 4: 24-Hour Consolidation

The effects of 24-hour consolidation are reported in Table 4 and Table 6. An overall significant improvement in accuracy was found in the delayed posttest (i.e., higher accuracy rates of lexicalization) for both English and Chinese L1 speakers (Table 4, Model C), consolidation $b = 0.22$, $SE = 0.06$, $p < .001$ (see also Figure 6 for visualization). However, the improvement in lexicalization was not observed for all L1 Chinese speakers. The L2 lexicalization model (Table 6) reported an interaction effect between L1 [ʋ]–[w] preference and consolidation, production × consolidation $b = 0.39$, $SE = 0.19$, $p = .04$. Solely the L1 [w]-only group had benefited from the 24-hour consolidation (see Figure 8).

In RTs, English L1 speakers reacted significantly more slowly in rejecting the target lure than the control lure in the posttest on Day 1 (Table 4, Model

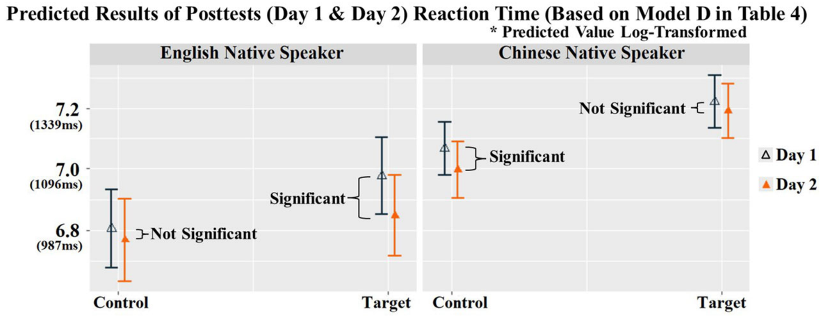

Figure 9 Lexicalization: reaction time improvement (based on Model D in Table 4). The error bars represent standard errors. The significance labels indicate the simple effect for consolidation within each speaker group after releveling the reference level (target and group) and rerunning Model D (see Appendix S11 for the releveled models).

D), target $b = 0.17$, $SE = 0.04$, $p < .001$, but their performance improved significantly after 24 hours, target × consolidation $b = -0.09$, $SE = 0.03$, $p = .002$, which might indicate that the /v/–/w/ contrast took more time for full consolidation of lexical representations in L1 English speakers than the very salient /k/–/g/ voicing contrast in the word-initial position. In contrast, Chinese L2 learners showed RT facilitation after 24 hours only in the control condition, $b = -0.07$, $SE = 0.03$, $p = .01$ (see Appendix S11 for the releveled model) and not in the target condition, $b = -0.03$, $SE = 0.03$, $p = .32$ (see Appendix S11 for the releveled model). These group differences across time are illustrated in Figure 9.

Discussion

In this study, we set out to investigate the effect of L1 language background on phonological categorization and lexicalization of a novel L2 contrast. In particular, we examined the role of phonotactically constrained allophonic variation in L1 production in forming new lexical representations in the L2. In this section, we will revisit the research questions formulated earlier in this paper and discuss how the findings of this study inform these questions. Discussion of working memory and other covariates can be found in the Supporting Information online.

Research Question 1: L2 Phonological Categorization

The first research question asked to what extent the English /v/–/w/ phonological contrast is auditorily confusable, given that there is only one /w/

phoneme in Chinese, and Chinese L2 learners vary in their use of one allophone ([w]) or two allophones ([ʋ] and [w]) of L1 /w/, with the two allophones phonetically close to the English /v/ and /w/. Overall, results demonstrate that even advanced-level Chinese L2 learners experienced certain difficulties in discriminating English /v/–/w/ (accuracy 84.72%), compared to their nativelike performance on the control contrast (accuracy 96.52%). Additionally, a wide range of variability (Figure 5) in Chinese L2 learners' target contrast discrimination indicates that some L2 learners indeed struggled with the perception of the English /v/–/w/ contrast. These findings increase our understanding of the role of phonological variables and the phonetic properties of phonemes in the L1 in shaping phonological categorization in a L2.

Research Question 2: L2 Lexicalization

The second research question asked to what extent Chinese L2 learners experience difficulties in the lexical encoding of the novel L2 /v/–/w/ contrast. Given the observed variability in [ʋ]–[w] pronunciation among our participants and the well-documented [ʋ]–[w] free allophonic variation in L1 Chinese to represent the single phoneme /w/ (e.g., Wang, 2007; Wiener & Shih, 2013), L2 learners were expected to experience problems with the encoding of the L2 phonological /v/–/w/ contrast. The results of accuracy and RT analyses provide unambiguous support for the conclusion that even advanced-level L2 learners experienced great difficulty in correctly lexically encoding and efficiently retrieving the novel L2 contrast, compared to the familiar /k/–/g/ contrast and to the outcomes of English native speakers. Importantly, all words were learned by both L1 English and Chinese participants to meet the 90% accuracy criterion in the exit test. Therefore, although the overall word form was learned, the problematic L2 contrast was imprecisely lexically encoded. It should be noted that L2 speakers' /v/–/w/ lexicalization accuracy was at chance, although systematic differences between the RTs of accurate versus inaccurate /v/–/w/ trials were observed. Conclusions based on accurate RT trials here should be treated carefully as only a secondary source of information that corroborates the accuracy results, as a large proportion of the L2 /v/–/w/ response (RTs) likely reflected noise pertaining to their predictive validity.

These findings provide empirical evidence in support of the fuzzy lexical representations hypothesis (Gor et al., 2021). A plausible explanation for the low accuracy and long RTs cannot rely solely on perceptual categorization difficulties: As discussed above, Chinese L2 learners were 84.72% accurate in discriminating the /v/–/w/ contrast. Given that phonological perception was controlled in the lexicalization model, it appears that unfaithful encoding (i.e., low accuracy) and slower retrieval (i.e., longer RTs) of newly established L2 representations involving the novel L2 contrast do not occur solely as a consequence of perceptual categorization difficulty. Rather, the locus of fuzziness in novel L2 lexical representations is also at the lexicalization level.

Research Question 3: L1 Influence on L2 Phonological Perception and Lexicalization

The third research question explored the influence of L1 phonology, including L1 free allophonic variation constrained by phonotactics in the L1 production of the study participants, on their success in L2 novel contrast perception and lexicalization. The direction of lexical confusion has also been examined. Our results extend the understanding of how L1 free allophonic variation can influence L2 phonological categorization and lexicalization: The interchangeable use of two allophones, [ʋ] and [w], in L1 Chinese to represent the phoneme /w/ makes it more difficult for Chinese L2 learners to accurately perceive and lexicalize the novel L2 /v/–/w/ contrast.

More specifically, consistent with one L1-Spanish L2-English study that reported a one-allophone L1 variety advantage over the two-allophone L1 variety with free allophonic variation in a scenario of L1–L2 allophonic split (López Velarde & Simonet, 2020), Chinese L1 speakers in our study who preferred to consistently use [w] rather than both allophones also enjoyed a clear advantage in correctly perceiving and lexicalizing the L2 /v/–/w/ contrast. Indeed, having one L1 phonemic category with a lot of variation makes it harder to separate that space into two separate L2 categories. Considering the particular case of L1 allophonic variation explored in this study, where both allophones represent the same /w/ phoneme and are used interchangeably in the same phonotactically legal positions, it is plausible that Chinese L1 speakers are used to categorizing both as /w/ even if they notice the phonetic differences (cf. the perceptual assimilation model; Best, 1995; Best & Tyler, 2007). Accordingly, it becomes difficult for them to consistently categorize and encode [v] and [w] as different phonemes in English.

Regarding L2 lexicalization, there were two exceptions to the systematic L1 [w]-only advantage discussed above: No significant [w]-only advantage was observed for phonotactically incongruent or new /v/-lures. The finding that the accuracy for rejecting phonotactically incongruent or new /v/-lures showed no L1 [ʋ]–[w] preference effect on lexicalization is in line with the following hypothesis: In the incongruent condition, Chinese L1 speakers will reject the use of [ʋ] over [w] in either L1 listening or production (Wang, 2007); therefore, no incorrect match can potentially occur for an incongruent [ʋ] with /w/ in Chinese. The same is true for the L1–L2 new condition: The consonant–vowel combinations unavailable in the L1 are unaffected by the mismatch introduced by L1 free allophonic variation. Consequently, the lexicalization of English /w/ in the incongruent or new condition is comparable across the two L1 allophonic varieties.

The findings regarding the directions of confusion reveal a more nuanced pattern of the influence of L1 phonotactics on L2 lexicalization. Among the L1 [ʋ]–[w]-mixed users, we observed a stronger competition effect caused by the dominant /w/-lure on the lexical encoding of /v/-words than by the nondominant /v/-lure on /w/-words (i.e., stronger dominant-to-nondominant than nondominant-to-dominant competition) in the incongruent and new conditions. The dominance of L2 /w/ for Chinese speakers was assumed, based on the phonemic status of /w/ (unlike [ʋ]) in Chinese, which is supported by orthography. This effect is consistent with the asymmetries in lexical access documented in previous research (e.g., Cutler et al., 2006; Darcy et al., 2013; Escudero et al., 2008): The lexical encoding of the nondominant L2 category that is unavailable in L1 phonology (/v/) is more likely to be incorrectly matched to the dominant category available in L1 phonology than vice versa, which leads to the asymmetries in lexical access found in our analysis.

At the same time, an opposing pattern was found for the congruent condition, in which the nondominant /v/-lure was less often rejected than the dominant /w/-lure (i.e., stronger nondominant-to-dominant than dominant-to-nondominant competition). A plausible explanation is that L1–L2 congruence leads to a disadvantage for discriminating /v/-lures from /w/-words because it makes it possible for two-allophone users to apply L1 phonotactic constraints licensing the interchangeable use of both allophones in the English words. Furthermore, for the L1 [w]-only users, the L1 [w]-only processing advantage offsets the effects of relying on L1–L2 congruence, both dominant-to-nondominant and nondominant-to-dominant, resulting in statistically comparable accuracies for the dominant and nondominant lures across the congruent, incongruent, and new conditions (see Figure 7). The overall accuracy level was much higher in [w]-only than in [ʋ]–[w]-mixed L1 allophone users—a major indicator of the advantage of consistent reliance on one phonetic allophone in L1 categorization. This pattern of competition for the dominant and nondominant lures critically extends our understanding of the asymmetries in lexical access. Our study has shown different patterns of asymmetry: from dominant to nondominant, nondominant to dominant, and no asymmetry (see also Barrios & Hayes-Harb, 2021). These patterns can be jointly predicted by the availability of allophonic variation in L1, the phonotactic constraints on L1 allophonic variation, and the preference for the L1 dominant phoneme (/w/) in L1 production.

Research Question 4: 24-Hour Consolidation

The fourth research question explores whether 24-hour consolidation can be observed in L2 lexicalization. Meanwhile, we asked a novel question about how L2 learners would deal with the lexical encoding of a difficult L2 contrast after a period of consolidation. Overall, the accuracy in /v/–/w/ and /k/–/g/ lexicalization for both L1 groups improved, without further exposure and practice, after 24 hours. This finding aligns well with earlier reports on the role of 24-hour consolidation in lexical integration (Dumay & Gaskell, 2007; Tamminen et al., 2017).

In particular, solely the L1 [w]-only users consolidated the initially unstable /v/–/w/ representations into more robust ones ($M_{\text{posttest}} = 53.69\%$; $M_{\text{delayed posttest}} = 61.86\%$), whereas the L1 [ʋ]–[w]-mixed users continued to struggle ($M_{\text{posttest}} = 35.33\%$; $M_{\text{delayed posttest}} = 36.83\%$). In addition to the perception and lexicalization advantage seen for the one-allophone L1 variety, this research highlights the possibility that consolidation in lexicalization may also be exclusive to that variety. Moreover, the low accuracy level among the [ʋ]–[w]-mixed users ($M = 36.08\%$) may indicate consistent misidentification patterns predicted by the type of lure and phonotactic constraints (see Figure 7). These two original findings have broadened the current understanding of fuzziness in lexicalization: Despite an overall improvement, consolidation may fail to take place in L2 learners with a L1 phoneme allowing free allophonic variation and thereby introducing fuzziness in L2 phonolexical encoding. Such L2 learners may even lexicalize the incorrect representations.

The accuracy rates were designed to measure the success of lexicalization, whereas the RT analysis provided a measure of efficiency in processing the contrast—a secondary source of observation that corroborates the accuracy data, as they show similar effects across the conditions and participant groups (see Table 4). A significant reduction in RTs that arguably accompanies successful lexicalization after 24 hours was observed in Chinese L1 speakers only for the familiar L2 control contrast, which indicates that further lexical consolidation (i.e., efficiency in lexical processing) only happens when the contrast is not fuzzy. Interestingly, this efficiency improvement was also present in the native English speaker RT data with the /v/–/w/ contrast (alongside a potential ceiling effect for /k/–/g/ processing, as the /k/–/g/ contrast is more salient). Moreover, the magnitude of improvement was larger than the relatively small improvement observed only for the control contrast in the Chinese L2 learner

RT data (see Figure 9). As the target contrast is not fuzzy for native speakers of English, that is, both phonemes in the contrast have robust phonological representations, the native speaker RT data further support our claim that more stable phonological representations produce lexical representations that are better consolidated than the fuzzy ones observed in L2s.

Limitations and Future Directions

Before discussing the broader implications of the current study, some limitations should be mentioned. The first concerns the potentially unequal treatment, in terms of amount of input, of those participants ($n = 8$) who failed to achieve the 90% pass rate on the exit test at the first attempt. They were given a second chance to review the training materials before retaking the exit test to meet the threshold requirement. Future studies increasing the sample size could help to minimize this unequal treatment problem. Additionally, a vocabulary training study is limited in the number of new words that participants can learn in one session. In the current experiment, in order to maximize efficiency and minimize fatigue, we included only 16 trials for each L1 phonotactic constraint condition (eight matches and eight mismatches). We hope that future research will improve on ours by expanding the study to a longer timeframe to deal with this limitation.

Conclusion

Despite these limitations, the present study deepens our understanding of the encoding and memorization of newly formed L2 lexical representations in several ways. First, it explores the impact of L1 allophonic variation on L2 phonological perception and lexicalization. The impact on perception has previously been demonstrated mainly in research on allophonic variation in complementary distribution (Barrios et al., 2016; Eckman et al., 2003; Shea, 2014), whereas the present study addresses free allophonic variation; and the impact on lexicalization has not previously been explored, to the best of our knowledge. This study demonstrates that L1 allophonic variation involving allophones that are acoustically close to the members of a novel L2 contrast does not guarantee either near-native perceptual categorization or successful L2 lexicalization, even for advanced L2 learners. Crucially, after controlling for perceptual categorization, L2 learners' lexical encoding of the L2-specific phonological contrast remains unfaithful, or fuzzy.

In addition to providing empirical support for the fuzzy lexical representations hypothesis (Gor et al., 2021), the findings from this research expand our understanding of the pattern of mapping from phonology to lexical representations. Multilevel modeling results show a three-way interaction between L1 allophonic variation, L1 phonotactics, and lure type, thereby supporting the view that patterns of asymmetries in lexical access are a joint product of multiple variables (Barrios & Hayes-Harb, 2021). Furthermore, this study is also original in reporting the role of L1 allophonic variation in shaping newly formed L2 lexical representations: Learners with no variability in their L1 phonemic inventory, that is, one-allophone users, are relatively more successful in L2 /v/–/w/ perception and lexicalization than two-allophone users. Additionally, L1 phonotactic constraints are also predictive of the degree of fuzziness in L2 lexicalization: When the vowel context is congruent across the L1 and L2, the two-allophone L1 users experience greater difficulty in the lexicalization of the dominant L1 category; and conversely, when the context is incongruent or new (involving a consonant–vowel combination that is novel in the L1), the two-allophone users experience greater difficulty in lexically encoding the nondominant L1 category. Finally, overall, 24-hour memory consolidation takes place in both L1 and L2 participants; however, only learners with no variability in L1 allophonic use consolidate the initially unstable representations, and further enhancement of lexical encoding (as assessed by increased processing speed) is only available for a L2 contrast also existing in the L1, and not for the critical /v/–/w/ contrast.

Final revised version accepted 6 March 2023

Open Research Badges

This article has earned an Open Materials badge for making publicly available the components of the research methods needed to reproduce the reported procedure. All materials that the authors have used and have the right to share are available at https://osf.io/un9m7. All proprietary materials have been precisely identified in the manuscript.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Barrios, S., & Hayes-Harb, R. (2021). L2 processing of words containing English /æ/-/ɛ/and /l/-/ɹ/ contrasts, and the uses and limits of the auditory lexical decision task for understanding the locus of difficulty. *Frontiers in Communication*, *6*, Article 689470. https://doi.org/10.3389/fcomm.2021.689470

Barrios, S. L., Namyst, A. M., Lau, E. F., Feldman, N. H., & Idsardi, W. J. (2016). Establishing new mappings between familiar phones: Neural and behavioral evidence for early automatic processing of nonnative contrasts. *Frontiers in Psychology*, *7*, Article 995. https://doi.org/10.3389/fpsyg.2016.00995

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). York Press.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech. In O. S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James EmilFlege* (pp. 13–34). John Benjamins. https://doi.org/10.1075/lllt.17.07bes

Bordag, D., Gor, K., & Opitz, A. (2022). Ontogenesis model of the L2 lexical representation. *Bilingualism: Language and Cognition*, *25*(2), 185–201. https://doi.org/10.1017/S1366728921000250

Bürki, A., Welby, P., Clément, M., & Spinelli, E. (2019). Orthography and second language word learning: Moving beyond "friend or foe?" *The Journal of the Acoustical Society of America*, *145*(4), 265–271. https://doi.org/10.1121/1.5094923

Cook, S. V., & Gor, K. (2015). Lexical access in L2: Representational deficit or processing constraint? *The Mental Lexicon*, *10*(2), 247–270. https://doi.org/10.1075/ml.10.2.04coo

Cook, S. V., Pandža, N. B., Lancaster, A. K., & Gor, K. (2016). Fuzzy nonnative phonolexical representations lead to fuzzy form-to-meaning mappings. *Frontiers in Psychology*, *7*, Article 1345. https://doi.org/10.3389/fpsyg.2016.01345

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

Darcy, I., Daidone, D., & Kojima, C. (2013). Asymmetric lexical access and fuzzy lexical representations in second language learners. *The Mental Lexicon*, *8*(3), 372–420. https://doi.org/10.1075/ml.8.3.06dar

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

Duanmu, S. (2007). *The phonology of standard Chinese*. Oxford University Press.

Dumay, N., & Gaskell, M. G. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science*, *18*(1), 35–39. https://doi.org/10.1111/j.1467-9280.2007.01845.x

Eckman, F. R., Elreyes, A., & Iverson, G. K. (2003). Some principles of second language phonology. *Second Language Research*, *19*(3), 169–208. https://doi.org/10.1191/0267658303sr2190a
Escudero, P., Hayes-Harb, R., & Mitterer, H. (2008). Novel second-language words and asymmetric lexical access. *Journal of Phonetics*, *36*(2), 345–360. https://doi.org/10.1016/j.wocn.2007.11.002
Escudero, P., Simon, E., & Mulak, K. E. (2014). Learning words in a new language: Orthography doesn't always help. *Bilingualism: Language and Cognition*, *17*(2), 384–395. https://doi.org/10.1017/S1366728913000436
Flege, J. E. (1995). Second language speech learning: Theory, findings and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). York Press.
Flege, J. E., & MacKay, I. R. (2004). Perceiving vowels in a second language. *Studies in Second Language Acquisition*, *26*(1), 1–34. https://doi.org/10.1017/S0272263104026117
Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods*, *35*(1), 116–124. https://doi.org/10.3758/BF03195503
Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition*, *89*(2), 105–132. https://doi.org/10.1016/S0010-0277(03)00070-2
Gor, K., & Cook, S. V. (2020). A mare in a pub? Nonnative facilitation in phonological priming. *Second Language Research*, *36*(1), 123–140. https://doi.org/10.1177/0267658318769962
Gor, K., Cook, S., Bordag, D., Chrabaszcz, A., & Opitz, A. (2021). Fuzzy lexical representations in adult second language speakers. *Frontiers in Psychology*, *12*, Article 732030. https://doi.org/10.3389/fpsyg.2021.732030
Havas, V., Taylor, J. S. H., Vaquero, L., de Diego-Balaguer, R., Rodríguez-Fornells, A., & Davis, M. H. (2018). Semantic and phonological schema influence spoken word learning and overnight consolidation. *Quarterly Journal of Experimental Psychology*, *71*(6), 1469–1481. https://doi.org/10.1080/17470218.2017.1329325
Liu, J., & Wiener, S. (2020). Homophones facilitate lexical development in a second language. *System*, *91*, 102–249. https://doi.org/10.1016/j.system.2020.102249
López Velarde, M., & Simonet, M. (2020). The perception of postalveolar English obstruents by Spanish speakers learning English as a foreign language in Mexico. *Languages*, *5*(2), Article 27. https://doi.org/10.3390/languages5020027
Jamieson, D. G., & Morosan, D. E. (1986). Training non-native speech contrasts in adults: Acquisition of the English /ð/-/θ/contrast by francophones. *Perception & Psychophysics*, *40*(4), 205–215. https://doi.org/10.3758/bf03211500
Kliegl, R., Masson, M. E., & Richter, E. M. (2010). A linear mixed model analysis of masked repetition priming. *Visual Cognition*, *18*(5), 655–681. https://doi.org/10.1080/13506280902986058

Kroll, J. F., & Sunderman, G. (2003). Cognitive processes in second language learners and bilinguals: The development of lexical and conceptual representations. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 104–129). Blackwell. https://doi.org/10.1002/9780470756492.ch5

Lindsay, S., & Gaskell, M. G. (2013). Lexical integration of novel words without sleep. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 608–622. https://doi.org/10.1037/a0029243

Llompart, M. (2021). Lexical and phonetic influences on the phonolexical encoding of difficult second-language contrasts: Insights from nonword rejection. *Frontiers in Psychology*, *12*, Article 659852. https://doi.org/10.3389/fpsyg.2021.659852

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/and /l/: A first report. *The Journal of the Acoustical Society of America*, *89*(2), 874–886. https://doi.org/10.1121/1.1894649.

Lüdecke, D. (2020). *sjPlot: Data visualization for statistics in social science* (R Package, Version 2.8.5) [Computer software]. https://CRAN.R-project.org/package=sjPlot

Masoura, E. V., & Gathercole, S. E. (2005). Contrasting contributions of phonological short-term memory and long-term knowledge to vocabulary learning in a foreign language. *Memory*, *13*(3–4), 422–429. https://doi.org/10.1080/09658210344000323.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

Pallier, C., Colomé, A., & Sebastián-Gallés, N. (2001). The influence of native-language phonology on lexical access: Exemplar-based versus abstract lexical entries. *Psychological Science*, *12*(6), 445–449. https://doi.org/10.1111/1467-9280.00383

R Core Team. (2018). *R: A language and environment for statistical computing* (Version 3.5.1) [Computer software]. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org

Shea, C. E. (2014). Second language learners and the variable speech signal. *Frontiers in Psychology*, *5*, Article 1338. https://doi.org/10.3389/fpsyg.2014.01338

Stone, J. M., & Towse, J. (2015). A working memory test battery: Java-based collection of seven working memory tasks. *Journal of Open Research Software*, *3*(1), Article e5. https://doi.org/10.5334/jors.br

Tamminen, J., Rastle, K., Darby, J., Lucas, R., & Williamson, V. J. (2017). The impact of music on learning and consolidation of novel words. *Memory*, *25*(1), 107–121. https://doi.org/10.1080/09658211.2015.1130843

Vaden, K. I., Halpin, H. R., & Hickok, G. S. (2009). *Irvine Phonotactic Online Dictionary* (Version 2.0) [Data set with tools]. http://www.iphod.com

Wang, F. (2007). Putonghua lingshengmu w-yin-v-du diyufenbu guilu ji xingchengyuanyin [The geographic distribution rules and reasons of zero initial "w" pronounced as "v" in Mandarin]. *Journal of Central South University (Social Science)*, *13*(4), 484–490.
Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, *50*(1), 1–25. https://doi.org/10.1016/S0749-596X(03)00105-0
Weber, A., & Cutler, A. (2006). First-language phonotactics in second-language listening. *The Journal of the Acoustical Society of America*, *119*(1), 597–607. https://doi.org/10.1121/1.2141003
Wiener, S., & Shih, Y. T. (2013). Evaluating the emergence of [ʋ] in modern spoken Mandarin. In Z. Jing-Schmidt (Ed.), *Increased empiricism: Recent advances in Chinese linguistics* (pp. 171–187). John Benjamins. https://doi.org/10.1075/scld.2
Ylinen, S., Uther, M., Latvala, A., Vepsäläinen, S., Iverson, P., Akahane-Yamada, R., & Näätänen, R. (2010). Training the brain to weight speech cues differently: A study of Finnish second-language users of English. *Journal of Cognitive Neuroscience*, *22*(6), 1319–1332. https://doi.org/10.1162/jocn.2009.21272

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Accessible Summary
Appendix S1. Background Questionnaire.
Appendix S2. Participant Inclusion Criteria Based on Experience Related to Language Learning.
Appendix S3. Phonological Neighborhood Density Information.
Appendix S4. Visual Materials: 24 Line Drawings.
Appendix S5. Noun Frequency Information Relating to the 24 Line Drawings.
Appendix S6. The Digit-Span Task and Discussions on Working Memory as a Covariate.
Appendix S7. AX Discrimination Task Item List.
Appendix S8. Chinese Read-Aloud Task Paragraphs.
Appendix S9. L1 [ʋ]–[w] Production Among L1 Chinese Participants.
Appendix S10. Centering in Multilevel Data Analysis.
Appendix S11. Additional Model Outputs (Releveling and Refitting).
Appendix S12. Reaction Time Data Transformation: Comparisons and Discussions.
Appendix S13. Age of Acquisition: An Extended Discussion.
Appendix S14. Extended Discussion for RQ 3: Reaction Time Model and Missing Data.