Short title
The role of phonology in non-native word learning.

Full title
The role of phonology in non-native word learning: Evidence from cross-situational statistical learning.

Authors and affiliations
Yuxin Ge[1, 2], Padraic Monaghan[1] & Patrick Rebuschat[1, 3]
[1] Lancaster University, UK
[2] Linguistics Research Centre, NOVA University Lisbon, Portugal
[3] University of Tübingen, Germany

Competing interests: The author(s) declare none.

Author note: P. M. and P. R. contributed equally to the supervision of this project and are joint senior authors in this report.

Address for correspondence
Yuxin Ge
Department of Linguistics and English Language
Lancaster University, Lancaster LA1 4YL
United Kingdom
Email address: y.ge4@lancaster.ac.uk

Abstract

Adults often encounter difficulty perceiving and processing sounds of a second language (L2). In order to acquire word-meaning mappings, learners need to determine what the language-relevant phonological contrasts are in the language. In this study, we examined the influence of phonology on non-native word learning, determining whether the language-relevant phonological contrasts could be acquired by abstracting over multiple experiences, and whether awareness of these contrasts related to learning. We trained English- and Mandarin-native speakers with pseudowords via a cross-situational statistical learning task (CSL). Learners were able to acquire the phonological contrasts across multiple situations, but similar-sounding words (i.e., minimal pairs) were harder to acquire, and words that contrast in a non-native suprasegmental feature (i.e., Mandarin lexical tone) were even harder for English-speakers, even with extended exposure. Furthermore, awareness of the non-native phonology was not found to relate to learning.

Keywords & key phrases

Implicit learning, statistical learning, cross-situational word learning, adult language learning, non-native phonology, lexical tone, minimal pairs,

Introduction

Learning new words is a continuous process throughout our lifetime. Starting from our first words in early childhood, we keep accumulating vocabulary in our native language (L1) and any additional language we learn (Davies, Arnell, Birchenough, Grimmond & Houlson, 2017). Child and adult learners can rapidly pick up new words, most of the time without explicitly being taught. This is impressive given the highly variable environment in which language learning happens. As illustrated by the classic Gavagai problem in word learning (Quine, 1960), upon the first encounter with a new word, it is often hard to define the appropriate referent as the word could refer to anything in the environment, and more often than not the learner does not get explicit instruction on the word-referent mapping. Similar situations arise when second or foreign language (L2) learners hear new words outside of the language classroom. Recent research on statistical learning has found a potential solution to this problem: child and adult learners can keep track of the linguistic information across multiple situations to aid word learning (known as cross-situational learning, CSL) (e.g., Escudero, Smit & Mulak, 2022; Monaghan, Schoetensack & Rebuschat, 2019; Rebuschat, Monaghan & Schoetensack, 2021; Suanda & Namy, 2012). That is, when the word occurs repeatedly over time, learners can follow the pattern across contexts and identify the always-co-occurring referent. In the classic CSL paradigm used in most studies (e.g., Yu & Smith, 2007), referential ambiguity was created by presenting multiple objects together with multiple pseudowords, with no clear indication of the word-referent mappings. This can be seen as a simplified representation of the real-life situation, as in the real world, there are usually more potential referents in the environment.

However, in learning a novel language, the challenge is more complex. In addition to referential uncertainty, in naturalistic language learning conditions, numerous words sound similar but have contrasting meanings (e.g., *bag vs. beg* in English; *pāo vs. gāo* in Mandarin).

Learners need to accurately perceive and discriminate these unfamiliar non-native sound contrasts to learn words, which is an ability that starts diminishing during infancy (Kuhl, Stevens, Hayashi, Deguchi, Kiritani & Iverson, 2006; Werker & Tees, 1984). In the bilingualism literature, this perceptual issue has not been well examined and little research has directly investigated how non-native sounds interfere with word learning (for exceptions, see Chandrasekaran, Sampath & Wong, 2010; Silbert, Smith, Jackson, Campbell, Hughes & Tare, 2015; Wong & Perrachione, 2007). Our current study will address this gap by exploring the effect of phonology on non-native word learning using a CSL paradigm. It also provides insights into the role of awareness in statistical learning.

Statistical learning of non-native vocabulary

Although learners of non-native languages usually have already developed sophisticated representations of various conceptual meanings, they face similar challenges as children in connecting these concepts to the appropriate forms. Thus, understanding how language learners deal with this referential uncertainty problem is not only an important topic in early word learning literature (e.g., Markman, 1990; Tomasello & Barton, 1994; Smith & Yu, 2008), but also has implications for second and foreign language research (e.g. Monaghan, Ruiz & Rebuschat, 2021; Smith, Smith & Blythe, 2011; Walker, Monaghan, Schoetensack & Rebuschat, 2020). One influential approach is the statistical learning account, which shows that learners can extract statistical regularities from the linguistic contexts to facilitate language learning (e.g., Maye & Gerken, 2000 and Maye, Werker & Gerken, 2002 for sound discrimination; Saffran, Aslin & Newport, 1996 for word segmentation; see Isbilen & Christiansen, 2022; Siegelman, 2020; Williams & Rebuschat, 2022, for reviews). For word learning specifically, a classic cross-situational statistical learning paradigm has been widely explored（Smith & Yu, 2008; Yu & Smith, 2007). CSL

proposes that learners can extract and accumulate information about word-referent co-occurrences across multiple ambiguous encounters to eventually identify the correct referents.

There has been extensive evidence on the effectiveness of CSL for both children (e.g., Childers & Pak, 2009; Smith & Yu, 2008; Suanda, Mugwanya & Namy, 2014; Yu & Smith, 2011) and adults (e.g., Gillette, Gleitman, Gleitman & Lederer, 1999; Smith, Smith & Blythe, 2009, 2011; Yurovsky, Smith & Yu, 2013). For example, in an early study, Yu and Smith (2007) created referentially ambiguous learning conditions for adult learners, presenting multiple words and pictures at the same time, and tested whether learners made use of the word-picture co-occurrence information across learning events to acquire the appropriate mappings. It was found that after only six minutes of exposure, learners were able to match pictures to words at above chance levels even in highly ambiguous conditions with four words and four pictures presented in each learning event. Monaghan et al. (2019) extended the CSL settings and presented participants with motions rather than referent objects. The results showed that participants were able to extract syntactic information from cross-situational statistics and acquire words from different syntactic categories (i.e., nouns, verbs). And more recently, it has been reported that CSL can also drive syntactic acquisition of word order (Rebuschat, Monaghan & Schoetensack, 2021).

However, most of the CSL literature left aside the important impact of phonology on word learning. There are two potential issues related to this. First, in most CSL studies, the word or pseudoword stimuli used were phonologically distinct (e.g., pseudowords such as *barget, chelad* in Monaghan et al., 2019). However, as reported by Escudero, Mulak and Vlach (2016), the degree of phonological similarity between words can affect learning outcomes. Escudero and colleagues found that minimal pairs that differ in only one vowel (e.g., DEET-DIT) were harder to identify after cross-situational learning than consonant

minimal pairs (e.g., BON-TON) and non-minimal pairs (e.g., BON-DEET). Thus, to better resemble natural learning conditions, it is necessary to examine the effects of both phonologically similar and distinct words in CSL and the first aim of our study is to provide further evidence for this.

Second, previous research has largely included pseudowords that contained phonemes that were familiar to the participants (in the sense that they existed in their native languages) and phoneme combinations that followed the phonotactics of their native language(s) (e.g., Escudero et al., 2016; Monaghan & Mattock, 2012; Monaghan et al., 2019; see Hu, 2017, and Junttila & Ylinen, 2020, for an exception). In other words, CSL studies tended to create a situation for learning additional words in L1. Naturally, the use of familiar phonemes and phoneme combinations could make the discrimination between these pseudowords less challenging. To extend the results to second language research, it is important to consider the phonological difficulties associated with non-native sounds (e.g., Dupoux, Sebastián-Gallés, Navarrete & Peperkamp, 2008; Iverson et al, 2003; Rato, 2018; Rato & Carlet, 2020; Takagi & Mann, 1995; Wong & Perrachione, 2007). Tuninetti, Mulak and Escudero (2020) trained Australian English speakers with novel Dutch and Brazilian Portuguese vowel minimal pairs in a CSL setting. The vowel pairs were classified into perceptually difficult or easy pairs based on acoustic measurements (Escudero, 2005). The perceptually easy minimal pairs contained vowel contrasts that could be mapped to two separate L1 vowel categories, and the perceptually difficult ones had no clear corresponding L1 contrasts (Escudero, 2005 – Second Language Linguistic Perception model (L2LP); Best and Tyler, 2007 – Perceptual Assimilation-L2 model (PAM-L2)). It was found that learners performed the best in non-minimal pair trials, followed by perceptually easy pairs and then perceptually difficult pairs, suggesting the role of L1-L2 phonetic and phonological similarity in CSL. A more recent study by Escudero, Smit and Mulak (2022) directly compared cross-situational word learning

by L1 and L2 speakers of English. The authors presented the same set of English pseudowords as in Escudero et al. (2016) to English-native and Mandarin-native speakers, either in a consonant, vowel or non-minimal pair condition. Overall, the English group performed better in identifying word-picture mappings in all minimal pair conditions than the Mandarin group, though the Mandarin group also showed some degree of learning.

These previous CSL studies provided evidence for the crucial role of phonology in the acquisition of novel, non-native words. However, there are several gaps in our knowledge of how non-native cues affect learning. Firstly, previous studies focused primarily on segmental contrasts (i.e., vowels and consonants), leaving aside the suprasegmental cues (e.g., tone). Suprasegmental development can diverge from segmental development in L2 acquisition (e.g., Hao & Yang, 2018; Sun, Saito & Tierney, 2021), and the integration of suprasegmental and segmental features can be challenging for beginner learners (Zou, Chen & Caspers, 2017). It is thus important to explore how suprasegmental cues affect cross-situational learning of non-native words. Furthermore, previous research looked at the reconfiguration of phonological features (phonemes) from L1 to the novel language, and the perceptual difficulty and learning depended on L1-L2 phonemic differences (e.g., Tuninetti et al., 2020). But in natural word learning, there also exist phonological features that, in the learners' L1s, are not used contrastively at the lexical level at all. In such cases, perception and learning are not only affected by L1-L2 phonemic differences, but also depend on learners tuning in to these novel features in the first place. Our study specifically addresses these issues by exploring how English-native speakers with no prior experience in learning tonal languages develop their ability to use lexical tones in word learning.

Another important aspect of our study design is that we presented only one word per trial together with multiple referents. This mirrors natural language learning situations more closely as it requires learners to keep track of the minimal pairs throughout learning. Previous

CSL studies, following the paradigm used by Yu and Smith (2007), usually presented several words together with several referents in one trial. This means that minimal pairs were presented to participants in a single situation during training, which might make the phonological differences more salient to learners (Escudero et al., 2016, 2022; Tuninetti et al., 2020). However, in natural language learning settings, minimal pairs tend not to occur in immediate proximity but have to be acquired by uncovering the contrastive property of certain phonological features across situations. This raises the question of how it is possible for learners to distinguish minimally contrasting words when the contrast is not explicitly available during learning, but must be extracted from correspondences that occur in the wider communicative environment.

Research questions and predictions

The current study explored how non-native phonology influences cross-situational word learning. The following research questions are addressed:

RQ1: Do minimal pairs pose difficulty during cross-situational learning compared to phonologically distinct words?

RQ2: Do minimal pairs that differ in non-native phonological contrasts pose further difficulty compared to minimal pairs with contrasts that are similar to native sounds?

RQ3: Does learners' non-native sound perception develop during cross-situational learning?

We predicted that minimal pairs would be more difficult to learn compared to non-minimal pairs even when those minimal pairs are presented across multiple experiences of the language as in natural language learning (RQ1). Moreover, minimal pairs with non-native phonological contrasts would generate the greatest difficulty in learning (RQ2). We also hypothesized that the learning process would lead to non-native phonological advances, and

learners would improve in their performance on the non-native minimal pairs over time (RQ3).

To compare the performance on native versus non-native contrasts, we created a pseudoword vocabulary based on Mandarin Chinese and recruited Mandarin-native and English-native speakers to take part. Mandarin Chinese is a tonal language employing syllable-level pitch changes to contrast word meanings, which is particularly difficult for learners whose native languages lack such prosodic cues (e.g., Chan & Leung, 2020; Francis, Ciocca, Ma & Fenn, 2008; So & Best, 2010). In the tonal perception literature, many studies have reported that Mandarin Tone 1 vs Tone 4 is hard for non-native listeners when tested in monosyllables (e.g., Kiriloff, 1969; So & Best, 2010, 2014). However, in Mandarin Chinese, over 70% of the vocabulary consists of multi-syllabic words (two or more syllables), and learners encounter tones more often in di- or multi-syllables rather than isolated monosyllables (Jin, 2011). Thus, the previous work on monosyllabic perception may not be representative in the case of Mandarin word learning. In our design, we decided to use disyllabic words to better reflect the real Mandarin word-learning situation. In disyllabic structures, the prosodic positions (initial vs final syllable) and tonal contexts (the preceding and following tones) play a role in perception as well (Chang & Bowles, 2015; Ding, 2012; Hao, 2018). There are relatively few studies taking into account this tonal environment effect, but according to Hao (2018), English-native learners of Mandarin can identify T1 and T4 at word-initial positions better compared to T2 and T3. Thus, we decided to use T1 and T4 as they are likely to be easier for non-native listeners in the disyllabic environment. We wanted the tones to be relatively easily captured by the non-native (English) participants before learning because previous studies have found that better tonal word learning outcome is associated with better pre-learning tonal perception (e.g., Cooper & Wang, 2013; Wong &

Perrachione, 2007). Since our learning task is short (~10 min), the use of the easier tones might allow us to observe clearer learning effects.

We predicted that for English-native speakers, minimal pairs that contrast in lexical tones would be the most difficult (i.e., with lowest accuracy), followed by minimal pairs that differ in consonants and vowels. The non-minimal pairs would be relatively easy to learn. For Mandarin-native speakers, previous studies suggested that tonal language speakers rely more on segmental than tonal information in word processing (e.g., Cutler & Chen, 1997; Sereno & Lee, 2015; Yip, 2001). Thus, we predicted that learning of tonal pairs would still be lower than that in consonant/vowel pairs, but Mandarin speakers would learn tonal minimal pairs better than English speakers. It was also hypothesized that English-native speakers' performance on tonal contrasts would improve across the task.

Experiment 1: Learning non-native sound contrasts from cross-situational statistics

The study was preregistered on the Open Science Framework (OSF) platform. The preregistration, the materials, anonymized data and R scripts are available at: https://osf.io/2j6pe/.

Method

Participants

Fifty-six participants were recruited through either the Department of Psychology at Lancaster University (N=28) or the social media platform WeChat (N=28). To estimate the sample size needed for expected effects, we ran power analyses for the interaction effect of language group, learning trial type and block with Monte Carlo simulations of data. (The power analysis R script can be found on the OSF site referred to above.). All participants were university students (aged 18~30) and spoke either English or Mandarin Chinese as a native language. The L1 English participants had no previous experience learning any tonal

languages before taking part in the study. Thirteen participants in the L1 English group reported knowing more than one language or language variety[1] (Arabic, Dutch, French, German, Korean, Russian, Spanish,) at beginner, intermediate or advanced levels[2.] Twenty-four L1 Mandarin participants reported speaking more than one language (English, French, Indonesian, Italian, Japanese, Korean, other Chinese varieties), among which 22 participants spoke English as a second/foreign language. Participation was voluntary and the Psychology Department participants received credits for their university courses.

Materials

Cross-situational learning task. The CSL task involved learning 12 pseudoword-referent mappings. All pseudowords were disyllabic, with CVCV structure, which satisfies the phonotactic constraints of both Mandarin Chinese and English. The pseudowords contained phonemes that were similar between the two languages. This made the pseudowords sound familiar to both groups of participants. Each syllable in the pseudowords carried a lexical tone which is either Tone 1 (high) or Tone 4 (falling) in Mandarin Chinese, which created a simplified lexical tone system.

Six different consonants /p, t, k, l, m, f/ and four different vowels /a, i, u, ei/ were combined to form eight distinct base syllables (/pa, ta, ka, li, lu, lei, mi, fa/), which were

[1] A comparison between learning performance of English L1 participants with and without foreign language experience was conducted, as learning more than one language was found to be associated with better tonal statistical learning abilities (e.g., Wang & Saffran, 2014) and cognitive functions (see Adesope, Lavin, Thompson & Ungerleider, 2010, for review). However, adding FL experience (with or without) as a fixed effect in our model did not significantly improve model fit ($\chi^2(1) = 0.168$, $p = .682$), nor did the interaction between block, trial type and FL experience ($\chi^2(1) = 7.968$, $p = .336$). Thus, for the main analyses, we will not include FL experience as a factor. The bi/multilingualism effect in CSL had mixed findings in previous research as well, with some reporting a bilingual advantage (Escudero, Mulak, Fu & Singh, 2016) and some observing similar performance among monolinguals and bilinguals (Poepsel & Weiss, 2016).

2 To further disentangle the bi/multilingualism effect, we tested if participants with different proficiency levels in their FLs perform differently. We contrasted participants with no FL experience, beginner-level FLs, and those with intermediate/advanced-level in at least one FL. However, adding the FL proficiency effect did not improve model fit ($\chi^2(2) = 1.484$, $p = .476$), not the interaction between proficiency, block and trial type ($\chi^2(11) = 7.624$, $p = .747$). Therefore, for the main analyses, we will not include this effect.

further paired to form six minimally distinct base words (/pami, tami, kami, lifa, lufa, leifa/). Three of the base pseudowords differed in the consonant of the first syllable (/pami, tami, kami/), which were assigned to the consonantal set; and the other three differing in the vowel of the first syllable were assigned to the vocalic set (/lifa, lufa, leifa/). The second syllables in the pseudowords were held constant in each set to ensure that the words in each set were minimal pairs. These base words were then superimposed with lexical tones. The first syllable of each of the six base words was paired with either T1 or T4, and the second syllable always carried T1. This resulted in an additional tonal minimal pair contrast (e.g., /pa1mi1/ vs /pa4mi1/) among the pseudowords. Therefore, a total of 12 pseudowords were created (full list shown in Table 1). The pseudowords (with their corresponding referent objects) were later paired to create consonantal, vocalic, tonal, and non-minimal pair trials, and each pseudoword-referent mapping could occur in different trial types based on the paired foil. All pseudowords have no corresponding meanings in English or Mandarin Chinese, though the base syllables are phonotactically legal in the languages. The audio stimuli were produced by a female native speaker of Mandarin Chinese. The mean length of the audio stimuli was 800ms.

<Insert Table 1 about here>

Twelve pictures of novel objects were selected from Horst and Hout's (2016) NOUN database and used as referents. The pseudowords were randomly mapped to the objects, and we created four lists of word-referent mappings to minimize the influence of a particular mapping being easily memorisable. Each participant was randomly assigned to one of the mappings.

Background questionnaire. We collected information on participants' gender, age and history of language learning. The questionnaire was adapted from Marian, Blumenfeld and Kaushanskaya's (2007) Language Experience and Proficiency Questionnaire (LEAP-Q).

Participants were asked to specify their native languages and all non-native languages they have learned, including the age of learning onset, contexts of learning, lengths of learning, and self-estimated general proficiency levels.

Debriefing questionnaire. After the CSL task, participants were given a debriefing questionnaire to elicit retrospective verbal reports about their awareness of the phonological patterns of the pseudowords and whether they noticed the tonal contrasts in the language. The questionnaire was adapted from Rebuschat, Hamrick, Riestenberg, Sachs and Ziegler (2015) and Monaghan et al. (2019). It contained seven short questions ordered in a way that gradually provided more explicit information about the language, which reduced the possibility that participants learn about the explicit patterns of the language from questions. The first three questions were general questions about the strategies used when choosing referents. The next two questions narrowed down the scope and asked if participants noticed any patterns or rules about the artificial language and the sound system. The final two questions explicitly asked if participants noticed the lexical tones.

Experimental design and procedure

All participants were directed to the experiment platform Gorilla to complete the tasks. After providing informed consent, participants completed the background questionnaire, followed by the CSL task. The latter took approximately 10 minutes to complete and consisted of a 2-alternative forced-choice task, where learners selected the referent for a spoken word from two objects. There were four types of trials in CSL – consonantal, vocalic, tonal and non-minimal pair trials. We manipulated the target and foil objects in each trial to create the different trial types. Each target object was paired with different foils according to the trial type. For instance, the target object for *pa1mi1* was paired with the (foil) object for *ta1mi1* in a consonantal minimal pair trial; and the same object for

pa1mi1 was paired with the (foil) object for *pa4mi1* in a tonal minimal pair trial. Taking an example of a consonantal minimal pair trial, participants saw two objects – object A for *pa1mi1* and object B for *ta1mi1* – and heard the word *pa1mi1*. They needed to select object A and reject object B. The labels of these two objects only differ in the first consonant, and hence participants had to be able to distinguish *pa1mi1* from *ta1mi1*, as well as learned the associations between each of these words and the object to which they are paired, in order to make the correct selection. Similarly, in vocalic minimal pair trials, the labels of the two objects differed in one vowel (e.g., *li1fa1* vs *lu1fa1*), and in tonal minimal pair trials, the labels of the two objects differed in the lexical tone (e.g., *pa1mi1* vs *pa4mi1*). The non-minimal pair trials contained objects that were mapped onto phonologically more distinct words (e.g., *pa1mi1* vs *li4fa1*). Choosing the correct referent object was expected to be harder if participants were not able to distinguish the labels associated with the two objects. For example, English-native participants may have difficulty distinguishing the tonal pairs such as *pa1mi1* vs *pa4mi1.* And when they see two objects referring to *pa1mi1* and *pa4mi1* and hear the word *pa1mi1*, they may not be able to select the corresponding object. This manipulation allowed us to explore whether and to what extent minimal pairs cause difficulty in CSL, and if non-native minimal pairs such as the tonal pairs pose even greater difficulty for English-native speakers. And, more importantly, whether adult learners improve in the perception of non-native sounds (i.e., tones in this study) through a short CSL session.

The occurrence of each trial type was controlled in each block and throughout the experiment. There were six CSL blocks, with 24 trials each, resulting in 144 trials in total. Each of the four trial types occurred six times in one block, leading to a total of 36 trials across the experiment. Within each learning block, each of the 12 pseudowords was played twice, and each of the novel objects was used as the target referent twice (in two different trial types). The foil object was randomly selected from all the possible minimal pairs using

the randomization function in excel. Hence, in each block, each pseudoword occurred twice with the target object, and once each with two other foil objects. Throughout the experiment, each pseudoword occurred 12 times with the target object, and no more than three times with each of the six possible foils. Thus, the associations between pseudowords and their targets were strengthened over the co-occurrences, and the associations between pseudowords and foil objects remained low. Additionally, the correct referent picture was presented on the left side in half of the trials and the position of the target was determined by the randomization function as well. There were four types of word-referent mappings randomly created, and each participant was randomly allocated to one of the mapping types. Participants' accuracy at selecting the correct referent was recorded throughout the experiment, and their response time in each trial was measured.

After the CSL task, participants completed the debriefing questionnaire, in the question sequence outlined above. Only one question was presented on the screen each time.

Trial procedure

In each CSL trial, participants first saw a fixation cross at the centre of the screen for 500ms to gather their attention. They were then shown two objects on the screen, one on the left side and one on the right, and were played a single pseudoword. After the pseudoword was played, participants were prompted to decide which object the pseudoword referred to. They were instructed to press 'Q' on the keyboard if they thought the picture on the left was the correct referent of the word and 'P' for the picture on the right. The objects remained on the screen during the entire trial, but the pseudoword was only played once. The next trial only started after participants made a choice for the current trial. No feedback was provided after each response. Figure 1 provides an example of a CSL trial.

The keyboard response recorded participants' answers in each trial and was used to calculate accuracy. It also allowed us to measure reaction time more accurately than mouse clicking on the pictures, as it avoided interference from the time taken to move the cursor.

<Insert Figure 1 about here>

Data analysis

We excluded participants who failed to successfully complete the initial sound check or failed to complete the CSL task within one hour. We also excluded individual responses that lasted over 30 seconds. This was because these participants failed to follow the instructions to respond as quickly and accurately as possible. After excluding these data points, we visualized the data using R for general descriptive patterns. We then used generalized linear mixed effects modelling for statistical data analysis. Mixed effects models were constructed from null model (containing only random effects of item and participant) to models containing fixed effects. We tested if each of the fixed effects improved model fit using log-likelihood comparisons between models. A quadratic effect of block was also tested for its contribution to model fit, as block may exert a quadratic rather than linear effect. The planned analyses were explained in our preregistration.

Results

Performance on cross-situational learning task

Accuracy. Figure 2A presents the overall percentage correct responses of the L1 English and L1 Mandarin groups. Both groups showed learning effects – with improvements in accuracy from chance level to 66.8% (L1 English group) and 70.5% (L1 Mandarin group) at the end of the learning. For the different minimal pair trials (as in Figure 2B & 2C), there was a common pattern across groups that accuracy was the highest in non-minimal pair trials.

For the L1 English group, the learning of tonal minimal pair trials was not clear, with participants performing at around chance level throughout the task. But there seemed to be improvement in the vocalic (block 6 accuracy 66.1%) and consonantal (block 6 accuracy 56.5%) trials, as the mean accuracies showed an increasing pattern throughout the experiment. For the L1 Mandarin group, the accuracies in the tonal, vocalic and consonantal trials were all above chance at the end of the CSL session.

<Insert Figure 2 about here>

As outlined in our preregistration, to investigate whether learning was different across language groups and trial types, we ran generalized linear mixed effects models to examine performance accuracy across learning blocks. We started with a model with the maximal random effects that converge, which included item slope for learning block, language group and trial type, and participant slope for learning block and trial type. Then we added fixed effects of learning block, language group, trial type and the 3-way interaction to test if they improve model fit. We also tested for a quadratic effect for block.

Compared to the model with only random effects, adding the fixed effect of learning block improved model fit significantly ($\chi^2(1) = 5.478$, $p = .020$), adding English versus Mandarin language group did not significantly improve fit ($\chi^2(1) = 0.072$, $p = .789$), adding trial type (consonant, vowel, tone, non-minimal pair) improved model fit further ($\chi^2(3) = 32.246$, $p < .001$) as well as the 3-way interaction ($\chi^2(7) = 26.847$, $p < .001$). The quadratic effect for block did not result in a significant difference ($\chi^2(8) = 9.740$, $p = .284$). The best-fitting model is reported in Table 2.

<Insert Table 2 about here>[3] [4]

[3] The table shows the summary of the best-fitting model, however, these statistics were not reported in detail as the primary focus of our analysis (as in our pre-registration plan) was to compare models, which we reported in the text.

[4] Table 2 shows the model with non-minimal pair trial as the reference level. In supplementary materials, Table S2 present models with other trial types as reference levels respectively.

Exploratory analyses. We carried out exploratory analyses to examine the effect of block and language group on each trial type separately. For tonal trials, adding the fixed effect of language group ($\chi^2(1) = 4.2111$, $p = .040$) and block ($\chi^2(1) = 3.8967$, $p = .048$) significantly improved fit, whereas the interaction effect did not improve model fit further ($\chi^2(1) = 0.0012$, $p = .973$). In Table 3 we presented the best-fitting model for T trials. The L1 English group scored significantly lower than the L1 Mandarin group in tonal trials, but both groups of learners showed overall improvement across blocks. In all other trial types, language group did not significantly improve model fit (consonantal $\chi^2(1) = 0$, $p = 1$; vocalic $\chi^2(1) = 0.1928$, $p = .661$; non-minimal pair $\chi^2(1) = 0.7839$, $p = .376$) and learning block did improve fit (consonantal: $\chi^2(1) = 15.606$, $p < .001$; vocalic: $\chi^2(1) = 5.7728$, $p = .016$; non-minimal pair: $\chi^2(1) = 15.452$, $p < .001$). Adding the language group by block interaction significantly influenced the model fit for consonantal ($\chi^2(1) = 5.0314$, $p = .025$) and non-minimal pair trials ($\chi^2(1) = 4.4963$, $p = .034$), but not for vocalic trials ($\chi^2(1) = 0.8722$, $p = .350$).

<Insert Table 3 about here>

To disentangle the performance of the two language groups in each trial type, we ran separate mixed-effects models on the Mandarin-native and the English-native dataset. For the Mandarin-native group, adding the effect of block ($\chi^2(1) = 11.01$, $p <.001$), trial type ($\chi^2(3) = 18.576$, $p <.001$) and block by trial type interaction ($\chi^2(3) = 22.067$, $p <.001$) significantly improved model fit. The Mandarin-native participants performed best in non-minimal pair trials, followed by consonant/vowel trials, and then tonal trials (as illustrated in Table S3). A similar pattern was observed for the English-native group (Table S4).

Reaction time. There was a general tendency of reducing reaction time across learning blocks for both groups of participants, especially from Block 1 to the following blocks (Figure S1). But no clear relationship between trial type and response time can be observed.

As reaction time is not our focus here, all figures are presented in supplementary materials. To investigate whether the fixed effects of block, language group and trial type affected participants' reaction time, we used generalized mixed effects models with a log-link Gamma function, as the raw reaction time data were positively skewed. The inclusion of block ($\chi^2(1) = 24.159$, $p < .001$) and language group ($\chi^2(1) = 9.881$, $p = .002$) significantly improved model fit. The effects of trial type ($\chi^2(3) = 6.221$, $p = .101$) and the 3-way interaction ($\chi^2(7) = 4.436$, $p = .728$) did not further improve fit. The best-fitting model can be found in Table S5. There were significant effects of learning block and language group on participants' reaction time. L1 English participants reacted significantly faster than L1 Mandarin participants.

Retrospective verbal reports

Participants' answers to the debriefing questions were coded to explore if awareness or explicit knowledge of the pseudoword phonology predicts performance on the CSL task. We focused primarily on the awareness measure of the English-native speakers, as the Mandarin-native speakers were all expected to be aware of the tonal differences.

The awareness coding followed the guidance of Rebuschat et al.'s (2015) coding scheme, ranking from full awareness to complete unawareness. Participants who reported using lexical tones to distinguish words strategically were considered "full awareness" (Q1~3), those who mentioned lexical tones in response to the questions on patterns of the language or the sound system were considered "partial awareness" (Q4~5), and those who only mentioned that they noticed lexical tones after the question explicitly asked so were coded as "minimal awareness" (Q6~7). Participants who reported that they did not think lexical tones contrast word meanings were deemed "unaware". All participants who reported minimal, partial or full awareness were included as "aware" participants and others as

"unaware". Two researchers independently coded the retrospective verbal reports to ensure consistency and agreement on criteria.

Proportion of aware and unaware participants. Following the criteria outlined above, we found that no learners developed full awareness of the tonal cues. Participants reported no specific strategy and simply guessed (e.g., *I guessed some with how similar it was to the word in English*) at the beginning of the study. Twenty-one participants reported at least noticing the pitch-related change, with wording differing among tone, intonation, pitch, and high/low sound (e.g., *One of the syllables changed tone*). The remaining seven participants reported no awareness of pitch-related changes. Among the aware learners, we observed different degrees of awareness. Following Schmidt (1990, 1995), eight participants were classified as being aware at the level of *understanding* as they specifically mentioned that tones change meanings. The remaining thirteen participants were classified as being aware at the level of *noticing* as they perceived the tonal changes but did not link them to meaning changes. However, we did not find significant differences between the noticing and understanding groups in an exploratory analysis, and hence the two groups were pooled as a single 'aware' group in further analyses.

Performance of aware and unaware participants in CSL task. As shown in Figure 3, the learning trajectories of aware and unaware participants are not significantly different. There was an unexpected drop in accuracy for the unaware participants at learning block 6, specifically in the tonal and vocalic minimal pair trials.

<Insert Figure 3 about here>

To explore the influence of awareness on learning performance for the L1 English group, we constructed models with fixed effects of block, trial type, awareness status (aware vs unaware), and the 3-way interaction in order. The inclusion of trial type ($\chi^2(3) = 10.770$, $p = .013$) and block ($\chi^2(1) = 11.925$, $p < .001$) led to better model fit. Awareness ($\chi^2(1) = 0$, $p =$

1) and the interaction effect ($\chi^2(7) = 5.172$, $p = .639$) did not further influence model fit significantly. Table 4 summarizes the final model.[5]

<Insert Table 4 about here>

Exploratory analysis. We investigated if aware and unaware participants differ at the end of the CSL task. The results showed that only trial type ($\chi^2(3) = 13.943$, $p = .003$) significantly improved fit, but not awareness ($\chi^2(1) = 3.037$, $p = .081$) nor the interaction ($\chi^2(3) = 1.897$, $p = .594$). The best-fitting model is provided in Table S7. Considering only the most challenging tonal minimal pair trials in the last block, we found that the aware participants performed significantly better than the unaware ones ($t(26) = 2.2193$, $p = .035$), with an average accuracy of 0.55 and 0.38 respectively.

Discussion

Experiment 1 confirmed that adults can learn non-native words by keeping track of cross-situational statistics (Escudero et al., 2016, 2022; Tuninetti et al., 2020), and this was possible even when those minimal pairs were not immediately apparent and available within a single learning trial. The experiment also showed that the presence of minimal pairs and non-native speech sounds can interfere with learning outcomes. As predicted, we found that phonologically distinct items (non-minimal pairs) resulted in better learning than phonologically similar items (RQ1). Additionally, learners' familiarity with the phonological contrasts influenced learning as words with non-native contrasts (tonal minimal pairs) were less accurately identified (RQ2). It is worth noting that Mandarin participants' performance in tonal trials was also lower than that in consonant/vowel trials, despite lexical tone being in their native phonology. This is consistent with our prediction and previous studies, as Mandarin speakers might weigh segmental information greater than tonal information.

[5] Additional Table S6 presents models with other trial types as reference levels.

The three-way interaction between trial type, language group, and learning block showed that learners' language background and knowledge of the new phonology are critical in how they perform in the CSL task. Specifically, the English-native speakers were significantly less accurate in tonal trials compared to the Mandarin-native speakers but were comparable in all other types of trials. Although these non-native contrasts resulted in more difficulties, we found that learners improved on these challenging contrasts after CSL (RQ3). The block effect and language group effect (without interaction) on tonal trials means that both L1 English and L1 Mandarin groups improved in tonal minimal pairs over time. However, the learning effect was still small, especially for L1 English participants. Their performance on the tonal trials was not significantly above chance after six learning blocks. One possible explanation is that the amount of exposure was insufficient. The CSL task took, on average, less than 10 minutes to complete. Thus, the training might be too minimized for participants to capture a subtle non-native cue, especially when this non-native tonal cue was embedded in minimal pairs, and learning required a highly accurate perception of the acoustic contrast. Therefore, we carried out Experiment 2 to explore if doubled exposure to the same materials can lead to improved learning outcomes.

Regarding participants' awareness of the phonological properties of the words, we did not observe the effect of awareness among L1 English participants across learning blocks, though at the final block (Block 6), aware participants scored significantly higher than unaware participants in tonal trials. However, this difference resulted from a drop in unaware participants' performance in the final block, rather than a rise in aware learners' performance. Thus, it is unlikely that being aware of the tones benefited the learning outcomes. Rather, as shown in Figure 3, the unaware learners showed an accuracy decline in all trial types at the final block, which might reflect a loss of attention (e.g., due to distraction or fatigue) towards

the end of the task. In Experiment 2, we further investigated if awareness would play a role after a longer learning exposure.

Experiment 2: The effect of extended training on learning

Method

Participants

Twenty-eight participants were recruited through the Department of Psychology at Lancaster University for course credits. This sample size matched the group size in Experiment 1. One participant was excluded because their native language was Cantonese. The remaining 27 participants were university students (aged 18~26) who spoke English as a native language and had no previous experience learning tonal languages. Eleven participants reported knowing more than one language[6].

Materials and procedure

Auditory and visual stimuli were the same as in Experiment 1. The procedure replicated Experiment 1, except with twice the amount of CSL trials (i.e., participants went through the Experiment 1 CSL task twice, 12 blocks in total). Experiment 2 was preregistered on OSF: https://osf.io/2m4nw/.

Results

Performance on cross-situational task

Accuracy. Figure 4A presents the overall performance of participants across the 12 learning blocks. There is a clear improvement in accuracy from chance level to 70.5% at the end of the learning. Like Experiment 1, the L1 English participants performed best in non-

[6] We had technical issues with the language history dataset, so the exact foreign languages were unknown.

minimal pair trials, followed by clear learning in consonantal and vocalic trials. However, learning in tonal trials was still not observed (Figure 4B).

<Insert Figure 4 about here>

To be comparable to Experiment 1, we ran similar mixed effects models to examine the effect of learning block and trial types. We included a comparison between L1 English participants in Experiment 1 and participants in Experiment 2 to test the effect of short versus long (doubled) exposure. The fixed effect of learning block ($\chi^2(1) = 3.394$, $p = .065$) and exposure ($\chi^2(1) = 0.656$, $p = .418$) did not significantly improve model fit. But adding trial type ($\chi^2(3) = 29.146$, $p < .001$) and the 3-way interaction ($\chi^2(7) = 42.022$, $p < .001$) led to significant improvement. The quadratic effect for block did not result in a significant difference ($\chi^2(8) = 14.274$, $p = .075$). The best-fitting model is reported in Table 5[7].

<Insert Table 5 about here>

Exploratory analysis. We further ran separate models to test if exposure played a role in any particular trial type. The results showed that exposure effect was not significant in all trial types.

Reaction time measurement. Participants' reaction time for correct responses showed a similar decreasing tendency as in Experiment 1 (Figure S2). The generalized mixed effect models revealed that adding exposure ($\chi^2(1) = 0$, $p = 1$) did not improve fit, but the effect of trial type ($\chi^2(3) = 9.193$, $p = .027$) and block ($\chi^2(1) = 38.15$, $p < .001$) and the 3-way interaction ($\chi^2(7) = 28.852$, $p < .001$) all improved model fit significantly. The best-fitting model is provided in Table S9.

Retrospective verbal reports

[7] Additional Table S8 presents models with other trial types as reference levels.

Proportion of aware and unaware participants. Three participants were coded as fully aware as they reported using tonal cues strategically without being explicitly asked so (e.g., *…after I loosely assigned words to pictures, I more listened out for the differences in the tones of the words…*). A further eighteen participants reported that they noticed the tone/pitch difference in the language when explicitly asked so (e.g., *The tones of the words did change, which is how I correlated the word to the picture*). The remaining six participants reported no awareness of the tonal difference. The total number of aware participants was the same in Experiment 1 and 2, though in Experiment 2 a few participants developed full awareness of the tones but none in Experiment 1.

Performance of aware and unaware participants in CSL task. As in Experiment 1, the aware and unaware participants shared similar learning trajectories (Figure 5).

<Insert Figure 5 about here>

Since the aware and unaware subgroups did not differ in general accuracy, we ran mixed effects models for tonal trials specifically to explore if participants who noticed the existence of tones performed better. The results showed that none of the fixed effects improve model fit compared to a random effect model (learning block: $\chi^2(1) = 3.3854$, $p = .066$; exposure: $\chi^2(1) = 0.107$, $p = .744$; awareness: $\chi^2(1) < .001$, $p = .976$; 3-way interaction: $\chi^2(3) = 1.2278$, $p = .746$). In Experiment 1, we found a significant difference between aware and unaware participants in tonal trials at the end of the CSL task, but in Experiment 2, no such difference was detected ($t(25) = 0.57781$, $p = .569$).

Discussion

Experiment 2 revealed a significant overall learning effect for L1 English participants, even when the words involved unfamiliar sounds and were phonologically overlapping. Also, minimal pairs led to greater difficulty in learning. That is, when participants were presented

with two objects that were associated with two phonological overlapping words (minimal pairs), their performance (accuracy) was reduced. These confirm the findings from Experiment 1. However, we did not find the expected exposure effect. Critically, participants did not improve significantly in tonal trials with doubled exposure, suggesting that the lack of improvement in tonal trials in Experiment 1 is not merely a lack of input exposure. Furthermore, we did not observe the effect of awareness on learning outcomes, either in overall accuracy or in the tonal trials. In Experiment 1, we observed better performance among aware participants in tonal trials at the last learning block, but this difference was not found in Experiment 2. This observation supports our explanation above that the different performances between aware and unaware learners in Experiment 1 might result from factors (e.g., attention loss due to distraction or fatigue) other than awareness of the tones. Simply being aware of the tonal difference may not be sufficient for learners to accurately use the tonal cue in word learning. Mapping spoken tonal words to meanings requires categorical perception of tones and forming representations of tonal words in the mental lexicon. To summarize, Experiment 2 confirmed the findings in Experiment 1 but did not provide further evidence for the learning of the tonal contrast.

General Discussion

In this study, we explored the impact of phonology on non-native vocabulary learning using a cross-situational learning paradigm which combines implicit and statistical learning research (see Monaghan et al., 2019). We found evidence that CSLis effective when words contain non-native suprasegmental features. Furthermore, we manipulated the phonological similarity between words and generated different (non)minimal pair types to assimilate the natural language learning situation. Learners' performance was significantly influenced by

how similar the words sounded, thus suggesting that future word learning research needs to take into account the role of phonology more fully.

RQ1: Do minimal pairs pose difficulty during cross-situational learning compared to phonologically distinct words? As predicted (and outlined in our preregistration), in both experiments, learners performed better in non-minimal pair trials as compared to other minimal pair trials. One explanation is that, in non-minimal pair trials, learners can rely on several phonological cues (e.g., consonants, vowels, tones) to activate the corresponding referent; but in minimal pair trials, most of the cues are uninformative and activate both objects, with only one informative cue indicating the correct referent. Our finding is consistent with Escudero et al.'s (2016) results of lower performance for minimal pairs, though we included not only segmental but also suprasegmental minimal pairs. Our study tested effects of minimal pairs in disyllabic words without context, but for acquiring a larger vocabulary under more naturalistic circumstances, the learner is likely to be affected by other properties of the language. For instance, Thiessen (2007) found that infants could distinguish and learn minimal pairs more easily after being exposed to the specific phonemic contrasts in dissimilar contexts, hence the prevalence of minimal pairs may play a role. Therefore, in real-life word learning, though minimal pairs are widespread in natural language vocabularies (e.g., in CELEX, Baayen, Piepenbrock, & van Rijn, 1993), 28% of English word types have a neighbour with one letter different, and in Mandarin, most words have at least one neighbour with only tonal differences (Duanmu, 2007)), context can provide information about the likely meaning of the word to support identification (e.g., Levis & Cortes, 2008).

RQ2: Do minimal pairs that differ in non-native phonological contrasts pose further difficulty compared to minimal pairs with contrasts that are similar to native sounds? As predicted, in both experiments, English-native speakers' accuracy in tonal minimal pair trials was lowest, as compared to consonantal and vocalic minimal pair trials. It is also worth

noting that in Experiment 1, only in the tonal trials did L1 English participants score lower than L1 Mandarin participants, whereas in all other trials, the two groups were comparable. This finding is important when we extend the CSL paradigm to L2 acquisition research, where difficulty in non-native sound perception may impede learning. Our results also provide insights into more immersive learning situations, such as living abroad, in which learners are not explicitly pre-trained with the phonological and phonetic details of the new language and are required instead to divine the important phonemic distinctions from exposure to the language. In our study, minimal pairs were not immediately available to the participant in a learning trial (in contrast to the methods used by Escudero et al., (2016, 2022), and Tuninetti et al.,2020), but, as in natural language, emerged as a result of experience of phonologically overlapping words across contexts. Under these conditions, we found that it may be harder for learners to pick up words incidentally from the environment when they contain such minimal pair contrasts.

RQ3: Does learners' non-native sound perception develop during cross-situational learning? Contrary to our predictions, no significant improvement was found in L1 English participants' performance in tonal trials across learning. Learners' difficulty in dealing with non-native contrasts remained after implicit-statistical learning, and simply increasing exposure to stimuli was not greatly facilitative. It is worth noting that in a previous statistical learning study, Nixon (2020) did observe successful learning of non-native tonal words. This is likely due to the differences in experimental settings. For example, Nixon's (2020) Experiment 1 involved feedback during training, but it is critical in our CSL paradigm that no feedback is given throughout. In Nixon's Experiment 2, participants learned the word-picture mappings in an unambiguous way – one word and one picture were presented in each trial, whereas our CSL paradigm involved ambiguous learning trials. Moreover, Nixon (2020) presented words and referents in a sequential order to enable learning from prediction and

prediction error, whereas we presented words and referents simultaneously. This could potentially provide evidence for the role of error-driven learning (Rescorla & Wagner, 1972). One follow-up is that we could replicate the current study with a sequential presentation of words and referents, and compare the results with simultaneous presentation to discern the effect of cue order in learning.

There are multiple possible explanations for this lack of improvement in L1 English participants' tonal trial performance. Firstly, the training task in our experiments was relatively short, with only one CSL session of 10 to 20 minutes. In the classic L2 speech learning studies that target non-native sound acquisition, the length and number of training sessions are typically much greater than our design and sometimes run over several days (e.g., Cheng, Zhang, Fan & Zhang, 2019; Fuhrmeister & Myers, 2020; Godfroid, Lin & Ryu, 2017; Iverson & Evans, 2009). Thus, despite the qualitative difference in the training processes (i.e., explicitness of training), the quantity of input exposure in our design is not as intensive as in previous studies, which may account for the minimal improvement in our results.

Secondly, our CSL task involves different levels of lexical tone processing rather than simply discrimination. Some participants reported that they noticed but intentionally ignored the tones to avoid confusion. The ignoring of tonal cues results from the interpretive narrowing process in early native language development (Hay, Graf Estes, Wang & Saffran, 2015). Infants with non-tonal native languages learn to constrain the type of acoustic details used in word learning and learn not to attend to the pitch contour information, as variations in pitch are mostly irrelevant at the lexical level. This process happens as early as around 17 months old, which leads to difficulty in interpreting tonal cues as meaningful in word learning (Hay et al, 2015; Liu & Kager, 2015). However, at the same age, infants can still discriminate the tonal differences. This suggests stages in the decreasing tonal processing

ability among non-tonal infants – interpretation of tones reduces greatly before perception of tones. When it comes to learning a tonal language, the challenge, therefore, may not be the perception but the referential use of lexical tones. Therefore, it is possible that our learners were able to discriminate the acoustic details between the tonal contrasts after learning, but they could not use them contrastively in learning. For non-tonal language speakers to learn a tonal language, it may be more important to restore their interpretation of tones than perceptual training. The presentation of minimal pairs, like in our design, may serve this purpose well, as it creates ambiguity if tones are not interpreted referentially and hence leads listeners to pay attention to tones. But the minimal pair training paradigm may need to last longer and be more focused on tones. In our study, we introduced different minimal pair trials, and this may reduce the emphasis on tones.

Additionally, we did not observe a relationship between tonal awareness and learning performance. This contradicts previous CSL findings that learners aware of the linguistic features start to improve earlier in the learning process (Monaghan et al., 2019). One possibility is that awareness affects different aspects of language learning differently. Monaghan et al. (2019) examined the acquisition of morphosyntactic rules, where explicit knowledge of the rules can lead to direct application of the rules in processing. However, as for phonological development, even the advanced learners of tonal languages who performed well at tone discrimination showed difficulty in tone processing at a lexical level (Pelzl, Lau, Guo & DeKeyser, 2019). Thus, merely being aware of the unfamiliar phonological feature may not allow learners to explicitly make use of the cues in word learning.

Limitations and further directions

We tested learners' vocabulary and phonological development with a single accuracy measure in the CSL task. However, as discussed, it is possible that English-native

participants' tonal perception ability improved in terms of acoustic discrimination of tones, which, using the CSL task, cannot be separated from their vocabulary knowledge. Future studies can incorporate direct tests of sound perception and discrimination before and after the CSL task to explore more precisely how CSL interferes with perceptual abilities (for pre-registered study, see: https://osf.io/kqagx). It would also be interesting to examine learners' categorical perception of lexical tones after learning sessions to investigate at which level (acoustic, phonological, or lexical) the difficulties arise. Furthermore, not many studies have explicitly compared perception and production training in lexical tone acquisition. One relevant study by Lu, Wayland and Kaan (2015) reported no significant benefit of adding a production component in explicit lexical tone training. However, it is not clear whether there could be an interaction between training type (explicit/implicit) and training mode (perception/production). One potential follow-up on the current design is that we could add a production task to the perceptual CSL task. Imitation of the tonal stimuli may direct more attention to the tonal contrast and facilitate learners' understanding of tonal use. Lastly, we noticed that there was great variation among L1 English participants' performance in tonal trials, especially in Experiment 2 where some learners reached an accuracy of over 80% after learning. We will carry out further individual difference studies to investigate the various predictors that contribute to better word learning outcomes, from auditory processing (Saito, Sun & Tierney, 2020), working memory, to implicit and explicit language aptitudes.

Supplementary materials

Table S1. List of minimal pairs in the four trial types.

Table S2. Best fitting models for accuracy in Experiment 1, with consonantal (A), vocalic (B), and tonal (C) minimal pair trials as the reference level, respectively.

Table S3. Best fitting model for accuracy for L1 Mandarin group in Experiment 1, with non-minimal pair (A), consonantal (B), vocalic (C), and tonal (D) minimal pair trials as the reference level, respectively.

Table S4. Best fitting model for accuracy for L1 English group in Experiment 1, with non-minimal pair (A), consonantal (B), vocalic (C), and tonal (D) minimal pair trials as the reference level, respectively.

Table S5. Best fitting model for reaction time in Experiment 1, showing fixed effects.

Table S6. Best fitting model for accuracy for the L1 English group in Experiment 1, testing awareness effect, with consonantal (A), vocalic (B), and tonal (C) minimal pair trials as the reference level, respectively.

Table S7. Best fitting model for accuracy in Block 6 for the L1 English group in Experiment 1, testing awareness effect.

Table S8. Best fitting model for accuracy in Experiment 2, with consonantal (A), vocalic (B), and tonal (C) minimal pair trials as the reference level, respectively.

Table S9. Best fitting model for reaction time in Experiment 2, showing fixed effects.

Figure S1. Experiment 1: Mean reaction time for correct responses in each learning block – overall (A) and in different trial types (B & C).

Figure S2. Experiment 2: Mean reaction time for correct responses in each learning block - overall (A) and in different trial types (B).

References

Adesope, OO, Lavin, T, Thompson, T and Ungerleider, C (2010) A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research* 80(2), 207-245.

Baayen, H, Piepenbrock, R and van Rijn, H (1993) The CELEX lexicaldatabase (CD-ROM). University of Pennsylvania, Philadelphia: Linguistic Data Consortium.

Chan, RK and Leung, JH (2020) Why are lexical tones difficult to learn?: insights from the incidental learning of tone-segment connections. *Studies in Second Language Acquisition* 42(1), 33-59.

Chandrasekaran, B, Sampath, PD and Wong, PC (2010) Individual variability in cue-weighting and lexical tone learning. *The Journal of the Acoustical Society of America* 128(1), 456-465.

Chang, CB and Bowles, AR (2015) Context effects on second-language learning of tonal contrasts. *The Journal of the Acoustical Society of America* 138(6), 3703-3716.

Cheng, B, Zhang, X, Fan, S and Zhang, Y (2019) The role of temporal acoustic exaggeration in high variability phonetic training: A behavioral and ERP study. *Frontiers in Psychology* 10, 1178.

Childers, JB and Pak, JH (2009) Korean- and English-speaking children use cross-situational information to learn novel predicate terms. *Journal of Child Language* 36, 201– 224.

Cooper, A and Wang, Y (2013) Effects of tone training on Cantonese tone-word learning. *The Journal of the Acoustical Society of America* 134(2), EL133-EL139.

Cutler, A and Chen, H-C (1997) Lexical tone in Cantonese spoken-word processing. *Perception and Psychophysics* 59, 165–179.

Davies, RA, Arnell, R, Birchenough, JM, Grimmond, D and Houlson, S (2017) Reading through the life span: Individual differences in psycholinguistic effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 43, 1298-1338. https://doi.org/10.1037/xlm0000366

Ding, H (2012) Perception and production of Mandarin disyllabic tones by German learners. In *Speech Prosody 2012*.

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

Dupoux, E, Sebastián-Gallés, N, Navarrete, E and Peperkamp, S (2008) Persistent stress 'deafness': The case of French learners of Spanish. *Cognition* 106(2), 682-706.

Escudero, P. (2005). *Linguistic Perception and Second Language Acquisition: Explaining the Attainment of Optimal Phonological Categorization*. PhD thesis, LOT Dissertation Series 113, Utrecht University.

Escudero, P, Mulak, KE, Fu, CS and Singh, L (2016) More limitations to monolingualism: Bilinguals outperform monolinguals in implicit word learning. *Frontiers in Psychology* 7, 1218.

Escudero, P, Mulak, KE and Vlach, HA (2016) Cross-situational learning of minimal word pairs. *Cognitive Science* 40(2), 455-465.

Escudero, P, Smit, EA and Mulak, KE (2022) Explaining L2 Lexical Learning in Multiple Scenarios: Cross-Situational Word Learning in L1 Mandarin L2 English Speakers. *Brain Sciences* 12(12), 1618.

Francis, AL, Ciocca, V, Ma, L and Fenn, K (2008) Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *Journal of Phonetics* 36(2), 268-294.

Fuhrmeister, P and Myers, EB (2020) Desirable and undesirable difficulties: Influences of variability, training schedule, and aptitude on nonnative phonetic learning. *Attention, Perception, & Psychophysics* 82(4), 2049-2065.

Gillette, J, Gleitman, H, Gleitman, L and Lederer, A (1999) Human simulations of vocabulary learning. *Cognition* 73(2), 135-176.

Godfroid, A, Lin, CH and Ryu, C (2017) Hearing and seeing tone through color: An efficacy study of web-based, multimodal Chinese tone perception training. *Language Learning* 67(4), 819-857.

Hao, YC (2012) *Journal of Phonetics* 40(2), 269-279.

Hao, YC (2018) Contextual effect in second language perception and production of Mandarin tones. *Speech Communication* 97, 32-42.

Hao, YC and Yang, CL (2018) The role of orthography in L2 segment and tone encoding by learners at different proficiency levels. In *Proceedings of TAL2018, Sixth International Symposium on Tonal Aspects of Languages* (pp. 247-251).

Hay, JF, Graf Estes, K, Wang, T and Saffran, JR (2015) From flexibility to constraint: The contrastive use of lexical tone in early word learning. *Child development* 86(1), 10-22.

Horst, JS and Hout, MC (2016) The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behavior Research Methods* 48(4), 1393–1409.

Hu, CF (2017) Resolving referential ambiguity across ambiguous situations in young foreign language learners. *Applied Psycholinguistics* 38(3), 633-656.

Hummel, KM (2021) Phonological Memory and L2 Vocabulary Learning in a Narrated Story Task. *Journal of Psycholinguistic Research* 50(3), 603-622.

Isbilen, ES and Christiansen, MH (2022) Statistical Learning of Language: A Meta-Analysis Into 25 Years of Research. *Cognitive Science* 46(9), e13198.

Iverson, P and Evans, BG (2009) Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *The Journal of the Acoustical Society of America* 126(2), 866-877.

Iverson, P, Kuhl, PK, Akahane-Yamada, R, Diesch, E, Kettermann, A and Siebert, C (2003) A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition* 87(1), B47-B57.

Jin, W (2011) A statistical argument for the homophony avoidance approach to the disyllabification of Chinese. In *Proceedings of the 23rd North American Conference*

on Chinese Linguistics, edited by Z. Jing-Schmidt (University of Oregon, Eugene, OR), Vol. 1, pp. 35–50.

Junttila, K and Ylinen, S (2020) Intentional training with speech production supports children's learning the meanings of foreign words: a comparison of four learning tasks. *Frontiers in Psychology* 11, 1108.

Kachlicka, M, Saito, K and Tierney, A (2019) Successful second language learning is tied to robust domain-general auditory processing and stable neural representation of sound. *Brain and Language,* 192, 15-24.

Kiriloff, C (1969) On the auditory perception of tones in Mandarin. *Phonetica* 20(2-4), 63-67.

Kuhl, PK, Stevens, E, Hayashi, A, Deguchi, T, Kiritani, S and Iverson, P (2006) Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science* 9(2), F13-F21.

Levis, J and Cortes, V (2008) Minimal pairs in spoken corpora: Implications for pronunciation assessment and teaching. *Towards adaptive CALL: Natural language processing for diagnostic language assessment*, 197208.

Liu, L and Kager, R (2018) Monolingual and bilingual infants' ability to use non-native tone for word learning deteriorates by the second year after birth. *Frontiers in Psychology* 9, 117.

Lu, S, Wayland, R and Kaan, E (2015) Effects of production training and perception training on lexical tone perception–A behavioral and ERP study. *Brain Research* 1624, 28-44.

Marian, V, Blumenfeld, HK and Kaushanskaya, M (2007) The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research* 50, 940-967.

Markman, EM (1990) Constraints children place on word meanings. *Cognitive Science* 14, 57– 77.

Martin, KI and Ellis, NC (2012) The roles of phonological short-term memory and working memory in L2 grammar and vocabulary learning. *Studies in Second Language Acquisition* 34(3), 379-413.

Maye, J and Gerken, L (2000) Learning phonemes without minimal pairs. In *Proceedings of the 24th annual Boston university conference on language development* (Vol. 2, pp. 522-533).

Maye, J, Werker, JF and Gerken, L (2002) Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82, 101–111.

Monaghan, P and Mattock, K (2012) Integrating constraints for learning word-referent mappings. *Cognition* 123, 133–143. https://doi.org/10.1016/j.cognition.2011.12.010.

Monaghan, P, Ruiz, S and Rebuschat, P (2021) The role of feedback and instruction on the cross-situational learning of vocabulary and morphosyntax: Mixed effects models reveal local and global effects on acquisition. S*econd Language Research* 37(2), 261-289.

Monaghan, P., Schoetensack, C and Rebuschat, P (2019) A single paradigm for implicit and statistical learning. *Topics in Cognitive Science* 11(3), 536-554.

Pelzl, E, Lau, EF, Guo, T and DeKeyser, R (2019) Advanced second language learners' perception of lexical tone contrasts. *Studies in Second Language Acquisition* 41(1), 59-86.

Poepsel, TJ and Weiss, DJ (2016) The influence of bilingualism on statistical word learning. *Cognition* 152, 9-19.

Quine, WVO (1960) *Word and object.* Cambridge, MA: MIT Press.

Rato, A. (2018). Perceptual categorization of English vowels by native European Portuguese speakers. *Revista Linguíftica,* 14(2), 61-80.

Rato, A., & Carlet, A. (2020). Second language perception of English vowels by Portuguese learners: The effect of stimulus type. *Ilha do Desterro*, 73, 205-226.

Rebuschat, P, Hamrick, P, Riestenberg, K, Sachs, R and Ziegler, N (2015) Triangulating measures of awareness: A contribution to the debate on learning without awareness. *Studies in Second Language Acquisition* 37(2), 299-334.

Rebuschat, P, Monaghan, P and Schoetensack, C (2021) Learning vocabulary and grammar from cross-situational statistics. *Cognition* 206, 104475.

Saffran, JR, Aslin, RN and Newport, EL (1996) Statistical learning by 8-month-old infants. *Science* 274(5294), 1926-1928.

Saito, K, Kachlicka, M, Sun, H and Tierney, A (2020) Domain-general auditory processing as an anchor of post-pubertal L2 pronunciation learning: Behavioural and neurophysiological investigations of perceptual acuity, age, experience, development, and attainment. *Journal of Memory and Language*, 115.

Saito, K, Sun, H and Tierney, A (2020) Brief report: Test-retest reliability of explicit auditory processing measures. bioRxiv.

Schmidt, R (1990) The role of consciousness in second language learning. *Applied Linguistics* 11, 129-158.

Schmidt, R (1995) Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R Schmidt (eds.), *Attention and awareness in foreign language learning* (Vol.9, pp. 1-63). Second Language Teaching & Curriculum Center, University of Hawaii.

Sereno, JA and Lee, H (2015) The contribution of segmental and tonal information in Mandarin spoken word processing. *Language and Speech* 58(2), 131-151.

Siegelman, N (2020) Statistical learning abilities and their relation to language. *Language and Linguistics Compass* 14(3), e12365.

Silbert, NH, Smith, BK, Jackson, SR, Campbell, SG, Hughes, MM and Tare, M (2015) Non-native phonemic discrimination, phonological short term memory, and word learning. *Journal of Phonetics* 50, 99-119.

Smith, L and Yu, C (2008) Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition* 106*,* 1558-1568.

Smith, K, Smith, AD and Blythe, RA (2009) Reconsidering human cross-situational learning capacities: A revision to Yu and Smith's (2007) experimental paradigm. In N Taatgen and H van Rijn (eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2711– 2716). Austin, TX: Cognitive Science Society.

Smith, K, Smith, AD and Blythe, RA (2011) Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science* 35(3), 480-498.

So, CK and Best, CT (2010) Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences*. Language and speech* 53(2), 273-293.

So, CK and Best, CT (2014) Phonetic influences on English and French listeners' assimilation of Mandarin tones to native prosodic categories. *Studies in Second Language Acquisition* 36(2), 195–221.

Suanda, SH, Mugwanya, N and Namy, LL (2014) Cross-situational statistical word learning in young children. *Journal of Experimental Child Psychology* 126, 395-411.

Suanda, SH and Namy, LL (2012) Detailed behavioral analysis as a window into cross-situational word learning. *Cognitive Science* 36(3), 545-559.

Sun, H, Saito, K and Tierney, A (2021) A longitudinal investigation of explicit and implicit auditory processing in L2 segmental and suprasegmental acquisition. *Studies in Second Language Acquisition* 43(3), 551-573.

Takagi, N and Mann, V (1995) The limits of extended naturalistic exposure on the perceptual mastery of English /r/ and /l/ by adult Japanese learners of English. *Applied Psycholinguistics* 16(4), 380-406.

Thiessen, ED (2007) The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language* 56(1), 16-34.

Tomasello, M and Barton, ME (1994) Learning words in nonostensive contexts. *Developmental Psychology* 30, 639-650.

Tuninetti, A, Mulak, KE and Escudero, P (2020) Cross-situational word learning in two foreign languages: effects of native language and perceptual difficulty. *Frontiers in Communication* 5, 602471.

Walker, N, Monaghan, P, Schoetensack, C and Rebuschat, P (2020) Distinctions in the acquisition of vocabulary and grammar: An individual differences approach. *Language Learning* 70(S2), 221-254.

Wang, T and Saffran, JR (2014) Statistical learning of a tonal language: The influence of bilingualism and previous linguistic experience. *Frontiers in Psychology* 5, 953.

Williams, JN and Rebuschat, P (2022) Implicit learning and SLA: a cognitive psychology perspective. In A Godfroid and H Hopp (eds.), *The Routledge handbook of second language acquisition and psycholinguistics*. Taylor & Francis.

Werker, JF and Tees, RC (1984) Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development* 7, 49– 63.

Wong, PC & Perrachione, TK (2007) Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics* 28(4), 565-585.

Yip, M (2001) Phonological priming in Cantonese spoken-word processing. *Psychologia* 44, 223–229.

Yu, C and Smith, L (2007) Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science* 18, 414–420.

Yu, C and Smith, L (2011) What you learn is what you see: using eye movements to study infant cross-situational word learning. *Developmental Science* 14(2), 165-180.

Yu, C, Zhong, Y & Fricker, D (2012) Selective attention in cross-situational statistical learning: evidence from eye tracking. *Frontiers in Psychology*, 3, 148.

Yurovsky, D, Smith, L and Yu, C (2013) Statistical word learning at scale: The baby's view is better. *Developmental Science* 16(6), 959-966.

Zou, T, Chen, Y and Caspers, J (2017) The developmental trajectories of attention distribution and segment-tone integration in Dutch learners of Mandarin tones. *Bilingualism: Language and Cognition* 20(5), 1017-1029.

Tables and Figures

Table 1. *Pseudowords in the consonantal set and the vocalic set*

Consonant set		Vocalic set	
pa1mi1	pa4mi1	li1fa1	li4fa1
ta1mi1	ta4mi1	lu1fa1	lu4fa1
ka1mi1	ka4mi1	lei1fa1	lei4fa1

Note. Numbers "1" and "4" refer to the lexical tones T1 and T4 carried by the syllables

Table 2. *Best fitting model for accuracy in Experiment 1, showing fixed effects*

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	0.269	0.141	1.913	.056
block	0.093	0.043	2.146	.032 *
langgroupEnglish	-0.080	0.135	-0.589	.556
MPtypeC	-0.383	0.162	-2.363	.018 *
MPtypeT	-0.178	0.180	-0.986	.324
MPtypeV	-0.078	0.187	-0.417	.677
block:langgroupMandarin:MPtypeN	0.244	0.068	3.572	<.001***
block:langgroupEnglish:MPtypeN	0.071	0.046	1.526	.127
block:langgroupMandarin:MPtypeC	0.153	0.064	2.396	.017 *
block:langgroupEnglish:MPtypeC	0.020	0.046	0.446	0.655
block:langgroupMandarin:MPtypeT	0.018	0.059	0.308	0.758
block:langgroupEnglish:MPtypeT	-0.088	0.045	-1.938	0.053
block:langgroupMandarin:MPtypeV	0.113	0.055	2.046	0.041 *

Number of observations: 8038, Participants: 56, Item, 12. AIC = 10025.3, BIC = 10367.9, log-likelihood = -4963.7.

R syntax: glmer(acc ~ block + langgroup + MPtype + langgroup:MPtype:block + (1 + block + langgroup + MPtype | item) + (1 + block + MPtype | subjectID), family = binomial, data = fulld, glmerControl(optCtrl=list(maxfun=2e5), optimizer = "nloptwrap", calc.derivs = FALSE)).

Table 3. *Best fitting model for accuracy in tonal trials in Experiment 1, showing fixed effects*

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	0.260	0.121	2.149	.032 *
langgroupEnglish	-0.458	0.170	-2.689	.007 **
block	0.064	0.033	1.969	.049 *

Number of observations: 2008, Participants: 56, Item, 12. AIC = 2732.9, BIC = 2822.6, log-likelihood = -1350.4.

R syntax: glmer(acc ~ langgroup + block + (1 + langgroup + block + langgroup:block | item) + (1 + block | subjectID), family = binomial, data = ttrials, glmerControl(optCtrl=list(maxfun=2e5), optimizer = "nloptwrap", calc.derivs = FALSE))

Table 4. *Best fitting model for accuracy for the L1 English group in Experiment 1, testing awareness effect*

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	0.542	0.154	3.518	<.001***
block	0.116	0.026	4.453	<.001***
MPtypeC	-0.630	0.135	-4.651	<.001***
MPtypeT	-0.849	0.195	-4.345	<.001***

MPtypeV	-0.487	0.166	-2.929	.003 **

Number of observations: 4025, Participants: 28, Item, 12. AIC = 5383.5, BIC = 6171.0, log-likelihood = -2566.7.

R syntax: glmer(acc ~ block + MPtype + (1 + block + awareness + MPtype + block:awareness:MPtype | item) + (1 + block + MPtype | subjectID), family = binomial, data = fulld.awareness, glmerControl(optCtrl=list(maxfun=2e5), optimizer = "nloptwrap", calc.derivs = FALSE))

Table 5. *Best fitting model for accuracy in Experiment 2, showing fixed effects*

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	0.674	0.129	5.235	<.001 ***
block	0.118	0.036	3.275	.001 **
exposureshort	-0.153	0.110	-1.392	.164
MPtypeC	-0.592	0.153	-3.866	<.001 ***
MPtypeT	-0.741	0.189	-3.912	<.001 ***
MPtypeV	-0.419	0.181	-2.311	.021 *
block:exposurelong:MPtypeN	0.012	0.042	0.288	.773
block:exposureshort:MPtypeN	0.010	0.043	0.240	.810
block:exposurelong:MPtypeC	0.010	0.040	0.257	.797
block:exposureshort:MPtypeC	-0.013	0.044	-0.290	.772
block:exposurelong:MPtypeT	-0.098	0.038	-2.601	.009 **
block:exposureshort:MPtypeT	-0.051	0.041	-1.233	.218
block:exposurelong:MPtypeV	-0.013	0.034	-0.391	.696

Number of observations: 11793, Participants: 55, Item, 12. AIC = 14100.7, BIC = 14462.1, log-likelihood = -7001.4.

R syntax: glmer(acc ~ block + exposure + MPtype + exposure:MPtype:block + (1 + block + exposure + MPtype | item) + (1 + block + MPtype | subjectID), family = binomial, data = fulld, glmerControl(optCtrl=list(maxfun=2e5), optimizer = "nloptwrap", calc.derivs = FALSE))

Figure 1. Example of cross-situational learning trial. Participants were presented with two novel objects and one spoken word (e.g., pa1mi1). Participants had to decide, as quickly and accurately as possible, if the word refers to the object on the left or right of the screen.

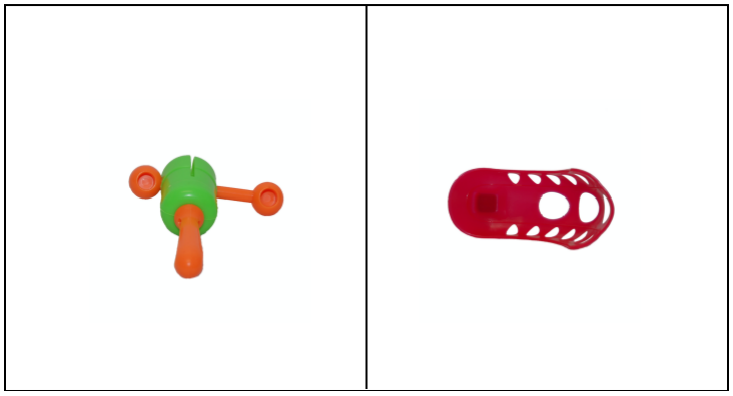
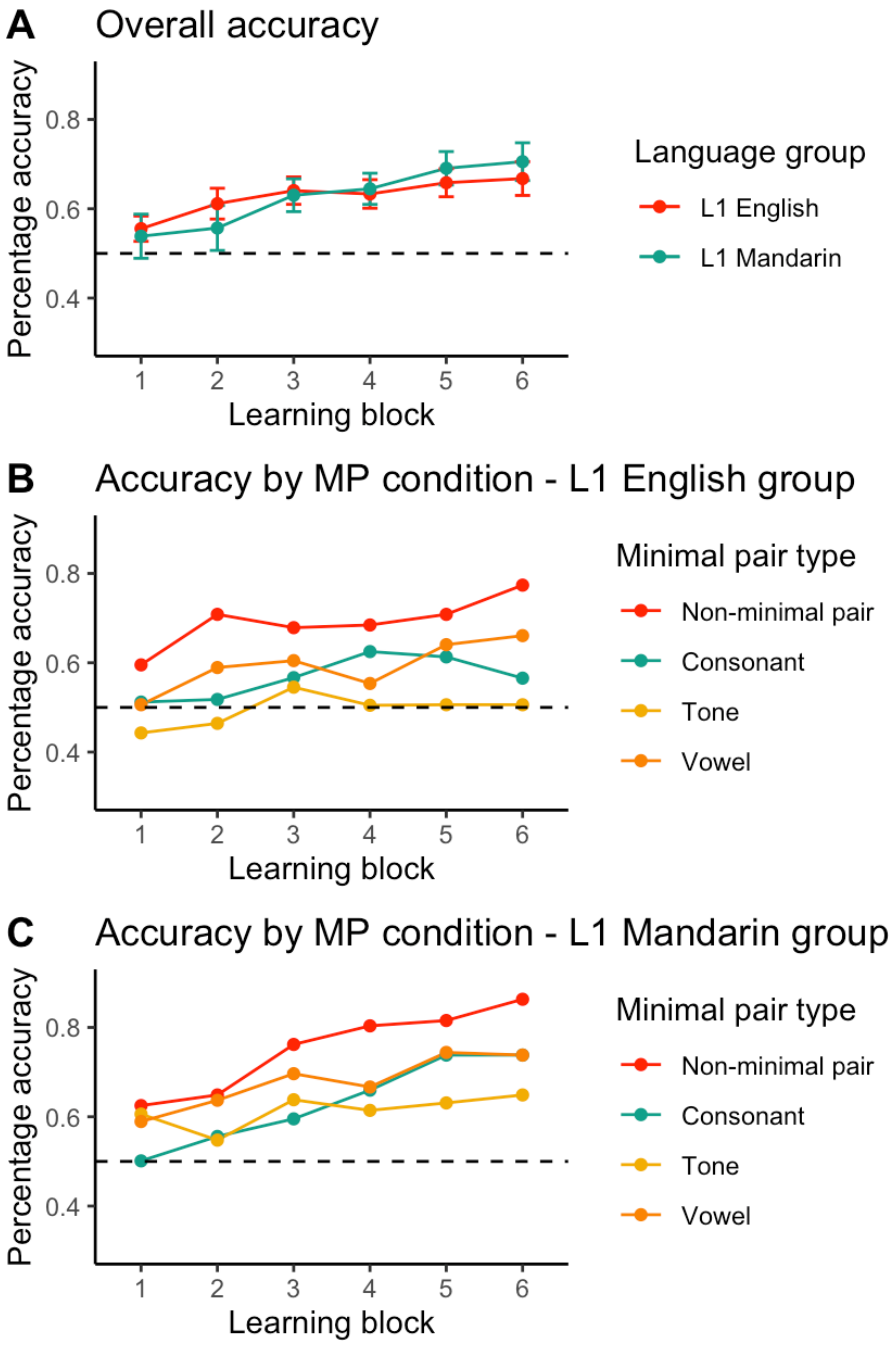
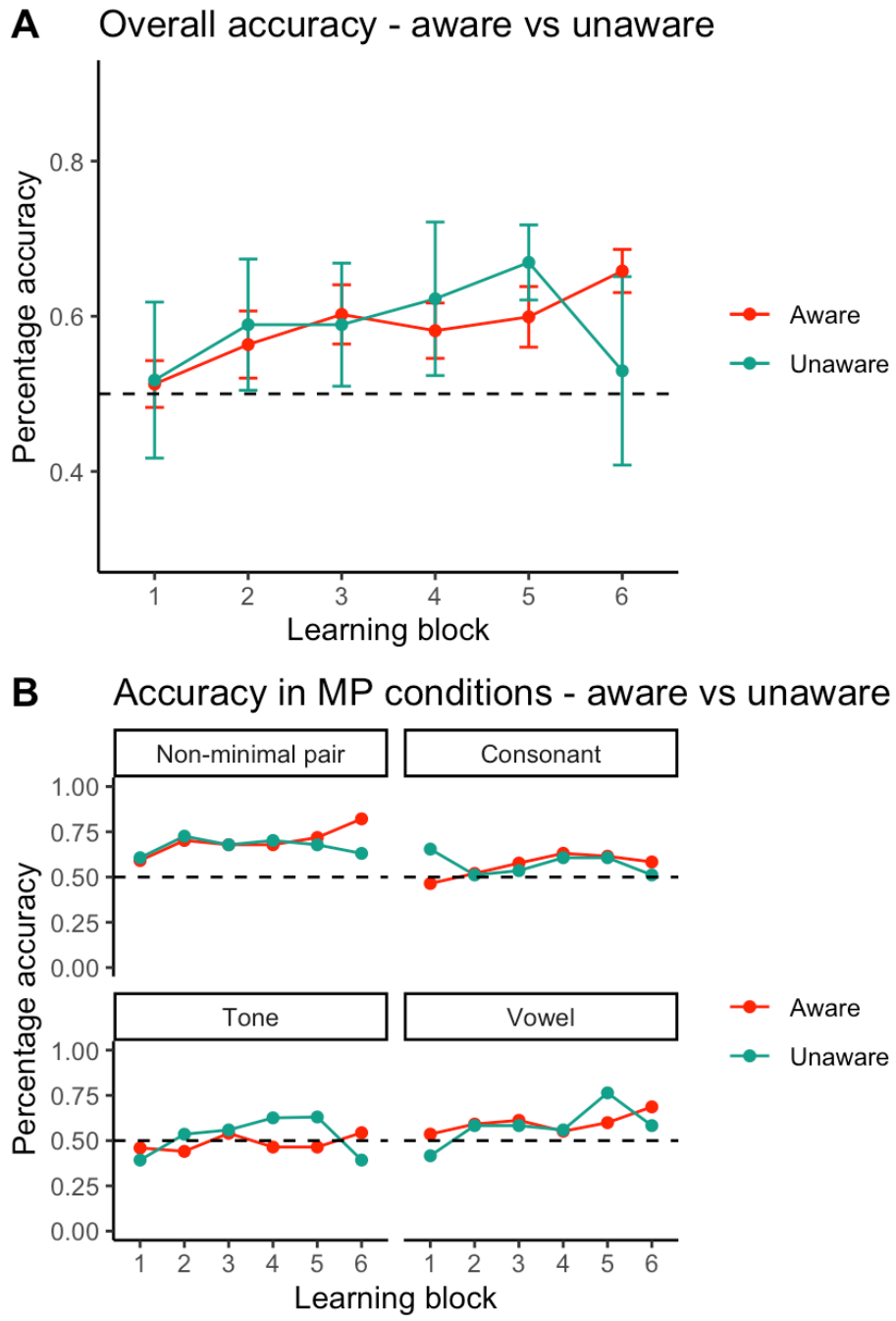


Figure 2. Experiment 1: Mean proportion of correct pictures selected in each learning block - overall (A) and in different trial types (B & C).



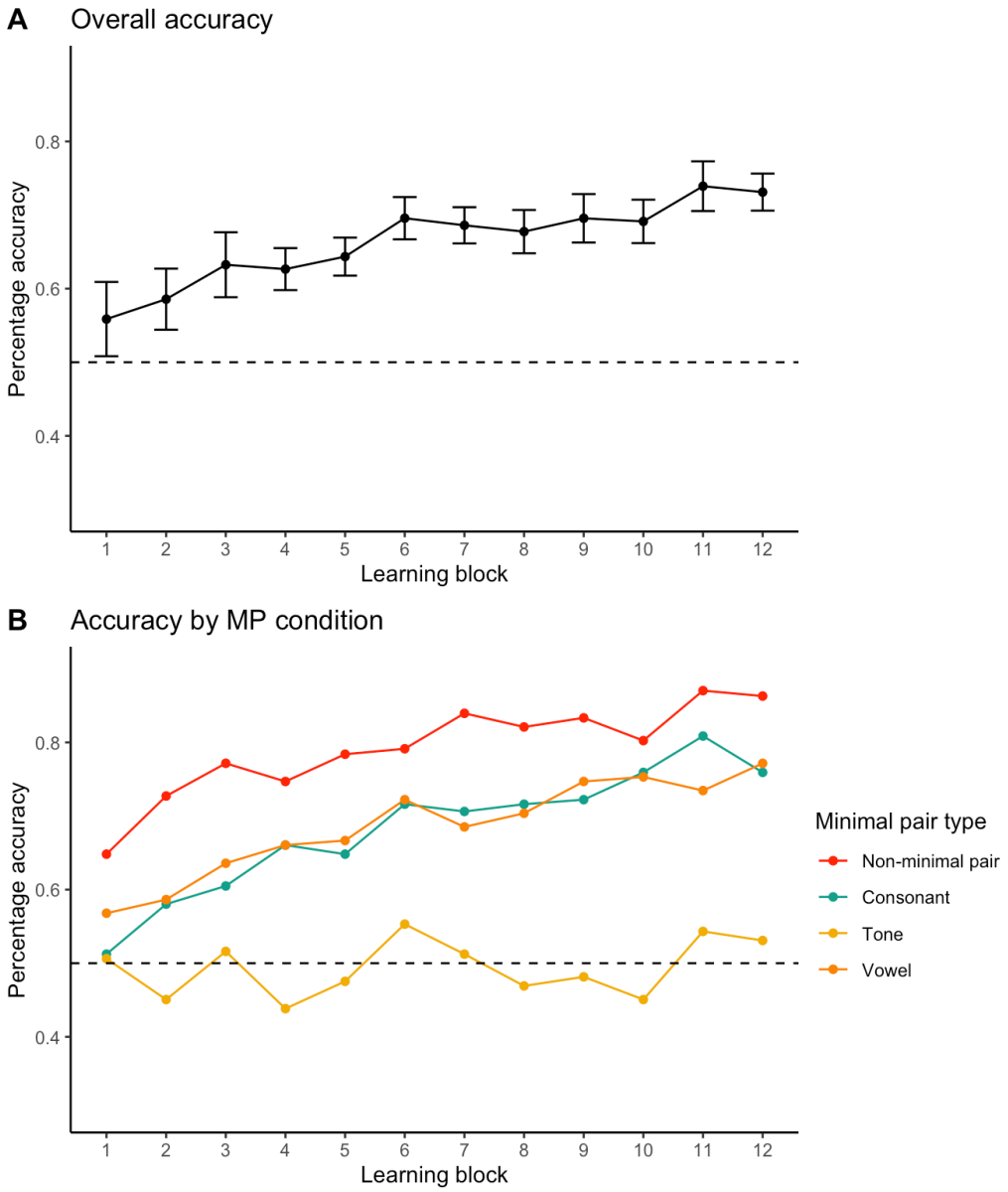
Note. Error bars represent 95% Confidence Intervals.

Figure 3. Experiment 1: Proportion of correct responses in each learning block for aware and unaware participants (L1 English group only) – overall (A) and in different trial types (B).



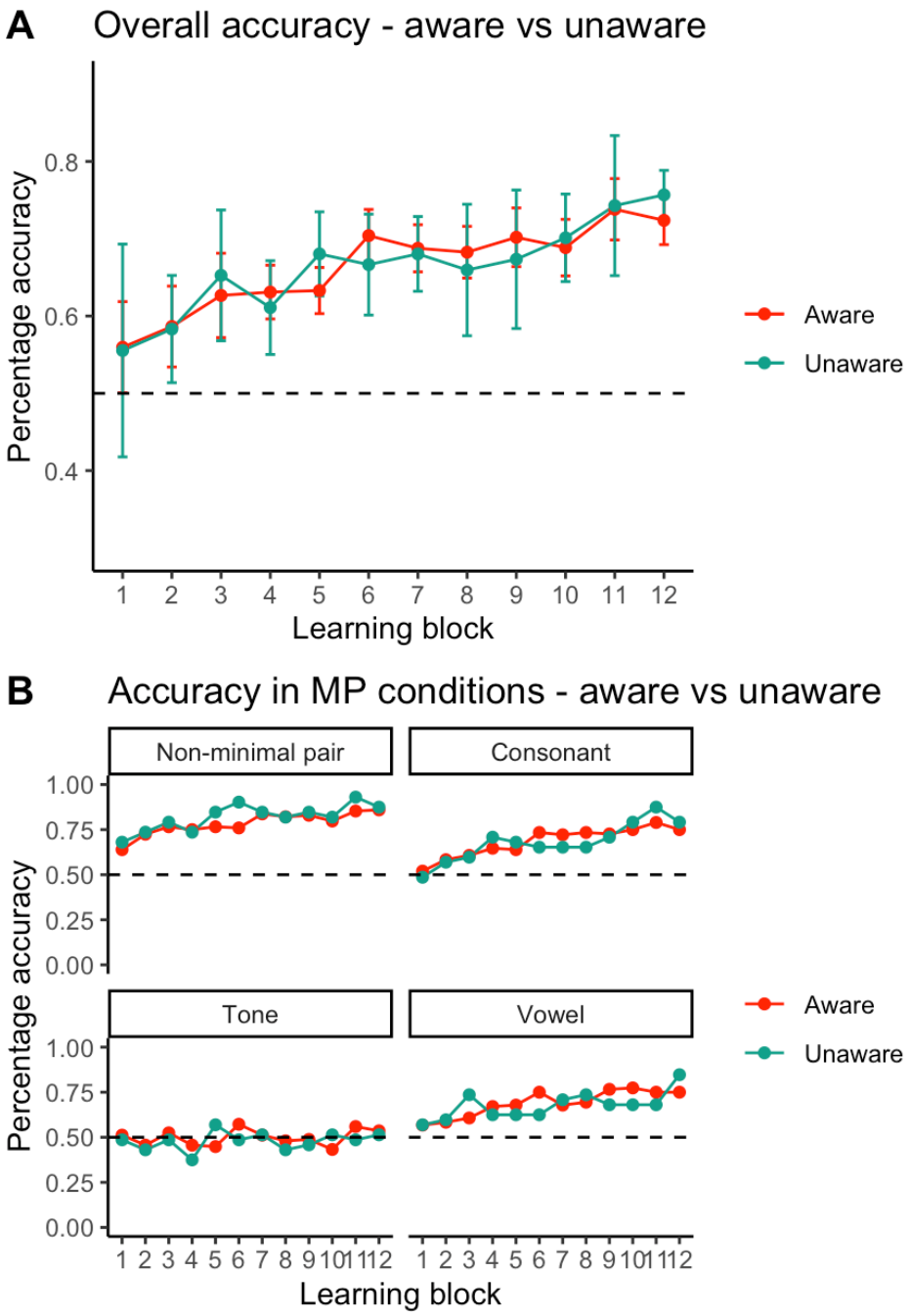
Note. Error bars represent 95% Confidence Intervals.

Figure 4. Experiment 2: Mean proportion of correct pictures selected in each learning block - overall (A) and in different trial types (B).



Note. Error bars represent 95% Confidence Intervals.

Figure 5. Experiment 2: Proportion of correct responses in each learning block for aware and unaware participants - overall (A) and in different trial types (B).



Note. Error bars represent 95% Confidence Intervals.

Data availability statement

Data availability: the data that support the findings of this study are openly available in Open Science Framework at https://osf.io/2j6pe/ (for Experiment 1) and https://osf.io/2m4nw/ (for Experiment 2).

Publishing ethics

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.