Influence of phonology and individual differences in adults' statistical word learning

Yuxin Ge

This thesis is submitted for the degree of

Doctor of Philosophy in Linguistics

Department of Linguistics and English Language

Lancaster University

June 2024

Acknowledgement

I am deeply grateful to many individuals who have supported and guided me throughout my PhD study.

First and foremost, I would like to express my heartfelt thanks to my supervisors, Professor Patrick Rebuschat and Professor Padraic Monaghan. Their constant support, insightful feedback, and invaluable guidance have been instrumental in both my academic development and personal growth. It has been such an honour and a privilege to work with them both. Their encouragement and mentorship have made this journey incredibly rewarding and fulfilling.

I would like to extend my gratitude to all my co-authors, Dr. Anabela Rato, Dr. Magdalena Kachlicka, and Prof Kazuya Saito. Their invaluable insights and collaborative spirit have significantly enriched our research. Working with them has been an inspiring and educational experience.

Special thanks to Dr. Susana Correia for the opportunity to collaborate on the phonetic training projects; her dedication and passion for research have deeply motivated me. Sincere gratitude to Dr. Aina Casaponsa for her patience and expertise in guiding me through the development of an EEG experiment.

To the members of our lovely language learning lab and good friends, thank you all for making the past three years such a wonderful and memorable experience. I have thoroughly enjoyed our lab meetings, the stimulating discussions, and of course, the snacks we shared.

The camaraderie and support have created a nurturing and enjoyable environment that I will always cherish.

Last but certainly not least, I would like to thank my parents, Mrs Xiaofeng Ge (葛晓凤) and Mr Honglin Fan (范红林). Their unwavering support and love have been my foundation throughout my studies abroad. Their belief in me has been a constant source of strength and motivation. I am deeply grateful for everything they have done to help me achieve my goals.

Thank you all for being a part of this amazing journey.

Abstract

Language learners often acquire vocabulary rapidly from their environment, usually without explicit instruction. One explanation for this process is that learners track input statistics to map words to meanings, a mechanism known as cross-situational statistical word learning (e.g., Escudero et al., 2022; Monaghan et al., 2019; Rebuschat et al., 2021). Despite its efficacy, limited research has explored how the phonological properties of words interact with the statistical learning process, especially in second language acquisition. This thesis presents a series of studies investigating the effects of phonological overlap and learners' native languages on the statistical learning of novel, non-native words. Additionally, it takes into account the factors that predict individual differences in word learning performance.

In Study 1, English-native and Mandarin-native participants were trained with Mandarin tonal pseudowords via a cross-situational, statistical word learning (CSWL) task (Yu & Smith, 2007). The CSWL task contained ambiguous word-referent mappings: participants were presented with two referent pictures and one pseudoword in each trial, and they had to decide which picture the spoken word referred to by tracking the word-picture co-occurrences across trials, with no feedback provided. It was found that similar-sounding words (i.e., minimal pairs) were harder to acquire, and words that contrast in a non-native phonological feature (i.e., tonal minimal pairs for English-native speakers) were even harder.

Study 2 extended the CSWL task by doubling the number of trials to test whether extended training facilitated learning. Results suggested that doubled exposure did not significantly improve English-native speakers' word learning performance.

Study 3 targeted the heritage speaker population and explored whether early exposure to Mandarin tones promotes tonal minimal pair learning later in life. It was observed that heritage speakers of Mandarin, who were exposed to Mandarin at an early age but then acquired and became dominant in English, also showed difficulty with tonal minimal pair learning.

Study 4 examined whether the difficulty with non-native minimal pairs may be modulated by individual differences in lower-order, domain-general auditory processing ability (encoding and reproducing fundamental acoustic features; Mueller et al., 2012). Results indicated that more precise auditory processing (pitch discrimination and melody reproduction) was associated with better learning of non-native tonal words.

Overall, the findings demonstrated the significant influence of phonology in implicit, statistical word learning. Additionally, variations in learning outcomes can be partially explained by individual differences in auditory processing ability.

Author's declaration

This thesis is my own work and has not been submitted in substantially the same form for the award of a higher degree elsewhere. This thesis includes three published or publishable papers. The following pages contain the authorship statements signed by all co-authors of each paper.

Authorship statement

The article “The role of phonology in non-native word learning: Evidence from cross-situational statistical learning” has been published in Bilingualism: Language and Cognition with authors Yuxin Ge, Padraic Monaghan and Patrick Rebuschat. Yuxin Ge was responsible for writing up the article, carrying out the research, conceptualization and design, and formal analysis. Padraic Monaghan and Patrick Rebuschat were responsible for supervision, conceptualization and design, guidance on formal analysis, and providing comments on the text.

Authorship statement

The article “Constraints on novel word learning in heritage speakers” has been published in Frontiers in Psychology with authors Yuxin Ge, Anabela Rato, Patrick Rebuschat and Padraic Monaghan. Yuxin Ge was responsible for writing up the article, carrying out the research, conceptualization and design, and formal analysis. Anabela Rato was responsible for conceptualization and design, and providing comments on the text. Patrick Rebuschat was responsible for supervision, conceptualization and design, funding acquisition, and providing comments on the text. Padraic Monaghan was responsible for supervision, guidance on formal analysis, and providing comments on the text.

Authorship statement

I, Yuxin Ge, have written the article and carried out the research behind it in collaboration with Magdalena Kachlicka, Kazuya Saito, Patrick Rebuschat and Padraic Monaghan. I was responsible for writing up the article, carrying out the research, conceptualization and design, and formal analysis. Magdalena Kachlicka was responsible for analyzing the auditory processing test scores and providing comments on the text. Kazuya Saito was responsible for conceptualization and providing comments on the text. Patrick Rebuschat and Padraic Monaghan were responsible for supervision, conceptualization and design, guidance on formal analysis, and providing comments on the text.

Table of contents

List of Tables

List of Figures

1. Introduction

Adults frequently encounter challenges in perceiving and processing sounds of an additional language (L2) (e.g., Flege & MacKay, 2004; Iverson et al., 2003; So & Best, 2010). This can cause problems when acquiring new words, as learners need to accurately encode the non-native phonological contrasts during the process. In natural languages, many words tend to sound similar but have contrasting meanings (e.g., *bag vs beg* in English; *pāo vs gāo* in Mandarin), which makes the need for precise sound perception and processing more crucial. However, in the second language acquisition literature, relatively limited research has directly investigated the relationship between non-native phonology and adults' word learning outcomes (see Chandrasekaran et al., 2010; Silbert et al., 2015; Wong & Perrachione, 2007). In a series of four studies, I addressed this gap and examined whether and how the phonological properties of words influence word learning in an implicit, statistical learning environment (Monaghan et al., 2015, 2019).

Furthermore, the extent to which word learning is influenced by the phonological contrasts it comprises can vary significantly across learners (e.g., Wong & Perrachione, 2007). Even if the same quantity and quality of exposure or training is provided, learners' ultimate attainment differs. Hence, individual difference factors that may contribute to this learner variation were taken into account in the current studies. Specifically, I investigated two factors that are potentially related to learners' perception of the target words: the domain-general auditory processing ability and the experience and usage of the target language.

In the following chapter, I will first summarize previous empirical research and theoretical frameworks on L2 speech sound perception, before discussing how speech sound perception is linked to L2 word learning. I will then explain the critical research gaps addressed in the current studies of the dissertation.

2. Literature review

2.1. L2 speech sound perception

Adult second language learners, who possess a well-developed native (L1) sound system, face challenges in non-native sound perception. This is because their existing perceptual space for L1 may not be able to accommodate the new speech sounds (Iverson et al., 2003). In the following sections, I will present empirical evidence demonstrating this cross-linguistic and cross-feature perceptual difficulty in L2 acquisition. I will then review three of the widely examined L2 perception models, which provide theoretical accounts of the perceptual difficulties. Moreover, these perceptual issues have drawn considerable research interest in developing training methods that facilitate and promote L2 perceptual development. As such, the final subsection will briefly introduce the phonetic training methods and their focus on different aspects of speech sound learning.

2.1.1. Perceptual difficulties in L2 speech sound acquisition

When investigating the perceptual difficulties associated with non-native speech sounds, one of the most important factors is the learners' L1. Specifically, difficulties arise when learners perceive and map L2 sounds onto L1 categories. This is because L2 learners have already developed a sophisticated perceptual space for their native sounds, and they use the native sound system as a reference when processing new sounds. To understand how non-native sounds are perceived and processed, it is thus important to compare learners' L1 and L2 speech perception and explore how they assimilate L2 sounds to L1 counterparts.

Research on L2-to-L1 perceptual assimilation has designed and widely used sound identification tasks and similarity judgement measures (e.g., early studies by Best et al., 2003; Guion et al., 2000; Strange et al., 2004, 2005). Listeners are typically presented with L2 sound exemplars and asked to identify the L1 categories to which they are best mapped.

Additionally, listeners can be asked to rate how well the exemplars fit with the L1 category (i.e., goodness-of-fit ratings). The perceptual assimilation tasks enabled researchers to explore the relative perceptual ease/difficulty of language-specific sound contrasts for different L2 learner groups. Previous studies have long revealed substantial variations in L2 learners' perception of different non-native sound contrasts (e.g., Bohn et al., 2011; Best et al., 2003; Goto, 1971; Iverson et al., 2003; Kubo & Akahane-Yamada, 2006; Matthews, 2000; So, 2005; So & Best, 2010, 2014; Wayland & Guion, 2003). For learners with a specific L1 background, different sound contrasts from the same L2 can pose different degrees of difficulty. For example, the /ɹ/-/l/ contrast in English has been largely reported to cause perceptual issues for Japanese-native learners, as the contrast may be perceived as a single /r/ sound in Japanese (e.g., Goto, 1971; Iverson et al., 2003). However, the /b/-/v/ and /s/-/θ/ contrasts from the same language (English) can be discriminated at higher accuracy by Japanese-native speakers, although these contrasts also contain unfamiliar phonemes (/i.e. /v/ and /θ/) (e.g., Kubo & Akahane-Yamada, 2006; Matthews, 2000). For the same sound contrasts, the perceptual difficulty also varies significantly across different learner groups. Take the same example of the English /ɹ/-/l/ contrast, Danish-native speakers exhibited a categorical perception of the contrast similar to the English-native speakers. This indicated that the same contrast was not as perceptually challenging for Danish learners of English as for Japanese natives, potentially because a similar (though not identical) /r/-/l/ contrast exists in Danish (Bohn & Best, 2012).

Such variations in perceptual difficulty were not only observed in consonants but also in vowels (e.g., Best et al., 2003; Flege & MacKay, 2004; Souza et al., 2017). In vowel perception, L2 learners tend to weigh the acoustic cues (e.g., temporal and spectral cues) differently from native speakers of the language. For instance, L2 English learners' perception of the /i/-/ɪ/ contrast has been found to rely more on the temporal cue over the

spectral cue, while native English speakers' perception depends primarily on spectral differences (e.g., Escudero & Boersma, 2004; Ylinen et al., 2010). This divergence in cue weighting, however, also varies with learners' L1. Souza et al. (2017) observed that Danish-native learners showed a more native-like perception of the /i/-/ɪ/ contrast with high identification accuracy and significant reliance on the spectral cue, whereas Portuguese-, Catalan- and Russian-native speakers had greater identification difficulty and more reliance on the temporal differences. It again indicates that the perceptual space of learners' L1 can largely impact the relative perceptual ease/difficulty of a particular sound contrast.

In addition to the segmental features, L2 suprasegmental perception also entails varying degrees of perceptual difficulty. For example, lexical tone, as a critical feature in tonal languages, has become an important aspect of L2 perception research (e.g., Hallé et al., 2004; Hao, 2018; Pelzl et al., 2019; So & Best, 2010, 2014; Zou et al., 2012). Cross-linguistic perception of L2 tones has been found to be influenced by the constraints of learners' L1 prosodic/tonal systems, though the overall degree of L1 tonality (i.e., tonal vs non-tonal L1s) may not be the key factor (So & Best, 2010; Wang, 2006). For example, So and Best (2010) compared Cantonese-native (a tonal language), Japanese-native (a pitch accent language) and English-native (a non-tonal language) speakers' perception of L2 Mandarin tones. English-native participants only showed a lower identification than the other groups for Tone 4 (falling), but not for other tones. This suggests that the L1 influence is more complicated than a tonal versus non-tonal dichotomy. Instead, the contrast-specific difficulties revealed an effect of learners' L1 tonal inventory. Cantonese-native speakers showed more mistakes identifying T4 as T1 (high level) and T2 (rising) as T3 (falling rising) compared to the Japanese-native and English-native speakers, indicating that the errors were associated with perceptual assimilation to the native tonal contrasts.

To summarize, there has been extensive research showing the language-specific and contrast-specific difficulties in L2 speech sound perception. The relative perceptual ease/difficulty depends largely on learners' native categories and how L2 sounds are perceived and mapped to the L1 system. In the following section, I will illustrate how these perceptual difficulties are accounted for in three of the well-known speech perception models in second language acquisition.

2.1.2. L2 speech perception models

Theoretical accounts of L2 speech perception have offered potential explanations for the different degrees of perceptual difficulties associated with different non-native contrasts. One of the models that directly address this question is the Perceptual Assimilation Model (PAM) (Best, 1994, 1995). PAM aims to account for the perception performance associated with different types of non-native contrasts, primarily at the beginning of L2 development (i.e., naïve L2 perception). It is hypothesized that when naïve listeners hear non-native sounds, they map the sound to the closest native category. Thus, the prediction on the perceptual ease/difficulty of a particular sound contrast depends on how the two contrasting sounds are perceived and assimilated to the listeners' native categories. Table 2.1 summarizes the different types of assimilation patterns predicted by PAM, as well as their relative difficulty for naïve L2 listeners.

The types of assimilation that are predicted to receive the highest perceptual accuracy are the *Two Category* and the *Uncategorized-Categorized* assimilations. Two Category assimilation involves direct assimilated mapping of the two sounds in a contrast to two separate native categories, such as Australian English listeners' perception of Danish /œ/-/u/ to their native categories /ɜː/ and /ʊ/ (Faris et al., 2018). As in the case of Uncategorized-

Categorized assimilation, one of the non-native sounds in a pair can be well assimilated to a native exemplar, whereas the other cannot be perceived as close to any single L1 category (i.e., ambiguous exemplar that falls between L1 categories). For example, Japanese listeners perceive Australian English /ɜː/ as uncategorized in their native language and /ʉː/ as similar to Japanese /uː/, and hence /ɜː/-/ʉː/ makes a Uncategorized-Categorized contrast for Japanese listeners (Bundgaard-Nielsen et al., 2011). These L2-to-L1 assimilations are expected to be easy because they can be clearly distinguished even based on listeners' L1 perceptual space.

If, however, a non-native pair are perceived as similar to one single phonemic category, the perceptual difficulty depends on how well each of the sounds is assimilated to the L1 category. If one of the sounds is perceived as a better exemplar of the L1 category than the other, such as Greek listeners' perception of Southern British English /æ/ as a better fit to native /a/ compared to Southern British English /ʌ/ to /a/ (Lengeris, 2009; Lengeris & Hazan, 2007), then the perception of the contrast is predicted to be moderately good (*Category Goodness*). If the two sounds are perceived as similarly fit exemplars of the L1 category, they fall into the *Single Category* assimilation and it leads to severe perceptual difficulty, as in the case of Mandarin listeners' assimilation of Thai Tone45 and Tone315 to the same native Tone35 (Chen et al., 2020). As for the cases where both non-native phones are perceived as uncategorized (*Uncategorized-Uncategorized*), PAM predicts moderate perceptual performance (e.g., Australian English /əʉ/-/oː/ for Japanese listeners; Bundgaard-Nielsen et al., 2011).

Lastly, PAM also takes into account *Non-Assimilable* phones that fall out of the listeners' perceptual space and are perceived as non-speech sounds (e.g., click sounds in Bantu language; Best et al., 1988). Such sounds may be perceived fairly well because they do not interfere with the native categories. These predictions, however, are subject to individual

variations even within the same listener group. In other words, the same contrast may be perceived and assimilated in different ways by individual listeners with the same L1 (Tyler et al., 2014).

Table 2.1 PAM predictions on the perception of different types of non-native contrasts (Best, 1994, 1995)

Assimilation Type	*Level of perception performance*	*Example*
Two Category (TC)	Very good to excellent	Danish /œ/-/u/ to Australian English /ɜː/-/ʊ/ (Faris et al., 2018)
Uncategorized-Categorized (UC)	Very good	Australian English /ɜː/-/ʉː/ to Japanese listeners (Bundgaard-Nielsen et al., 2011)
Category Goodness (CG)	Moderate to good	Southern British English /æ/-/ʌ/ to Greek /a/ (Lengeris & Hazan, 2007)
Non-Assimilable (NA)	Moderate to good	Zulu click sounds to English listeners (Best et al., 1988)
Uncategorized-Uncategorized (UU)	Moderate	Australian English /əʉ/-/oː/ to Japanese listeners (Bundgaard-Nielsen et al., 2011)
Single Category (SC)	Poor	Thai Tone45 and Tone315 to Mandarin Tone35 (Chen et al., 2020)

PAM makes predictions on how naïve listeners perceive non-native sounds but does not capture the L2 developmental process. To account for L2 perceptual development and changes, Best and Tyler (2007) extended the original PAM to PAM-L2. PAM-L2 introduces further hypotheses on the relative ease/difficulty for L2 learners to acquire and develop new L2 categories. It is predicted that L2 phones that can be well assimilated to corresponding L1 categories will cause difficulties in developing new L2 categories. This applies to the Two Category case because the two L2 sounds are already well distinguishable at the phonemic level based on the assimilations. Another situation is Single Category assimilation, as learners are less likely to discriminate and form new categories for the two sounds if they perceive the sounds as different realizations (variants) of the same L1 category. For a new L2 category to be formed, it is hypothesized that the L2 phones need to have less well (or less similar) L1 counterparts. For example, in Category Goodness assimilation, learners are likely to develop a new L2 category for the less-fit exemplar of the L1 category while keeping the L1 category for the better-fit exemplar. This also applies to the uncategorized sounds, where a novel L2 category may be formed for the uncategorized phones that are assimilated to the same set of L1 phonemes. For the non-assimilable sounds, it is possible that L2 learners will, over time, learn and add them to the speech sound perceptual space as uncategorized sounds.

Whereas PAM and PAM-L2 consider the initial and developmental stages of L2 perception, the Speech Learning Model (SLM, Flege, 1995) takes a different perspective and focuses on ultimate L2 learning attainment. Although SLM was originally proposed as a model for L2 speech production and pronunciation, its hypotheses can be extended to L2 perception as it assumes that production is 'guided' by perceptual similarities between L2 and L1 sounds. Different from PAM which always takes into account sound pairs or contrasts, SLM inspects individual L2 phones and makes predictions based on their (dis)similarities to the learners' L1. It predicts that for adult L2 learners, the perception of L2 sounds that have

identical L1 counterparts should be highly accurate (i.e., successful L2 learning) due to direct L1 transfer. When the L2 sounds are not identical to any L1 sound, a new phonetic category is needed and the relative ease/difficulty of forming the new category depends on how well learners perceive the L1-L2 phonetic differences. If the L2 sound is phonetically distinct from any L1 sound (i.e., a *new* sound), learners should be able to detect the phonetic difference easily. Thus, a new phonetic category will be formed for the L2 sound and the ultimate perception and production of the sound is expected to be good. However, if the L2 sound shares similarities with (but not identical to) any L1 sound, learners may encounter difficulty capturing the more trivial phonetic differences and forming a new L2 category, which leads to less native-like perception and production of the sounds. Overall, SLM's predictions on L2 perception (and production) success of specific phones depend on the individual similarity ratings between L1 and L2 sounds. One limitation is that it does not make direct predictions about L2 learners' perception of sound contrasts and hence may be less informative when examining the learning of minimally different sounds/words.

A later model that has been widely tested is the Second Language Linguistic Perception (L2LP) model (Escudero, 2005; Van Leussen & Escudero, 2015). Different from PAM and SLM, which mainly account for a particular state of L2 learning (i.e., initial or attainment), L2LP makes predictions on the entire L2 perception development process. Similar to the previous models, it proposes that L1 categories contribute to the starting point of L2 speech perception, and the perceptual differences between the L2 and L1 sounds determine learning difficulty. Three different learning scenarios were hypothesized based on the L2-to-L1 sound associations. Table 2.2 presents the different scenarios and compares these hypothesized scenarios with the PAM assimilation types. L2LP defines the *new* scenario similarly to PAM's Single Category assimilation, where there are fewer L1 categories than that is needed for optimal perception of L2 sounds (e.g., two L2 sounds are

perceived as similar to one single L1 category). Thus, a new L2 category needs to be formed to accommodate the L2 contrast and this is predicted to be a difficult scenario. If, on the contrary, there exist more L1 categories than needed, one L2 sound may be perceptually associated with more than one L1 category and it creates a *subset* scenario (similar to PAM *uncategorized* sounds). This scenario is expected to cause less difficulty than the *new* scenario, as it does not depend on the formation of a new L2 category. The third *similar* scenario is where there is a match in the number of L1 categories and desired L2 categories. Similar to PAM's prediction on Two Category assimilation, this situation is considered less difficult as learners only need to adjust their L1 categorical boundaries to fit the L2 sounds. It is worth noting that although the terminology seems to be the same, the *new* and *similar* scenarios in L2LP are different from SLM's proposal of *new* and *similar* phones. L2LP (and PAM) specifically looks at L2 sound contrasts and their mappings onto learners' L1, whereas SLM compares individual L2 sounds to L1 phonetic categories. Therefore, the hypotheses of L2LP/PAM are not directly comparable to SLM, even though they all predict on the relative perceptual and learning difficulties of L2 sounds.

Table 2.2 L2LP predictions on the perception of different types of non-native contrasts (Escudero, 2005)

Scenario	*Learning difficulty*	*Equivalent PAM assimilation type*
New	Most difficult	Single Category
Subset	Medium difficulty	Uncategorized
Similar	Less difficult	Two Category

Overall, the theoretical frameworks on L2 speech perception provide detailed predictions on the potential perceptual difficulty associated with non-native sounds. Although

they tend to have different theoretical focuses, there is general agreement that perceptual and learning difficulties arise from the imperfect (or even problematic) mapping of L2 sounds to L1 categories. There are, certainly, also other factors that influence L2 speech perception, from L1 orthography (e.g., Escudero & Wanrooji, 2010), cognitive factors (e.g., attention effect in Guion & Pederson, 2007), to socio-interactive factors (e.g., age of learning onset in Stölten et al., 2014; L2 experience effect in Bohn & Flege, 1992). The combined impact on L2 perception is significant, and hence, treatment or training may be necessary to help L2 learners cope with the challenge. In the next section, I will briefly review the phonetic training methods emerging from second language acquisition research.

2.1.3. Effect of phonetic training

A number of phonetic training methods have been proposed to improve L2 speech sound perception, differing in the design of stimuli and training schedule, modality of training, and types of feedback provided (e.g., Fouz-González & Mompean, 2021; Grenon et al., 2019; Godfroid et al., 2017; Iverson et al., 2012; Lee & Lyster, 2016; Lim & Holt, 2011; Nishi & Kewley-Port, 2007; Thomson, 2012; Wiener et al., 2021). One of the most famous approaches, High Variability Phonetic Training (HVPT), takes into account the nature of authentic speech input and presents listeners with natural variability in stimuli (e.g., Iverson et al., 2005; Iverson et al., 2012; Lively et al., 1993; Logan et al., 1991; Thomson, 2012). Logan et al.'s (1991) seminal study founded the HVPT paradigm by showing that context and speaker variabilities in speech stimuli improved Japanese listeners' perception of English /ɹ/ and /l/ in both trained and novel stimuli. Since then, HVPT has become a standardized method and has been widely used with different target L2 features.

However, the effectiveness of HVPT has recently been questioned by a large-scale replication of the original seminal study. Brekelmans et al. (2022) reported that, with a

similar methodology to the original studies and a large sample size (N=166), they did not observe a clear difference between high-variability and low-variability training groups in generalizing the target /ɹ/-/l/ perception to novel stimuli. These controversial results indicate that HV input may not be necessary to boost the training effect, though there is a clear need for more work to determine the optimal types and degrees of variability in phonetic training. Moreover, researchers have also investigated whether the size of the stimuli set plays a role in the training effect. For example, Nishi and Kewley-Port (2007) demonstrated that training Japanese-native learners with the full American English (AE) vowel set was more beneficial than concentrating on three of the most difficult AE vowels, alerting phonetic training research to consider the full perceptual space rather than biasing learners to a subset of the categories.

Along with the manipulation of stimuli design, there have been attempts to adjust the training procedure to promote the training effect. Given the significant individual differences in L2 speech learning, it is hypothesized that an adaptive training program would be facilitative (e.g., Grenon et al., 2019; Qian et al., 2018; Yang et al., 2021). Traditionally, researchers pre-determine the training procedure and presentation of the stimuli before training begins, but this may fail to account for individual learners' specific learning patterns and may not be equally effective for all learners. An adaptive training program, instead, provides a way to target learners' individual progress and performance. For instance, Grenon et al. (2019) set up eight levels of difficulty in the training program by gradually introducing more variations in speech rate and stimuli complexity. Learners in the training program started from Level 1 and could only proceed to the next level after they reached the accuracy thresholds determined by the researchers. Thus, each learner can receive just sufficient training based on their individual performance in the training process.

Another widely employed technique is multimodal perception training, providing both auditory and visual cues to aid learning. The visual information could be orthographic forms of the speech stimuli (e.g., Bhide et al., 2020; Escudero, 2015) or non-orthographic symbols or semantic cues (e.g., Godfroid et al., 2017). For instance, Bhide et al. (2020) reported an improvement in the perception of Marathi /d̪/-/ɖ/ and /t̪/-/ʈ/ contrasts, which were found to be particularly challenging for English native listeners (Polka, 1991), after receiving Marathi orthographic input together with the auditory stimuli. The visual aid in L2 perceptual training also extended to the prosodic domain. In Godfroid et al.'s (2017) examination of English listeners' acquisition of Mandarin tones, it was observed that the use of pitch contours and numbers to explicitly mark lexical tones benefited listeners' tonal perception, though colour-marking of lexical tones did not generate similar benefits.

In addition to the manipulations of training materials, the role of feedback has been considered a significant contributor to learning success (e.g., Hardison, 2003; Lee & Lyster, 2015, 2016). In their 2015 study, Lee and Lyster trained Korean learners of English on the /i/-/ɪ/ vowel contrast, either with instruction only or with instruction plus corrective feedback (CF). The results suggested a facilitative role of corrective feedback, as listeners performed better in vowel identification after CF training. The authors furthered these results in their 2016 study, showing the more significant effect of auditory CF compared to visual CF on improving English vowel contrast perception. This suggests potential interactions between training modality and feedback on training outcomes.

In summary, phonetic training methods aiming at improving L2 speech perception have provided fruitful results and diverse perspectives. Most of the training methods involve prolonged training, which typically lasts for several hours over multiple days. It again emphasizes the unneglectable challenges associated with L2 sounds. These challenges can further lead to difficulties in other aspects of L2 acquisition, such as word learning and

processing. Section 2.2. will discuss how and to what extent the perceptual issues extend to the lexical level.

2.2. Linking sounds to meanings: the role of non-native sounds in early word learning

In language development, a crucial milestone involves linking acoustic signals to higher-level semantic information, laying the foundation for constructing meaningful communications. Yet, in L2 acquisition, this process can be limited by the perceptual challenges discussed earlier. If the phonetic information cannot be represented accurately, assigning meanings to it can be problematic. In this section, I will explain how non-native sound perception influences the acquisition of new words. To understand this question, I will focus on empirical evidence from two classic word learning paradigms, one exploring one-to-one sound-meaning mappings (*paired-associate word learning*) and the other considering word learning in ambiguous contexts (*statistical word learning*).

2.2.1. Paired-associate word learning

Laboratory research on word learning has extensively used an explicit training paradigm, *paired-associated word learning.* In this paradigm, learners are typically presented with novel word forms alongside the corresponding referents (e.g., Gupta et al., 2004) or translations (e.g., Krepel et al., 2021), and are instructed to learn the meanings of the new words. Critically, one word is presented together with one referent/translation, allowing for explicit, unambiguous form-to-meaning mappings. However, in L2 word learning, learners may encounter novel word forms that contain non-native sounds, and the word learning outcomes may be greatly influenced. As discussed in the previous sections, L2 learners may have difficulties perceiving and discriminating non-native sounds. This leads to issues in

forming the appropriate phonological categories for the non-native sounds and eventually inaccurate phonological representations for the words.

Although a number of studies have examined L2 sound perception and word learning separately, comparably fewer (but a growing body of research) have directly connected the two areas. One early study that directly investigated this 'phonetic-phonological-lexical continuity' was Wong and Perrachione's (2007) work on L2 tonal word learning. In this study, English-native participants who had no tonal experience learned pseudowords that contained Mandarin tonal features via a paired-associate paradigm. During the training session, learners listened to spoken words while being presented with the corresponding referent pictures. They were then tested on their learning success with a word-meaning mapping task. Importantly, a pitch pattern identification task was employed to examine learners' sensitivity to the tonal features before any tonal training. A correlation was observed between pitch pattern identification and word learning performance, indicating that better tonal identification at the non-lexical level was associated with better word learning for naïve learners. This link between tonal perception and tonal word learning was replicated more widely by later studies examining different tone types (i.e., pitch contours/patterns) and L1 backgrounds (Bowles et al., 2016; Chandrasekaran et al., 2010; Cooper & Wang, 2012; Laméris et al., 2023; Laméris & Post, 2023).

In addition to L2 tonal features, the non-native segmental contrasts have also been found to interfere with word learning. For example, Silbert et al. (2015) examined participants' perception of nine non-native contrasts, differing either in voicing, place of articulation or tone, and the use of the contrasts in word learning. Results revealed a feature-specific influence of perceptual ability on word learning outcomes: better discrimination of a particular feature predicted better learning of the words that utilized the feature. That is, for instance, the discrimination of the place of articulation contrasts (e.g., Igbo bilabial vs labio-

velar stops) was associated with word learning that contained the place contrasts, but it did not predict voicing or tone word learning.

The above-mentioned studies measured and observed a relationship between the perception of specific L2 features/contrasts and the acquisition of words that involved the corresponding features/contrasts, thus bridging L2 perception and word learning based on learners' processing of a specific contrast at different levels. Furthermore, there were studies that focused on the lexical level and assessed how well learners encode various perceptually challenging non-native sounds in word learning (e.g., Escudero et al., 2008; Hayes-Harb & Masuda, 2008; Llompart & Reinisch, 2020). For example, Escudero et al. (2013) examined Spanish-native participants' performance in learning Dutch minimal pair pseudowords while manipulating the perceptual difficulties associated with the minimal pairs. The perceptually difficult minimal pairs (e.g., /piχ/-/pɪχ/) differed in one non-native contrast that caused perceptual issues for Spanish listeners, such as the /i/-/ɪ/ contrast which may be perceived as a single /i/ category in their L1. This fits into the more difficult *Single Category* assimilation in PAM's prediction or the *new* scenario in the L2LP model (see section 2.1.2 for details). The perceptually easy minimal pairs, by contrast, involved contrasts such as /i/-/a/, which could be mapped to two separate Spanish categories (i.e., *Two Category* assimilation in PAM or *similar* scenario in L2LP). Participants were then trained to map these pseudowords with referent pictures in an audiovisual paired-associate learning task and tested on their learning success. A clear divergence was found between participants' performance on the easy and difficult minimal pair words – higher accuracy and shorter response time were observed for the easy pairs. This direct between-contrast comparison provided further evidence for the perceptual-lexical relationship, as word learning was reduced at the group level when perceptually difficult contrasts were present.

However, the continuity observed between L2 perceptual and lexical development needs to be addressed and interpreted with caution. Crucially, good perceptual ability does not necessarily entail good word learning attainment or lexical processing among L2 learners at different proficiency levels (e.g., Llompart, 2021; Pelzl et al., 2019; Sebastián-Gallés & Díaz, 2012; Simonchyk & Darcy, 2017). Take again L2 tonal learning as an example. Pelzl et al. (2019) reported that English-native learners of Mandarin Chinese who reached an advanced level could successfully identify and categorize the L2 tones at near-native accuracy. However, these advanced learners still exhibited issues in a lexical decision task where they needed to reject tonal non-words. Similarly, in a study with English-native learners of Russian, Simonchyk and Darcy (2017) found that accurate discrimination between the Russian plain/palatalized contrasts was not always associated with good processing of the contrasts at the lexical level in a word-meaning mapping task. However, it was also important to note that there were no participants who showed poor discrimination of the contrasts and good lexical processing, indicating that perceptual sensitivity can be a prerequisite of lexical processing.

To summarize, empirical evidence suggests that non-native sound perception can largely impact explicit L2 word learning, as perceptual ability (i.e., discrimination and identification of target contrasts) predicts early word learning performance in unambiguous (paired-associate) learning conditions. Moreover, the relative perceptual difficulty of non-native contrasts as proposed by the L2 speech models can be extended to make predictions on explicit word learning performance. However, the pre-lexical perception of L2 sounds does not guarantee successful lexical processing and accurate lexical representations in the later stages of L2 learning. It is more appropriate to consider perceptual abilities as a fundamental element rather than a determining factor of lexical learning/processing.

2.2.2. Statistical word learning

The paired-associate learning paradigm offered an explicit method to investigate word learning via one-to-one word-meaning mappings. This is more similar to the classroom learning situations where learners are provided with unambiguous mappings between new words and their concepts and are required to memorize the associations. Despite the effectiveness of explicit learning methods (Ellis, 2015), it is not representative of all the word learning scenarios in the real world. Language learners can also link new words with meanings implicitly, that is, without being explicitly taught. This is clearly evident in children's word learning, where their early vocabulary development can effectively emerge from implicit interactions and exposure (e.g., Dickinson et al., 2019). The question then arises as to how learners pick up and acquire the meanings of novel words from the environment, given the highly variable and often ambiguous input around us.

The statistical learning approach provides potential explanations for this learning scenario, suggesting that learners can keep track of the statistical information in the input (e.g., word-referent co-occurrences) over time and across encounters (e.g., Ellis, 2015; Williams & Rebuschat, 2022). One paradigm that has developed from the statistical learning perspective is the *cross-situational word learning* (CSWL) paradigm, which attempts to mimic the ambiguous, immersive learning environment in natural language learning. Since all studies in this dissertation project are based on the CSWL paradigm, I will first introduce in detail how the paradigm works, and then discuss how CSWL has been extended to examine speech sound perception and word learning.

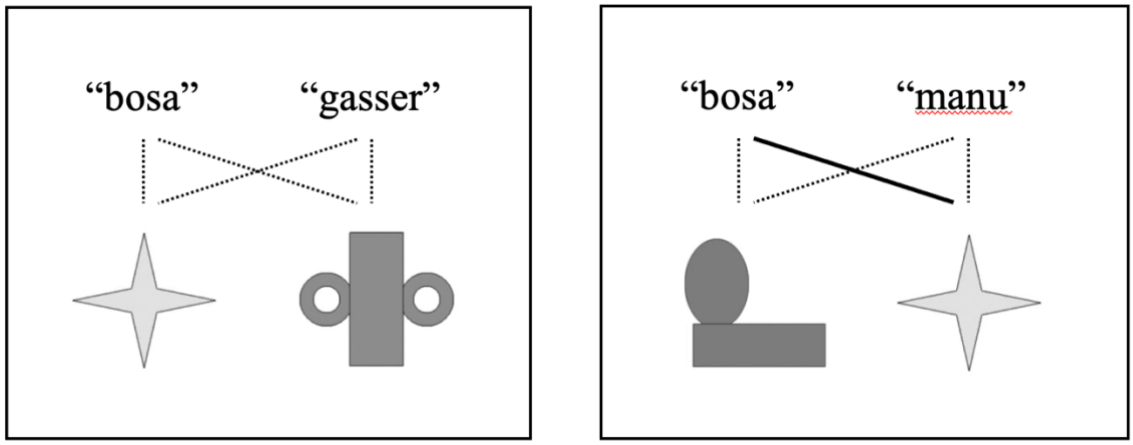
The first empirical application of CSWL was carried out by Akhtar and Montague (1999), where children aged two to four years learned novel adjectives from ambiguous naming contexts. In each individual learning (or naming) context, children heard a novel label (e.g., 'This is a *modi* one.') for an object, with the label referring to one feature of the

object (either shape or texture). The meaning of the new label is ambiguous in one single learning event; however, children can accumulate information on the label-feature co-occurrence over multiple learning events with different objects. They were able to apply the newly acquired adjective to novel objects that shared the same feature.

This early demonstration of the effectiveness of statistical tracking in word learning was further examined after a more controlled laboratory design was proposed by Yu and Smith (2007; Smith & Yu, 2008). In their design, learners were exposed to experimental trials in which the within-trial ambiguity was carefully controlled. In each trial, learners were presented with multiple spoken words and their referent pictures simultaneously but were not instructed on the individual word-referent mappings (Figure 2.1a presents an example, dotted lines indicated potential mappings). And the degree of ambiguity can be manipulated by increasing or reducing the number of co-occurring words and referents. This thus resembles the real-life learning situations in a laboratory setting, where referential ambiguity is introduced in word learning (as opposed to paired-associate learning). From each individual trial, it is impossible for learners to infer the word-referent associations, as each word can refer to any of the two referents (Figure 2.1a). Instead, they need to store information across trials, and when they encounter the same word-referent combination again in another trial (Figure 2.1b, "bosa" and the star shape), they will start to form the corresponding associations.

Figure 2.1 Illustration of CSWL trials based on Smith and Yu's (2008) stimuli.

1a 1b



This cross-situational, statistical learning of words has been found to be successful for different word categories (e.g., nouns - Yurovsky et al., 2013; Yu & Smith, 2007; verbs – Monaghan et al., 2019; Zhang et al., 2021; adjectives – Akhtar & Montague, 1999; Rebuschat et al., 2021) and for different learner groups (e.g., infants – Smith & Yu, 2008; Vlach & Johnson, 2013; children – Suanda et al., 2014; younger adults - Yu & Smith, 2007; older adults – Bulgarelli et al., 2021). However, most of these CSWL studies explored the acquisition of additional words in learners’ native languages as the stimuli typically involved native pseudowords (i.e., non-words that follow learners’ native phonotactics). This design is likely because these studies using CSWL, or statistical learning in general, aimed to investigate the mechanisms underpinning language acquisition, and thus has an initial focus on first language development. This is different from the paired-associate paradigm in that paired-associate learning naturally fits into the more explicit L2 learning conditions. Nevertheless, more recently, a few studies started to extend CSWL to second or foreign language acquisition, exploring whether and how L2 learners deal with non-native sounds in addition to resolving referential ambiguities (Hu, 2017; Junttila & Ylinen, 2020 for child learners; Escudero et al., 2022; Ge et al., under review; Tuninetti et al., 2020 for adult learners).

Tuninetti et al. (2020) was the first study that directly addressed the question of how non-native phonological contrasts interact with CSWL among adults. To compare word learning performance with the presence of different non-native contrasts, English-native speakers were presented with Dutch and Brazilian Portuguese pseudowords that contained different word-middle vowels. The CSWL learning phase contained 2*2 word-picture associations – each trial included the presentation of two words with two pictures. Importantly, some of the words were minimally different by only one non-native vowel contrast (e.g., Dutch pseudowords /piχ/-/pyχ/, Brazilian Portuguese pseudowords /fɛfe/-/fefe/). Thus, the learning of these words depends on accurate discrimination and encoding of the respective vowel contrasts at the lexical level. Moreover, according to the L2LP and PAM models, some of the non-native vowel pairs were expected to be perceptually easier as they had separate L1 counterpart categories, whereas others were more difficult if they were perceived as one single L1 sound. Learning performance was measured via a 2-alternative forced-choice task, where learners chose the correct referent from two pictures for every spoken word. It was observed that, during testing, if the two pictures were associated with two non-minimal pair words, decisions were made accurately. But when the pictures corresponded to two minimal pair words, learners' performance was influenced by the perceptual difficulty of the target contrasts – higher accuracy was achieved for perceptually easy minimal pairs. This study demonstrated that CSWL was effective in the L2 context, but learning outcomes were linked to the perceptual difficulty of the non-native sounds present in the words, which is consistent with Escudero et al.'s (2013) findings from a paired-associate training task.

Another study that examined cross-situational learning of L2 words was Escudero et al. (2022), where Mandarin-native speakers were trained with English pseudowords. The pseudowords were paired to form minimally different words that contrast either in a

consonant (e.g., /bɔn/-/pɔn/) or a vowel (e.g., /dit/-/dɪt/). Based on the predictions of L2LP and PAM, it was hypothesized that the vowel minimal pairs would cause more issues for Mandarin-native learners because the vowel contrasts could be assimilated to one L1 category (e.g., /i/-/ɪ/ to Mandarin /i/), whereas the consonant contrasts would be perceptually easier and would lead to better word learning performance. However, results showed no such difference between consonant and vowel minimal pair learning, and Mandarin-native participants' performance in all (non)minimal pair conditions was significantly below that of English-native participants. This seems to contradict the hypothesis that perceptual issues with non-native sound contrasts influence performance at the lexical level. The authors offered a potential explanation for the findings, attributing the lack of a perceptual-lexical link to the nature of the stimuli. As they employed stimuli produced with infant-directed speech that contained great pitch variations, it was proposed that this might mislead Mandarin-native speakers' attention to the irrelevant pitch cue and hence impede learning. Overall, although this study did not provide evidence that perceptually more difficult L2 contrasts would reduce word learning outcomes, we can draw from the English vs Mandarin group comparison that CSWL tends to be affected by the presence of non-native sounds in general.

Moreover, a recent study by Ge et al. (under review) reported that even with phonetic (perceptual) training on the target non-native contrasts, cross-situational learning of non-native minimal pairs still caused substantial difficulty. In their study, English-native speakers were provided with perceptual discrimination training on Portuguese consonant and vowel contrasts (e.g., /l/-/ʎ/, /n/-/ɲ/, /e/-/ɛ/, /o/-/ɔ/) before learning pseudowords that contain these contrasts (e.g., /paʎu/, /dɛtu/) via CSWL. The trained contrasts exist in Portuguese but not in English, and they are considered challenging for English-native speakers because of the perceived phonetic similarity (Macedo, 2015; Rato, 2019). It was observed that the

perceptual training improved learners' discrimination of the non-native contrasts, but they still had difficulty identifying newly acquired non-native minimal pair words (e.g., /palu/ vs /paʃu/). The results further confirmed the difficulty associated with non-native speech sounds in CSWL, and it suggested that the speech sound discrimination ability may not directly transfer to statistical word learning.

In addition to the three studies that investigated the direct impact of non-native contrasts on CSWL, Li and Benitez (2023) provided indirect evidence for the topic by exploring CSWL under bilingual conditions. To mimic a bilingual input situation, participants were trained to map two novel words to one single referent via a CSWL task. In some mappings, the two words can be distinguished based on a tonal cue (i.e., one word carried lexical tone and the other did not). It was found that Chinese-English bilinguals performed better than English monolinguals and Spanish-English bilinguals in utilizing the tonal cue to aid bilingual word learning. It can be interpreted that a familiar phonetic cue facilitates statistical word learning, which indirectly supports the link between perceptual and lexical processing.

In summary, there has been scant investigation into L2 word learning in the CSWL literature, with notably fewer studies exploring the integration of L2 speech perception and word learning. The broad aim of the studies in this dissertation is to address this research gap and provide further empirical evidence for the phonetic-phonological-lexical relationship from a statistical learning perspective. In the following section, I will discuss the limitations of the previous research and how the present studies contribute to the existing research gaps in the CSWL literature.

2.2.3. Research gap

Previous findings have shown that adult learners' acquisition and processing of L2 vocabulary can be substantially affected by their perception of the L2 contrasts. However, these observations are constrained in their applicability to naturalistic language learning contexts because of the predominant reliance on explicit, paired-associate training methods that resemble classroom teaching. Hence, the inference we can derive is that the perceptual accuracy of L2 contrasts is associated with the explicit pairing of novel L2 words to their respective meanings. To accommodate the more immersive word learning contexts (beyond classroom settings), it is crucial for empirical research to utilize a learning task that closely mirrors the implicit acquisition of words through exposure. The CSWL paradigm is well-suited for this purpose. Therefore, the studies in this dissertation employ the CSWL task to examine the impact of non-native phonological contrasts on word learning.

The few studies that explored L2 word learning using the CSWL paradigm have several limitations in their scope. Firstly, Tuninetti et al. (2020), Escudero et al. (2022) and Ge et al. (under review) focused on non-native segmental contrasts and segmental minimal pair words, setting aside the suprasegmental features such as lexical tones. Although Li and Benitez's (2023) design included a tonal cue, the primary purpose was to investigate the role of lexical tone in distinguishing and acquiring two languages from bilingual input, rather than comparing the learnability of different non-native tonal contrasts/words. Thus, the role of non-native lexical tones in CSWL has not been thoroughly tested. Yet, lexical tone received significant research attention in the paired-associate learning literature, where a continuity across the phonetic, phonological and lexical domains was observed (e.g., Wong & Perrachione, 2007). Additionally, the processing of tonal and segmental features has been found to be separate for L2 learners, especially at the beginner level (e.g., Zou et al., 2017), which further leads to challenges in integrating the two dimensions. This indicates that our

understanding of L2 segmental contrast acquisition may not be directly extendable to tonal acquisition. The present studies, therefore, address this gap and investigate how non-native tonal contrasts influence the cross-situational, statistical learning of novel words.

Another limitation of the previous CSWL research is that it tested the potential word learning difficulties resulting from L2-to-L1 sound assimilation (e.g., Tuninetti et al., 2020), but did not look into the learning of L2 sound features that are not used contrastively at all in learners' L1 phonological inventory (e.g., pitch or duration cues). In such instances, learning depends not only on how learners assimilate the L2 sounds to L1 categories, but also on their attentional distribution to the relevant L2 cues. An examination of such features allows us to understand how L2 learners incorporate and encode novel phonetic dimensions in speech perception and word learning.

Lastly, although a group-level impact of non-native contrasts on CSWL was reported (Escudero et al., 2022; Tuninetti et al., 2020), it is not yet clear how individual learners perform in the tasks. As discussed in section 2.2.1., paired-associate learning research has revealed that learners' individual perceptual sensitivity to the specific non-native contrasts predicts their use of the contrasts in explicit word-meaning mapping (e.g., Silbert et al., 2015; Wong & Perrachione, 2007). The current studies examine whether these individual variations are also present when learning from cross-situational statistics. Moreover, several different individual difference (ID) factors are taken into account in addition to L2 speech sound perception ability. The next section will introduce the ID measures involved in the studies and the rationale behind them.

Overall, the current studies add to the existing literature by testing the link between non-native sound perception and word learning under a statistical learning framework, particularly emphasizing the relatively understudied suprasegmental tonal features and performance at both individual and group levels.

2.3. Individual difference factors in word learning

Despite the general impact of non-native sounds on word learning, there exist potential individual variations in the extent to which learners are influenced. This individual variation may arise from a range of factors, from cognitive abilities (e.g., working memory, Wen & Jackson, 2022) to conative factors such as motivation (Dörnyei, 2005, Chapter 4). The current studies closely examined two ID factors, target language experience and usage, and auditory processing ability, one examining a usage-based predictor and the other focusing on the domain-general cognitive processing of acoustic signals. In the following sections, I will illustrate why these two factors are of particular interest.

2.3.1. Target language experience and usage – the heritage speaker population

The first individual difference factor examined is experience with and usage of the target language. Since unfamiliar, non-native sound contrasts lead to perceptual and lexical learning difficulties, one further question to be answered is whether and to what extent experience with the target contrasts facilitates learning. There are two different populations whose language profiles can help address this question – the L2 learners and the heritage speakers of the language. Both groups entail potentially large variations in the levels of experience and usage of the language, allowing for the individual difference investigation. In the present studies, I focus on the heritage speakers because their early contact with the target language (and speech contrasts) is of particular interest. Heritage speakers have exposure to a minority (heritage) language in the family context during early childhood, but they begin to learn a different societal/majority language upon starting school and become proficient and dominant in the societal/majority language thereafter. The early exposure may allow heritage speakers to develop sensitivity to the target speech contrasts in the first years of life, during

which humans are tuned to perceive their native sounds (Kuhl, 2004). However, they may have relatively limited experience and usage of the target contrasts later in life if the contrasts are not relevant in their dominant language (e.g., lexical tones in the English-dominant environment). It thus allows us to investigate not only the target language experience effect on word learning, but also how early exposure shapes learners' later language learning performance.

If the heritage speakers perform similarly to the Mandarin-native speakers at the group level, it emphasizes the importance of early contact with the speech contrasts on learning attainment. If, on the other hand, the individual variations are more striking, it implies the significant role of experience and usage in word learning. It is important to highlight that, within the context of heritage speakers, the speech sounds and words are considered 'native', as heritage speakers are acknowledged as bilingual speakers of both the societal and the heritage language. Therefore, I refrained from using the term 'non-native' when describing the heritage language. Instead, I opted for the term 'target language' and referred to what was previously called 'non-native contrasts' as 'contrasts that are irrelevant in one's dominant language'.

2.3.2. Auditory processing ability

Auditory processing refers to listeners' ability to encode and reproduce the spectral and temporal details of nonspeech sounds (e.g., Surprenant & Watson, 2001). This basic auditory perception has been found to be a potential predictor of different aspects of language learning and processing (e.g., Kachlicka et al., 2019, for speech perception and syntactic processing; Kempe et al., 2012, for speech perception; Mueller et al., 2012, for linguistic rule learning; Li & DeKeyser, 2017, for vocabulary learning). Within the context of the current studies, auditory processing is closely relevant for two main reasons. First, the question I am

addressing connects two areas of second language learning/processing – perception and lexical representation. As for the perception aspect, previous evidence primarily came from the exploration of language-specific or feature-specific perception, all of which belonged to the linguistic domain. However, the perceptual ability should also take into account the lower-order, domain-general processing of all acoustic signals (speech and nonspeech). The general sensitivity to certain acoustic dimensions may contribute to the processing of the dimension in speech sounds. For example, Kempe et al. (2012) observed that the perception of non-native sounds with a vowel length contrast was associated with participants' temporal auditory processing of nonspeech sounds. Therefore, it is interesting to explore how the domain-general perception interacts with the domain-specific (speech) perception in determining individuals' performance in word learning.

Secondly, the feature under examination in the current studies is lexical tone, which is highly reliant on the pitch dimension. Specifically, lexical tones are realized via changes in pitch patterns (i.e., fundamental frequency) and hence may be heavily influenced by listeners' acuity to trivial pitch variations. Moreover, better pitch acuity may enable non-native listeners to focus their attention on the tonal feature more effectively during the learning process. Thus, assessing individuals' pitch-related auditory processing may offer further insight into how well non-tonal speakers learning a tonal language adapt to integrate the pitch cue in lexical learning.

2.4. Research questions

In a series of four studies, this dissertation project aims to timely address the following research questions:

1. How do non-native phonological contrasts (or contrasts that are irrelevant in one's dominant language) influence adults' statistical word learning?

2. Which individual difference factors predict word learning outcomes when such contrasts are present?

For each study, a set of more detailed but interconnected research questions are taken into account. Study 1 set up the basic statistical word learning task for all studies following the CSWL paradigm (Smith & Yu, 2008; Yu & Smith, 2007). In particular, English-native and Mandarin-native participants learned Mandarin tonal pseudowords. The key research questions are whether minimal pairs pose difficulty CSWL compared to phonologically distinct words and additionally, whether minimal pairs that differ in non-native phonological contrasts (i.e., lexical tone) pose further difficulty compared to minimal pairs with contrasts similar to native sounds. Moreover, I examined if learners' non-native speech sound perception develops during CSWL.

Study 2 extended the CSWL task employed in Study 1 by doubling the number of learning trials. It addressed one further question: Does extended exposure to cross-situational statistics improve learning outcomes?

Study 3 explored the individual differences in target language experience and usage and its impact on word learning. It also demonstrated whether early exposure to the target feature is sufficient to promote the use of that feature in word learning later in life. By examining heritage speakers' performance in the CSWL task, I addressed the question of whether minimal pairs and phonological contrasts that do not exist in heritage speakers' majority language (i.e., lexical) pose difficulty during CSWL. Additionally, I explored whether the degree of heritage language experience and usage influence learning outcomes.

Study 4 investigated the individual difference measure of auditory processing abilities. It additionally involved comparing an online eye-tracking and an offline accuracy measure of learning performance. I tested if learners' auditory processing abilities predict statistical learning of non-native words, and if learners perceive and discriminate between

tonal differences before and after statistical learning of tonal words. Furthermore, I compared whether online eye-tracking and offline accuracy measures show similar learning performance patterns in CSWL.

3. Published paper 1: The role of phonology in non-native word learning: Evidence from cross-situational statistical learning.

Page number: 47-114

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 could be related to learning. We trained English- and Mandarin-native speakers with pseudowords via a cross-situational statistical learning task (CSWL). 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 et al., 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 word learning, CSWL) (e.g., Escudero et al., 2022; Monaghan et al., 2019; Rebuschat et al., 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 CSWL 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 et al., 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 et al., 2010; Silbert et al., 2015; Wong & Perrachione, 2007). Our current study will address this gap by exploring the effect of phonology on non-native word learning using a CSWL 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 to those children face 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; L. Smith & Yu, 2008; Tomasello & Barton, 1994), but also has implications for second and foreign language research (e.g. Monaghan et al., 2021; K. Smith et al., 2011; Walker et al., 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 et al., 2002 for sound discrimination; Saffran et al., 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 (L. Smith & Yu, 2008; Yu & Smith, 2007). CSWL 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 CSWL for both children (e.g., Childers & Pak, 2009; L. Smith & Yu, 2008; Suanda et al., 2014; Yu & Smith, 2011) and adults (e.g., Gillette et al., 1999; K. Smith et al., 2009, 2011; Yurovsky et al., 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 CSWL 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 CSWL can also drive syntactic acquisition of word order (Rebuschat et al., 2021).

However, most of the CSWL literature left aside the important impact of phonology on word learning. There are two potential issues related to this. First, in most CSWL studies, the stimuli (words or pseudowords) used were phonologically distinct (e.g., pseudowords such as *barget, chelad* in Monaghan et al., 2019). However, as reported by Escudero et al. (2016b), 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 CSWL 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., 2016b; Monaghan & Mattock, 2012; Monaghan et al., 2019; see Hu, 2017, and Junttila & Ylinen, 2020, for an exception). In other words, CSWL 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 et al., 2008; Iverson et al., 2003; Rato, 2018; Rato & Carlet, 2020; Takagi & Mann, 1995; Wong & Perrachione, 2007). Tuninetti et al. (2020) trained Australian English speakers with novel Dutch and Brazilian Portuguese vowel minimal pairs in a CSWL 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 & 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 CSWL. A more recent study by Escudero et al. (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. (2016b) 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 CSWL 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 et al., 2021), and the integration of suprasegmental and segmental features can be challenging for beginner learners (Zou et al., 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 CSWL 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., 2016b, 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 et al., 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,

[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 et al., 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

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 CSWL 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 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

bi/multilingualism effect in CSWL had mixed findings in previous research as well, with some reporting a bilingual advantage (Escudero et al., 2016a) 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.

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 3.1). 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.

Table 3.1 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

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 et al.'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 CSWL 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 et al. (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 CSWL 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 CSWL – 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 to learn 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 CSWL, 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 CSWL session.

The occurrence of each trial type was controlled in each block and throughout the experiment. There were six CSWL 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 CSWL task, participants completed the debriefing questionnaire, in the question sequence outlined above. Only one question was presented on the screen each time.

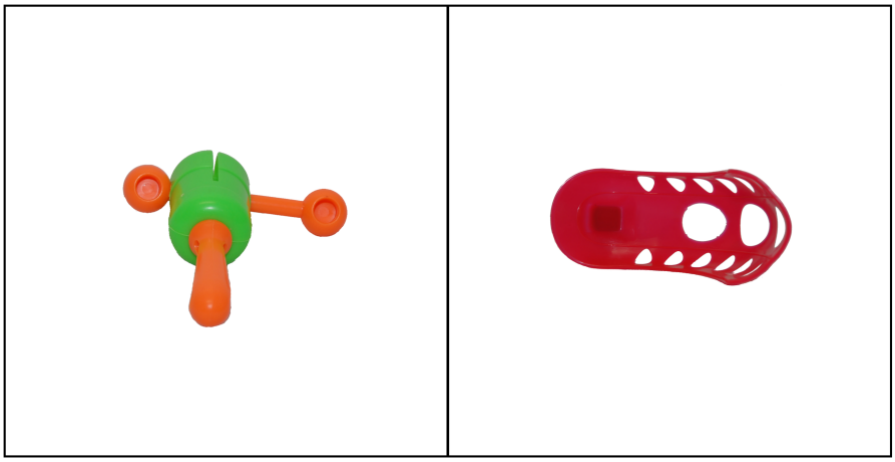
Trial procedure

In each CSWL 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 3.1 provides an example of a CSWL 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.

Figure 3.1 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.



Data analysis

We excluded participants who failed to successfully complete the initial sound check or failed to complete the CSWL 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

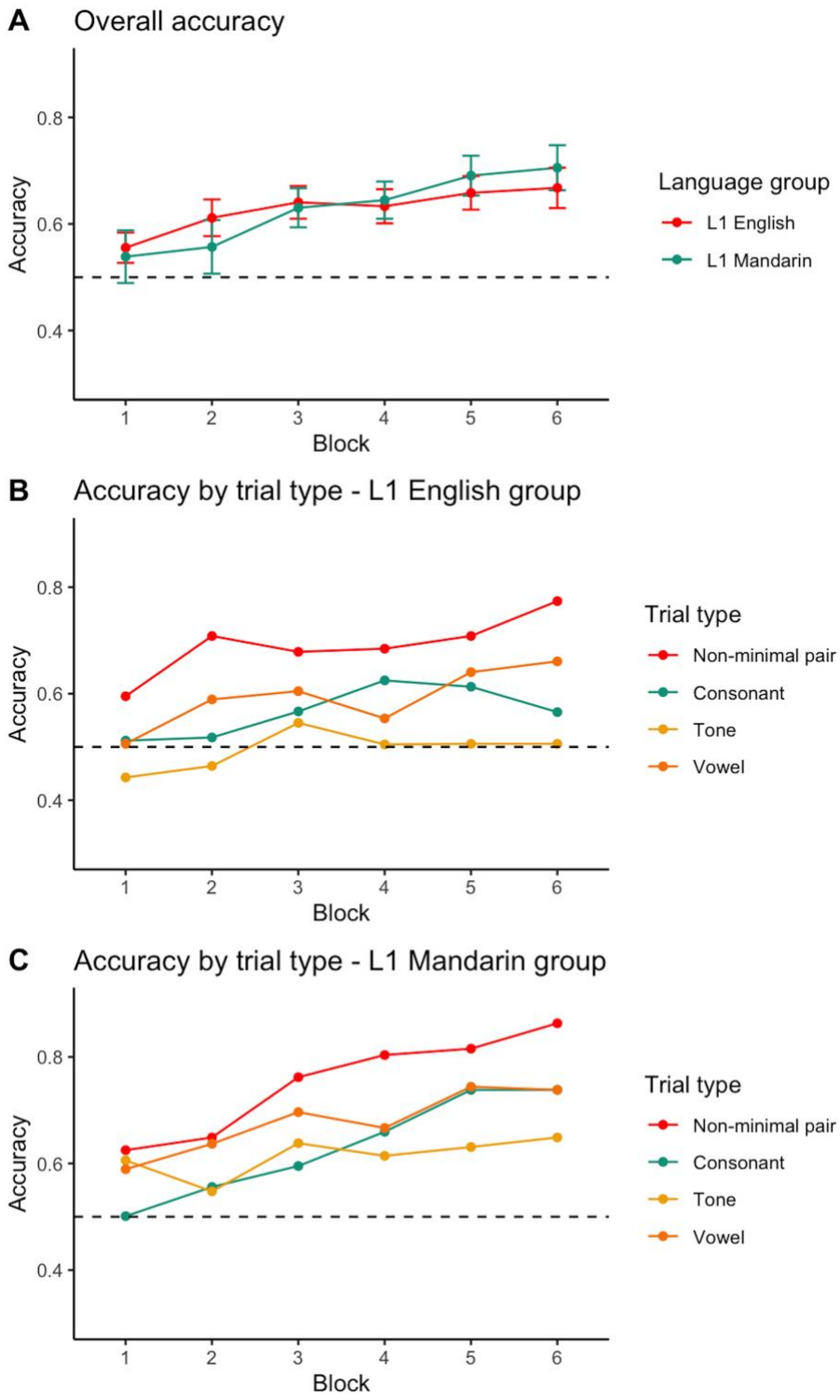
were constructed from null model (containing only random effects of item and participant) to models containing fixed effects. We tested if each of the fixed effects improved model fit using log-likelihood comparisons between models. A quadratic effect of block was also tested for its contribution to 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 3.2A 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 3.2B & 3.2C), 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 CSWL session.

Figure 3.2 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.

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 ($\chi^2(1) = 5.478$, $p = .019$), 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 3.2[3] [4].

Table 3.2 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

[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 3.2 shows the model with non-minimal pair trial as the reference level. In supplementary materials, Table S3.2 present models with other trial types as reference levels respectively.

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)).

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 3.3 we presented the best-fitting model for tonal 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 ($\chi^2(1) = 0.8722$, $p = .350$).

Table 3.3 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))

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 S3.3). A similar pattern was observed for the English-native group (Table S3.4).

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 S3.1). 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 S3.5. 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 CSWL 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.

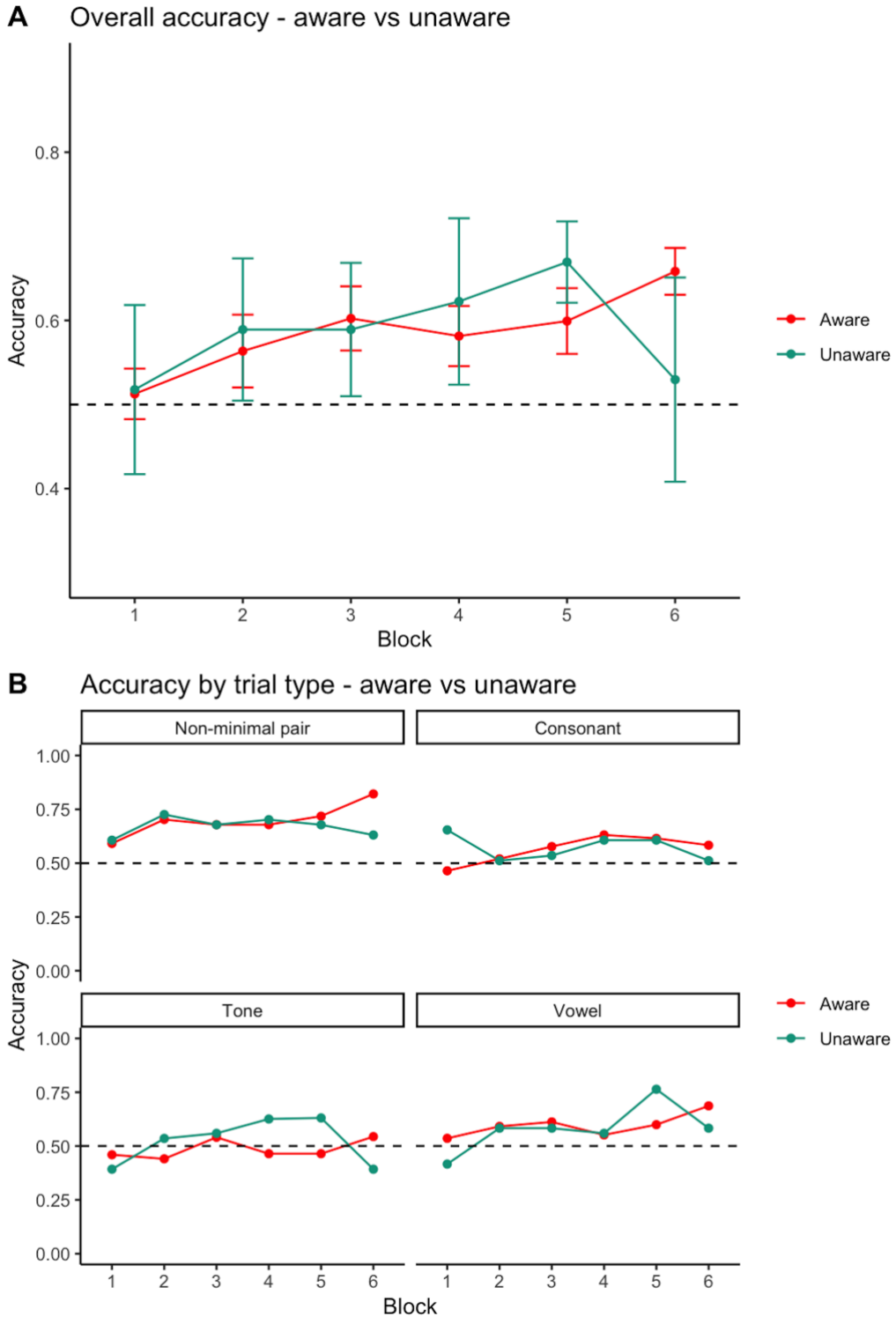
The awareness coding followed the guidance of Rebuschat et al.'s (2015) coding scheme, ranking from full awareness to complete unawareness. Participants who reported using lexical tones to distinguish words strategically were considered "full awareness" (Q1~3), those who mentioned lexical tones in response 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 was explicitly asked 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 CSWL task. As shown in Figure 3.3, 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.

Figure 3.3 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.

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 3.4 summarizes the final model.[5]

Table 3.4 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))

[5] Additional Table S3.6 presents models with other trial types as reference levels.

Exploratory analysis. We investigated if aware and unaware participants differ at the end of the CSWL 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 S3.7. 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., 2016b, 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.

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 CSWL 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 CSWL (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 CSWL 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 3.3, 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 CSWL trials (i.e., participants went through the Experiment 1 CSWL task twice, 12 blocks in total). Experiment 2 was preregistered on OSF: https://osf.io/2m4nw/.

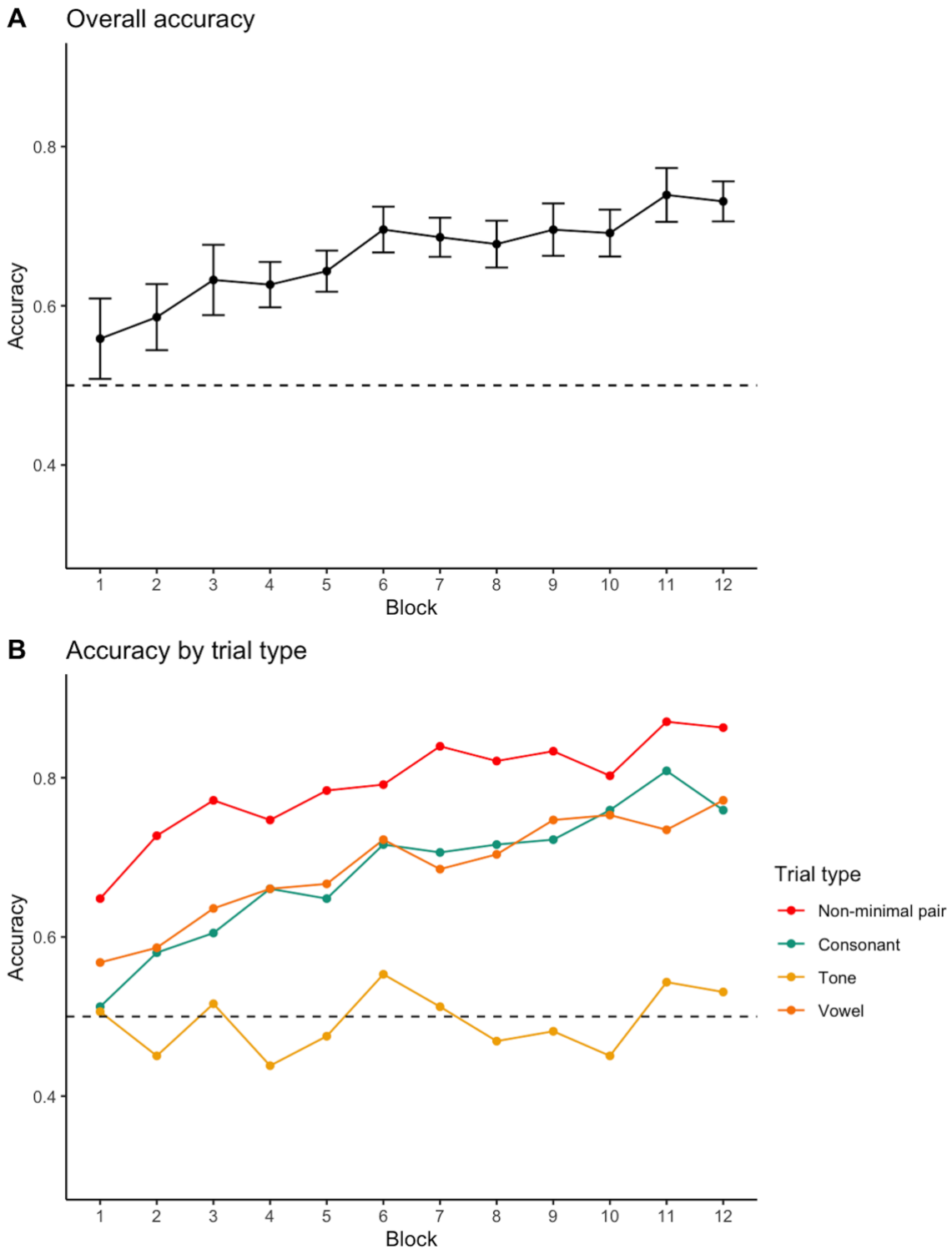
Results

Performance on cross-situational task

Accuracy. Figure 3.4A 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-minimal pair trials, followed by clear learning in consonantal and vocalic trials. However, learning in tonal trials was still not observed (Figure 3.4B).

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

Figure 3.4 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.

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 3.5[7].

Table 3.5 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

[7] Additional Table S3.8 presents models with other trial types as reference levels.

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))

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 S3.2). 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 S3.9.

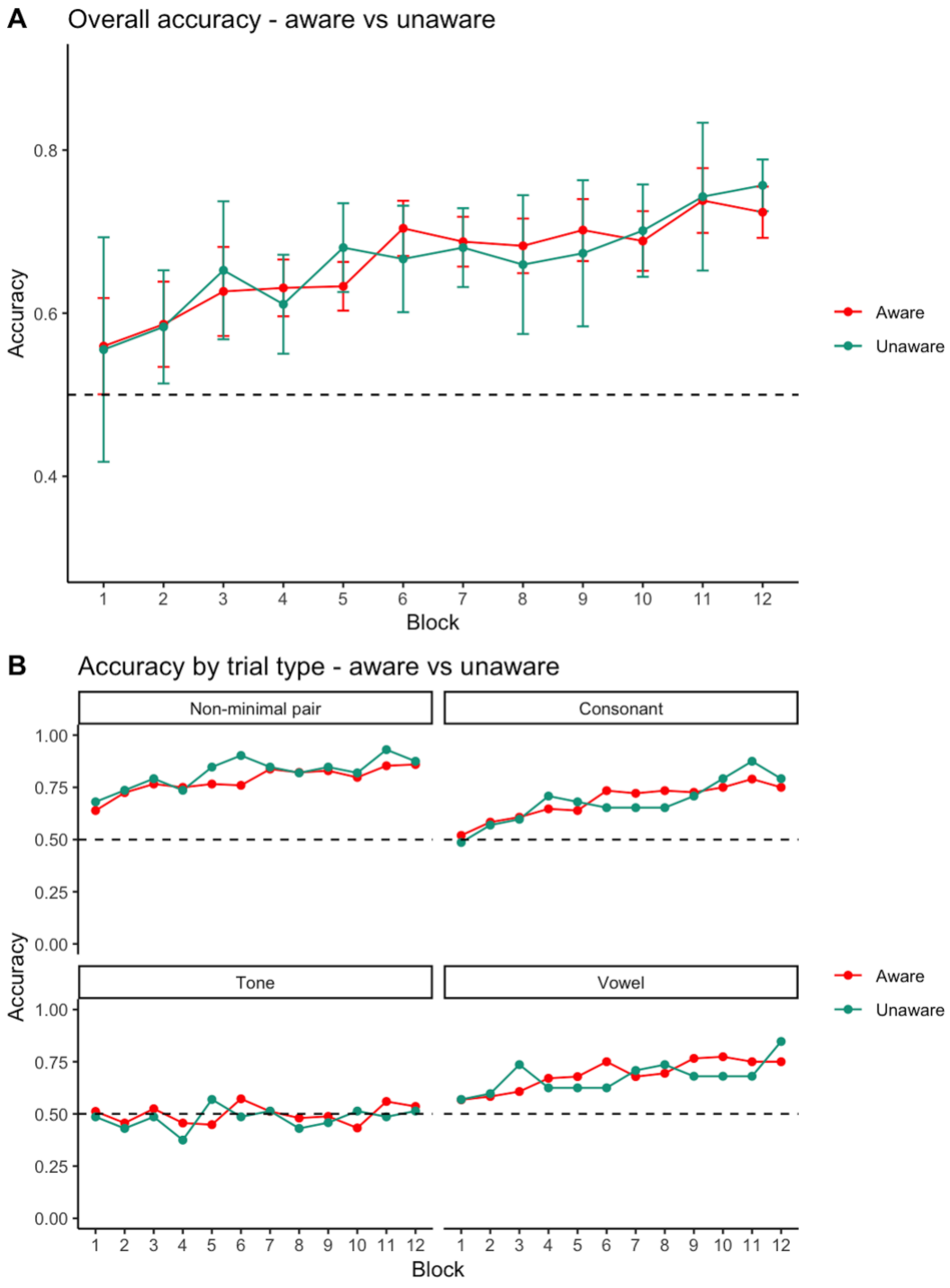
Retrospective verbal reports

Proportion of aware and unaware participants. Three participants were coded as fully aware as they reported using tonal cues strategically without being explicitly asked (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 (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 CSWL task. As in Experiment 1, the aware and unaware participants shared similar learning trajectories (Figure 3.5).

Figure 3.5 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.

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 CSWL 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 CSWL is 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. (2016b) 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 et al., 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 CSWL 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., 2016b, 2022; 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 CSWL 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 CSWL 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 CSWL 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 et al., 2019; Fuhrmeister & Myers, 2020; Godfroid et al., 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 CSWL 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 et al., 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, 2018). 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, as 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 CSWL 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 et al., 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 CSWL 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 CSWL task, cannot be separated from their vocabulary knowledge. Future studies can incorporate direct tests of sound perception and discrimination before and after the CSWL task to explore more precisely how CSWL 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 et al. (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 CSWL 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 et al., 2020a, 2020b), working memory, to implicit and explicit language aptitudes.

References

Adesope, O. O., Lavin, T., Thompson, T., & 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., & van Rijn, H. (1993). *The CELEX lexical database* (CD-ROM). University of Pennsylvania, Philadelphia: Linguistic Data Consortium.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In M. J. Munro & O-S. Bohn (Eds.), *Second Language Speech Learning: The Role of Language Experience in Speech Perception and Production*. Amsterdam: John Benjamins, pp. 13–34.

Chan, R. K., & Leung, J. H. (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, P. D., & Wong, P. C. (2010). Individual variability in cue-weighting and lexical tone learning. *The Journal of the Acoustical Society of America, 128*(1), 456-465.

Chang, C. B., & Bowles, A. R. (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., & 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, J. B., & Pak, J. H. (2009). Korean- and English-speaking children use cross-situational information to learn novel predicate terms. *Journal of Child Language, 36*(1), 201– 224.

Cooper, A., & 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., & Chen, H.-C. (1997). Lexical tone in Cantonese spoken-word processing. *Perception and Psychophysics, 59*(2), 165–179.

Davies, R. A., Arnell, R., Birchenough, J. M., Grimmond, D., & Houlson, S. (2017). Reading through the life span: Individual differences in psycholinguistic effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(8), 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., & 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*. [Doctoral dissertation, Utrecht University]. LOT Dissertation Series 113.

Escudero, P., Mulak, K. E., Fu, C. S., & Singh, L. (2016a). More limitations to monolingualism: Bilinguals outperform monolinguals in implicit word learning. *Frontiers in Psychology, 7*, 1218.

Escudero, P., Mulak, K. E., & Vlach, H. A. (2016b). Cross-situational learning of minimal word pairs. *Cognitive Science, 40*(2), 455-465.

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

Francis, A. L., Ciocca, V., Ma, L., & 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., & Myers, E. B. (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., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition, 73*(2), 135-176.

Godfroid, A., Lin, C. H., & 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, Y. C. (2018). Contextual effect in second language perception and production of Mandarin tones. *Speech Communication, 97*, 32-42.

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

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

Horst, J. S., & Hout, M. C. (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, C. F. (2017). Resolving referential ambiguity across ambiguous situations in young foreign language learners. *Applied Psycholinguistics, 38*(3), 633-656.

Isbilen, E. S., & Christiansen, M. H. (2022). Statistical Learning of Language: A Meta-Analysis Into 25 Years of Research. *Cognitive Science, 46*(9), e13198.

Iverson, P., & Evans, B. G. (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, P. K., Akahane-Yamada, R., Diesch, E., Kettermann, A., & 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 Z. Jing-Schmidt (Eds.), *Proceedings of the 23rd North American Conference on Chinese Linguistics* (Vol. 1, pp. 35–50), University of Oregon, Eugene, OR.

Junttila, K., & 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.

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

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & 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., & 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., & 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., & 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, H. K., & 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, E. M. (1990). Constraints children place on word meanings. *Cognitive Science, 14*(1), 57– 77.

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

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), 101–111.

Monaghan, P., & Mattock, K. (2012). Integrating constraints for learning word-referent mappings. *Cognition, 123*(1), 133–143. https://doi.org/10.1016/j.cognition.2011.12.010.

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

Nixon, J. S. (2020). Of Mice and Men: Speech Sound Acquisition as Discriminative Learning from Prediction Error, Not Just Statistical Tracking. *Cognition, 197*, 104081. doi:10.1016/j.cognition.2019.104081

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

Poepsel, T. J., & Weiss, D. J. (2016). The influence of bilingualism on statistical word learning. *Cognition, 152*, 9-19.

Quine, W. V. O. (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., & 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., & Schoetensack, C. (2021). Learning vocabulary and grammar from cross-situational statistics. *Cognition, 206*, 104475.

Rescorla, R. A., &Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory*, 64–99. Appleton-Century-Crofts.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926-1928.

Saito, K., Kachlicka, M., Sun, H., & Tierney, A. (2020a). 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*, 104168.

Saito, K., Sun, H., & Tierney, A. (2020b). 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, J. A., & 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, N. H., Smith, B. K., Jackson, S. R., Campbell, S. G., Hughes, M. M., & Tare, M. (2015). Non-native phonemic discrimination, phonological short term memory, and word learning. *Journal of Phonetics, 50*, 99-119.

Smith, K., Smith, A. D., & Blythe, R. A. (2009). Reconsidering human cross-situational learning capacities: A revision to Yu and Smith's (2007) experimental paradigm. In N. Taatgen & 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, A. D., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science, 35*(3), 480-498.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*(3)*,* 1558-1568.

So, C. K., & Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Language and Speech, 53*(2), 273-293.

So, C. K., & Best, C. T. (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, S. H., Mugwanya, N., & Namy, L. L. (2014). Cross-situational statistical word learning in young children. *Journal of Experimental Child Psychology, 126*, 395-411.

Suanda, S. H., & Namy, L. L. (2012). Detailed behavioral analysis as a window into cross-situational word learning. *Cognitive Science, 36*(3), 545-559.

Sun, H., Saito, K., & 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., & 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, E. D. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language, 56*(1), 16-34.

Tomasello, M., & Barton, M. E. (1994). Learning words in nonostensive contexts. *Developmental Psychology, 30*(5), 639-650.

Tuninetti, A., Mulak, K. E., & 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., & Rebuschat, P. (2020). Distinctions in the acquisition of vocabulary and grammar: An individual differences approach. *Language Learning, 70*(S2), 221-254.

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

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*(1)*,* 49– 63.

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

Wong, P. C., & Perrachione, T. K. (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., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18*(5), 414–420.

Yu, C., & 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.

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

Zou, T., Chen, Y., & 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.

Supplementary materials

Table S3.1 List of minimal pairs in the four trial types

Consonantal minimal pair		Vocalic minimal pair		Tonal minimal pair		Non-minimal pair	
pa1mi1	ta1mi1	li1fa1	lu1fa1	pa1mi1	pa4mi1	pa1mi1	li4fa1
pa1mi1	ka1mi1	li1fa1	lei1fa1	ta1mi1	ta4mi1	pa1mi1	lu4fa1
ta1mi1	ka1mi1	lu1fa1	lei1fa1	ka1mi1	ka4mi1	pa1mi1	lei4fa1
pa4mi1	ta4mi1	li4fa1	lu4fa1	li1fa1	li4fa1	ta1mi1	li4fa1
pa4mi1	ka4mi1	li4fa1	lei4fa1	lu1fa1	lu4fa1	ta1mi1	lu4fa1
ta4mi1	ka4mi1	lu4fa1	lei4fa1	lei1fa1	lei4fa1	ta1mi1	lei4fa1
						ka1mi1	li4fa1
						ka1mi1	lu4fa1
						ka1mi1	lei4fa1
						pa4mi1	li1fa1
						pa4mi1	lu1fa1
						pa4mi1	lei1fa1
						ta4mi1	li1fa1
						ta4mi1	lu1fa1
						ta4mi1	lei1fa1
						ka4mi1	li1fa1
						ka4mi1	lu1fa1
						ka4mi1	lei1fa1

Table S3.2 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 S3.2 (A)

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	-0.101	0.142	-0.710	.478
block	0.093	0.044	2.142	.032 *
langgroupEnglish	-0.080	0.137	-0.589	.556
MPtypeN	0.374	0.162	2.308	.021 *
MPtypeT	0.193	0.171	1.128	.259
MPtypeV	0.308	0.201	1.531	.126
block:langgroupMandarin:MPtypeC	0.153	0.064	2.374	.018 *
block:langgroupEnglish:MPtypeC	0.018	0.047	0.379	.705
block:langgroupMandarin:MPtypeN	0.244	0.069	3.555	<.001 ***
block:langgroupEnglish:MPtypeN	0.071	0.047	1.511	.131
block:langgroupMandarin:MPtypeT	0.018	0.059	0.307	.759
block:langgroupEnglish:MPtypeT	-0.086	0.045	-1.895	.058
block:langgroupMandarin:MPtypeV	0.112	0.056	2.023	.043 *

Number of observations: 8038, Participants: 56, Item, 12. AIC = 10023.5, BIC = 10366.1, log-likelihood = -4962.7.

Table S3.2 (B)

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	0.208	0.170	1.224	.221
block	0.006	0.039	0.142	.887
langgroupEnglish	-0.078	0.137	-0.569	.570

MPtypeC	-0.313	0.200	-1.563	.118
MPtypeN	0.060	0.186	0.324	.746
MPtypeT	-0.115	0.195	-0.590	.555
block: langgroupMandarin:MPtypeV	0.199	0.060	3.343	<.001 ***
block:langgroupEnglish:MPtypeV	0.086	0.045	1.902	.057
block: langgroupMandarin:MPtypeC	0.241	0.060	4.056	<.001 ***
block:langgroupEnglish:MPtypeC	0.106	0.045	2.376	.018 *
block: langgroupMandarin:MPtypeN	0.333	0.062	5.337	<.001 ***
block:langgroupEnglish:MPtypeN	0.159	0.047	3.356	<.001 ***
block: langgroupMandarin:MPtypeT	0.105	0.044	2.395	.017 *

Number of observations: 8038, Participants: 56, Item, 12. AIC = 10023.6, BIC = 10366.2, log-likelihood = -4962.8.

Table S3.2 (C)

Fixed Effects	Estimate	SD Error	z	*p*
(Intercept)	0.097	0.129	0.750	.454
block	0.160	0.045	3.530	<.001 ***
langgroupEnglish	-0.085	0.138	-0.616	.538
MPtypeV	0.113	0.195	0.580	.562
MPtypeC	-0.190	0.171	-1.112	.266
MPtypeN	0.179	0.180	0.992	.321
block: langgroupMandarin:MPtypeT	-0.051	0.060	-0.854	.393
block:langgroupEnglish:MPtypeT	-0.154	0.047	-3.238	.001 **
block: langgroupMandarin:MPtypeV	0.040	0.067	0.595	.552
block:langgroupEnglish:MPtypeV	-0.070	0.047	-1.499	.134

block: langgroupMandarin:MPtypeC	0.080	0.067	1.192	.233
block:langgroupEnglish:MPtypeC	-0.052	0.046	-1.137	.255
block: langgroupMandarin:MPtypeN	0.169	0.059	2.871	.004 **

Number of observations: 8038, Participants: 56, Item, 12. AIC = 10023.7, BIC = 10366.3, log-likelihood = -4962.8.

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 S3.3 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 S3.3 (A)

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	0.107	0.161	0.660	.510
block	0.348	0.060	5.790	<.001 ***
MPtypeV	0.072	0.234	0.310	.757
MPtypeT	0.138	0.241	0.573	.566
MPtypeC	-0.470	0.231	-2.034	.042 *
block:MPtypeV	-0.156	0.063	-2.493	.013 *
block:MPtypeT	-0.276	0.061	-4.524	<.001 ***
block:MPtypeC	-0.063	0.062	-1.009	.313

Number of observations: 4013, Participants: 28, Item, 12. AIC = 4828.1, BIC = 5067.4, log-likelihood = -2376.1.

Table S3.3 (B)

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	-0.362	0.171	-2.110	.035 *
block	0.284	0.056	5.028	<.001 ***
MPtypeN	0.469	0.231	2.034	.041 *
MPtypeV	0.546	0.240	2.277	.023 *
MPtypeT	0.605	0.234	2.590	.010 **
block:MPtypeN	0.065	0.062	1.050	.294
block:MPtypeV	-0.093	0.060	-1.560	.119

block:MPtypeT	-0.211	0.057	-3.680	<.001 ***

Number of observations: 4013, Participants: 28, Item, 12. AIC = 4826.5, BIC = 5065.8, log-likelihood = -2375.2.

Table S3.3 (C)

Fixed Effects	Estimate	SD Error	t value	*p*
(Intercept)	0.183	0.173	1.058	.290
block	0.191	0.057	3.351	<.001 ***
MPtypeC	-0.545	0.239	-2.281	.023 *
MPtypeN	-0.075	0.235	-0.321	.748
MPtypeT	0.062	0.257	0.241	.810
block:MPtypeC	0.093	0.060	1.557	.120
block:MPtypeN	0.158	0.063	2.520	.012 *
block:MPtypeT	-0.118	0.058	-2.042	.041 *

Number of observations: 4013, Participants: 28, Item, 12. AIC = 4826.4, BIC = 5065.7, log-likelihood = -2375.2.

Table S3.3 (D)

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	0.243	0.176	1.380	.168
block	0.073	0.054	1.344	.179
MPtypeC	-0.604	0.230	-2.624	.009 **
MPtypeN	-0.133	0.243	-0.549	.583
MPtypeV	-0.057	0.254	-0.227	.821
block:MPtypeC	0.211	0.057	3.685	<.001 ***

block:MPtypeN	0.276	0.061	4.520	<.001 ***
block:MPtypeV	0.117	0.058	2.030	.042 *

Number of observations: 4013, Participants: 28, Item, 12. AIC = 4827.1, BIC = 5066.4, log-likelihood = -2375.5.

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

Table S3.4 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 S3.4 (A)

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	0.555	0.151	3.667	<.001 ***
block	0.100	0.027	3.715	<.001 ***
MPtypeT	-0.859	0.198	-4.333	<.001 ***
MPtypeV	-0.431	0.182	-2.366	.018 *
MPtypeC	-0.590	0.134	-4.406	<.001 ***

Number of observations: 4025, Participants: 28, Item, 12. AIC = 5217.1, BIC = 5437.6, log-likelihood = -2573.6.

Table S3.4 (B)

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	-0.028	0.129	-0.219	.827
block	0.099	0.026	3.845	<.001 ***
MPtypeN	0.577	0.127	4.555	<.001 ***
MPtypeT	-0.281	0.176	-1.596	.111
MPtypeV	0.158	0.196	0.807	.420

Number of observations: 4025, Participants: 28, Item, 12. AIC = 5217.5, BIC = 5438.0, log-likelihood = -2573.8.

Table S3.4 (C)

Fixed Effects	Estimate	SD Error	t value	*p*

(Intercept)	0.135	0.186	0.728	.467
block	0.099	0.027	3.711	<.001 ***
MPtypeC	-0.172	0.201	-0.857	.391
MPtypeN	0.406	0.169	2.404	016 *
MPtypeT	-0.440	0.195	-2.260	.024 *

Number of observations: 4025, Participants: 28, Item, 12. AIC = 5216.0, BIC = 5436.5, log-likelihood = -2573.0.

Table S3.4 (D)

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	-0.304	0.140	-2.171	.030 *
block	0.099	0.027	3.708	<.001 ***
MPtypeC	0.267	0.185	1.443	.149
MPtypeN	0.846	0.196	4.316	<.001 ***
MPtypeV	0.441	0.194	2.269	.023 *

Number of observations: 4025, Participants: 28, Item, 12. AIC = 5216.0, BIC = 5436.5, log-likelihood = -2573.0.

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

Table S3.5 *Best fitting model for reaction time in Experiment 1, showing fixed effects*

Fixed Effects	Estimate	SD Error	t value	*p*
(Intercept)	7.493	0.118	63.323	<.001 ***
block	-0.168	0.023	-7.160	<.001 ***
langgroupEnglish	-0.456	0.161	-2.839	.005 **

Number of observations: 5042, Participants: 56, Item, 12. AIC = 78096.9, BIC = 78357.9, log-likelihood = -39008.5.

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

Table S3.6 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 S3.6 (A)

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	-0.083	0.141	-0.587	.557
block	0.116	0.026	4.445	<.001 ***
MPtypeN	0.626	0.135	4.629	<.001 ***
MPtypeT	-0.223	0.185	-1.205	.228
MPtypeV	0.138	0.194	0.711	.477

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

Table S3.6 (B)

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	0.053	0.173	0.307	.759
block	0.115	0.026	4.422	<.001 ***
MPtypeT	-0.358	0.181	-1.971	.049 *
MPtypeC	-0.138	0.194	-0.711	.477
MPtypeN	0.489	0.166	2.941	.003 **

Number of observations: 4025, Participants: 28, Item, 12. AIC = 5383.5, BIC = 6171.0, log-likelihood = -2566.8.

Table S3.6 (C)

Fixed Effects	Estimate	SD Error	Z	*p*

(Intercept)	-0.303	0.135	-2.251	.024 *
block	0.115	0.026	4.425	<.001 ***
MPtypeV	0.360	0.182	1.984	.047 *
MPtypeC	0.219	0.185	1.185	.236
MPtypeN	0.847	0.195	4.351	<.001 ***

Number of observations: 4025, Participants: 28, Item, 12. AIC = 5383.5, BIC = 6171.0, log-likelihood = -2566.8.

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 S3.7 Best fitting model for accuracy in Block 6 for the L1 English group in Experiment 1, testing awareness effect

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	1.618	0.322	5.020	<.001 ***
MPtypeC	-1.264	0.325	-3.885	<.001 ***
MPtypeT	-1.516	0.353	-4.295	<.001 ***
MPtypeV	-0.695	0.300	-2.321	.020 *

Number of observations: 671, Participants: 28, Item, 12. AIC = 873.2, BIC = 1003.9, log-likelihood = -407.6.

R syntax: glmer(acc ~ MPtype + (1 + awareness + MPtype | item) + (1 + MPtype | subjectID), family = binomial, data = awarenessblock6, glmerControl(optCtrl=list(maxfun=2e5), optimizer = "nloptwrap", calc.derivs = FALSE))

Table S3.8 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 S3.8 (A)

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	0.103	0.133	0.775	.438
block	0.112	0.035	3.164	.002 **
exposureshort	-0.161	0.112	-1.443	.149
MPtypeN	0.527	0.151	3.491	<.001 ***
MPtypeT	-0.155	0.165	-0.937	.349
MPtypeV	0.168	0.197	0.851	.395
block:exposurelong:MPtypeC	0.015	0.040	0.381	.703
block:exposureshort:MPtypeC	-0.008	0.044	-0.188	.851
block:exposurelong:MPtypeN	0.023	0.041	0.545	.586
block:exposureshort:MPtypeN	0.025	0.043	0.596	.551
block:exposurelong:MPtypeT	-0.095	0.037	-2.526	.012*
block:exposureshort:MPtypeT	-0.048	0.041	-1.170	.242
block:exposurelong:MPtypeV	-0.011	0.034	-0.325	.745

Number of observations: 11793, Participants: 55, Item, 12. AIC = 14092.9, BIC = 14454.3, log-likelihood = -6997.4.

Table S3.8 (B)

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	0.276	0.179	1.541	.123
block	0.063	0.033	1.908	.056
exposureshort	-0.163	0.111	-1.469	.142

MPtypeC	-0.172	0.201	-0.856	.392
MPtypeN	0.352	0.174	2.027	.043 *
MPtypeT	-0.325	0.185	-1.760	.078
block:exposurelong:MPtypeV	0.038	0.035	1.075	.282
block:exposureshort:MPtypeV	0.048	0.041	1.158	.247
block:exposurelong:MPtypeC	0.064	0.036	1.806	.071
block:exposureshort:MPtypeC	0.040	0.041	0.973	.330
block:exposurelong:MPtypeN	0.072	0.036	2.006	.045 *
block:exposureshort:MPtypeN	0.076	0.047	1.625	.104
block:exposurelong:MPtypeT	-0.046	0.028	-1.615	.106

Number of observations: 11793, Participants: 55, Item, 12. AIC = 14092.3, BIC = 14453.7, log-likelihood = -6997.1.

Table S3.8 (C)

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	-0.048	0.147	-0.325	.745
block	0.137	0.040	3.434	<.001 ***
exposureshort	-0.163	0.111	-1.458	.145
MPtypeV	0.325	0.184	1.763	.078
MPtypeC	0.155	0.167	0.930	.353
MPtypeN	0.680	0.187	3.640	<.001 ***
block:exposurelong:MPtypeT	-0.120	0.043	-2.777	.005 **
block:exposureshort:MPtypeT	-0.074	0.047	-1.582	.114
block:exposurelong:MPtypeV	-0.036	0.046	-0.787	.431
block:exposureshort:MPtypeV	-0.026	0.043	-0.598	.550

block:exposurelong:MPtypeC	-0.011	0.046	-0.228	.819
block:exposureshort:MPtypeC	-0.035	0.042	-0.838	.402
block:exposurelong:MPtypeN	-0.002	0.042	-0.058	.954

Number of observations: 11793, Participants: 55, Item, 12. AIC = 14091.6, BIC = 14453.0, log-likelihood = -6996.8.

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))

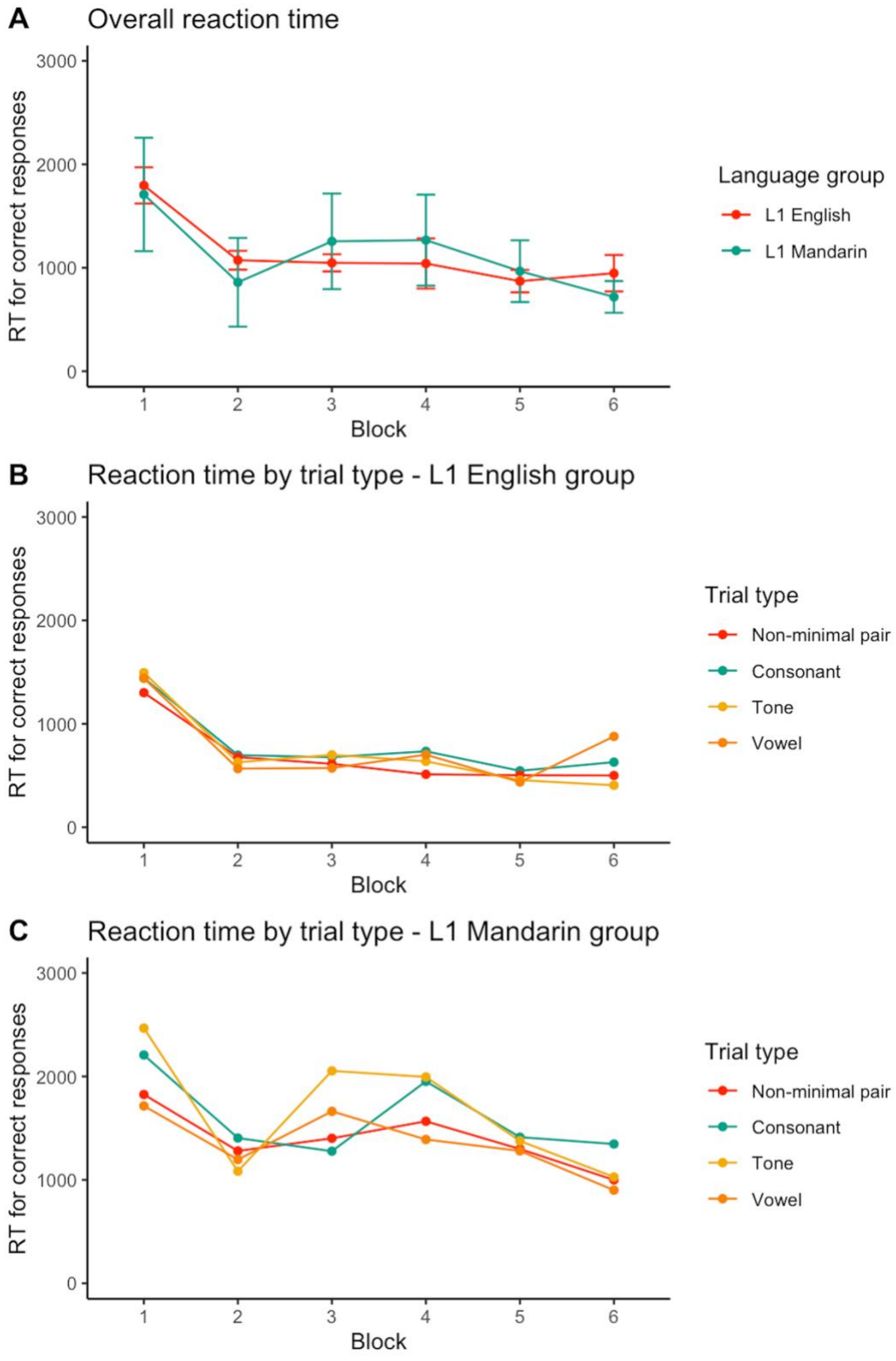
Table S3.9 Best fitting model for reaction time in Experiment 2, showing fixed effects

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	6.912	0.121	57.011	<.001 ***
block	-0.164	0.028	-5.878	<.001 ***
exposureshort	0.159	0.169	0.943	.346
MPtypeC	-0.210	0.095	2.204	.028 *
MPtypeT	0.076	0.098	0.776	.438
MPtypeV	-0.095	0.099	-0.964	.335
block:exposurelong:MPtypeN	0.087	0.033	2.606	.009 **
block:exposureshort:MPtypeN	-0.029	0.028	-1.039	.299
block:exposurelong:MPtypeC	0.072	0.034	2.154	.031 *
block:exposureshort:MPtypeC	-0.056	0.030	-1.858	.063
block:exposurelong:MPtypeT	0.083	0.034	2.408	.016 *
block:exposureshort:MPtypeT	-0.050	0.029	-1.704	.088
block:exposurelong:MPtypeV	0.094	0.031	3.066	.002 **

Number of observations: 7513, Participants: 55, Item, 12. AIC = 111970.3, BIC = 112316.5, log-likelihood = -55935.1.

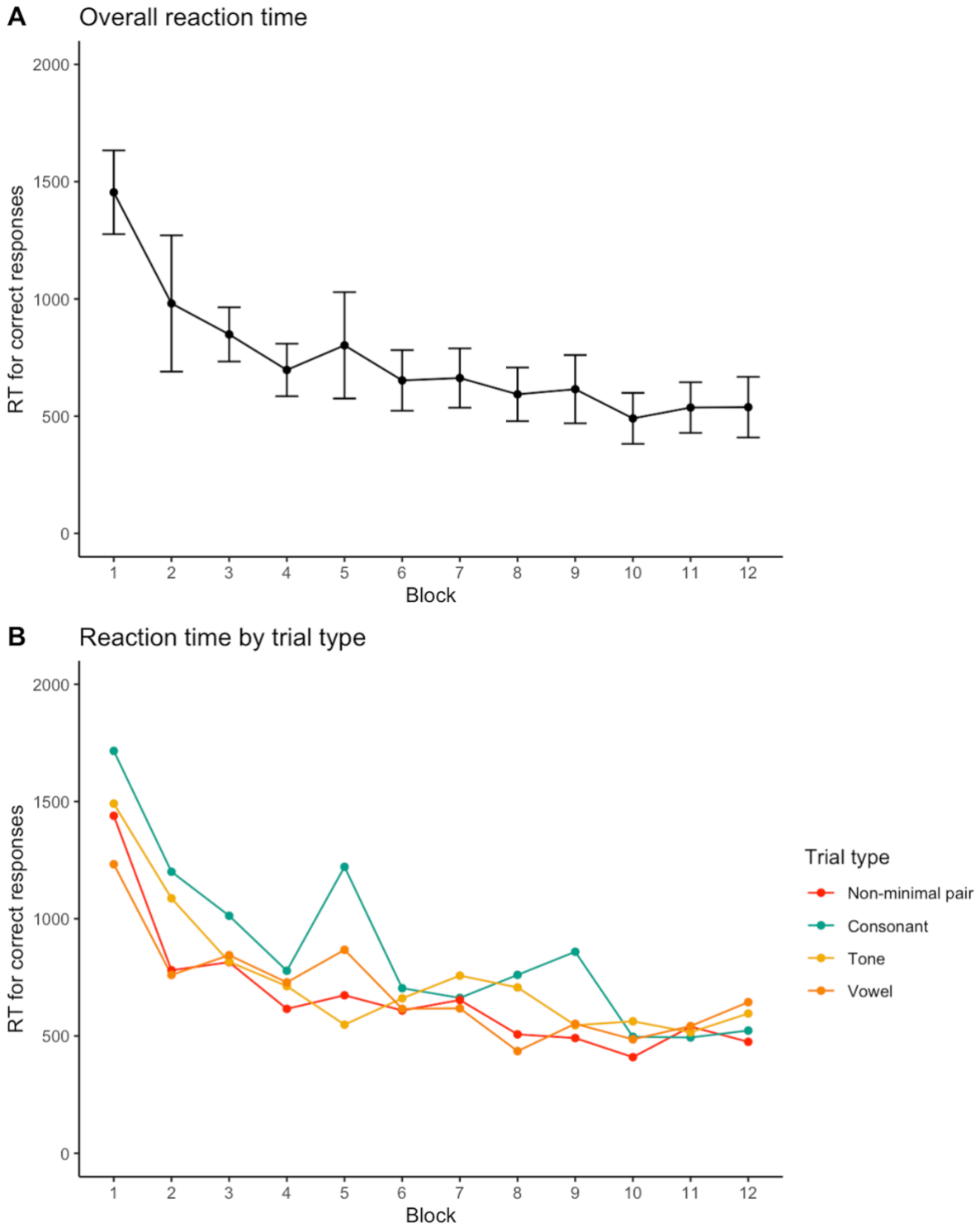
R syntax: glmer(RT ~ block + exposure + MPtype + exposure:MPtype:block + (1 + block + exposure + MPtype | item) + (1 + block + MPtype | subjectID), family = Gamma (link = "log"), data = fulld.correct, glmerControl(optCtrl=list(maxfun=2e5), optimizer = "nloptwrap", calc.derivs = FALSE))

Figure S3.1 Experiment 1: Mean reaction time for correct responses in each learning block – overall (A) and in different trial types (B & C).



Note. Error bars represent 95% Confidence Intervals.

Figure S3.2 Experiment 2: Mean reaction time for correct responses in each learning block - 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).

4. Published paper 2: Constraints on novel word learning in heritage speakers

Page number: 116-150

Abstract

Recent research on word learning has found that adults can rapidly learn novel words by tracking cross-situational statistics, but learning is greatly influenced by the phonological properties of the words and by the native language of the speakers. Mandarin-native speakers could easily pick up novel words with Mandarin tones after a short exposure, but English-native speakers had specific difficulty with the tonal components. It is, however, unclear how much experience with Mandarin is needed to successfully use the tonal cue in word learning. In this study, we explored this question by focusing on the heritage language population, who typically are exposed to the target language at an early age but then develop and switch to another majority language. Specifically, we investigated whether heritage Mandarin speakers residing in an English-speaking region and speaking English as a dominant language would be able to learn novel Mandarin tonal words from statistical tracking. It helps us understand whether early exposure to the target feature is sufficient to promote the use of that feature in word learning later in life. We trained 30 heritage Mandarin speakers with Mandarin pseudowords via a cross-situational statistical word learning task (CSWL). Heritage Mandarin speakers were able to learn the pseudowords across multiple situations, but similar-sounding words (i.e., minimal pairs) were more difficult to identify, and words that contrast only in lexical tones (i.e., Mandarin lexical tone) were distinguished at chance level throughout learning. We also collected information about the participants' heritage language (HL) experience and usage. We did not observe a relationship between HL experience/usage and performance in tonal word learning, suggesting that HL exposure does not necessarily lead to an advantage in learning the target language.

Keywords: statistical learning, cross-situational word learning, heritage speaker, heritage language phonology, lexical tone

Introduction

Language learners can rapidly pick up new words from the surrounding environment, most of the time without explicit instruction. This is impressive given the highly variable environment in which language learning happens. Quine (1960) illustrated this word learning challenge by referring to the well-known "Gavagai" conundrum. The first time a learner encounters a new word, the meaning is usually unclear because the word could refer to anything in the environment. Without any explicit information, the word-referent mapping is ambiguous. How do learners deal with this referential ambiguity problem in real life?

Research on statistical learning has found a potential solution to the Gavagai problem: child and adult learners can keep track of the linguistic information across multiple situations to aid word learning, an ability commonly referred to as *cross-situational word learning* (CSWL; e.g., Escudero et al., 2022; Monaghan et al., 2019; Rebuschat et al., 2021; Suanda and Namy, 2012). That is, when the same word occurs again, learners can track the always-co-occurring referent and, over time, form an association between the word and the referent. However, recent studies have shown that CSWL is greatly influenced by the phonological properties of the words (Escudero et al., 2016; Ge et al., in press; Tuninetti, Mulak and Escudero, 2020). Words that sound similar (e.g., phonological minimal pairs like *bag* vs. *beg* in English; *pāo* vs. *gāo* in Mandarin) generated difficulty in CSWL (e.g., Escudero et al., 2016), as well as the presence of non-native phonological features when adults learn an additional language (L2) via CSWL (e.g., Escudero et al., 2022; Ge et al., in press; Ge et al., under review). For example, L1 Mandarin speakers could learn Mandarin pseudowords from CSWL exposure regardless of the existence of tonal minimal pairs, but L1 English speakers had great difficulty with these non-native minimal pairs (Ge et al., in press). This is because Mandarin-native speakers had extensive experience with the Mandarin tonal feature since childhood and could make use of the tonal categories in identifying words, but English-native

speakers had no experience with tones and did not have the tonal representations. One question that arises is how much experience with the target feature would then be needed to develop the phonological representations and consequently use the feature in word learning.

To address this question, we targeted the heritage speaker population who are typically exposed to a minority (heritage) language at home in childhood, but start to rapidly acquire a different societal/majority language at the onset of school and become dominant in the societal/majority language. Specifically, we tested heritage speakers of Mandarin who were born to at least one Mandarin-speaking parent and resided in English-speaking countries from birth. These participants had early experience with the (Mandarin) tonal feature but then, later in life, had relatively limited use of lexical tones given that their majority language (English) is non-tonal. The performance of heritage speakers is particularly interesting because human sensitivity to sounds is largely shaped and tuned to their native languages at an early age, and hence experience with the target feature in early years might make a great difference even when exposure to the feature reduces later in life (Hartshorne et al., 2018; Kuhl, 2004). To summarize, in this study, we examined whether and how heritage speakers learn novel words from their heritage language (HL) via statistical tracking, and how they are affected by sounds that only exist in their HL but not in the majority language (i.e., lexical tones). Additionally, we tested whether the degree of HL experience and usage has an impact on word learning outcomes.

Statistical word learning

Language learners can extract statistical regularities of different aspects of the language from the linguistic input (e.g., Maye & Gerken, 2000 and Maye et al., 2002 for sound discrimination; Saffran et al., 1996 for word segmentation; see Isbilen and Christiansen, 2022; Siegelman, 2020; Williams and Rebuschat, 2022, for reviews). As for

word learning, this involves tracking word-referent co-occurrences across encounters. A cross-situational statistical learning paradigm has often been used to examine word learning under implicit learning conditions where there is ambiguity in words' referents (e.g., Escudero et al., 2022; Rebuschat et al., 2021; Smith and Yu, 2008; Suanda et al., 2014; Yu and Smith, 2007). For example, in Yu and Smith's (2007) seminal study, adult learners were first presented with multiple words and pictures in each learning trial, and then tested whether they could make use of the word-picture co-occurrence information across learning events to acquire the appropriate mappings. After only six minutes of exposure, learners could match pictures to words at above-chance level even in highly ambiguous conditions where four words and four pictures were presented in each learning trial.

However, this rapid learning effect has been found to reduce when there are phonological overlaps between words, which can be found in most vocabulary inventories (e.g., Escudero et al., 2016, 2022; Tuninetti et al., 2020). For example, when being presented with two pictures and two minimal pair words in each learning trial, Escudero et al. (2016) reported that learners' performance was inhibited – especially when the words were vowel minimal pairs (e.g., /dit/-/dɪt/) – compared to non-minimal pair presentations (e.g., /bɔn/-/dit/). This phonological similarity effect was even more profound when it came to L2 word learning. When the same CSWL task with English pseudo-minimal pairs (e.g., /dit/-/dɪt/, /bɔn/-/tɔn/) was presented to English-native and Mandarin-native speakers, it was observed that English-native speakers' overall word learning performance was better than the Mandarin-native speakers in different minimal pair types (Escudero et al., 2022). Thus, the existence of non-native English contrasts influenced Mandarin-native speakers' word learning outcomes. Similar evidence came from Australian English speakers learning Dutch and Brazilian Portuguese pseudo-minimal pairs (Tuninetti et al., 2020). Vowel minimal pairs were created based on Dutch and Brazilian Portuguese vowel inventories (e.g., /piχ/-/pyχ/,

/fɛfe/-/fefe/, respectively). As predicted, based on the Second Language Linguistic Perception model (L2LP - Escudero, 2005) and the Perceptual Assimilation-L2 model (PAM-L2 - Best and Tyler, 2007), some of the vowel pairs were defined as perceptually easier as they could be mapped to two separate Australian English vowel categories (e.g., Dutch /i/-/ɑ/ contrast might be mapped to AusEnglish /i/-/ɔ/), and some other vowel pairs were classified as perceptually difficulty as they had no clear corresponding Australian English contrasts (e.g., Dutch /i/-/y/ contrast). Learners performed better with perceptually easy pairs compared to the difficult pairs, indicating that the degree of perceptual cross-linguistic similarity associated with non-native segments influenced non-native statistical word learning.

Ge et al. (in press) found that the non-native phonology effect in CSWL was not only associated with segmental but also suprasegmental features. In addition to the segmental minimal pairs as in previous research (e.g., Escudero et al., 2022), Ge et al. (in press) involved tonal minimal pairs (i.e., two words that differ only in lexical tone: /pa1mi1/ vs /pa4mi1/ with numbers referring to Mandarin Tone 1 and Tone 4), which is a suprasegmental feature absent in non-tonal languages like English. A slightly different CSWL design is used to more closely resemble the minimal pairs learners encounter in the real world. Only one word was presented in each trial together with multiple referents, hence, minimal pairs were not presented side by side to participants in a single trial. This mirrors natural language learning situations in that minimal pairs tend not to occur in immediate proximity but need to be acquired by tracking the contrastive phonological features across situations. Through a short cross-situational exposure of ten minutes, participants who were English-native speakers successfully identified word-referent mappings in consonantal, vocalic and non-minimal pairs, as the segmental features in the stimuli were designed to be familiar to English speakers, but not in the tonal pairs. Participants who were Mandarin-native speakers, on the other hand, were able to identify words in the tonal pairs after the same amount of exposure.

These previous findings all suggest a significant role of phonology in statistical word learning and that L2 learners might encounter difficulty in picking up words from the environment because of the non-native sounds.

Such difficulty has been found even when specific phonetic (perceptual) training on the target non-native contrasts is included (Ge et al., under review). For example, in Ge et al., (under review), native speakers of English were provided with perceptual training on Portuguese consonant and vowel contrasts (e.g., /l/-/ʎ/, /n/-/ɲ/, /e/-/ɛ/, /o/-/ɔ/), and then trained on Portuguese pseudowords containing these contrasts via CSWL. The perceptual training did improve learners' perceptual discrimination of the non-native contrasts, but this improvement did not transfer to word learning – the English-native speakers still had difficulty with non-native minimal pairs in word learning. This finding indicates that L2 learners' difficulty comes from not simply perceptual issues, but also the lack of phonological representation of the novel sounds. As widely reported in infant speech development literature, during as early as the first year of life, humans start to tune in to their native sound system(s) and their sensitivity to non-native sounds and categories greatly reduces (e.g., Kuhl, 2004; Watson et al., 2014; Werker and Tees, 1984). This perceptual tuning persists into adulthood and might contribute to the difficulties in L2 word learning. Previous studies observed a phonetic-phonological-lexical continuity, indicating that categorical perception of non-native sounds was associated with performance in non-native word learning and processing (e.g., Laméris et al., 2023; Ling & Grüter, 2022; Wong & Perrachione, 2007). Hence, if the narrowing process in early years does play a significant role, one question that follows is whether exposure to the target language in early years would facilitate word learning (in the same language) later in life, as early exposure might allow learners to develop the necessary perceptual sensitivities and phonological categories.

A particular population that is perfect to study this research question is heritage speakers because of their special language profile. Like all native speakers of a language, heritage speakers have early exposure to the language, which would allow them to develop sensitivities to the language-specific phonological contrasts, but they switch to another dominant language after the early years and usually have limited HL use afterwards. It thus allows us to specifically test whether early exposure to the target language plays a role in later word learning. In other words, we explored whether heritage speakers' phonological representations that are developed early in life remain accessible and help them learn new words from their HL in adulthood.

Phonological advantages in heritage speakers

HL research has observed phonological advantages among heritage speakers in both speech perception and production compared to late L2 learners, and closer performance to native speakers in some dimensions (e.g., Chang, 2016; Lukyanchenko and Gor, 2011, for speech perception; Au et al., 2002; Chang et al., 2011, for speech production; Flores et al., 2017, for accentedness). For example, heritage Korean speakers who grew up in an English-speaking environment showed greater sensitivity to unreleased stops as it is an obligatory feature in Korean (Chang, 2016). Although unreleased final stops are present in American English, it is not considered the canonical form and English speakers rely more on released stops in word recognition. It was found that heritage Korean speakers' identification of the unreleased stops (in Korean and English) was comparable to L1 Korean speakers and was better than L1 English speakers. This suggests that early exposure to the phonological contrasts did persist into adulthood and facilitate sound recognition later in life. As for speech production, for instance, Chang et al. (2010) reported that compared to L2 Mandarin learners, heritage Mandarin speakers' back vowel production (e.g., Mandarin /u/) was closer to native

Mandarin speakers (though not the same). In addition to the segmental features, some research also found an advantageous performance in heritage speakers' suprasegmental realizations (e.g., Chang and Yao, 2016, 2019; Yang, 2015 for lexical tone; Kim, 2020 for lexical stress). Regarding lexical tone, for example, Yang (2015) examined the perception and production of Mandarin tones by native Mandarin speakers, heritage Mandarin speakers, and L2 learners. Heritage speakers' perception of tones lay in between the native and the L2 groups: heritage speakers exhibited a more stable categorical perception of the four tones than L2 learners, although they do not completely resemble native Mandarin speakers' perceptual patterns. Work on Mandarin speech production showed that heritage Mandarin speakers' production of tones also fell in the intermediate state between native and L2 speakers in general (Chang and Yao, 2016). In some dimensions, heritage speakers' tonal production resembles more native speakers (e.g., T3 low falling-rising tone turning point), whereas in some other dimensions, heritage speakers' production was in between the native and L2 groups (e.g., tone shortening in multisyllabic contexts). Overall, although heritage speakers do not pattern exactly the same as native speakers, much research evidence has shown that they are at least closer to native speakers in terms of speech perception and production than L2 learners are.

However, it is not clear if heritage speakers can make use of such phonological advantages at the lexical level to assist novel word learning in the HL. As discussed in the previous section, phonologically similar words pose difficulties for L2 learners when they lack the appropriate phonological representations. Here, we hypothesize that heritage speakers' advantages in speech perception and recognition would further facilitate their acquisition of phonologically overlapping words in the target language. In this study, we focus on a suprasegmental feature that has been found to be difficult for late L2 learners in word learning – lexical tones (Ge et al., in press). L2 learners of Mandarin were found to fail

in learning tonal minimal pair words from implicit exposure, whereas L1 Mandarin speakers could pick up novel tonal minimal pairs rapidly in the same situation. Our prediction is that heritage Mandarin speakers would be able to learn tonal minimal pairs to some extent because of their better categorical tonal perception, but whether they could match native speakers' performance largely depends on their individual HL experience.

Research questions and predictions

In the current study, we investigate the cross-situational learning of Mandarin pseudowords by adult heritage speakers of Mandarin who were born and reside in English-speaking countries. The following research questions are addressed:

RQ1: Do minimal pairs and phonological contrasts that do not exist in heritage speakers' majority language (i.e., the tonal contrasts) pose difficulty during cross-situational learning?

RQ2: Does the degree of heritage language experience and usage influence learning outcomes?

For RQ1, based on previous literature, we predicted that minimal pairs would be more difficult to learn compared to non-minimal pairs, and minimal pairs with phonological contrasts that do not exist in heritage speakers' dominant language would generate the greatest difficulty in learning (Escudero et al., 2022; Ge et al., in press). Specifically, we predicted that minimal pairs that contrast in lexical tones would be the most difficult (i.e., with the lowest accuracy), followed by minimal pairs that differ in consonants and vowels. The non-minimal pairs would be relatively easy to learn. However, we expected the heritage Mandarin speakers to show some degree of learning of the tonal minimal pairs.

For RQ2, we predicted that greater experience and usage of HL would be associated with better learning of the tonal minimal pairs, as participants with greater Mandarin

experience and usage would have more exposure to the tonal contrasts and might be more sensitive to the tonal minimal pairs.

Methods

Participants

Thirty bilingual speakers of Mandarin Chinese and English participated in this study. The sample size was inferred from Ge et al. (in press)[8], where the same stimuli and CSWL task were used and a significant learning effect was observed. Participants were recruited through email advertisements within university communities in Toronto, Canada, and through Prolific (www.prolific.com). Participants had to be at least 18 years old, bilingual speakers of English and Mandarin Chinese, and born in an English-speaking country (Canada or USA). An additional prerequisite was that participants needed to have at least one parent who was a native speaker of Mandarin Chinese. One participant was excluded because they were born in Hong Kong and only moved to an English-speaking country at the age of four. Thus, twenty-nine participants were included in the data analysis (11 F, 17M 1 preferred not to say). The mean age was 29.97 (SD = 8.60, ranging from 18 to 62 years). Regarding language background, 14 participants reported knowing additional languages/varieties other than Mandarin or English (e.g., Cantonese[9], French, Italian, Shanghainese, and Spanish). Nine participants reported having one Mandarin-native parent, and 20 participants with two Mandarin-native parents. Further details on participants' HL experience and use can be found in the results section.

[8] The power analysis of Ge et al.'s (in press) study with the same CSWL task is available at: https://osf.io/2j6pe/.

[9] Among these additional languages, Cantonese and Shanghainese are tonal. Thus, we carried out an analysis to test whether the eight participants who spoke additional tonal languages performed differently from the others who did not know other tonal languages. However, adding *additional tonal experience* as a fixed effect in our model on CSWL accuracy did not significantly improve model fit ($\chi^2(1) = 0$, $p = 1$), nor did the 3-way interaction between block, additional tonal experience and trial type ($\chi^2(7) = 11.177$, $p = .131$). Thus, for the main analyses, we will not include additional tonal experience as a factor.

Materials

Heritage Language Experience Questionnaire

We collected information about participants' HL (i.e., Mandarin) experience using Tomić, Rodina, Bayram and De Cat's (2023) Heritage Language Experience Questionnaire (HeLEx). The questionnaire was designed to capture the quantity and quality of HL exposure and use in different social contexts (e.g., family, external family (i.e., family outside the household), work, community, leisure). It also asked for participants' background information (e.g., gender, age, history of language learning, parents' language) and educational information (e.g., language used at different levels of schooling). Additionally, there were questions regarding participants' language attitudes and code-switching attitudes and behaviours, though we did not include these attitude-related questions in the analyses because language attitude is not the focus of the current study.

For the HeLEx data, we followed Tomić et al.'s (2023) instructions and derived a set of HL experience and usage measures, including HL experience (i.e., frequency of use) and proficiency[10] in four different modalities (reading, writing, speaking, listening), proportion of HL use in different social contexts (family, external family, work, community, leisure), language dominance, language entropy[11], proportion of HL use when accounting for actual time spent in each context (i.e., weighted HL use), and diversity of HL interlocutors (i.e., proportion of HL proficient and/or dominant interlocutors).

[10] HL experience was calculated from questions on frequency of HL use, for example, *how often do you speak it.* HL proficiency was based on questions such as *how well do you speak it.*

[11] Language entropy measures the level of language diversity in a particular context (e.g., family, external family, work, community, leisure) (Gullifer & Titone, 2020; Tomić et al., 2023). Higher language entropy in a given context means higher diversity in language use.

Cross-situational word learning task

The CSWL task involved 12 pseudowords and 12 referent pictures. All pseudowords were disyllabic, with CVCV structures, which satisfies the phonotactic constraints of both Mandarin Chinese and English. The pseudowords contained phonemes that were similar between the two languages. The choice of the phonotactics and phonemes ensured that the target feature, lexical tone, was the only feature that exist in participants' heritage language but not in the majority language. Each syllable in the pseudowords carried a lexical tone which was either Tone 1 (high-level) or Tone 4 (high-falling) in Mandarin Chinese, thus creating a simplified lexical tone system.

Six consonants /p, t, k, l, m, f/ and four vowels /a, i, u, ei/ were combined to form eight distinct base syllables (/pa, ta, ka, li, lu, lei, mi, fa/), which were 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/) and the other three differed in the vowel of the first syllable (/lifa, lufa, leifa/). 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 created additional tonal minimal pair contrasts (e.g., /pa1mi1/ vs /pa4mi1/). Therefore, a total of 12 pseudowords were created (full list shown in Table 4.1). 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 had no corresponding meanings in English or Mandarin Chinese. The audio stimuli were produced by a female native speaker of Mandarin Chinese. The mean length of the audio stimuli was 800ms.

Table 4.1 Pseudowords in the consonant 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

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.

The visual and auditory stimuli are available at: https://osf.io/q6354/.

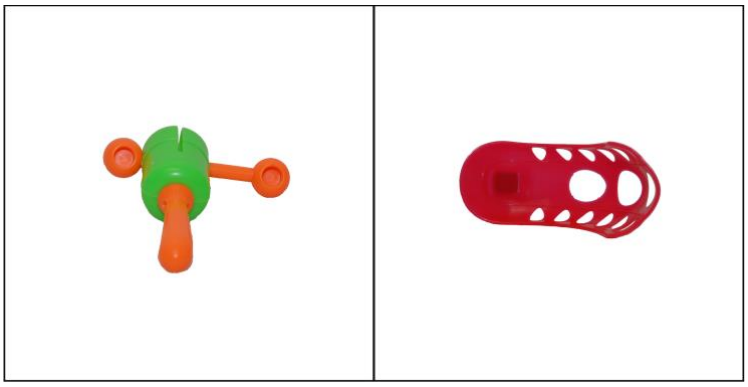
Procedure

All participants were directed to the experiment platform Gorilla (www.gorilla.sc) to complete the task and the questionnaire. After providing informed consent, participants completed the CSWL task, which took approximately 10 minutes. In the CSWL task, participants were told that they would hear one word and see two pictures of referent objects on the screen. Their task was to decide, as quickly and accurately as possible, which object the pseudoword referred to. They were instructed to press 'Q' on the keyboard if they thought the object on the left was the correct referent of the word and 'P' for the object on the right.

In each trial, participants first saw a fixation cross at the centre of the screen for 500ms. They were then presented with 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 enter their response on the keyboard (Q or P). 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 one. No feedback was provided after each response. We recorded the keyboard responses in each trial to calculate accuracy and response times. This allowed us to keep track of participants' performance throughout the CSWL task, and hence there were no separate training and testing phases. Figure 4.1 provides an example of a CSWL trial.

Figure 4.1 Example of cross-situational word learning (CSWL) trial. Participants were presented with two objects and played a single pseudoword. They had to decide if the pseudoword referred to the object on the left or the object on the right.



There were four types of CSWL trials. In non-minimal pair (non-MP) trials, the two objects presented on the screen referred to pseudowords that were phonologically distinct (e.g., /pa1mi1/ and /li4fa1/). In consonantal minimal pair (cMP) trials, the two objects on the screen referred to pseudowords that differed in only one consonant contrast (e.g., /pa1mi1/ and /ta1mi1/). In vocalic minimal pair (vMP) trials, the two objects referred to pseudowords that differed in only one vowel contrast (e.g., /li1fa1/ and /lu1fa1/). And in tonal minimal pair (tMP) trials, the two objects referred to pseudowords that differed only in lexical tone (e.g., /pa1mi1/ and /pa4mi1/). This manipulation allowed us to determine if and how phonological overlap between the pseudowords affected word learning. Each object was paired with different foils according to the trial type. For instance, the 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.

Each participant completed six CSWL blocks, with each pseudoword-object mapping occurring twice per block. There were thus 24 trials per block, and 144 trials in total. The four trial types (non-MP, cMP, vMP, tMP) occurred six times per block. The order of trials within each block was randomized for each participant as was the sequence in which the six blocks occurred. The correct referent picture was presented on the left side in half of the trials and on the right side in the other half of the trials.

After the CSWL task, participants completed the HeLEx questionnaire. When all tasks were completed, participants recruited from Prolific were directed back to the Prolific website and were granted compensation. Participants recruited through emailing received the vouchers via email.

Data analysis

We excluded participants who failed to successfully complete the initial sound check (one participant failed, and 30 participants passed the sound check). We also excluded individual responses that lasted over 30 seconds (11 out of 4176 individual responses were removed, leaving a total of 4165 data points for analysis). This was because they failed to follow the instruction to respond as quickly and accurately as possible. After excluding these data points, we visualized the data using R (R Core Team, 2022) for general descriptive patterns. We then used generalized linear mixed effects modelling for statistical data analysis. Mixed effects models were constructed from the null model (containing only random effects of item and participant) to models containing fixed effects, and the dependent variable was accuracy in the CSWL task. We tested if each of the fixed effects of trial type, block, and their interaction improved model fit using log-likelihood comparisons between models. A

quadratic effect of block was also tested for its contribution to model fit, as learning may have been non-linear over training. Additionally, we tested if adding the derived measures from the HeLEx questionnaire as fixed effect to the mixed-effect models improved model fit.

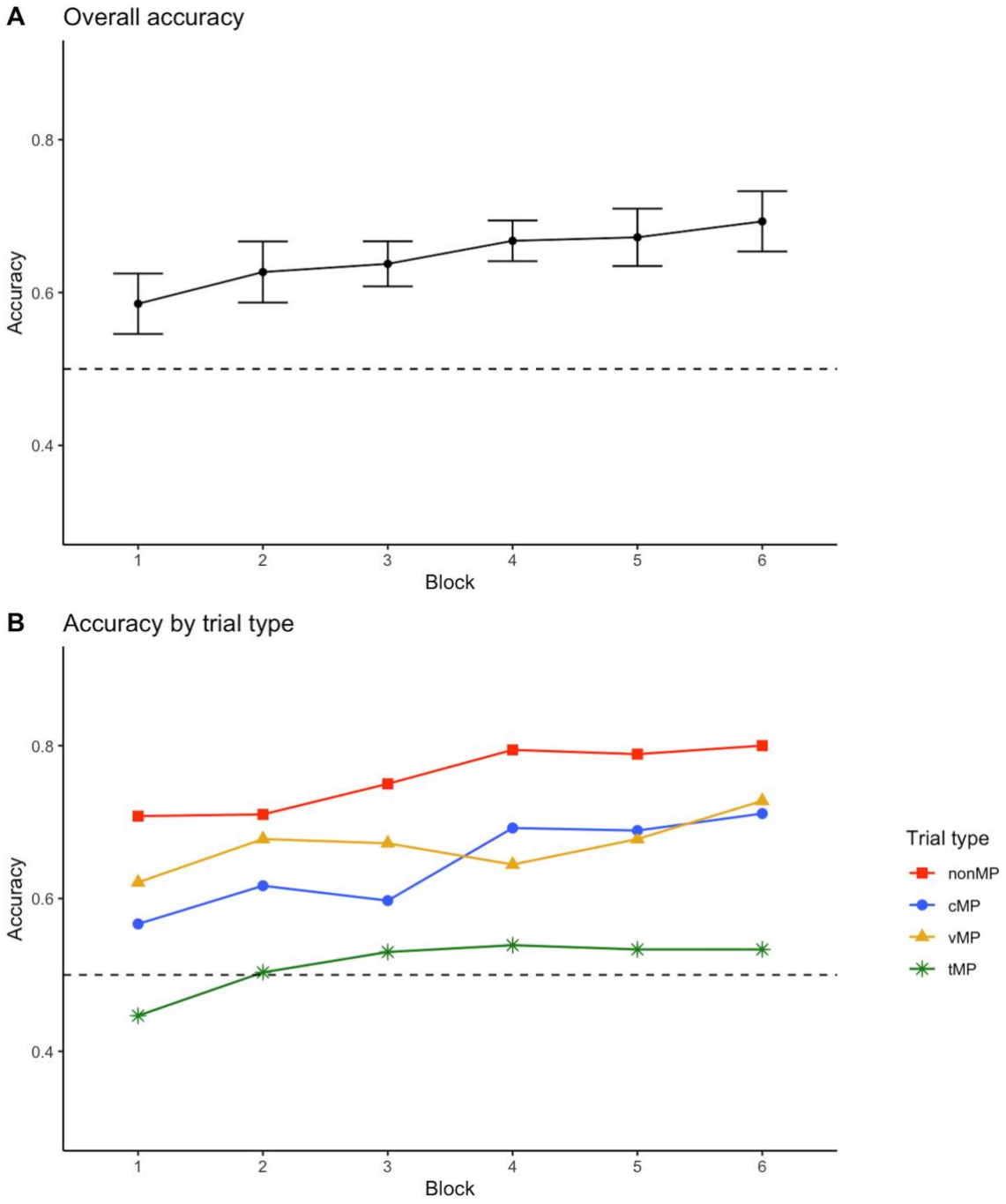
The anonymized data and R scripts are available at: https://osf.io/q6354/.

Results

Performance on the cross-situational word learning task

Figure 4.2A presents the overall proportion of correct responses in the CSWL task. Participants performed significantly above chance from Block 1 (mean accuracy = 0.59, t = 4.61, $p < .001$). For the different minimal pair trials (Figure 4.2B), accuracy was the highest in non-minimal pair trials, followed by consonantal and vocalic minimal pair trials. Performance in the tonal trials was the lowest and remained close to chance level (0.53) until the end of the CSWL task.

Figure 4.2 Mean proportion of correct pictures selected in each learning block - overall (2A) and in different trial types (2B).



Note. The dotted line represents chance level. Error bars represent 95% Confidence Intervals.

We ran generalized linear mixed effects models to examine performance accuracy across learning blocks. Compared to the model with only random effects, adding the fixed effect of learning block did not improve model fit significantly ($\chi^2(1) = 0.944$, $p = .331$). Adding trial type (consonant, vowel, tone, non-minimal pair) improved model fit ($\chi^2(3) = 28.298$, $p < .001$), but the block*trial type interaction ($\chi^2(3) = 4.365$, $p = .225$) did not improve fit further. This indicates that the overall performance differed significantly across trial types, but the learning trajectories (i.e., improvement across blocks) did not differ significantly in different trial types. The quadratic effect for block did not result in a significant difference ($\chi^2(4) = 2.109$, $p = .716$). The best-fitting model is reported in Table 4.2. Note that, whereas block did not contribute to explaining variance significantly when considered as a single fixed effect, it was significant in the model when trial type was also included (as shown in Table 4.2).

Table 4.2 Best fitting model for accuracy in CSWL, showing fixed effects. TrialTypeC refers to consonantal minimal pair trials, TrialTypeT refers to tonal minimal pair trials, TrialTypeV refers to vocalic minimal pair trials, with the reference being non-minimal pair trials.

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	0.934	0.176	5.298	< .001***
block	0.105	0.031	3.399	< .001***
TrialTypeC	-0.604	0.138	-4.383	< .001***
TrialTypeT	-1.217	0.179	-6.803	< .001***
TrialTypeV	-0.454	0.148	-3.078	.002**

Number of observations: 4165, Participants: 29, Item, 12. AIC = 5076.1, BIC = 5297.8, log-likelihood = -2503.1.

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

Heritage Language Experience Questionnaire

We computed a set of measures of HL use derived from the four modalities (reading, writing, speaking, hearing) and five contexts (family, external family, work, community, leisure) of language use. Tables 4.3 and 4.4 summarize the results.

Participants reported higher Mandarin proficiency and use in speaking and hearing compared to reading and writing. As for language dominancy, only one participant reported to be Mandarin-dominant in speaking and another participant being Mandarin-dominant in hearing/understanding. Overall, more participants were dominant in English in all modalities. In terms of the context of language use, participants reported more Mandarin use with families and external families, and relatively little Mandarin use in working conditions.

Table 4.3 Heritage language experience across four modalities.

	Reading	Writing	Speaking	Hearing	Scale
HL experience	3.97 (2.01)	2.97 (2.11)	5.48 (1.33)	5.83 (1.26)	1~7
HL proficiency	2.14 (0.95)	1.93 (0.96)	2.86 (0.64)	3.34 (0.67)	1~4
HL/SL dominance (experience-based)	0.57 (0.29)	0.43 (0.30)	0.79 (0.19)	0.83 (0.17)	1 = balanced Mandarin
HL/SL dominance (proficiency-based)	0.53 (0.24)	0.49 (0.24)	0.75 (0.20)	0.85 (0.18)	and English

Table 4.4 Heritage language (Mandarin) use in five contexts.

	Family	External family	Work	Community	Leisure
Proportion of HL use	0.64 (0.28)	0.64 (0.29)	0.10 (0.21)	0.12 (0.13)	0.18 (0.22)
Proportion of HL interaction	0.43 (0.28)	0.39 (0.34)	0.07 (0.17)	0.10 (0.20)	0.13 (0.23)
Proportion of HL use (weighted)	0.30 (0.22)	0.06 (0.08)	0.03 (0.08)	0.01 (0.02)	0.02 (0.04)
Proportion of HL proficient interlocutors	0.80 (0.31)	0.63 (0.46)	0.22 (0.42)	0.30 (0.47)	0.33 (0.46)
Proportion of HL dominant interlocutors	0.72 (0.36)	0.54 (0.46)	0.20 (0.39)	0.29 (0.46)	0.25 (0.42)
Language entropy	0.67 (0.34)	0.64 (0.37)	0.25 (0.30)	0.41 (0.34)	0.46 (0.39)

The relationship between heritage language background and CSWL

To investigate whether the proficiency and use of Mandarin influence the outcomes in learning novel tonal words (i.e., performance at the final block), we ran several sets of mixed-effect models with the derived measures from HeLEx as fixed effects.

For the measure of Mandarin use across modalities, we carried out three sets of analyses to explore the fixed effects of (1) Mandarin proficiency, (2) frequency of Mandarin usage, (3) usage-based and proficiency-based Mandarin dominance in the four modalities. ANOVA comparison between models containing fixed effects and the random effect model showed no significant differences, indicating that none of these fixed effects significantly explain variance in word learning outcomes.

As for the measures of Mandarin use in the five contexts, we ran four sets of analyses and tested if (1) the proportion of Mandarin use, (2) the proportion of Mandarin interaction, (3) language entropy, (4) the weighted proportion of Mandarin use (accounting for the actual time spent in each context) in the different contexts explained performance in the tonal trials. However, we did not find any significant predictors of performance from the derived measures.

Exploratory analyses. Since we did not observe any significant influence of the individual HeLEx measures on participants' learning outcomes in tonal trials, we carried out additional exploratory analyses based on other responses in the questionnaire. Firstly, we explored if having one or two Mandarin-native parent influences learners' performance, as having two Mandarin-native parents may provide a more Mandarin-dominant environment at home. Mixed-effects models containing parent language as a fixed effect showed no significant improvement compared to the random effect model ($\chi^2(1) = 0.0801$, $p = .78$). This means that the number of Mandarin-speaking parent did not explain variance in word learning outcome. Secondly, we coded whether or not participants used Mandarin at preschool, primary school, secondary school, post-secondary and post-graduate levels, and extracurricular Mandarin classes to test the effect of Mandarin schooling. Model comparisons revealed no significant effect of any of the variables.

Exploratory factor analysis

Given the large number of observed variables derived from the questionnaire, we decided to carry out an exploratory factor analysis and examine whether some of the variables could be grouped into a smaller number of factors for further analyses. We planned to run two rounds of factor analysis, one for the modality-related variables (see Table 4.3) and another for context-related variables (see Table 4.4). This is because mixing the variables

across modalities and the variables across contexts might make the resulting factors less interpretable.

For the modality-related variables, we first checked the correlations between HL experience and experience-based dominance measures, as well as between HL proficiency and proficiency-based dominance measures. The results suggested that the measures are very strongly correlated ($r > 0.90$), which was expected because they were derived from the same set of original questions. Thus, we took out the dominance measures and only entered HL experience and HL proficiency across modalities into the factor analysis. The exploratory factor analysis suggested three factors: Factor 1 relates to measures of written language experience and proficiency (i.e., reading/writing experience, reading/writing proficiency), Factor 2 relates to measures of oral language experience (i.e., speaking/hearing experience), and Factor 3 relates to measures of oral language proficiency (i.e., speaking/hearing proficiency). Table 4.5 summarizes the output factor loadings of each measure.

Table 4.5 Factor loadings for modality-related variables.

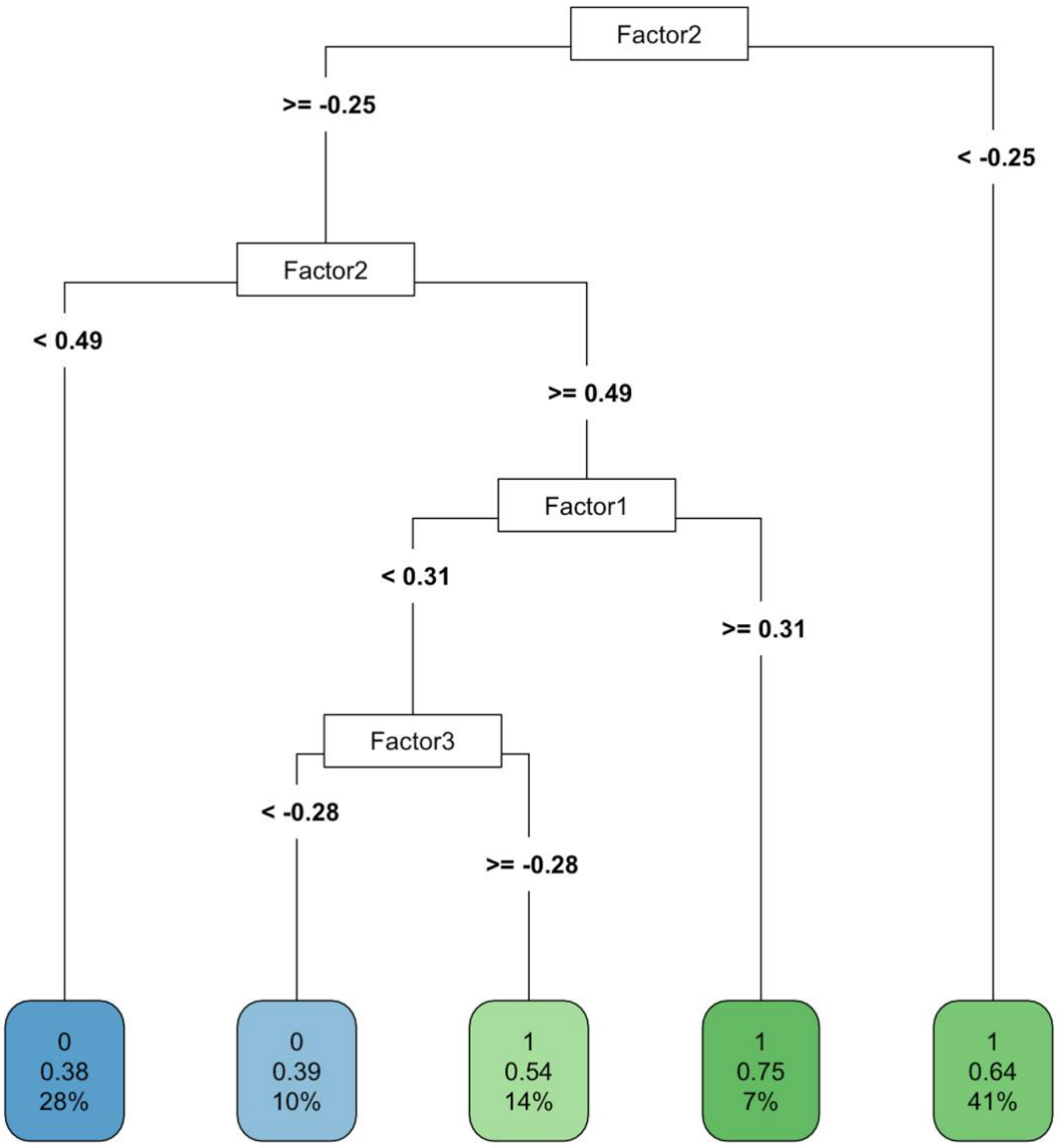
	Factor 1 (reading and writing)	Factor 2 (oral experience)	Factor 3 (oral proficiency)
Reading_ Experience	0.741		
Writing_ Experience	0.861		
Reading_Proficiency	0.869		
Writing_Proficiency	1.005		
Speaking_ Experience		0.911	
Hearing_Experience		0.996	
Speaking_Proficiency			0.933

Hearing_Proficiency	0.676

We then entered the three factors as fixed effects into the generalized mixed effect models mentioned above to explore if the grouped factors predicted participants' learning outcomes. Similar to our previous findings, ANOVA comparisons between models containing fixed effects of the three factors and the random effect model showed no significant differences, meaning that the three modality-related factors did not significantly explain variance in word learning outcomes.

In addition, we ran a decision tree analysis to explore and visualize the hierarchical contribution of the three factors to word learning outcomes. Figure 4.3 presents the results of the decision tree model. Higher Factor 2 score (oral experience) and Factor 1 score (written experience and proficiency) seemed to lead to a path to higher accuracy in tonal trials at the final block (when Factor 2 >= 0.49 and Factor 1 >= 0.31, accuracy = 0.75), though only a small proportion of data fell under this rule. Overall, however, the decision tree model did not provide clear relations between the factors and the tonal word learning outcomes.

Figure 4.3 Decision tree model based on the three modality-related factors.



We then tried to fit the same factor analysis and follow-up tests on the context-related measures. However, there was no good factor solution for the context-related measures (Kaiser-Meyer-Olkin test suggested that data was not suitable for factor analysis) – indicating that the individual measures of context of use should be kept separate. Thus, no further analyses based on the derived factors were conducted.

Comparison with English-native and Mandarin-native participants

To further understand Mandarin heritage speakers' word learning performance, we ran exploratory analyses combining data from the current study and data from Ge et al. (in press) since the two studies employed the same method and stimuli. This allowed us to compare Mandarin heritage speakers' learning trajectory with English-native participants (who had no tonal experience) and Mandarin-native participants (who had continuous, extensive tonal experience). Generalized linear mixed effects models revealed that, compared to the model with only random effects, adding the fixed effect of block ($\chi^2(1) = 21.012$, $p < .001$), trial type ($\chi^2(3) = 28.532$, $p < .001$), and the 3-way block*trial type*language group interaction ($\chi^2(11) = 42.459$, $p < .001$) significantly improve model fit. The effect of language group (English-native, Mandarin-native, Mandarin heritage) did not improve fit ($\chi^2(2) = 0.824$, $p = .662$).

We then explored the 3-way interaction in detail and ran separate mixed effects models for each trial type to test whether the group performances differed in any particular trial types. In the tonal trials, we observed a significant effect of language group ($\chi^2(2) = 6.851$, $p = .033$). The effect of block ($\chi^2(1) = 3.386$, $p = .066$) and the block*language group interaction ($\chi^2(2) = 0.020$, $p = .990$) was not significant. The best-fitting model summarized in Table 4.6 shows that the Mandarin-native group performed significantly better than the English-native group (the reference group) in tonal trials, whereas the Mandarin heritage group did not show significant divergence from the English-native group. This language group effect, however, was not significant in other trial types (consonantal $\chi^2(2) = 3.370$, $p = .185$; vocalic $\chi^2(2) = 2.254$, $p = .324$; non-minimal pair $\chi^2(2) = 3.149$, $p = .207$).

Table 4.6 Best fitting model for accuracy in tonal trials, combining data from the present study and data from Ge et al. (in press).

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	-0.188	0.143	-1.314	.189
block	0.061	0.026	2.345	.019 *
GroupMandarinL1	0.451	0.170	2.657	.008 **
GroupMandarinHeritage	0.066	0.116	0.569	.570

Number of observations: 3049, Participants: 85, Item, 12. AIC = 4186.6, BIC = 4355.2, log-likelihood = -2065.3.

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

Discussion

In this study, we explored how heritage speakers learn novel words from their HL via a cross-situational, statistical learning process and whether the degree of HL experience predicts learning outcomes. Heritage speakers could rapidly learn words that contain special phonological features which exist only in their HL but not in their dominant language (i.e., lexical tone for heritage Mandarin speakers residing in English-speaking environments). However, when this specific feature is the only informative cue to distinguish words (i.e., in the case of tonal minimal pairs), heritage speakers seem to encounter greater difficulties.

RQ1: Do minimal pairs and phonological contrasts that do not exist in heritage speakers' majority language pose difficulty during cross-situational learning?

Results suggested that learners' performance was greatly influenced by the presence of minimal pair words. As predicted, learners performed better in non-minimal pair trials as

compared to minimal pair trials, which is consistent with previous findings on CSWL of minimal pairs in other languages (e.g., Escudero et al., 2022). Moreover, we observed a difference in performance on segmental minimal pairs and tonal minimal pairs. Heritage Mandarin speakers' performance in tonal minimal pair trials was the lowest and remained at chance level throughout the experiment, whereas performance in consonantal and vocalic minimal pair trials improved over time. The lack of learning effect in tonal trials was contrary to our prediction that early exposure to Mandarin would allow the heritage speakers to develop tonal representations and be able to use tonal cues in word learning. Our combined data analysis with Ge et al. (in press) demonstrated that the Mandarin heritage speakers' learning pattern was similar to English-native speakers with no tonal experience, where tonal minimal pairs were particularly difficult, and performance in tonal trials was significantly lower than that of Mandarin-native speakers.

These findings could be explained from two perspectives – the nature of the stimuli and the participants' language profile. Firstly, the stimuli in the experiment were designed to have segments that are similar between English (the dominant language) and Mandarin (the heritage language), and also include a tonal feature that is specific to Mandarin. Since our participants were English-dominant, they might weigh more the segmental cues in their linguistic repertoire and attend more to the segmental features in the task. Previous research also suggested that even Mandarin-native speakers tend to rely more on segmental than tonal information in word processing (e.g., Cutler and Chen, 1997; Sereno and Lee, 2015; Yip, 2001). This might contribute to the divergence in the learning trajectories of segmental and tonal minimal pairs. Secondly, although the group of heritage speakers we recruited reported relatively high proficiency in Mandarin listening (rating 3.34 out of 4) and speaking (rating 2.86 out of 4), they were still significantly more dominant in English in all language modalities (see Table 4.3, HL dominance), and had very little Mandarin use outside of the

family (including external family) context (see Table 4.4). This might explain why their performance in the learning task at the group level resembles that of the English-native speakers in previous research (Ge et al., in press).

Furthermore, considering previous findings on heritage Mandarin speakers' perception and production of Mandarin tones (e.g., Chang and Yao, 2016, 2019), there is another possibility that derives from heritage speakers' distinct tonal representations. Although heritage speakers of Mandarin tend to possess categorical representations of tones that are closer to native Mandarin speakers, they are usually not entirely the same as native speakers (e.g., Yang, 2015). Therefore, even though the heritage Mandarin speakers in the experiment possess sensitivity to tonal variations, their categorization of the specific contrast (i.e., T1-T4) might be different from the native speakers in certain acoustic dimensions, resulting in the difficulty in tonal minimal pair learning. Additionally, the selection of the tones used in the stimuli was based on previous experiment testing English-native speakers' identification of Mandarin tones. Hao (2018) reported that English-native learners of Mandarin could identify T1 and T4 at word-initial positions better compared to T2 and T3, and hence these tones are likely to be easier in the disyllabic environment of this experiment. However, it is possible that the identification difficulty of the tones is different for heritage Mandarin speakers. Further research is needed to examine how tonal contexts (the preceding and following tones) affect heritage speakers' perception in particular.

RQ2: Does the degree of heritage language experience and usage influence learning outcomes?

According to the HeLEx questionnaire results, we did not find a clear relationship between participants' Mandarin experience or usage and their performance in the tonal word learning task. Specifically, the derived measures from the questionnaire did not predict how well participants respond to tonal minimal pairs. The questionnaire measures focused on how

much and how well participants use Mandarin in their daily communications, that is, the use of Mandarin in various contexts. When using Mandarin for communicative purposes, lexical tones are not the only focus because information from the context can be delivered even when lexical tones are not always correctly realized. However, in the word learning task, there was no contextual information and participants had to learn isolated words. For the tonal minimal pair trials in particular, a misperception of lexical tone would lead to failure in word identification. It is possible that heritage Mandarin speakers might rely more on contextual information in tonal perception than native speakers. Thus, a direct link between the questionnaire measures and the word learning outcomes was missing because they measured tonal abilities in different communicative situations.

Another noteworthy finding is that our factor analysis suggested a grouping of the derived measures of HL modality use, highlighting a distinction between written and oral language proficiency and use. Questionnaires like HeLEx usually contain a large number of measures to thoroughly record participants' language profiles. Our results suggested that some individual measures (even across the original categories) could be highly correlated and hence reasonably grouped into one single factor to facilitate further statistical analyses and predictions of the influence of HL on learning and behaviour.

Limitations and further directions

In the CSWL task, learning performance reflects the combined abilities at both the perceptual and lexical levels. Since we do not have a separate measure of tonal perception, it is unclear whether the difficulty comes from heritage Mandarin speakers' different tonal representations and categorizations. Thus, further studies could add tone identification tasks to examine whether more accurate identification would be associated with better word learning. It would also be interesting to test tone identification at both the pre-lexical level

(e.g., identification of isolated tonal syllables without meaning) and the lexical level (e.g., identification of tones in real words), since it indicates how well participants process tonal information when meanings are attached. Moreover, it would be worth testing whether greater HL experience and usage is directly linked to better tone identification ability.

Furthermore, it would be interesting to recruit participants from more diverse HL backgrounds. In our current sample, most participants were highly English-dominant. Future studies could compare whether heritage speakers who are more balanced in their English and Mandarin proficiency would perform differently and be more able to learn the tonal minimal pairs.

Conclusion

We found that heritage speakers of Mandarin learned Mandarin novel words in a similar pattern to English-native learners of Mandarin. They could pick up new words from a short exposure by tracking the statistics of input, but learning was reduced when minimal pairs were present. The greatest difficulty was associated with tonal minimal pairs. The degree of HL experience and usage did not seem to predict tonal word learning outcomes. Our results contribute to the understanding of heritage speakers' behaviours when learning and processing the target language. It suggests that heritage exposure does not necessarily lead to an advantage in learning the target language, and the amount of exposure may not be the key factor influencing learning outcomes, though further research into the role of diverse HL exposure is needed.

References

Au, T. K., Knightly, L. M., Jun, S.-A., Oh, J. S. (2002). Overhearing a language during childhood. *Psychological Science, 13*, 238–243.

Best, C. T., Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In M. J. Munro and O-S. Bohn (Eds.), *Second Language Speech Learning: The Role of Language Experience in Speech Perception and Production*. Amsterdam: John Benjamins, pp. 13–34.

Chang, C. B. (2016). Bilingual perceptual benefits of experience with a heritage language. *Bilingualism: Language and Cognition, 19*(4), 791-809.

Chang, C. B., Haynes, E. F., Yao, Y., Rhodes, R. (2010). The phonetic space of phonological categories in heritage speakers of Mandarin. In M. Bane, J. Bueno, T. Grano, A. Grotberg, and Y. McNabb (Eds.), *Proceedings from the 44th Annual Meeting of the Chicago Linguistic Society: The Main Session* (pp. 31-45). Chicago, IL: Chicago Linguistic Society.

Chang, C. B., Yao, Y. (2016). Toward an understanding of heritage prosody: Acoustic and perceptual properties of tone produced by heritage, native, and second language speakers of Mandarin. *Heritage Language Journal, 13*(2), 134-160.

Chang, C. B., Yao, Y. (2019). Production of neutral tone in Mandarin by heritage, native, and second language speakers. In S. Calhoun, P. Escudero, M. Tabain, and P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 2291-2295). Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Chang, C. B., Yao, Y., Haynes, E. F., Rhodes, R. (2011). Production of phonetic and phonological contrast by heritage speakers of Mandarin. *Journal of the Acoustical Society of America, 129*(6), 3964-3980.

Cutler, A., Chen, H.-C. (1997). Lexical tone in Cantonese spoken-word processing. *Perception and Psychophysics, 59*(2), 165–179.

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

Escudero, P., Mulak, K. E., Vlach, H. A. (2016). Cross-situational learning of minimal word pairs. *Cognitive Science, 40*(2), 455-465.

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

Flores, C., Rinke, E, Rato, A. (2017). Comparing the outcomes of early and late acquisition of European Portuguese: an analysis of morpho-syntactic and phonetic performance. *Heritage Language Journal, 14*(2). National Heritage Language Resource Center, UCLA.

Ge, Y., Monaghan, P., Rebuschat, P. (in press). The role of phonology in non-native word learning: Evidence from cross-situational statistical learning. *Bilingualism: Language and Cognition.*

Ge, Y., Correia, S., Fernandes, J., Hanson, K., Rato, A., Rebuschat, P. (under review). *Does Phonetic Training Benefit Word Learning?*

Gullifer, J. W., & Titone, D. (2020). Characterizing the social diversity of bilingualism using language entropy. *Bilingualism: Language and Cognition, 23*(2), 283-294.

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

Hartshorne, J. K., Tenenbaum, J. B., Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition, 177,* 263-277.

Horst, J. S., Hout, M. C. (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.

Isbilen, E. S., Christiansen, M. H. (2022). Statistical Learning of Language: A Meta-Analysis Into 25 Years of Research. *Cognitive Science, 46*(9), e13198.

Kim, J. (2020). Discrepancy between heritage speakers' use of suprasegmental cues in the perception and production of Spanish lexical stress. *Bilingualism: Language and Cognition, 23*(2), 233-250. doi:10.1017/S1366728918001220

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience, 5*(11), 831-843.

Laméris, T. J., Llompart, M., & Post, B. (2023). Non-native tone categorization and word learning across a spectrum of L1 tonal statuses. *Bilingualism: Language and Cognition*, 1–15. https://doi.org/10.1017/S1366728923000871

Ling, W., & Grüter, T. (2022). From sounds to words: The relation between phonological and lexical processing of tone in L2 Mandarin. *Second Language Research, 38*(2), 289-313.

Lukyanchenko, A., Gor, K. (2011). Perceptual correlates of phonological representations in heritage speakers and L2 learners. In: Danis, N., Mesh, K., Sung, H. (Eds.), *Proceedings of BUCLD 35*, vol. 2. Somerville: Cascadilla Press, 414–426.

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

Maye, J., Werker, J. F., Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), 101–111.

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

Quine, W. V. O. (1960). *Word and object.* Cambridge, MA: MIT Press.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

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

Saffran, J. R., Aslin, R. N., Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926-1928.

Sereno, J. A., 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.

Smith, L., Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*(3)*,* 1558-1568.

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

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

Tomić, A., Rodina, Y., Bayram, F., De Cat, C. (2023). Documenting heritage language experience using questionnaires. *Frontiers in Psychology, 14*, 1131374.

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

Watson, T. L., Robbins, R. A., Best, C. T. (2014). Infant perceptual development for faces and spoken words: An integrated approach. *Developmental Psychobiology, 56*(7), 1454-1481.

Werker, J. F., Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*(1), 49– 63.

Williams, J. N., 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 and Francis.

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

Yang, B. (2015). *Perception and production of Mandarin tones by native speakers and L2 learners.* Berlin, Germany: Springer Verlag.

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

Yu, C., Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18*(5), 414–420.

Data Availability Statement

The materials, anonymized datasets and R scripts for this study can be found in the Open Science Framework (OSF) platform: https://osf.io/q6354/.

5. Publishable paper 3: Auditory processing ability predicts statistical word learning

Page number: 152-219

Abstract

Infants and children can rapidly pick up novel words from the surrounding linguistic environment. However, this task is considerably more challenging for adults learning a non-native language (L2) because they may encounter sounds that do not exist in the inventory of their native language(s). Difficulty perceiving and representing L2 sounds has been found to interfere with word learning (Escudero et al., 2022; Ge et al., in press). However, there is also evidence that this domain-specific perceptual difficulty may be modulated by domain-general auditory processing abilities (i.e., the ability to encode acoustic features of sounds) (Saito et al., 2020; Wong & Perrachione, 2007). In this study, we investigate whether and how individual differences in auditory processing ability interact with non-native word learning. To better capture the word learning process and outcome, we employed an online eye-tracking measure in addition to an (offline) accuracy measure.

Fifty-three English-native speakers learned Mandarin pseudowords via cross-situational, statistical word learning (CSWL) (Monaghan et al., 2019). Each CSWL trial contained ambiguous word-picture mappings with one spoken pseudoword and two pictures. Participants had to decide which picture the word referred to by tracking the co-occurrences across trials. Their eye-gaze fixation during CSWL was recorded. Auditory processing was assessed through melody reproduction and pitch discrimination tasks (Saito & Tierney, 2022).

Results revealed successful word learning: participants were more likely to fixate on and select the correct referent. Melody reproduction ability and pitch discrimination ability predicted overall fixation at target and picture selection accuracy at the end of CSWL. Thus, more precise auditory processing (audio-motor integration and acoustic discrimination) was associated with better statistical learning of non-native words.

Introduction

Word learning is a fundamental process in both first (L1) and additional language (L2) acquisition. Although learning words through explicit vocabulary training is possible, we are unlikely to learn thousands of words this way. Instead, a large proportion of our lexical knowledge comes from repeated daily exposure. For example, when we first hear/see a word (e.g., *apple*), the meaning is ambiguous - the word can potentially refer to anything in the environment (e.g., shape/colour of the fruit, one specific part of the fruit, etc.). But over several encounters, we realize that one referent (the fruit itself) always co-occurs with the word, and we can establish an association between the word and the referent. This learning process is implicit and depends on learners keeping track of statistical information (i.e., word-referent co-occurrences) across encounters, which is known as cross-situational, statistical word learning (CSWL) (e.g., Angwin et al., 2022; Escudero et al., 2022; Monaghan et al., 2019; Rebuschat et al., 2021; Suanda & Namy, 2012; Yu & Smith, 2007). Although learning language from cross-situational statistics has been shown to be generally rapid (e.g., Escudero et al., 2016; Yu & Smith, 2007), there is substantial variation in the rate of language acquisition, observed both in laboratory-based learning studies (Li et al., 2022), as well as in naturalistic settings where children acquire their L1 (Frank et al., 2021; Jago et al., 2023) or L2 both within (Li et al., 2022) and outside (De Wilde et al, 2020) the classroom. Determining what are the drivers of these individual differences is a matter of theoretical and practical importance (Rebuschat, 2022; Williams, & Rebuschat, 2023). The current study contributes to this topic by exploring the individual learning variations in a cross-situational, statistical learning context.

Previous research has found that this implicit-statistical learning paradigm is effective for both children learning L1 words (e.g., Childers & Pak, 2009; Smith & Yu, 2008; Suanda et al., 2014; Yu & Smith, 2011) and adults learning non-native words from their L2 (e.g.,

Gillette et al., 1999; Smith et al., 2009, 2011; Yurovsky et al., 2013). For L2 learners, however, resolving ambiguity in word-referent mappings is not the only challenge. Research has found that non-native word learning is also greatly influenced by the presence of unfamiliar sounds that do not exist in the inventory of learners' native language(s) (Escudero et al., 2022; Ge et al, under review; Tuninetti et al., 2020). Specifically, learners' difficulty in perceiving and representing non-native sounds has been found to hinder statistical word learning (e.g., Escudero et al., 2022; Ge et al., in press). On the other hand, there is also evidence that this domain-specific perceptual difficulty in non-native word learning may be modulated by lower-order, domain-general auditory processing ability (i.e., the ability to encode and reproduce fundamental acoustic features of sounds) (Mueller et al., 2012; Saito et al., 2020a; Wong & Perrachione, 2007). This study directly examines this hypothesis and investigates whether and how individual differences in auditory processing ability interact with statistical learning of non-native words.

To better track and understand the implicit, statistical learning process, we employed an additional eye-tracking technique to compare online and offline behavioural measures (Rebuschat, 2013, 2022. Previous CSWL research mainly employed offline behavioural measures, tracking learners' progress through their accuracy in word-referent mapping tasks (for exception, see Angwin et al. (2022) for event-related potential measures; Yu & Smith (2011) and Yu et al. (2012) for eye-tracking measures). These offline tasks typically require learners to choose the correct referent for words in forced-choice tasks, which depend on explicit knowledge of the lexical items and explicit responses to stimuli, and involve decision making processes that may be external to the process of word learning (e.g., Isbilen et al., 2022). Therefore, the offline tasks may obscure, or at least not fully uncover, the implicit learning outcomes in CSWL. We addressed this issue by utilising an online eye-tracking measure, which tracks participants' eye gaze fixation throughout learning and does not

involve explicit decision-making, in addition to a standard explicit button press response to investigate learning. This provides a more implicit measure of word knowledge and contribute to the exploration of more reliable word learning measures.

The role of phonology in non-native word learning

A critical context for this study is that previous research has found an impact of phonology on non-native word learning, suggesting that unfamiliar, non-native sounds can negatively influence word learning performance (e.g., Havas et al., 2018; Kaushanskaya et al., 2013; Nora et al., 2015; Service & Craik, 1993). For example, in a paired-associate word learning task, Havas et al. (2018) reported that Spanish-native adults performed better in learning Spanish pseudowords compared to learning Hungarian words that contain unfamiliar phonemes. This influence of phonology was also observed in the cross-situational word learning task that better resembles the naturalistic word learning situations in the real world (e.g., Escudero et al., 2022; Ge et al., in press; Tuninetti et al., 2020). In the CSWL paradigm, learners are presented with ambiguous word-referent mappings in each learning trial, with multiple words and pictures co-occurring, and they need to keep tracking the co-occurrences over a number of trials to figure out the correct referent for each word. Through the CSWL task, Escudero et al. (2022) trained English-native and Mandarin-native learners with novel consonantal (e.g., /bɔn/-/tɔn/) and vocalic minimal pairs (e.g., /dit/-/dɪt/) that resembled English real words. Although the English-native learners showed greater difficulty identifying the word-referent mappings in the vocalic minimal pair condition than those in the consonantal and non-minimal pair (e.g., /bɔn/-/dit/) conditions, their overall performance was better than the Mandarin-native learners for all different (non)minimal pair types.

A similar non-native phonology impact was reported by Tuninetti et al. (2020) when Australian English speakers were trained with novel Dutch and Brazilian Portuguese vowel

minimal pairs (e.g., /piχ/-/pyχ/, /fɛfe/-/fefe/, respectively). The authors categorized the vowel minimal pairs into perceptually easy and perceptually difficult pairs based on the Second Language Linguistic Perception model (L2LP; Escudero, 2005) and the Perceptual Assimilation Model-L2 (PAM-L2; Best and Tyler, 2007). The Dutch/Brazilian Portuguese vowel contrasts that could be mapped onto two separate phonemic categories in Australian English were considered *perceptually easy*, whereas those without Australian English counterpart categories were marked as *perceptually difficult*. It was observed that learners better identified the non-native minimal pairs that were perceptually easy compared to those that were perceptually difficult, indicating a significant effect of non-native sound perception in word learning.

Furthermore, the non-native phonology effect is not only associated with segmental but also suprasegmental features. Ge et al. (in press) tested the effect of lexical tones within the CSWL paradigm, recruiting English-native speakers to learn Mandarin pseudowords with tonal differences. This study used a slightly different design than the previous research to more closely resemble the minimal pair situations in the real world. In previous studies of minimal pairs, learners heard the minimal pairs together, whereas in naturalistic learning situations minimal pairs tend to occur in different utterances. In Ge et al. (in press), learners did not hear the minimal pairs next to each other. Instead, they only heard one word in each learning trial and were asked to associate the word with one of the two pictures presented. In addition to the segmental (consonantal and vocalic) minimal pairs as in previous research (e.g., Escudero et al., 2022), this study involved tonal minimal pair trials where the two pictures shown were mapped to two words that differ only in lexical tones (e.g., /pa1mi1/ vs /pa4mi1/). To successfully distinguish and identify the word-picture mappings in this condition, learners need to develop tonal awareness and understand that lexical tones contrast meanings in the language by keeping track of the relations between words with minimal

phonological distinctions and their potential referents. Through a short cross-situational exposure, the English-native participants successfully identified word-referent mappings in consonantal, vocalic and non-minimal pair trials (as the segmental features in the stimuli were designed to be familiar to English speakers), but not in the tonal minimal pair trials, where the distinction was not part of the speakers' first language). This study added to the previous evidence that non-native phonological features, both segmental and suprasegmental, significantly affect the outcomes of statistical non-native word learning. Ge et al. (in press), however, also revealed that a few English-native participants *did* show improvement in identifying tonal minimal pair words during CSWL. This points towards potential individual differences in learning performance that are distinct from natural language experience. The critical question to be addressed in the present study is thus what possible individual differences might underlie this variation in the ability to detect and use unfamiliar phonological distinctions in implicit acquisition of a novel language. In this study, we focused on one perceptual aspect of individual differences, the domain-general processing of pitch variations in acoustic stimuli, as lexical tone realization is highly reliant on the pitch dimension and greater sensitivity to pitch changes in general might facilitate tonal perception and acquisition.

In addition, in the previous CSWL studies discussed above, it was not clear whether difficulties in using non-native contrasts for lexical distinctions were due to an inability to *perceive* the auditory differences, or whether the learners' difficulties were due to processing the different auditory signals as phonological distinctions. Thus, in the present study, we employed an additional perceptual discrimination task to test whether participants' lower performance in non-native minimal pair learning was associated with difficulty in basic auditory perception and discrimination.

Individual differences in language learning

As mentioned above, although the impact of phonology on non-native word learning has been largely found at the population level, there exist potentially large individual variations. Even when all learners receive the same quantity and quality of training (in laboratory settings), their learning outcomes may vary. Research on individual differences has explored extensively and conceptualized a set of cognitive learning abilities, collectively known as *language aptitude* (Carroll, 1981), that may contribute to the learning variation (see Li, 2016; Li et al., 2022 for detailed discussions). These cognitive abilities provide the basis for recognizing, analyzing and memorizing linguistic features. Despite decades of research, the construct of language aptitude is still undetermined. The originally hypothesised language aptitude components by Carroll (1981) involved phonemic coding, language analytic ability and rote memory, all of which are associated with explicit, instructed language learning in the classroom. Such a construct has been challenged in that it is not representative of the more naturalistic learning environments (Li, 2015; Skehan, 2012). A later hypothesis by Linck et al. (2013) thus introduced a measure of implicit aptitude (which specifically measures procedural memory) using a serial reaction time task, which aims to test learning beyond consciousness or awareness from contextual exposure.

Moreover, natural language learning outside the classroom relies primarily on auditory input from speech conversations, and hence the first step of processing involves decoding the sound streams. In the classic language aptitude constructs, this is usually accounted for by domain-*specific* auditory abilities such as phonemic discrimination and recognition (e.g., Carroll, 1981; Linck et al., 2013). However, the relationship between domain-specific auditory abilities and language learning attainment was not always clearly observed (Linck et al., 2013). More recently, Saito and colleagues (2020a,b) proposed that individual differences in domain-*general* auditory processing (e.g., non-linguistic sound

discrimination and reproduction) could also play a role in adult L2 learning. This hypothesis originated from the fact that L2 input is more limited in terms of quantity and quality, resulting in potential reliance on general auditory skills for the accurate initial processing of the acoustic input.

Auditory processing skills are typically measured by assessing participants' sensitivity to different acoustic dimensions of sounds, such as pitch, duration and formant frequency (e.g., Surprenant & Watson, 2001). There is empirical evidence that non-native speech perception and production correlates with the relevant auditory processing measures (e.g., Kempe, et al., 2012; Lengeris & Hazan, 2010). For example, Lengeris and Hazan (2010) examined the relationship between L2 vowel perception and learners' frequency discrimination acuity. After high variability phonetic training of English vowels, L1 Greek learners who had greater sensitivity to frequency differences in nonspeech sounds performed better at the discrimination, identification and production of English vowels.

In addition to L2 perception and production, Wong & Perrachione (2007) observed that non-native word learning was also linked to auditory processing skills. Seventeen English-native learners who had no tonal language experience were trained with pseudowords that contained Mandarin Chinese lexical tones. The training process involved one-to-one mapping of pseudowords and corresponding referents, with feedback provided. Results suggested that this explicit training was more successful for learners who exhibited a better ability to discriminate nonlexical pitch patterns, providing potential explanations for the individual variations in word learning attainment.

These studies, however, primarily investigated the predictive power of auditory processing ability in explicit training conditions, and the sample sizes were relatively small in order to examine individual differences comprehensively. Therefore, Kachlicka, Saito and Tierney (2019) extended the research to more naturalistic learning conditions by testing

variations in learner abilities in a living abroad context. Auditory processing (as reflected by both behavioural and neural measures) was found to predict L2 English learners' speech perception and grammatical abilities. This finding was also important as it demonstrated the role of auditory processing ability not merely at the phonetic and phonological level (e.g., in speech discrimination and identification), but also at a higher syntactic level of language processing (e.g., in grammaticality judgement). That is, greater sensitivity to general acoustic differences may facilitate the decoding of speech sounds, which further helps learners map sounds to meanings and look for structural patterns of the language.

To summarise, previous evidence showed that auditory processing ability is associated with variations in L2 learning performance in different learning conditions (e.g., instructed and naturalistic) and different linguistic domains (e.g., phonetic, phonological, lexical, syntactical). However, it is not yet clear whether such auditory processing ability can also explain individual differences in learners' ability to utilise novel phonological distinctions in statistical word learning. In the current study, we addressed this question and explored whether auditory processing ability accounts for (at least partially) why some learners performed better than others in distinguishing non-native minimal pair words , as observed in Ge et al. (in press). To examine this question, we employed a CSWL paradigm and included additional tests on participants' auditory processing ability.

Assessment of word learning via online eye-tracking

In order to better capture the individual differences in word learning performance, it is important to have sensitive measurements of what has been learned. As previously mentioned, the most widely used measures of language learning outcomes in laboratory settings are offline behavioural tasks, which usually involve a series of linguistic knowledge tests before and after treatment. For vocabulary learning studies, typical tests include, for

example, the Vocabulary Knowledge Scale test (Paribakht & Wesche, 1997) and form-meaning mapping tasks. These tasks are relatively sensitive to small vocabulary gains over short-term training or treatment (Kremmel, 2019). In cross-situational word learning research, these behavioural measures of word-meaning mappings have also been largely used in the format of forced-choice or judgment tasks.

One critical problem with these tests is that they only measure the accuracy of learners' explicit responses to word forms, but are unable to reflect learners' implicit knowledge and online processing of the words. For example, in CSWL, when learning L2 words with non-native sound contrasts, learners may start by developing implicit knowledge of the non-native contrast through repeated exposure. However, with no explicit instruction provided, this implicit knowledge may be insufficient for learners to form phonological representations of the contrasts and further make use of the contrasts to distinguish words. And when tested on word-meaning mappings that rely on explicit knowledge of the lexical representations, learners' responses may be affected (Isbilen et al., 2022). Therefore, the behavioural accuracy measure may not show the complete picture of what learners have acquired during the implicit, cross-situational word learning process.

To measure the more implicit word learning gains, one potential technique is to track learners' real-time eye movement throughout the learning session. The eye-tracking technique assumes that language processing requires attentional focus on linguistic features, and eye gaze location can be a reflection of the attentional focus (Reichle et al., 2006). Therefore, it allows us to track learners' moment-by-moment cognitive processing of linguistic input, and any changes in learners' eye movement patterns could be potential learning gains. In word learning research, there are two main streams of research designs in which eye-tracking is widely used – the visual-world paradigm (e.g., Spivey & Marian, 1999;

Weber & Cutler, 2004; Weighall, et al., 2017) and reading studies (e.g., Felser & Cunnings, 2011; Keating, 2009; Roberts, Gullberg & Indefrey, 2008).

The visual-world paradigm is typically employed to investigate L2 learning and processing at the lexical level. For example, Marian and Spivey (2003) used eye-tracking to test lexical competition among L1 Russian speakers who were proficient in English. When hearing an English target word (e.g., *speaker*), participants showed increased fixation at the distractor object *match* because the translation of *match* in Russian is *spichki*, which shares the initial phonemes with *speaker*. This illustrated the sensitivity of eye-tracking in capturing the prompt online processing of words.

Furthermore, in cross-situational word learning research, a few studies have also employed this eye-tracking measurement of learning (Yu & Smith, 2011; Yu et al., 2012). The rationale here is that if a learner can form an association between a word and a referent via CSWL, they will fixate on the referent among other co-occurring distracting objects when presented with the word. If the word-referent association is not well established in learners' mental representation, they will show attention shifts between the referents and other distracting objects. Yu and Smith (2011) investigated 14-month-old infants' cross-situational word learning and observed that the stronger learners (with better learning performance) showed more stable fixation at the correct referent after hearing a word. The weaker learners generated more gaze shifts between objects (see Dunn et al., 2024, for similar results). Similarly, Yu et al. (2012) found eye fixation differences among adults. The strong learners with higher word-referent mapping accuracies showed increasing fixation at the correct referent throughout the learning session. In contrast, the weak learners did not show any linear increase in the same time course. These previous findings provided evidence that eye-tracking could be a sensitive, reliable measure of word learning gains. Therefore, we utilized this measure in addition to the offline behavioural measure in the current study and explored

if the eye fixation information provided more fine-grained details on what has actually been learned in CSWL. Moreover, we decided to use the web-based eye-tracking function in Gorilla instead of the lab-based equipment because of its compatibility with our online data collection procedure and because sensitivity of viewing location was not critical in our task design. This is because we presented two pictures on screen in each trial, and the eye-tracking information we collected was whether the learners were looking at the left or right part of the screen. This required relatively less precise calibration for eye-tracking to provide useful information, and the web-based eye-tracking via device camera has the potential to capture this data. Our results, in addition, provide insights into the reliability of web-based eye-tracking techniques for this and related tasks.

Research questions and predictions

The following research questions were addressed:

RQ1: Do learners' auditory processing ability predict cross-situational learning of non-native words?

RQ2: Can learners perceive and discriminate between tonal differences before and after cross-situational learning of tonal words?

RQ3: Do online eye-tracking and offline accuracy measures show similar learning performance patterns in CSWL?

In terms of RQ1, we expected auditory processing ability to have a positive correlation with the learning outcomes (as measured by percentage fixation at the target and accuracy) at the end of the learning session. As for RQ2, participants were predicted to show above-chance tonal discrimination both before and after CSWL, as we expected their difficulty with tonal words to be associated with phonological representations rather than perceptual issues (Ge et al., in press). For RQ3, we predicted that online eye-tracking and

offline accuracy measures would show similar learning patterns across the (non)minimal pair trial types. However, we expected to find some extent of learning in the tonal trials from the eye-tracking measure (but not the accuracy measure), as it might be more sensitive to reflect the implicit processing of tonal minimal pairs.

With the eye-tracking technique, it was hypothesised that better learning could be reflected by longer eye fixation at the target, as it indicated that participants were more certain about the response and had fewer attention shifts between the objects. Based on previous results (Yu & Smith, 2011; Yu et al., 2012), we predicted that throughout learning, participants would gradually show greater fixation at the target object than the foil object. The percentage fixation at the target object would be the greatest when the target and the foil objects were associated with two words that sounded distinct (non-minimal pair trials). We expected the percentage fixation at the target to be slightly lower but significantly above chance when the two objects were mapped to two words that differed in only a consonant or a vowel (consonantal or vocalic minimal pair trials). When the two objects were associated with two tonal minimal pair words, the percentage fixation at target would be the lowest but still significantly above chance (50%) at the end of the CSWL. These predictions would be consistent with the accuracy measure, which revealed the highest accuracy in non-minimal pair trials, followed by consonantal and vocalic trials, and then tonal trials.

Methods

Participants

Sixty-five participants were recruited through the Department of Psychology participant recruitment pool at Lancaster University. Participation was voluntary, and participants were granted credits for their university courses. Participants had to be at least 18 years old, speak English as a native language (L1), and have no previous experience learning

any tonal languages before taking part in the study. Participants had normal or corrected-to-normal vision and hearing. Four participants were excluded from the data analysis because they spoke a language with tonal features (Mandarin, Punjabi, Somali). Thus, 61 participants were included in the final analysis.

The sample consisted of 46 females, 14 males and one non-binary participant. The average age was 20.11 years (SD = 5.19, range 18 to 51). All participants were native speakers of English, with three reporting having another native language apart from English (Russian, Urdu). Twenty-three participants reported knowing at least one foreign/second language (Bengali, Finnish, French, German, Italian, Portuguese, Spanish, Tamil, Welsh).

To estimate the sample size needed and the power for expected effects, we ran power analyses for two critical effects in the experiment with Monte Carlo simulations of data (full power analysis script available at https://osf.io/vhp75/). The first was the interaction effect of learning trial type and block, and the second considered the effect of auditory processing measures (pitch discrimination and melody reproduction) on participants' performance in the final CSWL block. Our target sample size was 62, with power > 0.85 for the two critical effects we wanted to measure. Note that the final sample size of 61 deviated from our preregistration plan. This was because we identified four participants as speaking a tonal feature language after data collection had closed. We preferred reducing the sample to 61, rather than reopening data collection for one additional participant, who would complete at a different time of year. Power for sample of 61 participants was still > 0.84 for both critical effects.

The study was approved by the ethics review panel of the Faculty of Arts and Social Sciences at Lancaster University and conducted in accordance with the provisions of the World Medical Association Declaration of Helsinki. The preregistration for this study can be

found on the Open Science Framework (OSF) website: https://osf.io/kqagx. The materials, anonymized data and R scripts are available at: https://osf.io/vhp75/.

Materials

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 about 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.

Tonal discrimination task. We presented a tonal discrimination task before and after the CSWL task to test whether participants could perceive the tonal differences without lexical contexts. The tonal discrimination task involved 16 Mandarin CV words. Four base syllables were chosen (tu, pi, wei, mao) because they form real Mandarin words with all four lexical tones, and the phonemes are familiar to English speakers. The four base syllables were superimposed with the four natural tones in Mandarin (T1 – high, T2 – rising, T3 – low dipping, T4 – falling) to form the 16 stimuli words. All four tones were included in the tonal discrimination task to examine participants' general sensitivity to the tonal feature rather than their discrimination of a specific tone pair.

Cross-situational word learning task. The CSWL task involved learning 16 pseudoword-referent mappings. Similar to Ge et al. (in press), all pseudowords were disyllabic, with CVCV structures. The pseudowords contained phonemes that were similar between Mandarin Chinese and English. This ensures that the English-native participants can easily distinguish the segmental contrasts, and hence the only non-native feature would be lexical tones. 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. We included T1 and T4 based on previous evidence that English-native learners of Mandarin could identify T1 and T4 at word-initial positions better than T2 and T3 (Hao, 2018). We did not include all four Mandarin tones in the CSWL stimuli to make the tonal system easier for English-native speakers who were naïve to lexical tones.

Seven different consonants /p, t, k, s, l, m, f/ and five different vowels /a, i, u, ei, ou/ were combined to form ten distinct base syllables (/pa, ta, ka, sa, li, lu, lei, lou, mi, fa/), which were further paired to form eight minimally distinct base words (pami, tami, kami, sami, lifa, lufa, leifa, loufa). Four of the base pseudowords differed in the consonant of the first syllable (pami, tami, kami, sami), which made up a consonant set; and the other four differed in the vowel of the first syllable (lifa, lufa, leifa, loufa), making up a vowel set. The second syllables in the pseudowords were held constant in each set, hence the words in each set were minimal pairs. These base words were then superimposed with lexical tones. The first syllable of the base words was paired with either T1 (high tone) or T4 (falling tone), and the second syllable always carried T1. This makes up a set of tonal minimal pair contrasts (e.g., pa1mi1 vs pa4mi1). The full list of pseudowords is shown in Table 5.1. The pseudowords (with their corresponding referent objects) were then paired to form consonantal, vocalic, tonal, and non-minimal pairs. All pseudowords had no corresponding meanings in English or Mandarin Chinese. The audio stimuli were produced by a female native speaker of Mandarin Chinese.

Table 5.1 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
sa1mi1	sa4mi1	lou1fa1	lou4fa1

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

Additionally, 16 pictures of novel objects were selected from Horst and Hout's (2016) NOUN database as referent objects. The pseudowords were randomly mapped to the objects. We created four lists of word-referent mappings to minimize the influence of a particular mapping being easily memorable. Participants were randomly assigned to one of the mappings.

Debriefing questionnaire. After the CSWL task, participants were given a debriefing questionnaire to elicit retrospective verbal reports about their awareness of the tonal contrasts in the language. The questionnaire was adapted from Rebuschat et al. (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. The full questionnaire can be found in Appendix A.

Auditory processing tasks. The auditory processing tasks were selected from Kachlicka et al.'s (2019) and Saito et al. (2020a)'s Auditory Processing Battery. Since our current study focused on the learning of lexical tones, we chose the pitch-related auditory processing measures from the battery – pitch discrimination and melody reproduction. The pitch discrimination task used 100 complex tone stimuli differing in frequency levels (F0 = 300-360 Hz, each differed by 0.3 Hz). The melody reproduction task involved ten melodies,

each consisting of 7 notes (300 ms per note). The notes differed in fundamental frequencies, and five different frequency levels were used (five different notes).

Experimental design

All participants were directed to the experiment platform Gorilla (www.gorilla.sc) to finish all tasks. The entire experiment took around one hour to finish. Participants started by completing the background questionnaire. They then completed the first tonal discrimination task. The CSWL task followed and lasted for approximately 25 minutes. After the CSWL task, participants completed the seven debriefing questions in order. Only one question was presented on the screen each time. Participants then completed the tonal discrimination task again, followed by the auditory processing tasks.

Tonal discrimination task. During each tonal discrimination trial, participants were presented with three stimuli, with either the first or the third being different from the other two. Participants had to indicate which stimulus was different by clicking on '1' or '3' on the screen. There were a total of 64 discrimination trials, 48 of which contained stimuli that differed only in tones (e.g., /tu1/-/tu1/-/tu2/), and the rest 16 trials contained stimuli that differed in both base syllables and tones (e.g., /tu1/-/tu1/-/wei2/).

Cross-situational word learning task. The CSWL task was 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 CSWL – consonantal, vocalic, tonal and non-minimal pair trials. We manipulated the target and foil objects in each trial to create these minimal pair trials. 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* could be 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 must be able to distinguish *pa1mi1* from *ta1mi1*, and learned the associations between each of these words and the object to which they were mapped, 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 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 might have difficulty distinguishing the tonal pairs such as *pa1mi1* vs *pa4mi1*. 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 and non-native sounds caused difficulty in CSWL.

The occurrence of each trial type was controlled in each block and throughout the experiment. There were 12 CSWL blocks, with 16 trials each, resulting in 192 trials in total. Each trial type (i.e., minimal pair type) occurred 4 times in one block, leading to a total of 48 trials across the experiment. Within each learning block, each of the 16 pseudowords was heard once, and each of the novel objects was used as the target referent once. The foil object was randomly selected from all the possible minimal pairs using the randomization function in Excel. Throughout the experiment, each pseudoword occurred 12 times with the target object, and no more than four times with each of the possible foil objects. 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

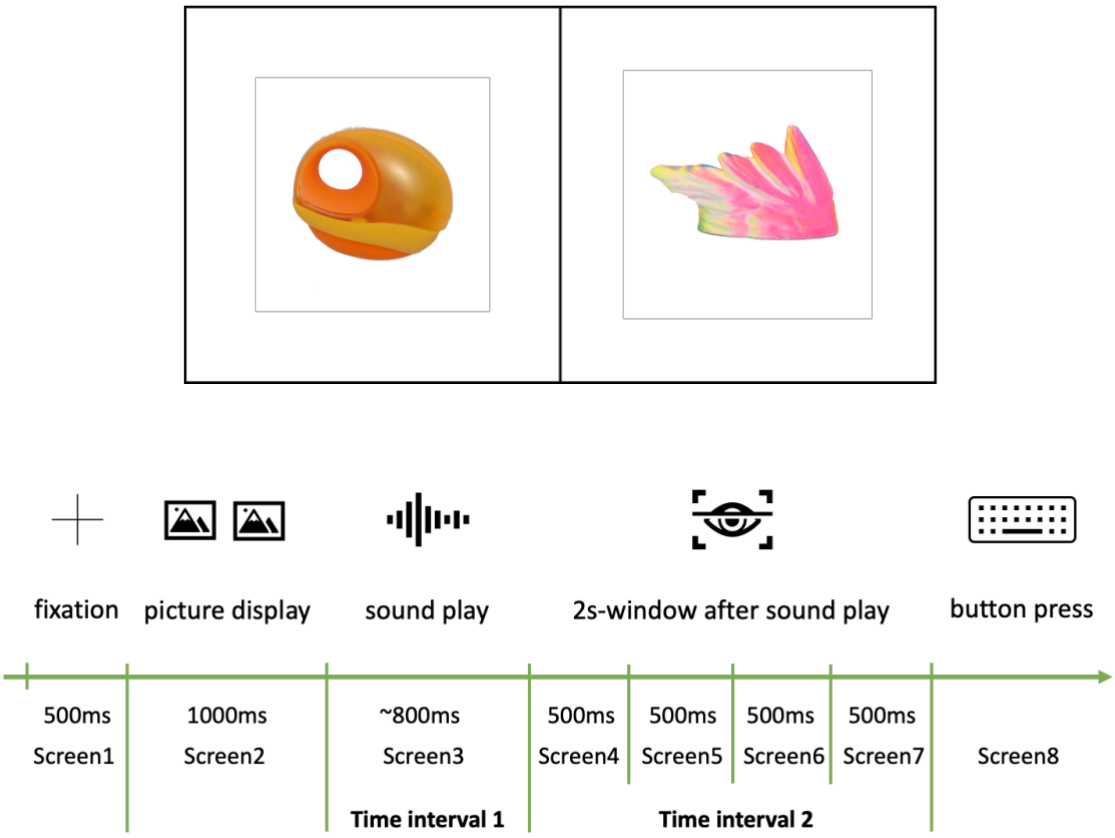
the position of the target was determined by the randomization function in Excel. There were four types of word-referent mappings randomly created, and each participant was randomly allocated to one of the mapping types. Participants' accuracy at selecting the correct referent was recorded throughout the experiment, and their response time in each trial was measured.

Throughout the CSWL task, participants were monitored with the eye-tracking function in Gorilla. This web-based eye-tracking technique was based on WebGazer.js (Papoutsaki et al., 2016). Participants first went through an initial calibration stage, during which they were instructed to follow and fixate their gazes on the points on the screen. This allowed the eye-tracker to use embedded models to predict participants' eye gaze locations in the experiment. If the accuracy of the calibration was not met, participants needed to retry the calibration process. After every two CSWL blocks (around 2-3 minutes), participants went through the calibration process again to improve the accuracy of eye-tracking. The webcam detected participants' faces, but the images used in the eye-tracking process were translated into coordinates on the computer screen, and only the coordinate data was collected.

In each CSWL trial, participants first saw a fixation cross at the centre of the screen for 500ms to gather their attention. They then saw two object pictures on the screen for 1000ms, after which the audio stimuli started playing. The mean length of the audio stimuli was 800ms, during which time the pictures were always present. After the audio stimuli, there was a 2000ms window, which was set to collect eye gaze location data after the stimuli display. The 2000ms window was further divided into four sub-intervals for data inspection and visualization. After the 2000ms window, participants saw a keyboard picture as a prompt to choose the correct referent for the word they heard. 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 next trial only started after participants made a choice for the current trial. No feedback was provided after each response. Figure 5.1 provides an

example and timeline of a CSWL trial. Each stage along the timeline was labelled by a 'screen', for example, 'screen 1' represents the fixation cross display.

Figure 5.1 Example and timeline of a CSWL trial. Participants were presented with two novel objects and one spoken word (e.g., pa1mi1). When they saw the keyboard prompt, they had to decide as quickly and accurately as possible if the word referred to the object on the left or right of the screen.



Auditory processing tasks. The pitch discrimination task was similar to the tonal discrimination task, in which participants had to choose a different sound from three stimuli. The different stimulus was either at the first or the third position. The task was programmed to decrease the discrimination difficulty (i.e., increase the F0 difference between stimuli)

after an incorrect response, and increase the difficulty after every three correct responses (i.e., reduce the F0 differences). The task stopped after participants had eight 'reversal' responses, that is, incorrect responses after a sequence of 'corrects' or correct responses after a sequence of 'incorrects'. A pitch discrimination score (0-100) was then computed for each participant by averaging the levels of the 'reversals' (see Kachilicka et al. 2019 for methodological details). A lower score meant that participants could distinguish between sounds with smaller F0 differences, hence indicating better pitch sensitivity.

In the melody reproduction task, participants listened to the melodies and were instructed to reproduce the melody by clicking on the five buttons (indicating five different notes) on the screen. The buttons were labelled 1-5 and vertically ordered, with 1 at the bottom and 5 at the top. A greater number indicated higher-frequency notes. Each melody was played three times. A melody reproduction score was calculated from the accuracies of the first seven buttons pressed.

Data analysis

For the offline behavioural measure, we excluded individual responses that lasted over 30 seconds as the extended response time indicated that participants failed to follow the instructions to respond as quickly and accurately as possible (one response was removed). Additionally, if a participant responded to the same side (e.g., pressing the left side button) for 90% or more of responses within a block, or showed a particular alternating pattern (e.g., left/right/left/right) for 90% or more within a block, then data for that block was omitted (one block from one participant was removed). We then computed accuracy in different trial types across the 12 blocks and visualized the data.

For eye-tracking data, individual trials or screens with missing eye-tracking information were excluded (1470 out of 58560 screens were excluded). After removing the

missing values, we computed percentage fixation at the target object during time interval 1 (sound play) and time interval 2 (2s window after sound play) in different trial types across blocks.

We used generalized linear mixed effects modeling for statistical data analysis (lme4 package, Bates et al., 2015). We first ran a set of models to investigate the effects of learning trial type, block, and their interaction on offline accuracy. Mixed effects models were constructed from the null model (containing only random effects) to the model containing fixed effects of block, learning trial type (with non-minimal pair as the reference level), and the block*trial type interaction. Random effects were item (target word) and participant. Slopes for item were trial type and block. Slopes for participant were trial type and block. A quadratic effect of block was also tested for its contribution to model fit, as block may exert a quadratic rather than linear effect. Additionally, based on participants' responses to the debriefing questions, we classified them into those who were *aware* of the tonal contrast and those who were *unaware*, and ran similar generalized linear mixed effects models with fixed effects of block, learning trial type, and awareness.

We then ran similar sets of mixed-effect models using percentage fixation at the target as the dependent variable for the eye-tracking data. In eye-tracking, there was an additional time interval variable. We extracted eye gaze information from two time intervals in each trial. The first interval (T1) was the time during which the audio was played, and the second interval (T2) was the 2000ms window after the audio offset. These two time intervals were entered into the models as a fixed effect with two levels (T1 and T2, categorical). This allowed us to explore if participants' percentage fixation at the target changed from T1 (during sound play) to T2 (after sound play) in response to the stimuli. Thus, the model included fixed effects of block , trial type, time interval, and the 3-way interaction.

For assessing the role of the auditory processing measures on learning, we tested the effect of auditory processing scores (pitch discrimination and melody reproduction) on participants' performance (accuracy and percentage fixation at target) in the final CSWL block. The model included fixed effects of trial type, pitch test score, melody test score, and the interactions between trial type and pitch/melody test scores. For the tonal discrimination measures, we tested if performance improved from pre-CSWL to post-CSWL and whether tonal discrimination accuracy was associated with CSWL performance.

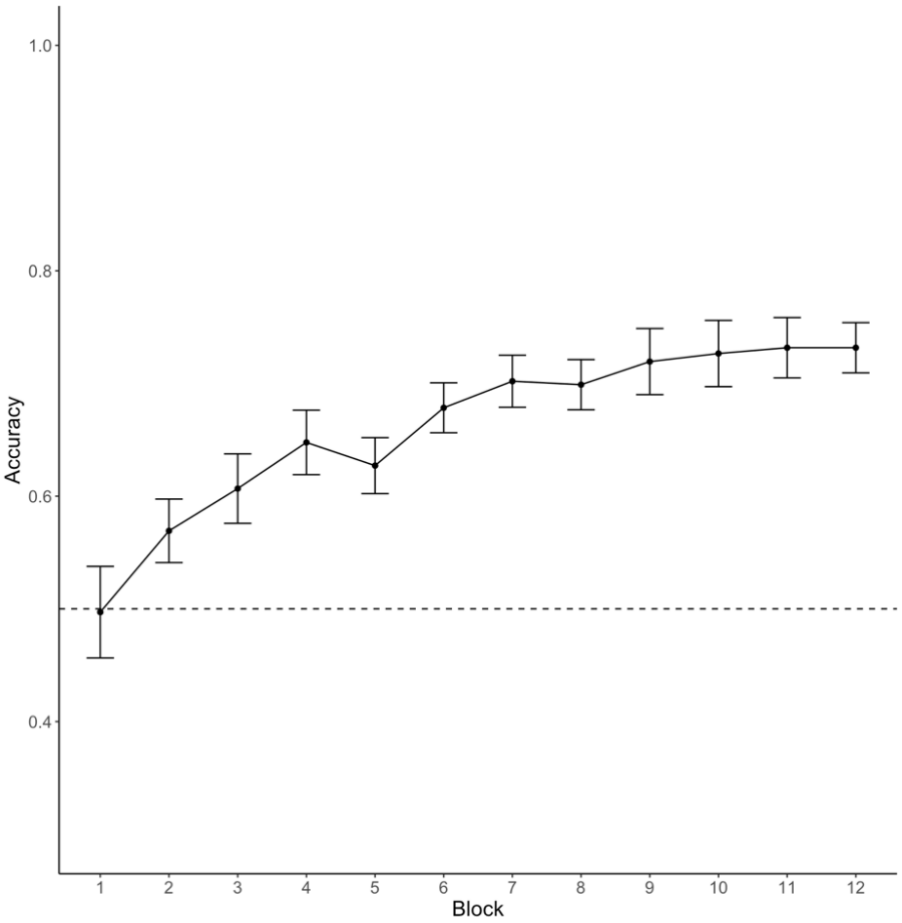
Results

Performance on the cross-situational word learning task

Accuracy measure. Figure 5.2 illustrates the participants' overall performance across the 12 blocks of the CSWL task. From the second block, participants started to score significantly above chance, which means that there is a clear learning effect in general. Figure 5.3 further demonstrated how performance was influenced by trial type. The learning effect was the greatest in the non-minimal pair trials, followed by the consonantal and vocalic trials. The accuracies in the tonal minimal pair trials remained at chance level throughout the CSWL task[12]. The summary of the mean accuracies across blocks in different trial types is presented in Supplementary Materials Table S5.1.

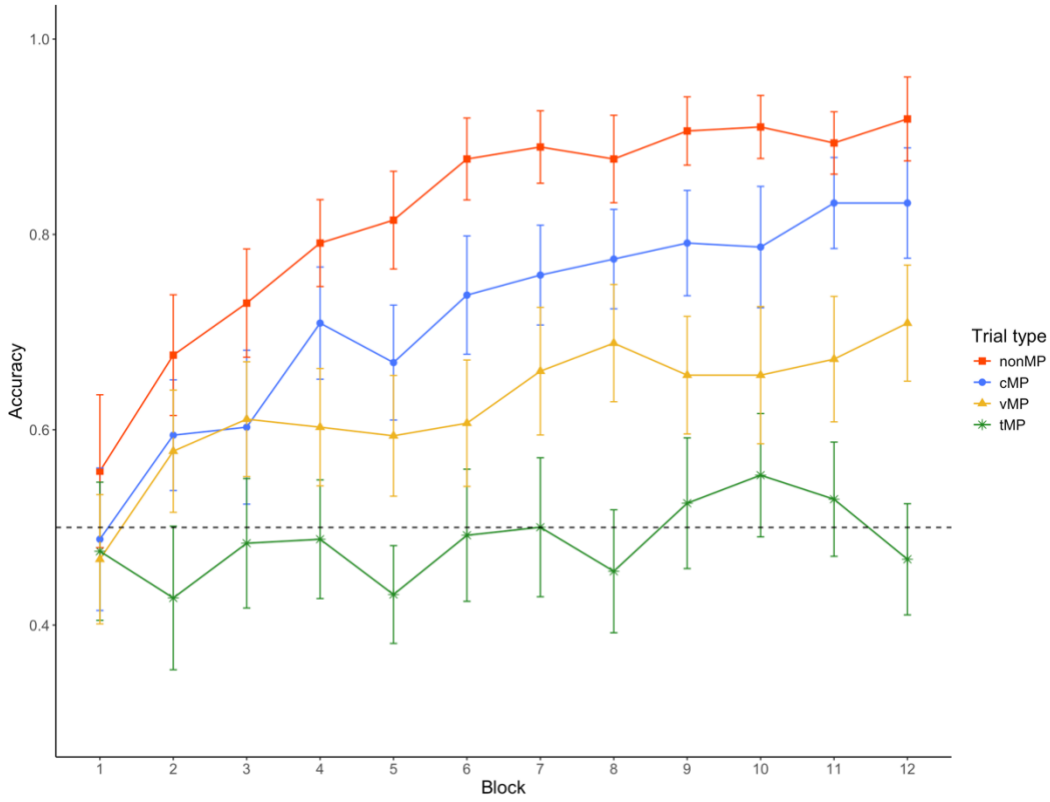
[12] Except for Block 10 where performance was slightly above chance ($t = 1.69$, $p = .048$). However, this above-chance performance was not retained afterwards in Block 11 and 12.

Figure 5.2 Mean proportion of correct pictures selected in each block of the CSWL task.



Note. The dotted line represents chance level. Error bars represent 95% Confidence Intervals.

Figure 5.3 Mean proportion of correct pictures selected in different trial types.



Note: nonMP refers to the non-minimal pair trials, cMP refers to the consonantal minimal pair trials, vMP refers to the vocalic minimal pair trials, and tMP refers to the tonal minimal pair trials.

As outlined in our preregistration, we ran generalized linear mixed effects models to test the effect of block, trial type and the block*trial type interaction on participants' accuracy in the CSWL task. We started with a null model including the maximal random effects structure that converged, which included random slopes for block and trial type for items, and random slopes for block and trial type for participants. Then we added fixed effects of block, trial type (with non-minimal pair as reference category) and the 2-way interaction to test if each of them significantly improved model fit. Finally, we tested the quadratic effect for the block to determine if learning was linear or non-linear over the training trials. ANOVA tests

on log-likelihood model fit were performed to examine if adding each fixed effect contributes significantly to explaining variance.

Compared to the model with only random effects, adding a single fixed effect of block did not significantly improve model fit (χ^2 (1) = 1.9762, p = .160), but adding trial type (χ^2 (3) = 52.113, p < .001) and the block*trial type interaction (χ^2 (3) = 136.17, p < .001) did improve fit. The quadratic effect of block and the quadratic block:trial type interaction resulted in a significant improvement in fit as well (χ^2(4) = 19.571, p < .001). The quadratic block effect indicated that learning performance improved more rapidly during the middle part of training and was asymptotic towards the end of training, and the interaction suggested that accuracy differences between trial types were the greatest during the middle part of training. The summary of the best-fitting model is shown in Table 5.2. The learning effect was significantly greater in the non-minimal pair trials compared to the minimal pair trials, and that in the tonal trials was the lowest[13]. Overall, the results replicated previous demonstrations that the phonology of native and non-native languages affects word learning from cross-situational statistics.

Table 5.2 *Best fitting model for offline accuracy measure in CSWL, showing fixed effects*

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	1.798	0.139	12.984	< .001 ***
poly(block, 2)1 (linear block effect)	82.073	7.340	11.182	< .001 ***
poly(block, 2)2 (quadratic block effect)	-22.992	5.734	-4.010	< .001 ***
TrialtypeC	-0.604	0.102	-5.945	< .001 ***

[13] This can be observed in Table S2 in Supplementary Material, where tonal minimal pair trial was treated as the reference level in the same model.

TrialtypeT	-1.815	0.137	-13.243	< .001 ***
TrialtypeV	-1.116	0.159	-7.014	< .001 ***
poly(block, 2)1:TrialtypeC	-16.046	7.890	-2.034	.042 *
poly(block, 2)1:TrialtypeT	-71.957	7.404	-9.718	< .001 ***
poly(block, 2)1:TrialtypeV	-49.902	7.629	-6.541	< .001 ***
poly(block, 2)2:TrialtypeC	14.533	7.493	1.940	.052 .
poly(block, 2)2:TrialtypeT	21.071	7.043	2.992	.003 **
poly(block, 2)2:TrialtypeV	18.358	7.209	2.547	.011 *

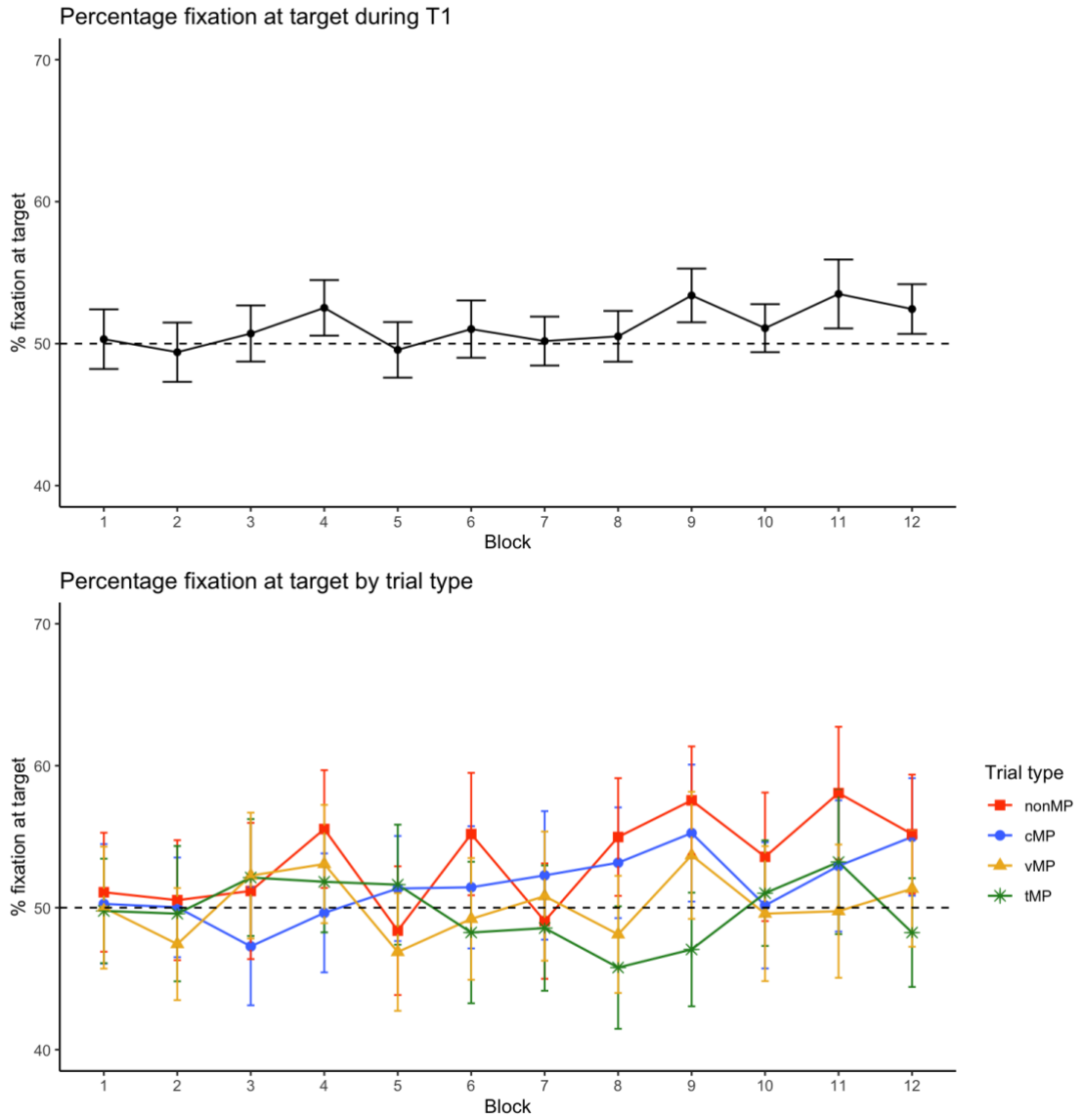
Number of observations: 11695, Participants: 61, Item, 16. AIC = 13233.1, BIC = 13542.5, log-likelihood = -6574.5.

R syntax: glmer(acc ~ poly(block, 2)+ Trialtype+ poly(block, 2):Trialtype+ (1 + block + Trialtype | item) + (1 + block + Trialtype | subjectID), family = binomial).

Eye-tracking measure. Figures 5.4 and 5.5 demonstrate participants' eye fixation at the target object during time interval 1 (during sound play) and time interval 2 (2s window after sound play). During time interval 1, participants' overall fixation at target was consistently above chance after block 11, indicating that participants started to respond to the stimuli during the sound play. When separating the different trial types, we observed that this early response to stimuli happened mainly in the non-minimal pair trials. In minimal pair trials, the percentage fixation at target during sound play remained around chance level. During time interval 2 (2s window after sound play), a clearer fixation pattern can be seen (Figure 5.5). Participants were consistently more likely to fixate at the target object from block 2. They showed greater fixation at target in non-minimal pair and consonantal minimal pair trials, followed by vocalic minimal pair trials. Fixation in tonal trials was around chance throughout the CSWL task. The summary of the mean percentage fixation at target across

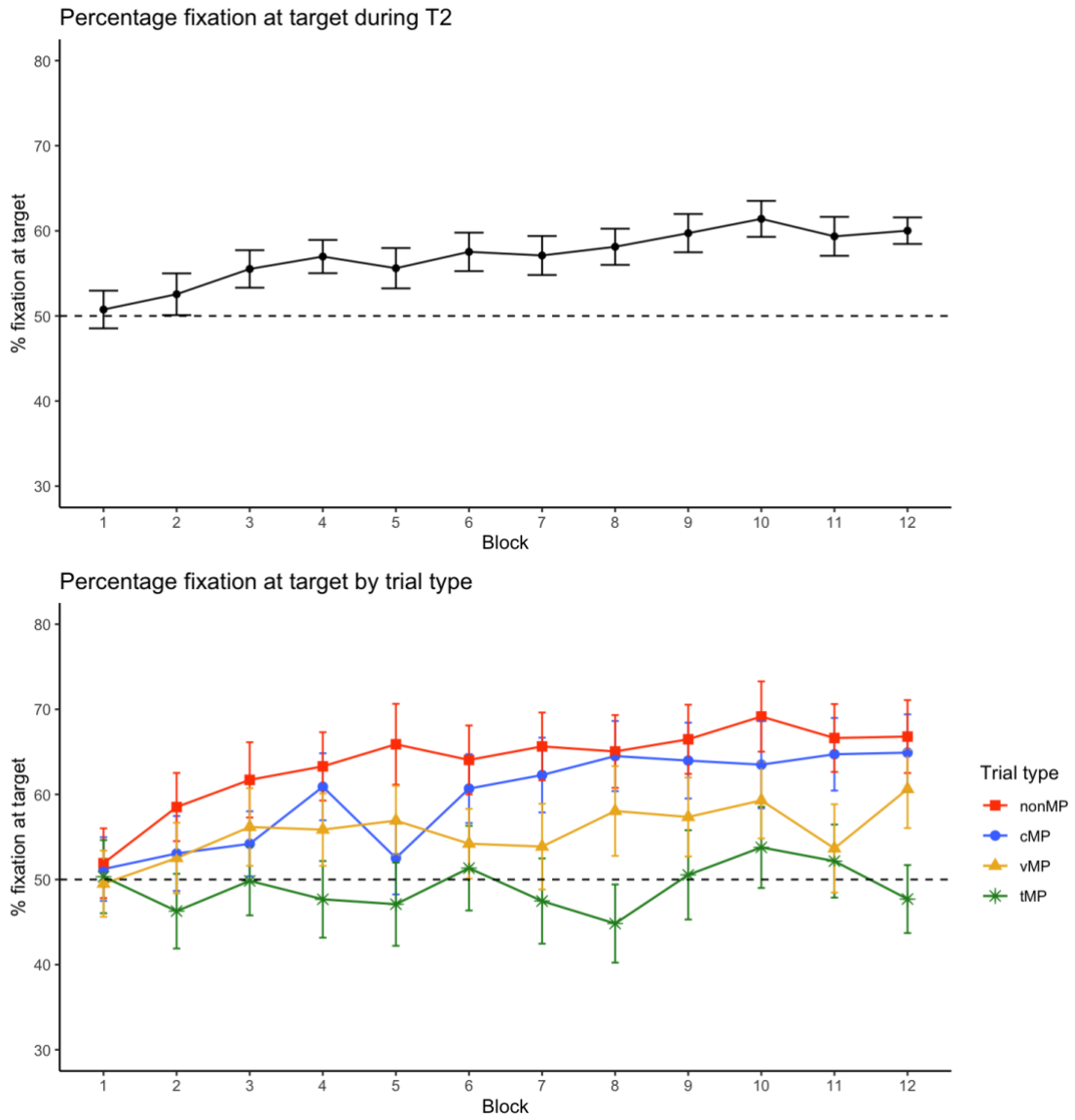
blocks in different trial types is presented in Supplementary Materials Table S5.2 (for time interval 1) and S5.3 (for time interval 2).

Figure 5.4 Percentage fixation at target in each block of the CSWL task during time interval 1.



Note: nonMP refers to the non-minimal pair trials, cMP refers to the consonantal minimal pair trials, vMP refers to the vocalic minimal pair trials, and tMP refers to the tonal minimal pair trials.

Figure 5.5 Percentage fixation at target in each block of the CSWL task during time interval 2.



Note: nonMP refers to the non-minimal pair trials, cMP refers to the consonantal minimal pair trials, vMP refers to the vocalic minimal pair trials, and tMP refers to the tonal minimal pair trials.

The inspection of the eye fixation data revealed a hybrid distribution. There was a large number of data points (42893 out of 57090) where percentage fixation at target equalled either 0 or 1, and the rest of the data points fell between 0 and 1. Given this special

distribution, we filtered the eye fixation dataset into two subsets. The first subset included all data points with either 0 or 1 fixation at target, which was then treated as a binomial distribution and formed the main analysis in line with the preregistration. We ran a binomial generalized linear mixed effects model to test the effect of block, trial type, time interval and the 3-way interaction on participants' fixation at target. The second subset contained data points with continuous percentage fixation at target between 0 and 1, and we conducted a generalized linear mixed effects model with the beta family function. As these results comprised a smaller subset of the data, they are reported in Supplementary Materials.

The model with binomial fixation data showed a significant effect of trial type (χ^2 (3) = 75.628, $p < .001$), time interval (χ^2 (1) = 16.022, $p < .001$), and the 3-way interaction (χ^2 (7) = 64.541, $p < .001$). The effect of block did not significantly improve model fit (χ^2 (1) = 0.5742, $p = .449$). Participants had overall more fixations at the target in non-minimal pair trials than the minimal pair trials, and during time interval 2 compared to time interval 1. The quadratic effect of block and the quadratic block:trial type interaction contributed to better model fit (χ^2 (8) = 25.697, $p = .001$), indicating a larger increase in fixation at target and larger fixation differences between trial types during the intermediate stages of training.

To better explore the 3-way interaction, we ran separate analyses on the two time intervals. Fixation at target during time interval 1 was significantly influenced by block (χ^2 (1) = 5.1744, $p = .023$) and trial type (χ^2 (3) = 24.128, $p < .001$). This indicated a general increase in fixation at target during time interval 1. Additionally, fixation at target in non-minimal pair trials and consonantal minimal pair trials was significantly higher than that in vocalic and tonal minimal pair trials. The effect of the block:trial type interaction (χ^2 (3) = 7.4302, $p = .059$) did not further improve model fit, nor did the quadratic block effect and the quadratic block:trial type interaction (χ^2 (4) = 2.0278, $p = .731$). Table 5.3 summarizes the best-fitting model.

For time interval 2, we observed significant effects of trial type (χ^2 (3) = 75.361, p < .001),the block:trial type interaction (χ^2 (3) = 53.42, p < .001), and the quadratic block:trial type effect (χ^2 (4) = 25.968, p < .001). As reflected by the model summary in Table 5.4, the increase (across blocks) in fixation at target was greater in non-minimal pair and consonantal minimal pair trials compared to that in vocalic and tonal trials.

Table 5.3 Best fitting model for fixation at target during time interval 1 in CSWL, showing fixed effects

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	0.272	0.105	2.597	.009 **
block	0.021	0.010	2.210	.027 *
TrialtypeC	-0.110	0.103	-1.067	.286
TrialtypeT	-0.456	0.101	-4.517	< .001 ***
TrialtypeV	-0.323	0.104	-3.108	.002 **

Number of observations: 3736, Participants: 61, Item, 16. AIC = 5111.9, BIC = 5155.5, log-likelihood = -2548.9.

R syntax: glmmTMB(fixation ~ block + Trialtype + (1 | item) + (1 | subjectID), family = binomial).

Table 5.4 Best fitting model for fixation at target during time interval 2 in CSWL, showing fixed effects

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	0.652	0.057	11.493	< .001 ***
poly(block, 2)1 (linear block effect)	37.454	6.143	6.097	< .001 ***

poly(block, 2)2 (quadratic block effect)	-19.708	4.279	-4.606	< .001 ***
TrialtypeC	-0.160	0.041	-3.879	< .001 ***
TrialtypeT	-0.676	0.056	-11.992	< .001 ***
TrialtypeV	-0.353	0.057	-6.148	< .001 ***
poly(block, 2)1:MPtypeC	10.523	6.411	1.641	.101
poly(block, 2)1:MPtypeT	-30.683	6.191	-4.956	< .001 ***
poly(block, 2)1:MPtypeV	-17.276	6.274	-2.753	.006 **
poly(block, 2)2:MPtypeC	11.419	6.024	1.896	.058 .
poly(block, 2)2:MPtypeT	23.464	5.893	3.982	< .001 ***
poly(block, 2)2:MPtypeV	18.056	5.945	3.037	.002 **

Number of observations: 39103, Participants: 61, Item, 16. AIC = 51509.7, BIC = 51766.9, log-likelihood = -25724.8.

R syntax: glmmTMB(fixation ~ poly(block, 2)+ Trialtype + poly(block, 2):Trialtype + (1 + block | item) + (1 + block + Trialtype | subjectID), family = binomial).

Individual differences in word learning and auditory processing ability

For the pitch discrimination test, a pitch discrimination score was computed for each participant. The average score was 15.37 (SD = 14.97, range = 3.4-75.5). Since the pitch discrimination scores were highly skewed, we used log-transformed pitch scores in the following analyses. For the melody reproduction test, we calculated the performance accuracy, with an average of 0.58 (SD = 0.24, range = 0.1-1.0), and used the raw accuracy values in the statistical models.

Auditory processing and the accuracy measure. To examine whether pitch discrimination and melody reproduction abilities predicted accuracy at the end of CSWL, and whether they predicted performance in any specific trial types, we entered pitch

discrimination score and melody reproduction accuracy (together with their respective interaction with trial type) as fixed effects in the mixed effects models. There was a significant main effect of pitch discrimination (χ^2 (1) = 4.7608, p = .029), but the interaction between trial type and pitch discrimination (χ^2 (3) = 1.9005, p = .593) was not significant. Table 5.5 presents the best-fitting model, illustrating that pitch discrimination score was negatively associated with overall accuracy in CSWL. Since a lower pitch discrimination score means better sensitivity to pitch changes, it indicated that participants who were better at discriminating trivial pitch differences showed higher accuracy at the end of CSWL. As for the melody reproduction measure, adding the effect of melody (χ^2 (1) = 6.0858, p = .014) improved model fit, but not the trial type:melody interaction (χ^2 (3) = 5.8149, p = .121). Table 5.6 shows the best-fitting model for the melody reproduction effect, suggesting a positive relationship between melody reproduction accuracy and picture selection accuracy in the CSWL task.

Table 5.5 Best fitting model for accuracy measure in CSWL, testing pitch discrimination effect

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	7.3690	1.1097	6.641	< .001 ***
TrialtypeC	-4.0027	1.0570	-3.787	< .001 ***
TrialtypeT	-6.7150	1.0701	-6.275	< .001 ***
TrialtypeV	-5.3585	1.0789	-4.966	< .001 ***
Log_pitch discrimination	-0.7228	0.3315	-2.180	.029 *

Number of observations: 976, Participants: 61, Item, 16. AIC = 980.6, BIC = 1102.7, log-likelihood = -465.3.

R syntax: glmer(accuracy ~ Trialtype + log_pitch discrimination + (1 + Trialtype | item) + (1 + Trialtype | subjectID), family = binomial).

Table 5.6 Best fitting model for accuracy measure in CSWL, testing melody reproduction effect

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	5.904	1.061	5.564	< .001 ***
TrialtypeC	-4.011	1.039	-3.863	< .001 ***
TrialtypeT	-6.630	1.055	-6.287	< .001 ***
TrialtypeV	-5.288	1.064	-4.972	< .001 ***
Melody reproduction	1.056	0.412	2.565	.010 *

Number of observations: 976, Participants: 61, Item, 16. AIC = 979.3, BIC = 1101.3, log-likelihood = -464.6.

R syntax: glmer(accuracy ~ Trialtype + melody reproduction + (1 + Trialtype | item) + (1 + Trialtype | subjectID), family = binomial).

Auditory processing and the eye-tracking measure. Similarly, we tested whether pitch discrimination and melody reproduction abilities predicted fixation at target[14] at the end of CSWL. The main effect of pitch discrimination did not lead to a significant improve in model fit (χ^2 (1) = 1.9856, p = .159), but the trial type:pitch discrimination interaction was significant (χ^2 (3) = 9.1759, p = .027). Table 5.7 shows the best-fitting model, demonstrating that when taking into account the different trial types, pitch discrimination score was negatively associated with fixation at target. That is, better pitch discrimination predicted

[14] The following statistical analyses reported were based on the binomial subset of the eye fixation data. The same analyses were carried out with the continuous subset of the eye fixation data, but no significant effect was observed.

greater fixation at the target object at the end of CSWL. Moreover, the association between pitch discrimination and fixation at target was greater in non-minimal pair and consonantal trials compared to vocalic and tonal trials (as can be observed in Figure 5.6).

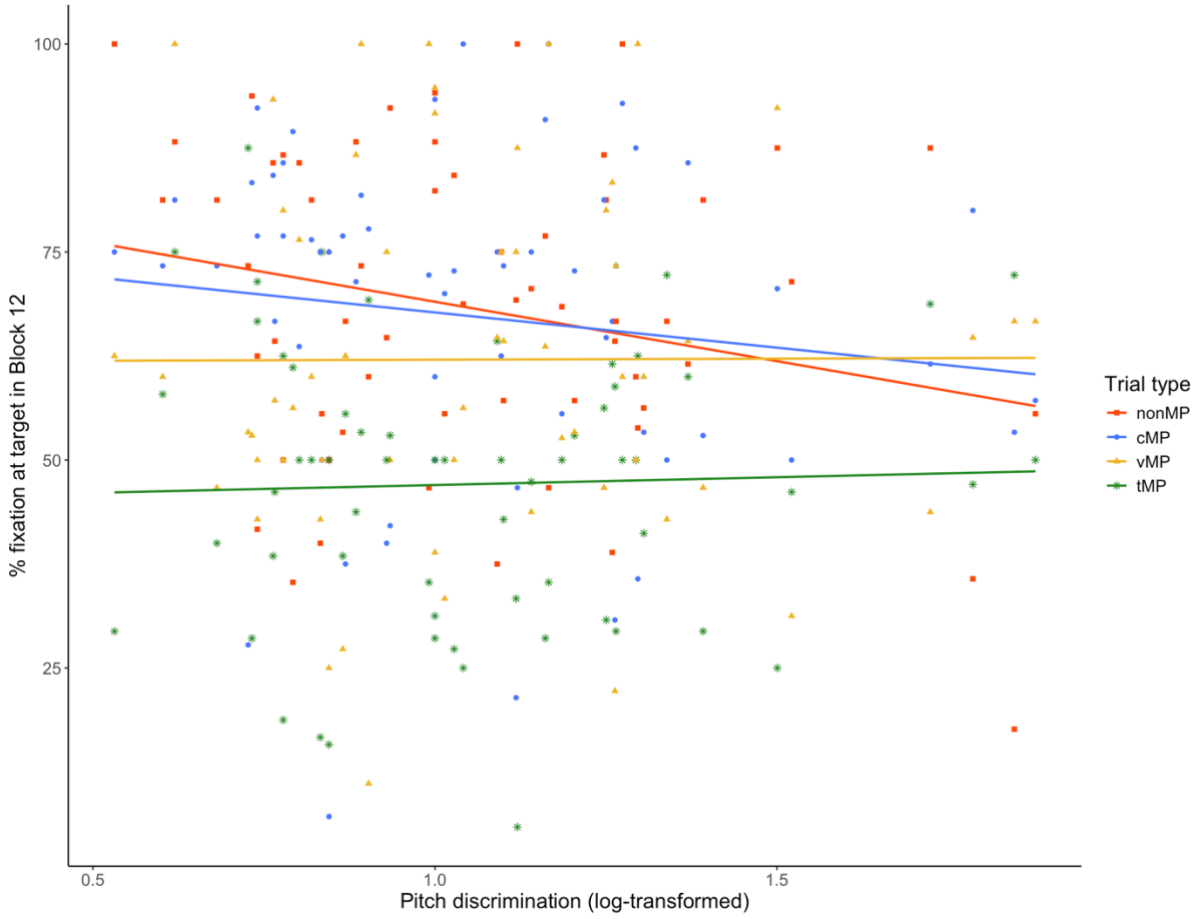
Table 5.7 *Best fitting model for fixation at target in CSWL, testing pitch discrimination effect*

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	1.571	0.304	5.176	< .001 ***
TrialtypeC	-0.386	0.379	-1.019	.308
TrialtypeT	-1.826	0.361	-5.062	< .001 ***
TrialtypeV	-1.049	0.366	-2.864	.004 **
Log_pitch discrimination	-0.741	0.263	-2.819	.005 **
TrialtypeC: Log_pitch discrimination	0.327	0.333	0.980	.327
TrialtypeT: Log_pitch discrimination	0.842	0.318	2.648	.008 **
TrialtypeV: Log_pitch discrimination	0.775	0.324	2.395	.017 *

Number of observations: 3689, Participants: 61, Item, 16. AIC = 4778.7, BIC = 4840.8, log-likelihood = -2379.3.

R syntax: glmmTMB(fixation ~ Trialtype + Log_pitch discrimination + Trialtype: Log_pitch discrimination + (1 | item) + (1 | subjectID), family = binomial).

Figure 5.6 Relationship between pitch discrimination and fixation at target in the final block of CSWL.



Similar analyses of the eye fixation data revealed significant effects of melody reproduction accuracy (χ^2 (1) = 5.7611, p = .016) and trial type:melody interaction (χ^2 (3) = 12.888, p = .005). More accurate melody reproduction was associated with greater fixation at target in CSWL, especially in the non-minimal pair and consonantal minimal pair trials. Table 5.8 summarizes the best-fitting model and Figure 5.7 visualizes the interaction effect.

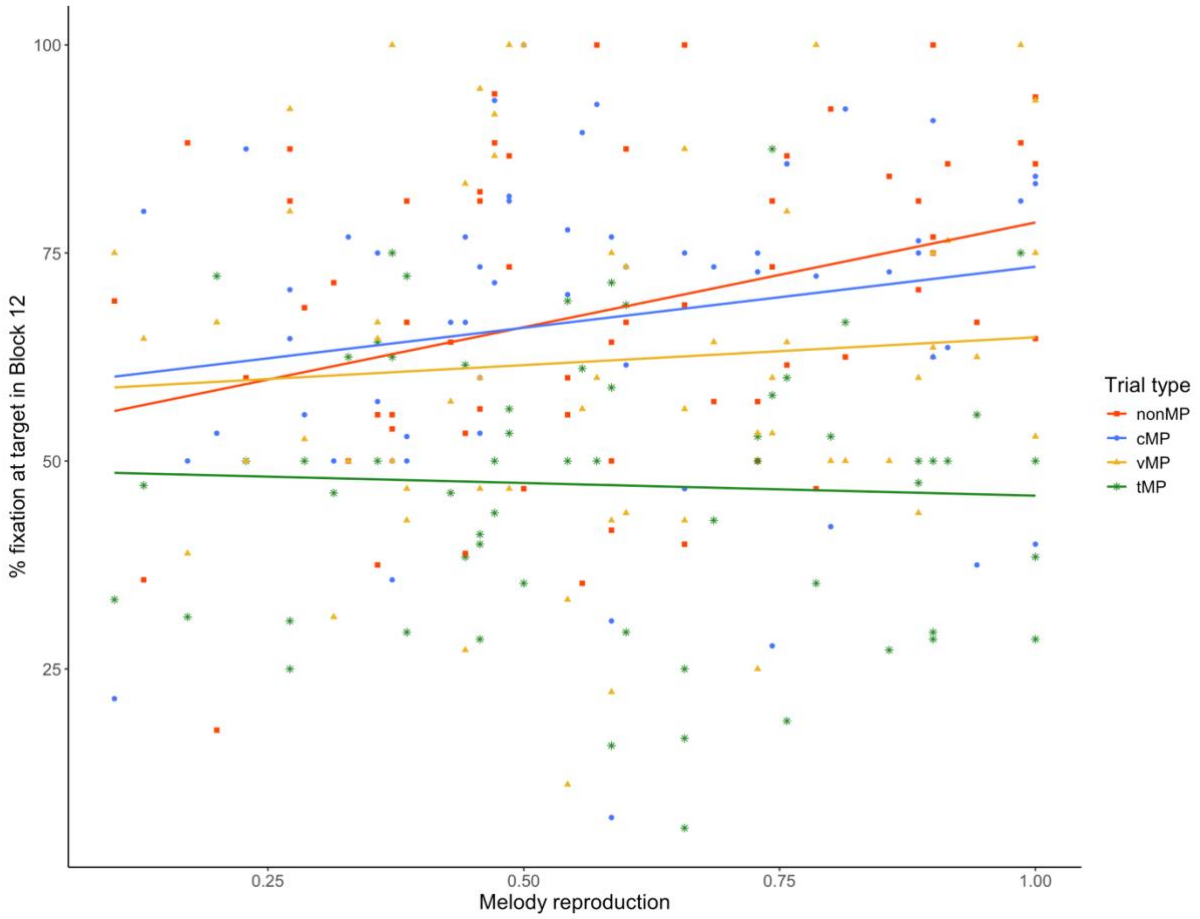
Table 5.8 Best fitting model for fixation at target in CSWL, testing melody reproduction effect

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	0.062	0.216	0.287	.774
TrialtypeC	0.238	0.266	0.896	.370
TrialtypeT	-0.115	0.259	-0.443	.658
TrialtypeV	0.347	0.265	1.310	.190
Melody reproduction	1.248	0.339	3.680	< .001 ***
TrialtypeC:melody reproduction	-0.476	0.433	-1.099	.272
TrialtypeT:melody reproduction	-1.405	0.419	-3.357	< .001 ***
TrialtypeV:melody reproduction	-0.988	0.422	-2.340	.019 *

Number of observations: 3689, Participants: 61, Item, 16. AIC = 4771.2, BIC = 4833.3, log-likelihood = -2375.6.

R syntax: glmmTMB(fixation ~ Trialtype + melody reproduction + Trialtype:melody reproduction + (1 | item) + (1 | subjectID), family = binomial).

Figure 5.7 Relationship between melody reproduction and fixation at target in the final block of CSWL.



Exploratory analysis. Since we were particularly interested in English-native participants' performance in the tonal minimal pair trials, we ran sub-analyses to test the effect of pitch discrimination and melody reproduction abilities in tonal trials. These analyses were pre-registered as exploratory analyses because they were additional to the critical analyses explained above. However, neither pitch discrimination (for accuracy χ^2 (1) = 0.7476, p = .387; for target fixation χ^2 (1) = 0.0012, p = .972) nor melody reproduction (for accuracy χ^2 (1) = 0.2273, p = .634; for target fixation χ^2 (1) = 0.0205, p = .886) predicted performance in tonal trials.

Furthermore, we tested an alternative model as exploratory, where participants' pitch discrimination and melody reproduction scores were coded as ‘normative’ versus ‘low’ based on the corresponding test thresholds. According to Saito and Tierney’s (forthcoming) large-scale ongoing research, the threshold score for the pitch discrimination task was 16.69. Participants who scored below 16.69 were considered normative listeners and those above 16.69 were considered having low pitch precision. For the melody reproduction test, an accuracy of 62.78% was the threshold performance, with participants performing above the threshold being the ‘normative’ group and below the threshold being the ‘low’ performance group. We then compared whether there were any group differences in word learning performance. This is because some previous studies (Perrachione et al., 2011; Ruan & Saito, 2023) have suggested that the effect of auditory processing could be dichotomous (normative vs. low) rather than continuous and that the predictive power of auditory processing could be more clearly observed especially among individuals with relatively low auditory processing (i.e., a lack of auditory precision can hinder learning).

When testing on the accuracy measure, we did not observe any pitch group difference or an interaction between trial type and pitch group, but there was a significant main effect of melody group (χ^2 (3) = 4.6329, p = .031). For the eye fixation measure, we found a significant interaction between trial type and melody group (χ^2 (3) = 12.943, p = .005), whereas the pitch group effect was again not significant. These results indicated that participants who performed below the melody test threshold showed greater difficulty (i.e., lower accuracy and fixation at target) in CSWL. However, it is worth noting that even the below-threshold melody groups performed above chance at the end of the CSWL (in t-test against chance level for accuracy, t = 10.229, p < .001; for target fixation, t = 9.3984, p < .001).

Tonal discrimination ability

We calculated participants' accuracies in the tonal discrimination tests before and after the CSWL task. We considered only the trials with stimuli differing in tones (e.g., tu1, tu1, tu2), as they reflected the ability to discriminate between tonal minimal pairs. Participants had overall high accuracy in distinguishing the tonal differences in both tonal discrimination tests (average accuracy 0.937 pre-CSWL, average accuracy 0.950 post-CSWL). The difference between the two tonal discrimination tests was found to be significant (χ^2 (1) = 8.7201, p = .003).

We further explored whether participants' tonal discrimination ability before the CSWL task predicted word learning outcomes. Results suggested that tonal discrimination accuracy predicted overall fixation at target in the final block of CSWL (χ^2 (1) = 4.353, p = .037), but not picture selection accuracy (χ^2 (1) = 1.1863, p = .276). Thus, better tonal discrimination ability was associated with greater fixation at target at the end of CSWL. When taking into account tonal trials in particular, we observed no significant effect of tonal discrimination on neither fixation at target (χ^2 (1) = 0.0237, p = .878) nor picture selection accuracy (χ^2 (1) = 0.1854, p = .667).

Retrospective verbal reports

Based on the debriefing questions, we coded participants' awareness of the tonal feature in the CSWL task and tested whether tonal awareness predicted word learning outcomes. The coding of awareness followed Rebuschat et al.'s (2015) and Ge et al.'s (in press) scheme. Participants who mentioned the use of tones/pitch/intonation differences to distinguish words as a strategy (Q1~3) were considered developing 'full awareness'. Participants who mentioned noticing the tones/pitch/intonation when being asked about the patterns of the language and the sound system were coded as 'partial awareness' (Q4~5).

Those who only responded 'yes' to the direct questions on the presence of tonal feature (Q6~7) were deemed 'minimal awareness'. The rest of the participants who noted that they thought there was no use of tone/pitch/intonation were coded as 'unaware'.

According to this coding scheme, five participants developed full awareness. They reported their strategy as, for example, "*started by guessing…I think a few were very close tonally to each other*". Fifteen participants were partially aware of the tonal cues, reporting the presence of different sound patterns from English (e.g., *It is different from English in that there are lots of similar sounding words with just a small tonal difference…*). Another 36 participants reported minimal noticing of the tonal feature (e.g., *It seemed like the tone changed the word meaning on some of them…*). Five participants were considered unaware of the tones. Since the number of unaware participants was low, we did not carry out a dichotomous comparison between aware and unaware participants. Instead, we compared whether different levels of awareness (i.e., unaware, minimal, partial, full, coded as an ordinal variable) influenced word learning outcomes.

We first constructed models with fixed effects of awareness status, trial type and the interaction on the picture selection accuracy at the end of CSWL. The single effect of awareness status (χ^2 (1) = 4.2524, p = .039) significantly improved model fit, but not the interaction between trial type and awareness (χ^2 (9) = 7.2011, p = .066). Table 5.9 presents the best-fit model, where awareness positively affected CSWL accuracy. This suggested that participants with higher level of tonal awareness performed more accurately in picture selection at the final block of CSWL (as shown in Figure 5.8). Similarly, we ran models on fixation at target during time interval 1 and 2 separately, but none of the awareness effects or the trial type:awareness interactions were significant.

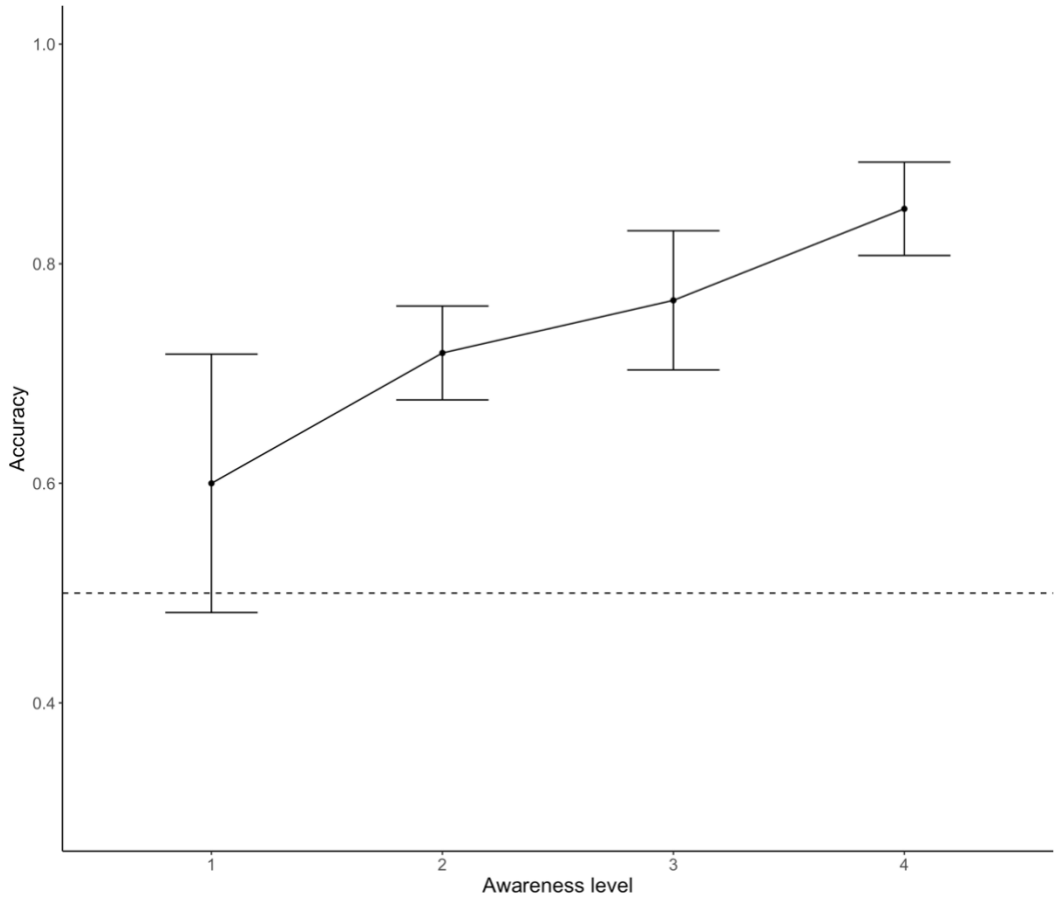
Table 5.9 Best fitting model accuract in CSWL, testing awareness effect

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	5.372	1.118	4.804	< .001 ***
Awareness	0.446	0.165	2.695	.007 **
TrialtypeC	-4.00	1.018	-3.928	< .001 ***
TrialtypeT	-6.522	1.029	-6.338	< .001 ***
TrialtypeV	-5.132	1.041	-4.931	< .001 ***

Number of observations: 3326, Participants: 61, Item, 16. AIC = 983.4, BIC = 1129.9, log-likelihood = -461.7.

R syntax: glmer(acc ~ awareness + Trialtype + (1 + awareness + Trialtype | item) + (1 + Trialtype | subjectID), family = binomial).

Figure 5.8 Mean proportion of correct pictures selected in the final block of the CSWL task by different awareness levels.



Note: 1 = unaware, 2 = minimal awareness, 3 = partial awareness, 4 = full awareness.

Discussion

The present study demonstrated that adults could learn non-native words via a short cross-situational exposure, as reflected by their word-referent mapping accuracy and their eye fixation during referent presentation. However, their word learning performance was greatly affected by the phonological properties of the words. Our results supported previous evidence that the presence of minimal pairs, especially those contrasting in a non-native feature, reduced learning outcomes (e.g., Escudero et al., 2016; Ge et al., in press). One critical finding of our study was that this cross-situational learning of non-native words was

modulated by learners' domain-general auditory processing ability and their domain-specific perception of the non-native contrasts.

RQ1: Do learners' auditory processing ability predict cross-situational learning of non-native words? Our investigation of the individual differences in participants' word learning outcomes revealed a significant role of auditory processing ability. Participants' sensitivity to pitch variations in non-linguistic sounds was, in general, associated with their cross-situational learning of novel tonal words. We employed two pitch-related measures of auditory processing ability, a pitch discrimination for the perceptual aspects of auditory processing and a melody reproduction measure for the cognitive and motoric aspects of auditory processing. The pitch discrimination task assessed participants' auditory acuity over fine-grained pitch differences via acoustic discrimination of non-linguistic sounds, whereas the melody reproduction task tested how well participants could listen to and reproduce a short melody, which was a reflection of audio-motor integration skills.

We found that participants' overall picture selection accuracy at the end of the CSWL task was associated with their performance in the pitch discrimination and melody reproduction tests. Better acoustic discrimination of pitch changes and audio-motor integration skills predicted higher accuracy in word-picture mapping. However, the auditory processing measures did not interact with any specific trial type, indicating a general relationship with statistical word learning (when taking into account the accuracy measure of learning outcomes). For the eye fixation measure, on the other hand, we observed interactions between auditory processing ability and trial type. Specifically, pitch discrimination and melody reproduction predicted fixation at target in the non-minimal pair and consonantal minimal pair trials to a greater extent than in the vocalic and tonal minimal pair trials. More precise auditory processing was linked to greater fixation at the target referent at the end of CSWL. Overall, these findings provided evidence that individual differences in auditory

processing of a specific acoustic dimension (e.g., F0) were linked to the acquisition of non-native speech sounds that utilize this specific dimension. The dimension-specific relation between auditory processing and learning agrees with Saito et al.'s (2022) observations, where Japanese-native speakers' proficiency on the English [r]-[l] contrast was associated with their individual sensitivity to the critical F3 and F2 acoustic dimensions.

It is important to note that although the relationships between auditory processing and the two word learning measures were generally consistent, only the eye fixation measure demonstrated an interaction with different trial types. This might indicate that the eye fixation measure was more sensitive in capturing both the individual differences in word learning performance (between-participant variation) and individual participant's different performance across trial types (within-participant variation). When considering the critical tonal minimal pair trials, however, we did not find any relationship between auditory processing and tonal minimal pair learning outcomes. One possible explanation is that in our study, participants were naïve to Mandarin tones and did not show much learning in the tonal minimal pair trials in general. Thus, the variations in learning performance in tonal trials might be too small to show a significant relationship with any potential predictor. In the non-minimal pair and consonantal minimal pair trials, on the contrary, participants exhibited an overall learning effect, and hence, it was more likely to observe individual differences in learning gains.

Furthermore, the threshold-based classification of pitch discrimination and melody reproduction groups (i.e., above or below the respective auditory processing test threshold) provided results which largely aligned with the continuous analyses of the auditory processing measures. The pitch groups did not differ significantly in word learning performance, but the below-threshold melody group did show a significantly lower level of fixation at target and picture selection accuracy than the above-threshold group. However,

even the below-threshold melody group showed above-chance performance at the end of CSWL. These findings indicated that auditory processing ability did not necessarily set a threshold for successful statistical word learning. Rather, it could be considered an important predictor of learning outcomes for all learners of various auditory abilities. This seems to contradict previous findings where auditory processing was found to be more dichotomous than continuous (e.g., Perrachione et al., 2011; Ruan & Saito, 2023). One potential explanation is that, in these previous studies, the learning/training paradigms were more explicit, with instructions and feedback provided. However, in our design, participants underwent a more implicit learning session. It is possible that the nature of the auditory processing factor depends on the type of treatment. In the explicit treatment methods, there might be a threshold below which it would be difficult for learners to process the explicit input/information and hence result in categorical performance, whereas in the implicit methods, such threshold might not be crucial.

To summarize, our findings suggested that auditory processing profiles could explain and predict (at least partially) learners' ability to acquire non-native words through statistical tracking. This insight could help detect learners who may benefit less from the statistical learning approach. However, having precise or normative auditory processing alone may not guarantee successful learning outcomes, as auditory processing is only one of many predictors of individual differences in language learning. Furthermore, the link between auditory ability and learning can be dimension-specific. The sensitivity to pitch variations, for instance, was related to the acquisition of tonal words that rely on pitch contours to distinguish meanings. This observation can be explained from the L2 speech learning perspective according to the Second Language Linguistics Perception (L2LP) framework (Escudero, 2005; Van Leussen & Escudero, 2015). In the case of English-native speakers learning tonal words, pitch is a critical acoustic dimension that is not categorized in the

learners' native language at all. Learners need to form new boundaries and eventually categories along the new dimension during the learning process. Thus, the domain-general sensitivity to the pitch may contribute to and facilitate the categorization along the pitch dimension in language learning.

RQ2: Can learners perceive and discriminate between tonal differences before and after cross-situational learning of tonal words? To better understand whether participants' difficulty in tonal trials was resulted from the lower-level acoustic perception of lexical tones, we tested participants' perceptual discrimination of lexical tones before and after the CSWL task. Results illustrated that in the discrimination tasks, participants were able to distinguish the tonal differences at a high accuracy rate even before the CSWL exposure. This implies that the lack of learning effect in tonal minimal pair trials was not merely a perceptual issue. Instead, participants might face challenges at the phonological level because they lack the appropriate representation of the tonal categories.

Moreover, we observed a relationship between participants' tonal discrimination before CSWL and their tonal word learning outcomes. Interestingly, this relationship was only found with the eye fixation measure of learning performance but not the accuracy measure. That is, better pre-CSWL tonal discrimination ability was associated with greater fixation at the target referent at the end of CSWL. This again suggested that the eye fixation measure may be more sensitive in recording participants' performance variations, as it did not involve participants' conscious or explicit responses. The current finding also provided evidence towards the phonetic-phonological-lexical continuity, where the perception of sounds was associated with higher level word learning ability (e.g., Chandrasekaran et al., 2010; Silbert et al., 2015; Wong & Perrachione, 2007). It is also worth noting that we did not observe any relationship between pre-CSWL tonal discrimination and participants' performance in tonal minimal pair trials specifically, indicating that the perceptual

discrimination of the tonal contrasts was not directly linked to the use of the contrasts in acquiring minimal pair words. This may be because the lexical encoding of tonal minimal pair words required more precise phonological representations of the tonal categories, but the auditory discrimination ability might not necessarily transfer to the formation of the required tonal categories during the short implicit learning process.

RQ3: Do online eye-tracking and offline accuracy measures show similar learning performance patterns in CSWL? To better capture participants' learning performance, we employed two measures of learning: online eye fixation and offline picture selection accuracy. Previous research on cross-situational word learning typically examined learners' knowledge via forced-choice tasks where participants mapped newly acquired (pseudo)words to referent pictures (e.g., Escudero et al., 2022; Yu & Smith, 2007). In the present study, we tested whether keeping track of participants' eye fixation throughout learning provided a more sensitive account of learning performance, especially in the more challenging minimal pair trials. It also enabled us to understand participants' online processing of words.

Overall, our results indicated that both the eye fixation and the picture selection measures successfully represented participants' learning patterns in the CSWL task. The two measures were largely consistent, demonstrating that learners performed better in non-minimal pair trials as compared to the minimal pair trials. Participants showed higher accuracy and greater percentage fixation at the target picture in non-minimal pair trials. The greater fixation at target reflected less confusion and hence fewer attention shifts between the two pictures, which in turn led to more accurate word-referent mapping (Yu & Smith, 2011; Yu et al., 2012). In contrast, participants' accuracy and fixation at target in the tonal minimal pair trials remained around chance and did not increase throughout the task. We observed from the online eye fixation measure that when hearing a word (e.g., pa1mi1) and seeing two pictures that were mapped to two tonal minimal pair words (e.g., pa1mi1 and pa4mi),

participants were more likely to look at both pictures throughout the stimuli presentation, suggesting greater uncertainty about the correct referent. The two measures consistently revealed the learning difficulty associated with similar-sounding words that contrast in a non-native phonological feature. It indicated that the lack of improvement in picture selection accuracy in the tonal trials did not result from the insensitivity of the offline measure but from the perceptual difficulty of tones.

However, the accuracy and eye fixation measures showed slightly different patterns when comparing the consonantal and vocalic minimal pair trials with the non-minimal pair trials. The accuracy measure showed that the proportion of correct responses in consonantal and vocalic trials over the CSWL blocks was significantly lower than that in the non-minimal pair trials, whereas the eye fixation results suggested that the learning trajectory in consonantal trials was comparable to that in non-minimal pair trials and higher than that in vocalic trials. In other words, the eye fixation measure seemed to show better learning performance in consonantal trials than the accuracy measure. This divergence between measures might reflect a transition between implicitly acquired information and more explicit knowledge. Participants gathered information implicitly from the repeated exposure in the CSWL task, as no instruction or feedback was provided. This information included not only the word-picture mappings, but also the fact that many words 'sounded similar' and only differed in one phoneme, either the consonant, the vowel or the tone. However, the conversion of such implicitly learned information to explicit knowledge might have different levels of difficulty. Once a few word-picture mappings were formed, participants could easily make conscious, strategic use of this knowledge to eliminate the incorrect answers in the non-minimal pair trials, as they could rely on several phonological cues to make the decisions. Meanwhile, in the minimal pair trials, only one cue was reliable and it might be harder for participants to explicitly represent the trivial phonological difference and reject the incorrect

answer in these situations. This might explain why the accuracy measure showed overall reduced learning in (consonantal and vocalic) minimal pair trials compared to non-minimal pair trials, as the explicit responses could depend more on explicitly available knowledge, but the knowledge of the precise minimal pairs was more challenging to be represented explicitly.

The eye fixation measure, on the contrary, did not rely on any explicit responses and hence might better reflect the implicit knowledge of participants. As mentioned above, the fixation at target in the consonantal trials was similar to that in the non-minimal pair trials, and better than that in vocalic trials. This pattern was in line with previous findings, where participants performed better in word-referent mapping of consonantal and non-minimal pairs than vocalic minimal pairs (Escudero et al., 2016; Mulak et al., 2019). There are several possibilities of this consonant-vowel learning difference. Firstly, as proposed by Nespor et al. (2003), consonants and vowels have different primary functions in language learning and processing, with consonants being more closely related to meanings. Thus, participants might pay more attention to the consonantal differences when mapping words to meanings. Secondly, in our stimuli, the consonantal contrasts were always in a prominent, word-initial position. Previous research has observed a stronger impact of word onset on lexical access compared to other segments of the words (e.g., Marslen-Wilson & Zwitserlood, 1989). Therefore, the stronger learning effect observed in consonantal than vocalic minimal pair trials in our design might come from participants' greater attention to the word onsets. Thirdly, neurophysiological studies have found that brain responses to consonant processing were more distinct from lexical tone processing, whereas the processing of vowels and lexical tones were more similar (Choi et al., 2017; Lee et al., 2012; Luo et al., 2006; Tong et al., 2014). This indicates that participants' processing and learning of the consonantal contrasts

might be less affected by the non-native tonal cues, and hence yielded better learning outcomes.

Awareness effect. The examination of participants' awareness of the tonal cue demonstrated that being more aware of the presence of tones was associated with higher overall accuracy at the end of CSWL, but awareness was not related to fixation at target. Additionally, there was not a specific association between tonal awareness and the tonal minimal pair learning. These findings suggested that increased awareness of the critical language feature may facilitate learning in general (e.g., Monaghan et al., 2019), and the awareness effect may be more obvious when learners need to make explicit responses to the stimuli (i.e., picture selection). This seems to diverge from previous observations where being aware of a non-native phonological feature was not directly linked to better use of the feature in distinguishing and learning novel words (Ge et al., in press; under review). This difference in the awareness effect may arise from the greater tonal exposure in the current experiment. For example, in Ge et al. (in press), participants only experienced the CSWL task with tonal words, whereas in this study, we presented a tonal discrimination test before the CSWL task. The extra tonal discrimination test may guide learners' attention to the tonal feature and lead to higher and more diverse levels of tonal awareness. This may explain why more learners developed partial and full awareness in our study than in Ge et al. (in press).

Limitations and further directions

Our current findings revealed a general link between auditory processing of pitch variations and the acquisition of non-native tonal words. However, we did not find a direct relationship between pitch processing and the acquisition of the tonal minimal pair words, possibly because our learners were naïve to tonal cues and showed overall little learning in tonal trials. For future studies, it would be interesting to recruit learners with different levels

of tonal experience (e.g., with different years of Mandarin learning) and examine whether the more advanced learners' performance on tonal minimal pairs is related to the auditory processing of pitch changes. Moreover, to better understand the dichotomous versus continuous nature of the auditory processing variable, future studies can directly compare explicit and implicit treatment methods (i.e., via a between-subject manipulation) and see if auditory processing ability predicts performance in explicit and implicit learning conditions differently.

In terms of the CSWL design, our stimuli included a simple structure of CVCV pseudowords, which might contribute to the better learning in the word-initial consonantal contrasts. Future research could employ different word structures, including vowel-initial ones, to provide a more representative view of the relative learning difficulties associated with non-native consonantal and vocalic contrasts. Another potential follow-up based on the current design is to introduce speaker variability to the pseudoword stimuli. Although previous research did not find a significant role of speaker variability in cross-situational learning of novel words from participants' native language (e.g., Crespo & Kaushanskaya, 2021; Crespo et al., 2024), it is worth investigating whether greater input variability facilitates CSWL of non-native words that contain unfamiliar phonological contrasts. Additionally, since tonal features can be largely influenced by speakers' F0 range in reality, it is interesting to explore whether and how learners deal with within-speaker and between-speaker F0 variations in CSWL. Moreover, speaker variability can be incorporated into the tonal discrimination task as well to examine whether the ability to discriminate between-speaker tonal productions predicts tonal word learning outcomes.

Conclusion

The current study demonstrated how individual differences in auditory processing influenced statistical learning of non-native words. In particular, learners with better auditory processing of pitch variations showed better learning of non-native tonal words. It provided evidence that statistical learning of non-native words can be modulated by domain-general sensitivity to the specific acoustic dimension involved in the non-native words. Our study has important implications for both individual difference research and non-native vocabulary learning practice. It helps us understand which factors predict successful non-native word learning from contextual exposure, which may further facilitate deciding on the appropriate training types (e.g., explicit vs. contextual) for different learners. It also contributes to the validation of auditory processing skills as a composite of language aptitude in second language acquisition.

Furthermore, the online and offline measurements of learning performance allowed for a more comprehensive representation of participants' learning behaviours and outcomes. Our results indicated that web-based eye-tracking techniques can be used reliably in picture-word mapping paradigms. And the implicit online measures may be vital for detecting subtle differences in how learners use information during language learning.

Supplementary materials

Table S5.1 Mean accuracy and standard deviations across the 12 blocks of the CSWL task in different trial types.

Trial type	Block											
	1	2	3	4	5	6	7	8	9	10	11	12
Overall	0.50 (0.16)	0.57 (0.11) ***	0.61 (0.12) ***	0.65 (0.11) ***	0.63 (0.10) ***	0.68 (0.09) ***	0.70 (0.09) ***	0.70 (0.09) ***	0.72 (0.11) ***	0.73 (0.11) ***	0.73 (0.10) ***	0.73 (0.09) ***
nonMP	0.56 (0.31)	0.68 (0.24) ***	0.73 (0.22) ***	0.79 (0.17) ***	0.81 (0.19) ***	0.88 (0.16) ***	0.89 (0.14) ***	0.88 (0.17) ***	0.91 (0.14) ***	0.91 (0.13) ***	0.89 (0.12) ***	0.92 (0.17) ***
cMP	0.49 (0.28)	0.59 (0.22) ***	0.60 (0.31) **	0.71 (0.22) ***	0.67 (0.23) ***	0.74 (024) ***	0.76 (0.20) ***	0.77 (0.20) ***	0.79 (0.21) ***	0.79 (0.24) ***	0.83 (0.18) ***	0.83 (0.22) ***
vMP	0.47 (0.26)	0.58 (0.24) **	0.61 (0.23) ***	0.60 (0.23) ***	0.59 (0.24) **	0.61 (0.25) ***	0.66 (0.26) ***	0.69 (0.23) ***	0.66 (0.24) ***	0.66 (0.27) ***	0.67 (0.25) ***	0.71 (0.23) ***
tMP	0.48 (0.28)	0.43 (0.29)	0.48 (0.26)	0.49 (0.24)	0.43 (0.19)	0.49 (0.26)	0.50 (0.28)	0.46 (0.25)	0.52 (0.26)	0.55 (0.25) *	0.53 (0.23)	0.47 (0.22)

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. (against chance level 0.5)
nonMP refers to the non-minimal pair trials, cMP refers to the consonantal minimal pair trials, vMP refers to the vocalic minimal pair trials, and tMP refers to the tonal minimal pair trials.

Table S5.2 Best fitting model for offline accuracy measure in CSWL, with tonal minimal pair trial as reference level

Fixed Effects	Estimate	SD Error	Z	*p*
(Intercept)	-0.016	0.087	-0.184	.854
poly(block, 2)1 (linear block effect)	9.754	5.801	1.681	.093
poly(block, 2)2 (quadratic block effect)	-1.836	4.086	-0.449	.653
TrialtypeN	1.817	0.136	13.337	< .001 ***
TrialtypeC	1.209	0.151	8.024	< .001 ***
TrialtypeV	0.702	0.146	4.817	< .001 ***
poly(block, 2)1:TrialtypeN	72.248	7.445	9.705	< .001 ***
poly(block, 2)1:TrialtypeC	56.078	6.635	8.452	< .001 ***
poly(block, 2)1:TrialtypeV	23.212	6.252	3.713	< .001 ***
poly(block, 2)2:TrialtypeN	-21.310	7.043	-3.025	.002 **
poly(block, 2)2:TrialtypeC	-6.781	6.328	-1.072	.284
poly(block, 2)2:TrialtypeV	-2.835	5.985	-0.474	.636

Number of observations: 11695, Participants: 61, Item, 16. AIC = 13233.8, BIC = 13543.2, log-likelihood = -6574.9.

R syntax: glmer(acc ~ poly(block, 2)+ Trialtype+ poly(block, 2):Trialtype+ (1 + block + Trialtype | item) + (1 + block + Trialtype | subjectID), family = binomial).

Table S5.3 Mean percentage (%) fixation at target during T1 and standard deviations across the 12 blocks of the CSWL task in different trial types.

Trial type	Block											
	1	2	3	4	5	6	7	8	9	10	11	12
Overall	50.31 (8.20)	49.40 (8.16)	50.71 (7.72)	52.52 (7.64) **	49.56 (7.64)	51.02 (7.89)	50.18 (6.73)	50.51 (6.99)	53.39 (7.38) ***	51.09 (6.59)	53.50 (9.48) **	52.43 (6.84) **
nonMP	51.09 (16.37)	50.53 (16.50)	51.18 (18.73)	55.54 (16.20) **	48.39 (17.70)	55.19 (16.83) **	49.06 (15.87)	54.98 (16.18) **	57.56 (14.84) ***	53.59 (17.67)	58.07 (18.24) ***	55.18 (16.36) **
cMP	50.27 (16.44)	50.03 (13.70)	47.28 (16.21)	49.64 (16.36)	51.35 (14.44)	51.44 (16.81)	52.28 (17.66)	53.17 (15.22)	55.26 (18.83) *	50.16 (17.35)	52.94 (18.05)	54.99 (16.16) **
vMP	50.01 (16.78)	47.44 (15.41)	52.26 (17.31)	53.08 (16.27)	46.88 (16.18)	49.21 (16.73)	50.82 (17.76)	48.12 (16.11)	53.69 (17.48)	49.58 (18.56)	49.76 (18.32)	51.34 (15.94)
tMP	49.77 (14.35)	49.58 (18.61)	52.13 (16.08)	51.83 (13.91)	51.62 (16.51)	48.25 (19.45)	48.57 (17.23)	45.79 (16.84)	47.06 (15.63)	51.02 (14.51)	53.22 (19.83)	48.25 (14.95)

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. (against chance level 50%)
nonMP refers to the non-minimal pair trials, cMP refers to the consonantal minimal pair trials, vMP refers to the vocalic minimal pair trials, and tMP refers to the tonal minimal pair trials.

Table S5.4 Mean percentage fixation at target during T2 and standard deviations across the 12 blocks of the CSWL task in different trial types.

Trial type	Block											
	1	2	3	4	5	6	7	8	9	10	11	12
Overall	50.75 (8.65)	52.55 (9.58) *	55.52 (8.60) ***	56.98 (7.62) ***	55.61 (9.25) ***	57.53 (8.82) ***	57.10 (8.95) ***	58.13 (8.30) ***	59.73 (8.78) ***	61.41 (8.26) ***	59.35 (8.93) ***	60.02 (6.08) ***
nonMP	51.92 (16.00)	58.52 (15.64) ***	61.71 (17.28) ***	63.29 (15.67) ***	65.89 (18.53) ***	64.05 (15.81) ***	65.64 (15.55) ***	65.05 (16.70) ***	66.48 (15.86) ***	69.15 (16.11) ***	66.63 (15.58) ***	66.79 (16.73) ***
cMP	51.23 (14.61)	53.06 (17.19)	54.19 (14.91) *	60.90 (15.38) ***	52.53 (16.69)	60.68 (15.78) ***	62.28 (17.14) ***	64.51 (16.10) ***	63.97 (17.40) ***	63.50 (20.01) ***	64.72 (16.67) ***	64.91 (17.59) ***
vMP	49.50 (15.17)	52.52 (16.20)	56.17 (17.81) **	55.86 (16.60) **	56.92 (15.95) ***	54.21 (16.01) *	53.87 (19.67)	58.05 (20.59) **	57.34 (18.10) **	59.30 (17.48) ***	53.65 (20.27)	60.63 (17.90) ***
tMP	50.34 (16.77)	46.28 (17.14)	49.84 (15.83)	47.67 (17.57)	47.10 (19.08)	51.34 (19.43)	47.47 (19.56)	44.83 (17.93)	50.55 (20.48)	53.77 (18.62)	52.16 (16.76)	47.71 (15.60)

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. (against chance level 0.5)
nonMP refers to the non-minimal pair trials, cMP refers to the consonantal minimal pair trials, vMP refers to the vocalic minimal pair trials, and tMP refers to the tonal minimal pair trials.

Summary of the sub-dataset with continuous percentage fixation at target

Adding the fixed effects of block (χ^2 (1) = 0.5564, p = .456), trial type (χ^2 (3) = 1.0482, p = .790), time interval (χ^2 (1) = 1.3983, p = .237) and the 3-way interaction (χ^2 (7) = 3.9651, p = .784) did not improve model fit compared to the model with only random effects. It suggested that the effects of the predictors were primarily reflected by the binomially distributed dataset.

References

Angwin, A. J., Armstrong, S. R., Fisher, C., & Escudero, P. (2022). Acquisition of novel word meaning via cross situational word learning: An event-related potential study. *Brain and Language, 229*, 105111.

Bates D, Mächler M, Bolker B, Walker S (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. doi:10.18637/jss.v067.i01.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In M. J. Munro & O-S. Bohn (Eds.), *Second Language Speech Learning: The Role of Language Experience in Speech Perception and Production*. Amsterdam: John Benjamins, pp. 13–34.

Caroll, J. B. (1981). Twenty-five years of research foreign language aptitude. In E. K. C. Diller (Ed.), *Individual differences and universals in language learning aptitude* (pp. 83-117). Rowley, MA Newbury House.

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

Childers, J. B., & Pak, J. H. (2009). Korean- and English-speaking children use cross-situational information to learn novel predicate terms. *Journal of Child Language, 36*(1), 201– 224.

Choi, W., Tong, X., Gu, F., Tong, X., & Wong, L. (2017). On the early neural perceptual integrality of tones and vowels. *Journal of Neurolinguistics, 41*, 11-23.

Crespo, K., & Kaushanskaya, M. (2021). Is 10 better than 1? The effect of speaker variability on children's cross-situational word learning. *Language Learning and Development, 17*(4), 397-410.

Crespo, K., Vlach, H., & Kaushanskaya, M. (2024). The effects of speaker and exemplar variability in children's cross-situational word learning. *Psychonomic Bulletin & Review*, 1-11.

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2020). Learning English through out-of-school exposure. Which levels of language proficiency are attained and which types of input are important?. *Bilingualism: Language and Cognition, 23*(1), 171-185.

Dunn, K. J., Frost, R. L., & Monaghan, P. (2024). Infants' attention during cross-situational word learning: Environmental variability promotes novelty preference. *Journal of Experimental Child Psychology, 241*, 105859.

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

Escudero, P., Mulak, K. E., & Vlach, H. A. (2016). Cross-situational learning of minimal word pairs. *Cognitive Science, 40*(2), 455-465.

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

Felser, C., & Cunnings, I. (2012). Processing reflexives in a second language: The timing of structural and discourse-level constraints. *Applied Psycholinguistics, 33*(3), 571-603.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The Wordbank project*. MIT Press.

Ge, Y., Correia, S., Fernandes, J., Hanson, K., Rato, A., & Rebuschat, P. (under review). *Does Phonetic Training Benefit Word Learning?*

Ge, Y., Monaghan, P., & Rebuschat, P. (in press). The role of phonology in non-native word learning: evidence from cross-situational statistical learning. *Bilingualism: Language and Cognition*, 1-16. doi:10.1017/S1366728923000986.

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

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

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.

Horst, J. S., & Hout, M. C. (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.

Isbilen, E. S., Frost, R. L., Monaghan, P., & Christiansen, M. H. (2022). Statistically based chunking of nonadjacent dependencies. *Journal of Experimental Psychology: General, 151*(11), 2623-2640.

Jago, L. S., Alcock, K., Meints, K., Pine, J. M., & Rowland, C. F. (2023). Language outcomes from the UK-CDI Project: can risk factors, vocabulary skills and gesture scores in infancy predict later language disorders or concern for language development?. *Frontiers in Psychology, 14*, 1167810.

Kachlicka, M., Saito, K., & 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.

Kaushanskaya, M., Yoo, J., & Van Hecke, S. (2013). Word learning in adults with second-language experience: Effects of phonological and referent familiarity. *Journal of Speech, Language, and Hearing Research,* 56, 667-678.

Keating, G. D. (2009). Sensitivity to violations of gender agreement in native and nonnative Spanish: An eye-movement investigation. *Language Learning, 59*(3), 503-535.

Kempe, V., Thoresen, J. C., Kirk, N. W., Schaeffler, F., & Brooks, P. J. (2012). Individual differences in the discrimination of novel speech sounds: effects of sex, temporal processing, musical and cognitive abilities. *PloS one, 7*(11), e48623.

Kremmel, B. (2019). Measuring vocabulary learning progress. In S. Webb (Eds.), *The Routledge handbook of vocabulary studies* (pp. 406-418). Routledge.

Lee, C. Y., Yen, H. L., Yeh, P. W., Lin, W. H., Cheng, Y. Y., Tzeng, Y. L., & Wu, H. C. (2012). Mismatch responses to lexical tone, initial consonant, and vowel in Mandarin-speaking preschoolers. *Neuropsychologia, 50*(14), 3228-3239.

Lengeris, A., & Hazan, V. (2010). The effect of native vowel processing ability and frequency discrimination acuity on the phonetic training of English vowels for native speakers of Greek. *The Journal of the Acoustical Society of America, 128*(6), 3757-3768.

Li, S. (2015). The associations between language aptitude and second language grammar acquisition: A meta-analytic review of five decades of research. *Applied Linguistics, 36*, 385– 408.

Li, S. (2016). The construct validity of language aptitude. *Studies in Second Language Acquisition, 38*, 801– 842.

Li, S., Hiver, P., & Papi, M. (Eds.). (2022). *The Routledge handbook of second language acquisition and individual differences.* New York: Routledge.

Linck, J. et al. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning, 63*(3), 530– 566.

Luo, H., Ni, J. T., Li, Z. H., Li, X. O., Zhang, D. R., Zeng, F. G., & Chen, L. (2006). Opposite patterns of hemisphere dominance for early auditory processing of lexical tones and consonants. *Proceedings of the National Academy of Sciences, 103*(51), 19558-19563.

Marian, V., Blumenfeld, H. K., & 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.

Marian, V., & Spivey, M. (2003). Bilingual and monolingual processing of competing lexical items. *Applied Psycholinguistics, 24*(2), 173-193.

Marslen-Wilson, W. D., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 576–585. https://doi.org/10.1037/0096-1523.15.3.576.

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

Mueller, J. L., Friederici, A. D., & Männel, C. (2012). Auditory perception at the root of language learning. *Proceedings of the National Academy of Sciences, 109*, 15953–15958.

Mulak, K. E., Vlach, H. A., & Escudero, P. (2019). Cross-situational learning of phonologically overlapping words across degrees of ambiguity. *Cognitive Science, 43*(5), e12731.

Nespor, M., Peña, M., & Mehler, J. (2003). On the different roles of vowels and consonants in speech processing and language acquisition. *Lingue e linguaggio, 2*(2), 203-230.

Nora, A., Renvall, H., Kim, J. Y., Service, E., & Salmelin, R. (2015). Distinct effects of memory retrieval and articulatory preparation when learning and accessing new word forms. *PLoS One, 10*(5), e0126652.

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer: Scalable webcam eye tracking using user interactions. *Proceedings of the 25th International Joint Conference on Artificial Intelligence* (IJCAI), 3839–3845

Paribakht, T. S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In Coady , J. and T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy,* (pp. 174-200).

Perrachione, T. K., Lee, J., Ha, L. Y., & Wong, P. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America, 130*(1), 461-472.

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning, 63*(3), 595-626.

Rebuschat, P. (2022). Implicit learning and language acquisition: Three approaches, one phenomenon. In A. S. Reber and R. Allen (Eds.) *The cognitive unconscious: The first half-century*. Oxford University Press.

Rebuschat, P., Hamrick, P., Riestenberg, K., Sachs, R., & 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., & Schoetensack, C. (2021). Learning vocabulary and grammar from cross-situational statistics. *Cognition, 206*, 104475.

Reichle, E. D., Pollatsek, A., & Rayner, K. (2006). E–Z Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research, 7*(1), 4-22.

Roberts, L., Gullberg, M., & Indefrey, P. (2008). Online pronoun resolution in L2 discourse: L1 influence and general learner effects. *Studies in Second Language Acquisition, 30*(3), 333-357.

Ruan, Y., & Saito, K. (2023). Less precise auditory processing limits instructed L2 speech learning: Communicative focus on phonetic form revisited. *System, 114*, 103020.

Saito, K., Kachlicka, M., Sun, H., & Tierney, A. (2020a). 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*, 1-15.

Saito, K., Kachlicka, M., Suzukida, Y., Petrova, K., Lee, B. J., & Tierney, A. (2022). Auditory precision hypothesis-L2: Dimension-specific relationships between auditory processing and second language segmental learning. *Cognition, 229*, 105236.

Saito, K., Sun, H., & Tierney, A. (2020b). Domain-general auditory processing determines success in second language pronunciation learning in adulthood: A longitudinal study. *Applied Psycholinguistics, 41*(5), 1083-1112.

Saito, K. & Tierney, A. (forthcoming). Aptitude-acquisition interaction in L2 speech learning: Establishing benchmarks for normative vs. low auditory precision. *Language Learning*.

Service, E., & Craik, F. I. M. (1993). Differences between young and older adults in learning a foreign vocabulary. *Journal of Memory and Language, 32*(5), 608-623.

Silbert, N. H., Smith, B. K., Jackson, S. R., Campbell, S. G., Hughes, M. M., & Tare, M. (2015). Non-native phonemic discrimination, phonological short term memory, and word learning. *Journal of Phonetics, 50*, 99-119.

Skehan, P. (2012). Language aptitude. In S. Gass and A. Mackey (Eds.), *Handbook of second language acquisition* (pp. 381– 395). New York: Routledge.

Smith, K., Smith, A. D., & Blythe, R. A. (2009). Reconsidering human cross-situational learning capacities: A revision to Yu and Smith's (2007) experimental paradigm. In N. Taatgen & 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, A. D., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science, 35*(3), 480-498.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*(3)*,* 1558-1568.

Spivey, M., & Marian, V. (1999). Crosstalk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science, 10*, 281–284.

Suanda, S. H., Mugwanya, N., & Namy, L. L. (2014). Cross-situational statistical word learning in young children. *Journal of Experimental Child Psychology, 126*, 395-411.

Suanda, S. H., & Namy, L. L. (2012). Detailed behavioral analysis as a window into cross-situational word learning. *Cognitive Science, 36*(3), 545-559.

Surprenant, A. M., & Watson, C. S. (2001). Individual differences in the processing of speech and nonspeech sounds by normal-hearing listeners. *The Journal of the Acoustical Society of America, 110*(4), 2085-2095.

Tong, X., McBride, C., & Burnham, D. (2014). Cues for lexical tone perception in children: Acoustic correlates and phonetic context effects. *Journal of Speech, Language, and Hearing Research, 57*(5), 1589-1605.

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

Van Leussen, J. W., & Escudero, P. (2015). Learning to perceive and recognize a second language: The L2LP model revised. *Frontiers in psychology, 6*, 103694.

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of memory and language, 50*(1), 1-25.

Weighall, A. R., Henderson, L. M., Barr, D. J., Cairney, S. A., & Gaskell, M. G. (2017). Eye-tracking the time-course of novel word learning and lexical competition in adults and children. *Brain and language, 167*, 13-27.

Williams, J. N. & Rebuschat, P. (2023). Implicit learning and SLA: A cognitive psychology perspective. In A. Godfroid and H. Hopp (Eds), *The Routledge Handbook of Second Language Acquisition and Psycholinguistics*. Routledge.

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

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18*(5), 414–420.

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

6. General discussion

6.1. Summary of key findings

This dissertation project investigated the influence of phonology on statistical word learning. Across four studies, I demonstrated that non-native phonological contrasts and contrasts unrelated to learners' dominant language affected the acquisition of words containing these contrasts. Specifically, I examined the acquisition of novel Mandarin tonal words by English-native, Mandarin-native, and Mandarin heritage speakers, and compared how learners' different experience with the lexical tones shaped their word learning performance. In addition to the population-level investigation, I further examined how individual differences in the domain-general auditory processing ability and the heritage language experience and usage impacted individual variations in learning outcomes.

In the first study, I designed a cross-situational word learning task following Yu and Smith's (2007) paradigm, and trained English-native and Mandarin-native participants to map Mandarin pseudowords to uncommon referents. After a short cross-situational exposure of around ten minutes, both groups of participants could identify the meanings of words at relatively high accuracy if no phonological overlap was involved (i.e., non-minimal pair words). When segmental overlaps were present (i.e., consonantal and vocalic minimal pair words), the accuracy was reduced, but all learners could still recognize the words at an above-chance level. However, if tonal overlaps were introduced (i.e., tonal minimal pair words), only the Mandarin-native participants successfully identified the words, whereas English-native participants performed at chance level. The findings suggest that learners can pick up novel words rapidly from statistical tracking, but their learning was significantly dependent on the phonological properties of the words.

Following up on the findings of Study 1, Study 2 hypothesized that the lack of the learning effect in tonal minimal pairs among English-native participants might result from the

relatively short exposure. Thus, it was tested whether providing doubled exposure would improve learning outcomes. Results indicated that doubling exposure in a single learning session did not make a substantial difference at this very early stage of non-native word learning. English-native participants who underwent this extended cross-situational learning session performed comparably to those who had less exposure in Study 1.

Study 3 examined a different group of participants, the Mandarin heritage speakers who spoke English as their dominant or societal language. The comparison between this group of participants and the English-native and Mandarin-native groups enabled us to explore the impact of experience with Mandarin tones on novel tonal word learning. It was found that the Mandarin heritage speakers performed more similarly to English-native participants, with significant difficulty with tonal minimal pairs. More importantly, I tested another individual difference factor, target language experience and usage. Interestingly, no correlation was observed between the quantity and quality of Mandarin experience and usage and word learning outcomes. The potential reasons for the lack of a relationship were discussed, possibly due to the constraints in participant profiles.

The above studies provided evidence for a group-level difficulty with the tonal minimal pairs for English-native participants. Study 4 further explored whether and how this group-level finding was subject to individual variations in auditory processing. The domain-general auditory processing ability was examined as a measure of participants' acuity to pitch variations in nonspeech sounds. It was observed that the individual differences in this lower-order processing of pitch differences predicted English-native participants' tonal word learning outcomes. Moreover, in Study 4, an eye-tracking measure was employed to reflect participants' online processing of the stimuli. Longer fixation at the target referent would indicate better learning of the words. The comparison between the eye-tracking measure and the offline picture-selection accuracy measure revealed similar learning patterns, as

participants were overall more likely to fixate at and select the target referent picture at the end of the learning session.

6.2. Statistical learning of non-native speech sounds and words

The present findings have both theoretical and practical implications for the L2 acquisition literature. Firstly, it was confirmed that statistical word learning is an effective learning method for adult learners of an additional language (Escudero et al., 2022; Tuninetti et al., 2020). The presence of an unfamiliar suprasegmental cue (i.e., the lexical tone) did not impede overall learning performance, as was evidenced by the similar performance among English-native and Mandarin-native participants in non-minimal pair, consonantal minimal pair and vocalic minimal pair trials. However, when the non-native phonological contrast is the only cue to differentiate words (i.e., tonal minimal pairs), learning was significantly reduced. Altogether, these observations suggest that implicit exposure can be a rapid way of word learning for beginner L2 learners, but they might have difficulty acquiring the minimal pair words that rely on a non-native contrast. Hence, more explicit training targeting the non-native minimal pairs might be necessary for learners to perceive and distinguish the phonetic differences and encode them in lexical processing. This training is especially important in the case of Mandarin learning, as most Mandarin words have more than one tonal neighbour (Duanmu, 2007).

Additionally, I provided evidence for the link between non-native speech perception and word learning from a statistical learning approach. Previous evidence demonstrating this relationship has primarily come from the paired-associate word learning research, where explicit one-to-one word learning was examined (e.g., Bowles et al., 2016; Silbert et al., 2015; Wong & Perrachione, 2007). The current studies extended the findings to the implicit learning environment and were likely to be generalizable to the immersive second language

learning situations in real life. Similar to the paired-associate learning results, it was observed that statistical learning of non-native words was influenced by the perception of the non-native contrasts involved. Only the perceptually difficult contrasts (i.e., tonal contrasts) led to reduced word learning performance, but not the perceptually easy contrasts (i.e., consonantal and vocalic contrasts). Also, better tonal perception and discrimination abilities predicted overall learning outcomes at the end of the learning session. Taking together findings from the two word learning paradigms, it suggests that the phonetic-phonological-lexical continuity upheld in different learning contexts. Future research can carry out a direct comparison between the paired-associate and statistical learning paradigms using the same set of non-native stimuli. This cross-paradigm examination has only been tested in the native word learning context thus far (e.g., Neveu & Kaushanskaya, 2023, 2024). It would be interesting to apply it in the L2 context as well, which enables us to investigate and compare to what extent explicit and implicit word learning performances are influenced by non-native phonology.

Furthermore, the current studies contributed to the L2 speech and word learning literature in that they provided relevant evidence for the theoretical hypotheses of the speech perception models. I examined English-native speakers' perception and learning of two Mandarin tones, a high level (T1) and a falling tone (T4). Since English is a non-tonal language and does not have any tonal categories, how English-native speakers perceive and assimilate the L2 tonal contrasts is more complicated than segmental (i.e., consonants and vowels) assimilations. One hypothesis is that non-tonal language speakers may perceive lexical tones as uncategorized speech sounds as there lacks a clear corresponding category in their L1 (Hallé et al., 2004). Alternatively, it has been proposed as an extension of the Perceptual Assimilation Model (PAM) that non-native tones might be assimilated into learners' L1 intonational categories if tonal categories do not exist in their L1 (PAM for

suprasegmental, PAM-s, So & Best, 2008, 2011, 2014). Specifically, So and Best (2008, 2011) examined English-native speakers' perceptual assimilation of Mandarin tones to four of the English intonations: *flat pitch*, *question*, *statement* and *exclamation*. It was found that Mandarin T1 and T4 were both perceived mainly as a *statement* intonation in isolated syllables. A similar assimilation pattern was observed when tones were presented in sentences (i.e., with contextual effects), where T1 and T4 were associated with *statement* expressions at the sentence-final positions (So & Best, 2014). According to this tone-to-intonation assimilation, it is possible that the learning difficulty of the tonal minimal pairs in the current studies arose from the perceptual assimilations of the two tonal contours. If naïve listeners assimilate both T1 and T4 contours to their native *statement* intonation, it belongs to the Single Category assimilation type as proposed by PAM, which is predicted to be perceptually challenging. Moreover, as hypothesized by PAM-L2, such Single Category assimilation causes issues in developing new L2 categories because both sounds are considered variants of one single category. This lack of appropriate categories can further influence the use of the contrasts in distinguishing and learning words. Therefore, although the current studies are not a direct test of the speech perception models, the findings agree with PAM and PAM-L2's predictions on the perceptual and learning difficulties associated with the target tonal contrasts.

The learning difficulties observed can also be explained within the Second Language Linguistic Perception (L2LP) framework (Escudero, 2005; Van Leussen & Escudero, 2015). L2LP posits that if the number of L1 phonological categories is smaller than the required number for L2 categories, learning will be considerably more difficult because new categories need to be formed to fit the L2 sounds. As in the case of non-tonal language speakers' learning of tonal contrasts and words, the target L2 sounds involve an auditory dimension (i.e., pitch) that was not categorized in learners' L1, and hence, there are no

corresponding native tonal categories at the phonemic level at all. Learners need to create new boundaries along the new dimension and map them to different phonological categories, which was predicted to be the most difficult among various learning scenarios.

In essence, PAM reflects on why Mandarin T1/T4 contrast might initially pose a perceptual challenge for non-tonal language speakers, whereas L2LP and PAM-L2 offer theoretical insights into the later learning difficulties from the perspective of new category development. However, it is important to note that while the current results are in line with the predictions of the L2 perception models, my assessment only focused on the learning of two lexical tones from one single language. Further investigation of other tones from Mandarin and different tonal languages is needed to generalize the findings to L2 tonal learning and test whether the models (e.g., PAM-S) predict learning performance in various tonal contexts. In addition, the current research concentrated on statistical learning of non-native tonal words by non-tonal language speakers. Further exploration of how tonal language speakers learn non-native tonal words from statistical tracking will also yield interesting results, revealing the L2-to-L1 tonal perception and acquisition process.

6.3. Statistical learning of native words

6.3.1. Mandarin-native speakers

The Mandarin-native participants in Study 1 enabled us to explore how native tonal language speakers learn novel words from their L1 through cross-situational statistics. Similar to previous findings regarding the statistical learning of English pseudowords among English-native speakers (Escudero et al., 2016, 2022), I observed a significant impact of phonological overlap. Minimal pair words that differed in only one phonological contrast were generally more difficult to learn and identify even for adult native speakers. This is also consistent with learning results from explicit word learning paradigms (e.g., paired-associate

learning), where either word-referent mappings were unambiguously presented or feedback on the correct referent was provided during training (e.g., Creel & Dahan, 2010; Escudero et al., 2013; Pająk et al., 2016). Such learning difficulty associated with similar-sounding L1 words has been noted as early as during infancy and early childhood (e.g., Archer et al., 2014; Creel & Frye, 2024; Tsui et al., 2019). For instance, in a meta-analysis that involved studies with 12- to 20-month-old infants, Tsui et al. (2019) reported that similar-sounding words led to smaller learning effect sizes compared to dissimilar words. It indicates that similar-sounding words could be learned at a young age, but they pose greater challenges in vocabulary development. Creel and Frye (2024) further noted that this phonological overlap effect might not quickly diminish in early childhood and could undergo long-term development. The authors found that three- to five-year-olds could still encounter substantial difficulties learning novel minimal pair words. It was proposed that children experience protracted development of minimal pair words, and they gradually achieve adult-like performance over the years with improving phonemic categorization (e.g., Hazan & Barett, 2000). Overall, in native word learning, evidence shows that the difficulties resulting from phonologically similar words can be tracked throughout the developmental stages and are not entirely resolved even in adulthood.

Furthermore, in the examination of Mandarin native speakers' learning of Mandarin pseudowords, I also observed an interesting divergence between segmental and tonal minimal pair learning. Consonantal and vocalic minimal pairs were easier to learn compared to tonal minimal pairs. This could potentially stem from Mandarin speakers' different processing of segmental and tonal information at the lexical level (e.g., Cutler and Chen, 1997; Sereno and Lee, 2015; Wiener & Turnbull, 2015; Zou et al., 2022). For example, Zou et al. (2022) examined Mandarin-native speakers' spoken word recognition with competitor words that shared consonant, rime or tone with the target word (e.g., *tang2*). It was reported that

consonant competitors (e.g., *ti1*) and rime competitors (e.g., *lang4*) were activated when hearing the target word, as participants showed greater fixation towards these competitor words, but not the tonal competitors (e.g., *niu2*). This suggested more reliance on the segmental information in Mandarin native speakers' lexical processing. Moreover, a competitor that matched with the target word in both consonant and rime (e.g., *tang1*) was activated to a greater extent, indicating that tonal minimal pair words were strong competitors in native speakers' lexical access due to the segmental overlap. In comparison, consonantal (e.g., *yang1*) or rime minimal pairs (e.g., *tou2*) were less activated in the process. This different contribution of segmental and tonal information in lexical processing might explain why Mandarin speakers could better distinguish and identify segmental minimal pairs than tonal minimal pairs in the current word learning task.

6.3.2. Heritage Mandarin speakers

The learning performance of the heritage speaker participants demonstrated that acquiring new vocabulary in a heritage language through statistical learning is indeed feasible. Yet, this learning process is greatly affected by phonological contrasts present in the heritage language but absent in the learners' dominant language. Mandarin heritage speakers growing up and residing in English-speaking regions exhibited similar difficulty with tonal minimal pairs as English-native speakers who had no tonal experience. However, I am not suggesting that Mandarin heritage speakers' tonal word learning is, in general, indifferent to L2 Mandarin learners. Previous evidence suggested that Mandarin heritage speakers showed more categorical perception of tones than L2 learners (e.g., Yang, 2015), and hence heritage speakers' word learning difficulty might not result from the lack of tonal categories but from potential divergence in category patterns from native speakers. Overall, I interpret the findings in that exposure to the target language and phonological feature during early

childhood does not guarantee successful use of the feature in later word learning during adulthood.

Although there is limited research directly exploring heritage speakers' word learning performance, studies on heritage language speech perception and lexical processing have provided evidence that aligns with this perspective (e.g., Kim, 2019; Ortín, 2022; Soo & Monahan, 2017, 2023). For instance, Ortín (2022) investigated heritage Spanish speakers' processing of stress and consonant minimal pairs via an ABX task. Participants were presented with two Spanish pseudowords that differed in either the first consonant, the stress pattern or both, and were required to decide if a third pseudoword was identical to the first or the second one. Results suggested different processing patterns between stress and consonant minimal pairs. Heritage Spanish speakers showed accurate processing of stress patterns only in ABB trials where the target word is in an adjacent position, whereas lower accuracy was observed in non-adjacent ABA trials. However, the processing of consonant minimal pairs was as accurate in both ABB and ABA trials. This potentially pointed towards an impact of the dominant language (i.e., English), as the consonant contrasts were present in both English and Spanish, whereas the stress patterns were unique to Spanish. In other words, heritage speakers' processing and retention of a phonological contrast that is missing in their dominant language might be reduced compared to contrasts that are actively used in both languages.

In terms of lexical tone processing, Soo and Monahan (2023) reported that heritage Cantonese speakers' encoding of tonal minimal pairs was not as accurate as native speakers and it was associated with the degree of English dominance. In a medium-term priming task, where the prime and the target were separated by eight to 20 trials, it was predicted that accurate processing of the tonal word pairs would lead to a priming effect on identity pairs (i.e., identical prime and target), but not when the prime and the target were tonal minimal

pairs. However, a priming effect was observed when heritage Cantonese speakers processed tonal minimal pairs with shared pitch contours (e.g., rising tones T2 and T5), indicating that they were perceiving and encoding these tonal minimal pairs as identical words. Importantly, this inaccurate processing of tonal minimal pairs increased with English dominance – the more English-dominant participants were more likely to show tonal minimal pair priming. Therefore, heritage speakers of a tonal language may have less accurate tonal representations at the lexical level compared to native speakers of the language. The current word learning results can be interpreted consistently with these findings, as heritage Mandarin speakers' tonal minimal pair learning might be constrained by their less robust encoding of the lexical tones.

6.4. Individual difference factors in word learning

In the current studies, I focused on two individual difference measures –target language experience and usage, and auditory processing ability. The target language experience measure was specifically relevant in the study with heritage language speakers because heritage speakers typically have a more diverse language profile, and hence, it is important to take into account individuals' unique heritage language experience and usage when interpreting word learning results. As for auditory processing ability, it was of particular interest because the research targets, lexical tone perception and tonal word learning, are closely related to individuals' processing of pitch changes in any speech and nonspeech sounds.

The findings revealed a significant role of domain-general auditory processing ability in non-native word learning. Consistent with prior research on tonal word acquisition, it was observed that individuals with more precise auditory processing of the pitch changes tended to demonstrate better tonal word learning outcomes (e.g., Cooper & Wang, 2012; Li &

DeKeyser, 2017). However, it is important to highlight that the current results contribute to the existing body of literature by showing that word acquisition in a more implicit, statistical learning task was similarly influenced by auditory processing ability as observed in explicit paired-associate learning tasks. The statistical word learning task more closely mirrors an immersive learning situation, where learners were exposed to the speech input without instructions or explicit feedback. Previous research has demonstrated that auditory processing is a good predictor of L2 perceptual learning success in immersive settings such as living abroad (e.g., Kachlicka et al., 2019; Saito et al., 2022; Sun et al., 2021), reasoning that learners with better auditory processing skills can better detect and keep track of the statistical distribution of auditory cues in speech, which further facilitates the formation and differentiation of non-native phonological categories in speech perception. In the case of the current learning task, although the naïve learners might not have developed the non-native categories yet, more accurate auditory processing can potentially help them notice the distribution of pitch variations and start to integrate pitch as a relevant cue in perception and word learning. Another importance of understanding individual differences in auditory processing is that it may influence the effectiveness of treatment or training materials in L2 learning practice. For example, Perrachione et al. (2011) showed that learners with good auditory processing skills benefited more from talker variability in input, whereas those with lower auditory processing ability learned better in a low-variability environment. Future studies can build upon this research direction and explore whether a similar interaction between auditory processing ability and training materials is present in statistical word learning tasks.

For the heritage speaker population, I investigated the impact of individual differences in target language experience and usage on word learning outcomes. Surprisingly, I did not find a link between heritage language experience or usage and statistical learning of novel

heritage vocabulary. This seems to contradict previous observations of a language experience or dominance effect on heritage language lexical processing (e.g., Ortín, 2022; Soo & Monahan, 2023). Ortín (2022) tested heritage Spanish speakers' experience with Spanish and English (the dominant language) and revealed that a higher degree of relative dominance in Spanish was associated with greater sensitivity to Spanish stress patterns. Similarly, Soo and Monahan (2023) reported a correlation between heritage Cantonese speakers' degree of English/Cantonese dominance and their tonal minimal pair encoding. The divergence in findings could arise from the distinct lexical processing tasks used and the complex language profiles of the heritage speaker population. For instance, Soo and Monahan's (2023) lexical priming test utilized real Cantonese words, and the encoding of real words in the mental lexicon might exhibit more individual variations related to language experience, compared to the novel words in the current word learning task. Moreover, to obtain a more comprehensive understanding of how heritage language experience impacts word learning, it would be essential to include heritage speakers with a broader range of experience and dominance profiles, potentially dividing them into different experience groups.

6.5. Methodological implications – incorporating web-based eye-tracking in behavioural research

In Study 4, I employed a web-based eye-tracking technique (WebGazer.js, Papoutsaki et al., 2016) to measure participants' online processing of the stimuli. This measurement provided highly consistent results with the offline accuracy measure, indicating that participants were likely to select a picture after fixating on it. Overall, the longer fixation at the target picture reflected better learning of the word-picture mappings. Moreover, in the analyses of the individual difference predictors, the fixation measurement successfully captured correlations that were not observed in the offline accuracy data. These results

provided further evidence that behavioural research that does not require very high spatial precision can reliably use the remote eye-tracking tool to promote more efficient data collection (Slim & Hartsuiker, 2024; Van der Cruyssen et al., 2023; Yang & Krajbich, 2021). For example, in the current learning task, there were two areas of interest (i.e., the left or right side of the screen) and the gaze estimation could distinguish the two areas relatively accurately. The combination of online and offline measures can be effective for entailing learning gains and variations in learning performance in such word-picture mapping designs.

However, there are constraints to utilizing the web-based eye-tracking tool. Studies comparing remote and in-lab eye-tracking have reported a potential decrease in the observed effect sizes (e.g., Van der Cruyssen et al., 2023). This might affect studies where the learning effect is originally small. Additionally, in practice, the implementation of the remote eye-tracking tool can be difficult for experimental tasks that take long to complete. In the current design, due to the duration of the task (i.e., around 25 minutes), I asked participants to go through a re-calibration every two to three minutes to improve data quality. However, this procedure greatly increased the dropout rate, with more than 50% of the participants who started the learning task failing to complete it. I also received frequent messages from participants requesting assistance with the calibration process, as some of them could not pass the calibration successfully. Therefore, for studies with extended duration, this measurement might lead to obstacles in maintaining participants and data quality.

6.6. Limitations and further directions

In the current studies, I explored how non-native speech sounds affected statistical word learning by focusing on participants who were naïve to the target language and target speech sounds. This, however, only represents the very beginning stage of learning an additional language. It cannot be generalized to the later developmental stages of L2 learning.

Therefore, one question that arises is how L2 learners of different proficiency levels are influenced by non-native sounds. It is possible that, with the development and formation of new phonological categories, L2 learners of higher proficiency in the target language will be able to learn the non-native minimal pairs better. Future research can target different groups of English-native, L2 learners of Mandarin and investigate if L2 proficiency predicts word learning outcomes. This enables us to further explore the processing of L2 lexical tones at different levels, as previous studies found that even for advanced learners, Mandarin tones can pose difficulty at the lexical level despite accurate tonal identification and categorization skills (e.g., Pelzl et al., 2019).

Another constraint of the current design is the lack of variability in the stimuli of the learning tasks, as all stimuli were generated by a single speaker. This restriction limits the extent to which the results can be generalized to the natural language learning contexts, where substantial within-speaker and between-speaker variabilities are present. Moreover, as observed by Perrachione et al. (2011), input variability might interact with learners' individual auditory processing abilities to influence learning outcomes. Introducing stimuli variability will allow us to test whether learners with more precise auditory processing of pitch variations can benefit from high-variability tonal word tokens in training. It is possible that for these learners, greater variabilities in input may attract attention to the tonal cues and promote the development of tonal categories. Additionally, in terms of limitations on the experiment design, although the current CSWL task aimed at creating an implicit learning situation that mirrors the natural language learning environment, this laboratory-based task was far simplified compared to natural word learning. For example, the number of potential referents was strictly controlled (i.e., two referents per trial), which greatly reduced the degree of ambiguity in the learning environment. Once learners figured out the mapping of one word-picture pair, they could strategically infer other mappings based on mutual

exclusivity and the process of elimination. Thus, it should be interpreted with caution how the results can be generalized to natural L2 learning.

Given the observed difficulties in non-native minimal pair learning, further studies can also look for effective phonetic training methods that improve learners' phonetic and phonological processing of the non-native contrasts. For example, Ge et al. (under review) examined whether perceptual discrimination training (via AX and oddity discrimination tasks) on the target contrasts facilitated later learning of the words that contained the contrasts. Although no significant training effect was found in this study, follow-up research can explore different types of training tasks (e.g., identification tasks) and different lengths of training to find the optimal treatment method.

Furthermore, other individual difference measures can be taken into account, including cognitive factors such as explicit and implicit memory (e.g., Walker et al., 2020; Wang, 2020), and biographical factors such as age (e.g., Bulgarelli et al., 2021). Previous studies examining these factors in statistical learning mainly used native novel words; hence, it would be interesting to extend the individual difference findings to the L2 context. Additionally, from a methodological perspective, braining-imaging techniques (e.g., electroencephalogram) can be used to explore how the newly acquired non-native words can be integrated into learners' mental lexicon (e.g., Angwin et al., 2022 for native novel word learning).

7. Conclusion

In a series of four studies, I presented evidence that statistical word learning is influenced by the phonological properties of the words. Recognizing newly acquired words were found to be more challenging when presented in minimal pairs. Moreover, increased difficulty was linked to the minimal pairs that differed in a non-native contrast or a contrast

irrelevant to learners' dominant language (such as lexical tone for English-native or English-dominant speakers). Despite these phonological influences, statistical learning effectively facilitated the acquisition of non-native vocabulary. Furthermore, variations in learners' domain-general auditory processing abilities were found to predict their learning success.

References

Akhtar, N., & Montague, L. (1999). Early lexical acquisition: The role of cross-situational learning. *First Language, 19*(57), 347-358.

Angwin, A. J., Armstrong, S. R., Fisher, C., & Escudero, P. (2022). Acquisition of novel word meaning via cross situational word learning: An event-related potential study. *Brain and Language, 229*, 105111.

Archer, S., Ference, J., & Curtin, S. (2014). Now you hear it: Fourteen-month-olds succeed at learning minimal pairs in stressed syllables. *Journal of Cognition and Development, 15*(1), 110-122.

Best, C. T. (1994). Learning to perceive the sound pattern of English. In C. Rovee-Collier & L. Lipsitt (Eds.), *Advances in infancy research*, Vol. 8 (pp. 217– 304). Hillsdale, NJ: Ablex Publishers.

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). Timonium, MD: York Press.

Best, C. T., Hallé, P. A., Bohn, O.-S., & Faber, A. (2003). Cross-language perception of non-native vowels: Phonological and phonetic effects of listeners' native languages. *Proceedings of the 15th International Congress of Phonetic Sciences*, 2889–2892.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for non-native speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 345–360. doi: 10.1037/0096-1523.14.3.345

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (Eds.),

Language Experience in Second Language Speech Learning: In Honor of James Emil Flege (pp. 13–34). Amsterdam: John Benjamins.

Bhide, A., Ortega-Llebaria, M., Fraundorf, S. H., & Perfetti, C. A. (2020). The contribution of orthographic input, phonological skills, and rise time discrimination to the learning of non-native phonemic contrasts. *Applied Psycholinguistics, 41*(3), 481-516.

Bohn, O.-S., Best, C. T., Avesani, C., & Vayra, M. (2011). Perceiving through the lens of native phonetics: Italian and Danish listeners' perception of English consonant contrasts. *Proceedings of the 17th International Congress of Phonetic Science*, 336–339.

Bohn, O.-S., & Best, C. T. (2012). Native-language phonological and phonetic influences on perception of English approximant contrasts by Danish and German listeners. *Journal of Phonetics, 40*, 109–128. doi: 10.1016/j.wocn.2011.08.002

Bohn, O.-S., & Flege, J. E. (1992). The production of new and similar vowels by adult German learners of English. *Studies in Second Language Acquisition, 14*, 131–158. doi: 10.1017/S0272263100010792

Bowles, A. R., Chang, C. B., & Karuzis, V. P. (2016). Pitch ability as an aptitude for tone learning. *Language Learning, 66*(4), 774-808.

Brekelmans, G., Lavan, N., Saito, H., Clayards, M., & Wonnacott, E. (2022). Does high variability training improve the learning of non-native phoneme contrasts over low variability training? A replication. *Journal of Memory and Language, 126*, 104352.

Bulgarelli, F., Weiss, D. J., & Dennis, N. A. (2021). Cross-situational statistical learning in younger and older adults. *Aging, Neuropsychology, and Cognition, 28*(3), 346-366.

Bundgaard-Nielsen, R. L., Best, C. T., & Tyler, M. D. (2011a). Vocabulary size is associated with second-language vowel perception performance in adult learners. *Studies in Second Language Acquisition, 33*, 433–461. doi: 10.1017/ S0272263111000040

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

Chen, J., Best, C. T., & Antoniou, M. (2020). Native phonological and phonetic influences in perceptual assimilation of monosyllabic Thai lexical tones by Mandarin and Vietnamese listeners. *Journal of Phonetics, 83*, 101013.

Cooper, A., & Wang, Y. (2012). The influence of linguistic and musical experience on Cantonese word learning. *The Journal of the Acoustical Society of America, 131*(6), 4756-4769.

Creel, S. C., & Dahan, D. (2010). The effect of the temporal structure of spoken words on paired-associate learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(1), 110.

Creel, S. C., & Frye, C. I. (2024). Minimal gains for minimal pairs: Difficulty in learning similar-sounding words continues into preschool. *Journal of Experimental Child Psychology, 240*, 105831.

Cutler, A., & Chen, H.-C. (1997). Lexical tone in Cantonese spoken-word processing. *Perception and Psychophysics, 59*(2), 165–179.

Dickinson, D. K., Nesbitt, K. T., Collins, M. F., Hadley, E. B., Newman, K., Rivera, B. L., Ilgez, H., Nicolopoulou, A., Golinkoff, R. M., & Hirsh-Pasek, K. (2019). Teaching for breadth and depth of vocabulary knowledge: Learning from explicit and implicit instruction and the storybook texts. *Early Childhood Research Quarterly, 47*, 341-356.

Dörnyei, Z. (2005). *The Psychology of the Language Learner: Individual Differences in Second Language Acquisition.* Taylor & Francis Group.

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

Ellis, N. C. (2015). Implicit and explicit language learning: Their dynamic interface and complexity. In P. Rebuschat (Eds.), *Implicit and explicit learning of languages* (pp. 1-24). John Benjamins.

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

Escudero, P. (2015). Orthography plays a limited role when learning the phonological forms of new words: The case of Spanish and English learners of novel Dutch words. *Applied Psycholinguistics, 36*, 7–22. doi: 10.1017/S014271641400040X

Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in second language acquisition, 26*(4), 551-585.

Escudero, P., Broersma, M., & Simon, E. (2013). Learning words in a third language: Effects of vowel inventory and language proficiency. *Language and Cognitive Processes, 28*(6), 746-761.

Escudero, P., Hayes-Harb, R., & Mitterer, H. (2008). Novel second-language words and asymmetric lexical access. *Journal of Phonetics, 36*(2), 345-360.

Escudero, P., Mulak, K. E., & Vlach, H. A. (2016). Cross-situational learning of minimal word pairs. *Cognitive Science, 40*(2), 455-465.

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

Escudero, P., & Wanrooij, K. (2010). The effect of L1 orthography on non-native vowel perception. *Language and speech, 53*(3), 343-365.

Faris, M. M., Best, C. T., & Tyler, M. D. (2018). Discrimination of uncategorised non-native vowel contrasts is modulated by perceived overlap with native phonological categories. *Journal of Phonetics, 70*, 1-19.

Flege, J. (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). Timonium, MD: York Press.

Flege, J. E., & MacKay, I. R. (2004). Perceiving vowels in a second language. *Studies in second language acquisition, 26*(1), 1-34.

Fouz-González, J., & Mompean, J. A. (2021). Exploring the potential of phonetic symbols and keywords as labels for perceptual training. *Studies in Second Language Acquisition, 43*(2), 297-328.

Ge, Y., Correia, S., Fernandes, J., Hanson, K., Rato, A., & Rebuschat, P. (under review). *Does Phonetic Training Benefit Word Learning?*

Godfroid, A., Lin, C. H., & 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.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "l" and "r." *Neuropsychologia, 9*(3), 317–323. https://doi.org/10.1016/0028-3932(71)90027-3

Grenon, I., Kubota, M., & Sheppard, C. (2019). The creation of a new vowel category by adult learners after adaptive phonetic training. *Journal of Phonetics, 72*, 17-34.

Guion, S. G., Flege, J. E., Akahane-Yamada, R., & Pruitt, J. C. (2000). An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *Journal of the Acoustical Society of America, 107*, 2711–2724. doi: 10.1121/1.428657

Guion, S. G., & Pederson, E. (2007). Investigating the role of attention in phonetic learning. *Language experience in second language speech learning*, 57-77.

Gupta, P., Lipinski, J., Abbs, B., Lin, P. H., Aktunc, E., Ludden, D., ... & Newman, R. (2004). Space aliens and nonwords: Stimuli for investigating the learning of novel word-meaning pairs. *Behavior Research Methods, Instruments, & Computers, 36*, 599-603.

Hallé, P. A., Chang, Y.-C., & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Chinese versus French listeners. *Journal of Phonetics, 32*, 395–421. doi: 10.1016/ S0095-4470(03)00016-0

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

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics, 24*, 495–522.

Hayes-Harb, R., & Masuda, K. (2008). Development of the ability to lexically encode novel second language phonemic contrasts. *Second Language Research, 24*(1), 5-33.

Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6–12. *Journal of phonetics, 28*(4), 377-396.

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

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English/r/-/l/to Japanese adults. *The Journal of the Acoustical Society of America, 118*(5), 3267-3278.

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

Iverson, P., Pinet, M., & Evans, B. G. (2012). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics, 33*(1), 145-160.

Junttila, K., & 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., & 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.

Kempe, V., Thoresen, J. C., Kirk, N. W., Schaeffler, F., & Brooks, P. J. (2012). Individual differences in the discrimination of novel speech sounds: effects of sex, temporal processing, musical and cognitive abilities. *PloS one, 7*(11), e48623.

Kim, J. Y. (2020). Discrepancy between heritage speakers' use of suprasegmental cues in the perception and production of Spanish lexical stress. *Bilingualism: Language and Cognition, 23*(2), 233-250.

Krepel, A., de Bree, E. H., Mulder, E., van de Ven, M., Segers, E., Verhoeven, L., & de Jong, P. F. (2021). Predicting EFL vocabulary, reading, and spelling in English as a foreign language using paired-associate learning. *Learning and Individual Differences, 89*, 102021.

Kubo, R., & Akahane-Yamada, R. (2006). Influence of aging on perceptual learning of English phonetic contrasts by native speakers of Japanese. *Acoustical science and technology, 27*(1), 59-61.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience, 5*(11), 831-843.

Laméris, T. J., Llompart, M., & Post, B. (2023). Non-native tone categorization and word learning across a spectrum of L1 tonal statuses. *Bilingualism: Language and Cognition*, 1–15. https://doi.org/ 10.1017/S1366728923000871

Laméris, T. J., & Post, B. (2023). The combined effects of L1-specific and extralinguistic factors on individual performance in a tone categorization and word identification task by English-L1 and Mandarin-L1 speakers. *Second Language Research, 39*(3), 833-871.

Lee, A. H., & Lyster, R. (2015). The effects of corrective feedback on instructed L2 speech perception. *Studies in Second Language Acquisition, 38*(1), 35-64.

Lee, A. H., & Lyster, R. (2016). Effects of different types of corrective feedback on receptive skills in a second language: A speech perception training study. *Language learning, 66*(4), 809-833.

Lengeris, A. (2009). Perceptual assimilation and L2 learning: Evidence from the perception of Southern British English vowels by native speakers of Greek and Japanese. *Phonetica, 66*(3), 169-187.

Lengeris, A., & Hazan, V. (2007). Cross-language perceptual assimilation and discrimination of southern British English vowels by Greek and Japanese learners of English. In *Proceedings of ICPhS* (Vol. 16, pp. 1641-1644).

Li, Y., & Benitez, V. L. (2023). Lexical tone as a cue in statistical word learning from bilingual input. *Bilingualism: Language and Cognition*, 1-15.

Li, M., & DeKeyser, R. (2017). Perception practice, production practice, and musical ability in L2 Mandarin tone-word learning. *Studies in Second Language Acquisition, 39*(4), 593-620.

Lim, S. J., & Holt, L. L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive science, 35*(7), 1390-1405.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English/r/and/l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the acoustical society of America, 94*(3), 1242-1255.

Llompart M (2021). Phonetic categorization ability and vocabulary size contribute to the encoding of difficult secondlanguage phonological contrasts into the lexicon. *Bilingualism: Language and Cognition, 24*, 481–496. https://doi.org/10.1017/S1366728920000656

Llompart, M., & Reinisch, E. (2020). The phonological form of lexical items modulates the encoding of challenging second-language sound contrasts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(8), 1590.

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.

Macedo, A. (2015). *Estudo da perceção de vogais e ditongos orais de alunos de PLNM, falantes de Inglês L1* [Unpublished Master's thesis]. University of Minho.

Matthews, J. (2000). The influence of pronunciation training on the perception of second language contrasts. In J. Leather & A. James (Eds.), *New sounds 1999*, 223–229, University of Klagenfurt Press.

Monaghan, P., Mattock, K., Davies, R. A., & Smith, A. C. (2015). Gavagai is as Gavagai does: Learning nouns and verbs from cross-situational statistics. *Cognitive science, 39*(5), 1099-1112.

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

Mueller, J. L., Friederici, A. D., & Männel, C. (2012). Auditory perception at the root of language learning. *Proceedings of the National Academy of Sciences, 109*(39), 15953-15958.

Neveu, A., & Kaushanskaya, M. (2023). Paired-associate versus cross-situational: How do verbal working memory and word familiarity affect word learning?. *Memory & Cognition, 51*(7), 1670-1682.

Neveu, A., & Kaushanskaya, M. (2024). The role of bilingualism in paired-associate and cross-situational word learning. *Bilingualism: Language and Cognition, 27*(1), 41-56.

Nishi, K., & Kewley-Port, D. (2007). Training Japanese listeners to perceive American English vowels: influence of training sets. *Journal of Speech, Language, and Hearing Research, 50*(6), 1496-1510.

Ortín, R. (2022). Spanish heritage speakers' processing of lexical stress. *International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral-2021-0187

Pajak, B., Creel, S. C., & Levy, R. (2016). Difficulty in learning similar-sounding words: A developmental stage or a general property of learning?. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(9), 1377.

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer: Scalable webcam eye tracking using user interactions. *Proceedings of the 25th International Joint Conference on Artificial Intelligence* (IJCAI), 3839–3845

Pelzl, E., Lau, E. F., Guo, T., & DeKeyser, R. (2019). Advanced Second Language Learners' Perception of Lexical Tone Contrasts. *Studies in Second Language Acquisition, 41*(1), 59-86.

Perrachione, T. K., Lee, J., Ha, L. Y., & Wong, P. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America, 130*(1), 461-472.

Polka, L. (1991). Cross-language speech perception in adults: Phonemic, phonetic, and acoustic contributions. *Journal of the Acoustical Society of America, 89*, 2961–2977. doi: 10.1121/1.400734

Qian, M., Chukharev-Hudilainen, E., & Levis, J. (2018). A system for adaptive high-variability segmental perceptual training: Implementation, effectiveness, transfer. *Language Learning & Technology, 22*(1), 69-96.

Rato, A. (2019). The predictive role of cross-language phonetic similarity in L2 consonant learning, Talk given at the *International Symposium on the Acquisition of Second Language Speech – New Sounds 2019*, Waseda University, Japan, August 30-September 1.

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

Saito, K., Sun, H., Kachlicka, M., Alayo, J. R. C., Nakata, T., & Tierney, A. (2022). Domain-general auditory processing explains multiple dimensions of L2 acquisition in adulthood. *Studies in Second Language Acquisition, 44*(1), 57-86.

Sebastián-Gallés, N., & Díaz, B. (2012). First and second language speech perception: Graded learning. *Language Learning, 62*, 131-147.

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

Silbert, N. H., Smith, B. K., Jackson, S. R., Campbell, S. G., Hughes, M. M., & Tare, M. (2015). Non-native phonemic discrimination, phonological short term memory, and word learning. *Journal of Phonetics, 50*, 99-119.

Simonchyk, A. & Darcy, I. (2017) Lexical encoding and perception of palatalized consonants in L2 Russian. In M. O'Brien & J. Levis (Eds). *Proceedings of the 8th Pronunciation in Second Language Learning and Teaching Conference*, ISSN 2380-9566, Calgary, AB, August 2016 (pp. 121-132). Ames, IA: Iowa State University.

Slim, M. S., & Hartsuiker, R. J. (2023). Moving visual world experiments online? A web-based replication of Dijkgraaf, Hartsuiker, and Duyck (2017) using PCIbex and WebGazer. js. *Behavior Research Methods, 55*(7), 3786-3804.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*(3)*,* 1558-1568.

So, C. K. (2005). The influence of L1 prosodic background on the learning of Mandarin tones: Patterns of tonal confusion by Cantonese and Japanese naïve listeners. In *Proceedings of the 2005 CLA Annual Conference.*

So, C. K., & Best, C. T. (2008). Do English speakers assimilate Mandarin tones to English prosodic categories?. *Interspeech*, 1120.

So, C. K., & Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Language and Speech, 53*(2), 273-293.

So, C. K., & Best, C. T. (2011). Categorizing Mandarin tones into listeners' native prosodic categories: The role of phonetic properties. *Poznań Studies in Contemporary Linguistics, 47*(1), 133.

So, C. K., & Best, C. T. (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. doi: 10.1017/S0272263114000047

Soo, R., & Monahan, P. J. (2017). Language exposure modulates the role of tone in perception and long-term memory: Evidence from Cantonese native and heritage

speakers. In *Proceedings of the 43rd Annual Meeting of the Berkeley Linguistics Society* (Vol. 2, pp. 47-54). Berkeley Linguistic Society.

Soo, R., & Monahan, P. J. (2023). Phonetic and lexical encoding of tone in Cantonese heritage speakers. *Language and speech, 66*(3), 652-677.

Souza, H. K. D., Carlet, A., Jułkowska, I. A., & Rato, A. (2017). Vowel inventory size matters: Assessing cue-weighting in L2 vowel perception. *Ilha do Desterro, 70*, 33-46.

Stölten, K., Abrahamsson, N., & Hyltenstam, K. (2014). Effects of age of learning on voice onset time: Categorical perception of Swedish stops by near-native L2 speakers. *Language and Speech, 57*(4), 425-450.

Strange, W., Bohn, O.-S., Trent, S. A., & Nishi, K. (2004). Acoustic and perceptual similarity of North German and American English vowels. *Journal of the Acoustical Society of America, 115*, 1791–1807. doi: 10.1121/1.1687832

Strange, W., Bohn, O.-S., Nishi, K. & Trent, S. A. (2005). Contextual variation in the acoustic and perceptual similarity of North German and American English vowels. *Journal of the Acoustical Society of America, 118*, 1751–1762. doi: 10.1121/1.1992688

Suanda, S. H., Mugwanya, N., & Namy, L. L. (2014). Cross-situational statistical word learning in young children. *Journal of experimental child psychology, 126*, 395-411.

Sun, H., Saito, K., & 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.

Surprenant, A. M., & Watson, C. S. (2001). Individual differences in the processing of speech and nonspeech sounds by normal-hearing listeners. *The Journal of the Acoustical Society of America, 110*(4), 2085-2095.

Thomson, R. I. (2012). Improving L2 listeners' perception of English vowels: A computer-mediated approach. *Language Learning, 62*(4), 1231-1258.

Tsui, A. S. M., Byers-Heinlein, K., & Fennell, C. T. (2019). Associative word learning in infancy: A meta-analysis of the switch task. *Developmental psychology, 55*(5), 934.

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

Tyler, M. D., Best, C. T., Faber, A., & Levitt, A. G. (2014). Perceptual assimilation and discrimination of non-native vowel contrasts. *Phonetica, 71*(1), 4-21.

Van der Cruyssen, I., Ben-Shakhar, G., Pertzov, Y., Guy, N., Cabooter, Q., Gunschera, L. J., & Verschuere, B. (2023). The validation of online webcam-based eye-tracking: The replication of the cascade effect, the novelty preference, and the visual world paradigm. *Behavior Research Methods,* 1-14.

Van Leussen, J. W., & Escudero, P. (2015). Learning to perceive and recognize a second language: The L2LP model revised. *Frontiers in psychology, 6*, 103694.

Vlach, H. A., & Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition, 127*(3), 375-382.

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

Wang, F. H. (2020). Explicit and implicit memory representations in cross-situational word learning. *Cognition, 205*, 104444.

Wang, X. (2006). Perception of L2 tones: L1 lexical tone experience may not help. In *Proceedings of speech prosody* (pp. 85-88).

Wayland, R., & Guion, S. (2003). Perceptual discrimination of Thai tones by naïve and experienced learners of Thai. *Applied Psycholinguistics, 24*(1), 113-129.

Wen, Z. E., & Jackson, D. O. (2022). Working memory. In S. Li, P. Hiver and M. Papi (Eds.), *The Routledge handbook of second language acquisition and individual differences* (pp. 54-66). Routledge.

Wiener, S., Ito, K., & Speer, S. R. (2021). Effects of multitalker input and instructional method on the dimension-based statistical learning of syllable-tone combinations: An eye-tracking study. *Studies in Second Language Acquisition, 43*(1), 155-180.

Wiener, S., & Turnbull, R. (2016). Constraints of tones, vowels and consonants on lexical selection in Mandarin Chinese. *Language and speech, 59*(1), 59-82.

Williams, J. N., & Rebuschat, P. (2022). Implicit learning and second language acquisition: A cognitive psychology perspective. In A. Godfroid and H. Hopp (Eds.), *The Routledge handbook of second language acquisition and psycholinguistics* (pp. 281-293). Routledge.

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

Yang, B. (2015). *Perception and production of Mandarin tones by native speakers and L2 learners.* Berlin, Germany: Springer Verlag.

Yang, X., & Krajbich, I. (2021). Webcam-based online eye-tracking for behavioral research. *Judgment and Decision making, 16*(6), 1485-1505.

Yang, R., Nanjo, H., & Dantsuji, M. (2021). Self Adaptive Phonetic Training for Mandarin Nasal Codas. *Computer-Assisted Language Learning Electronic Journal, 22*(1), 378-400.

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.

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18*(5), 414–420.

Yurovsky, D., Fricker, D. C., Yu, C., & Smith, L. B. (2014). The role of partial knowledge in statistical word learning. *Psychonomic bulletin & review, 21*, 1-22.

Zhang, Y., Amatuni, A., Cain, E., Wang, X., Crandall, D., & Yu, C. (2021). Human learners integrate visual and linguistic information cross-situational verb learning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43, No. 43).

Zou, T., Chen, Y., & 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.

Zou, T., Liu, Y., & Zhong, H. (2022). The roles of consonant, rime, and tone in mandarin spoken word recognition: an eye-tracking study. *Frontiers in Psychology, 12*, 740444.

Zou, T., Zhang, J., & Cao, W. (2012). A comparative study of perception of tone 2 and tone 3 in Mandarin by native speakers and Japanese learners. In *2012 8th International Symposium on Chinese Spoken Language Processing* (pp. 431-435). IEEE.

Appendices

Appendix A: Debriefing questionnaire

Debriefing Questions about the study

During the different trials of this study, you saw two pictures and heard one word. Your task was to choose which picture the word referred to.

1. How did you decide which picture was the correct referent? Did you just guess throughout the experiment or did you follow any particular strategies? If so, what strategies did you follow?
2. Do you think the way you made decisions on the pictures changed throughout the experiment?
3. If you knew the names for one of the pictures already, and they did not match the word you heard, how did you use that information?
4. Did you notice any particular patterns or rules about this new language (e.g. is it different from your native language)?
5. Did you notice any particular patterns or rules about the sound system of this new language in terms of pronunciations (e.g. is it different from your native language)?
6. Did you notice whether the language use tones to mark different word meanings or not (i.e. whether changing the tone would change the word meaning)?
7. If you think the language uses tones to contrast meanings, how many tones do you think the language has?

Appendix B: Background questionnaire (used in Study 1, 2 and 4)

Background information

Note: In accordance with Lancaster University's Research Ethics guidelines, all information provided in this questionnaire will be anonymized in order to protect your privacy.

Gender: ❑ female ❑ male ❑ non-binary ❑ other ❑ prefer not to specify

Age: _____

Language background

What are your native language(s)? Indicate all languages in which you are a native speaker.
_____
_____
_____
_____

What foreign languages you have learned? For each language, please also indicate at what age you started to learn this language, how you have learned the language, how many years you have learned it for, and what you estimate your proficiency level to be.

Language	At what age did you start to learn this language?	How did you learn it (e.g. school, study abroad, at home)?	How many years have you been learning the language for?	What is your estimated proficiency level (see below for options)?

Advanced proficiency:

- Able to converse about general matters of daily life and topics of one's specialty and grasp the gist of lectures and broadcasts. Able to read high-level materials, such as newspapers, and write about personal ideas.

Intermediate level:

- Able to converse about general matters of daily life. Able to read general materials related to daily life and write simple passages.

Beginner level:

- Able to give simple greetings using set words and phrases. Able to read simple sentences, grasp the gist of short passages, and to write a simple sentence in the foreign language.