Bilingualism: Language and Cognition

cambridge.org/bil

Phonological neighborhood density, phonetic categorization, and vocabulary size differentially affect the phonolexical encoding of easy and difficult L2 segmental contrasts

Brian Rocca[1], Miquel Llompart[2,3] and Isabelle Darcy[1,4]

[1]Indiana University, Department of Second Language Studies, Bloomington, Indiana, USA; [2]Universitat Pompeu Fabra, Department of Translation and Language Sciences, Barcelona, Spain; [3]Friedrich Alexander University Erlangen-Nuremberg, Department of English and American Studies, Erlangen-Nuremberg, Germany and [4]Université Grenoble-Alpes, Laboratoire de linguistique et didactique des langues étrangères et maternelles, Grenoble, France

Research Article

Cite this article: Rocca, B., Llompart, M. and Darcy, I. (2024). Phonological neighborhood density, phonetic categorization, and vocabulary size differentially affect the phonolexical encoding of easy and difficult L2 segmental contrasts. *Bilingualism: Language and Cognition*, 1–14
https://doi.org/10.1017/S1366728924000865

Received: 20 December 2023
Revised: 25 July 2024
Accepted: 06 October 2024

Keywords:
phonolexical encoding; second language acquisition; L2 phonology; L2 lexicon; phonological neighborhood density

Corresponding author:
Brian Rocca;
Email: brocca@iu.edu

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

Abstract

This study investigated the effect of phonological neighborhood density (PND) on the lexical encoding of perceptually confusable segmental contrasts and the extent to which the precision of encoding is modulated by phonetic categorization and vocabulary size. Korean learners of English and native speakers of American English completed an auditory lexical decision task that contained words and nonwords with /ɛ/, /æ/, /f/, and /p/ (/æ/ and /f/ do not exist in Korean), two phonetic categorization tasks (/ɛ/—/æ/ and /f/—/p/), and a vocabulary test. For the Korean group, participants' categorization of /f/—/p/ was the only significant predictor of /f/—/p/ nonword rejection. For /ɛ/—/æ/, nonword versions of high PND words were rejected more accurately than low PND. Additionally, vocabulary size and phonetic categorization significantly interacted so that as perception abilities improve, the benefits that come from having a large vocabulary grow as well.

Highlights

- Categorization scores predicted phonolexical encoding for easy contrasts (/f/—/p/)
- Vocabulary size predicted phonolexical encoding for difficult contrasts (/ɛ/—/æ/)
- For /ɛ/—/æ/, higher categorization scores magnified the effect of vocabulary size
- For /ɛ/—/æ/, high PND words had more precise representations
- For /f/—/p/, no effect of PND was found on encoding precision

1. Introduction

Building a second language (L2) lexicon is one of the most essential tasks adult L2 learners face. Compared to first language (L1) acquisition, this is an effortful and time-consuming task that is often done in classroom contexts. Learning a word means creating a long-term memory representation for it, which encodes information about the word's form, meaning, and use (Hulstijn, 2001). In this article, we focus on the phonological form of words in the mental lexicon (i.e., the *phonolexical representation*). These lexical representations need to be precise and phonologically distinct from each other for learners to be able to successfully recognize words while listening. The more times words are encountered, the stronger and more distinct representations become, enhancing the ability to discriminate among phonologically similar words (White et al., 2013); however, when L2 words contain confusable sounds, creating distinct and precise representations is extremely challenging (Barrios & Hayes-Harb, 2021). In this article, we focus on precision, and less on distinctiveness or target-likeness. We follow Barrios and Hayes-Harb (2021) for terminology.

One reason for this difficulty is that sound perception is more challenging in the L2 than the L1. While L1 speech perception is an accurate and automatic process, this is not always the case for L2 speech perception. When people begin learning an L2, they use their L1 speech categories to process L2 sounds. If multiple L2 categories are too similar to an L1 category, the L2 sounds will be assimilated into the L1 category (Best & Tyler, 2007), making it difficult to distinguish these L2 sounds. A classic example is L1 Japanese speakers' perception of English /l/—/ɹ/ (see Goto, 1971). The closest sound in Japanese is /ɾ/, which both English /l/ and /ɹ/ assimilate to. Because English /l/—/r/ is a confusable contrast for L1 Japanese speakers, they are also likely to lexically encode the phonological information of these words imprecisely or in a way that reflects L1 perception (Pallier et al., 2001; Ota et al., 2009). The process of encoding phonological information into

CAMBRIDGE UNIVERSITY PRESS

https://doi.org/10.1017/S1366728924000865 Published online by Cambridge University Press

lexical representations (regardless of whether it reflects L1 or L2 perception) is referred to as *phonolexical encoding.*

Research by Darcy et al. (2013) investigated whether accurate segmental discrimination (and the presence of separate categories for new sounds) is key to acquiring precise phonolexical representations. One experiment focused on L1 English-L2 German learners' perception and encoding of /o/—/ø/. Perception was measured using an ABX task and the precision of phonolexical encoding was measured using an auditory lexical decision task (LDT). In an auditory LDT, participants hear a string of sounds and indicate by button press whether they believe it corresponds to a real word or a nonword. This requires participants to search their mental lexicon to find a representation that matches the input. Stimuli were German words such as <Honig> /honɪç/ "honey" and < König> /kønɪç/ "king" as well as nonwords created from systematic mispronunciations: *H[ø]nig and *K[o]nig. If the contrast was encoded precisely, participants should reject nonwords. If the contrast was encoded imprecisely, participants should accept the majority of nonwords. The results showed that participants performed accurately on ABX tasks but had lower and asymmetrical accuracy in the LDT. Hence, this study shows that accurate segmental discrimination does not necessarily imply precise lexical representations.

Darcy et al.'s (2013) finding that discrimination accuracy does not predict precise encoding has been replicated with advanced L1 English-L2 Russian participants (Simonchyk & Darcy, 2017) and with bilingual Spanish-Catalan participants (Amengual, 2016). Limited evidence suggests the opposite pattern as well, where precise (or at least distinct) representations can occur even when discrimination is inaccurate or unreliable (Darcy et al., 2012; Weber & Cutler 2004). This raises the question of what other factors are involved in creating precise phonolexical representations if accurate perception[1] does not directly translate into phonolexical accuracy (Amengual, 2016; Darcy et al., 2012, 2013; Simonchyk & Darcy, 2017).

This article addresses this question by examining how factors beyond phonetic categorization – namely vocabulary size, and phonological neighborhood density (PND) – affect language learners' encoding of confusable segmental contrasts into phonolexical representations. The first subsection discusses why a larger vocabulary contributes to more precise phonolexical encoding, and the second subsection discusses why words with high PND are more likely to be encoded more precisely.

1.1 The role of vocabulary size in phonolexical encoding

To understand what factors other than perception may impact phonolexical encoding, researchers have begun examining the impact of variables such as vocabulary size. Daidone and Darcy (2021) examined L1 English-L2 Spanish learners' discrimination and encoding of /ɾ/—/r/, /ɾ/—/d/, /r/—/d/, and /f/—/p/. These contrasts either exist in English (/f/—/p/) or they are not confusable at the perceptual level[2] – instead, they are primarily confusable at the phonolexical level. This means English learners of Spanish can discriminate a contrast like /ɾ/—/r/ (henceforth tap-trill) quite well from the time they begin learning Spanish but have difficulty encoding these sounds in the correct words. In addition to discrimination accuracy, Daidone and Darcy (2021) assessed memory, attention, and vocabulary size to identify predictors of precise encoding. They found that vocabulary size was the only significant predictor of LDT accuracy for all four contrasts. This implies that learners who knew more words, in general, were able to represent those words with greater precision, whereas learners who knew fewer words displayed less precise encoding. The authors explained the results by drawing on Best and Tyler (2007) who theorized that acquiring minimal pairs may put pressure on the phonological system to begin to differentiate non-native sounds. Daidone and Darcy suggest that a similar process could be happening at the phonolexical level where acquiring enough minimal pairs gradually forces the phonolexical system to become more precise over time (we refer to this process as *updating).* This might be especially important if the number of minimal pairs for a given contrast is low, as is the case for Spanish tap-trill, which has a low functional load[3] with only approximately 30 minimal pairs in the lexicon (Willis & Bradley, 2008).

Bundgaard-Nielsen et al. (2011) also provide evidence that having a larger L2 vocabulary correlates with more accurate segmental discrimination. They proposed that knowing more words pushes learners to pay attention to articulatory, phonetic, and phonological details that are not meaningful in the L1. They found that, for L1 Japanese-L2 English participants, a larger vocabulary size correlated with improved L2 perception but only until vocabulary size reached a certain point. Beyond that, perception accuracy no longer correlated with improvements in vocabulary size. The authors suggested that once learners acquire enough words to function in the L2 adequately, vocabulary size no longer exerts pressure to further refine discrimination performance.

While Daidone and Darcy (2021) investigated segmental contrasts that are confusable at the lexical level, Llompart (2021a) did so for /ɛ/—/æ/, a confusable contrast for German learners of English at both the perceptual and lexical levels (Flege et al., 1997; Llompart & Reinisch, 2020). German has an /ɛ/ sound that is comparable to English /ɛ/ (Bohn & Flege, 1992) but no /æ/ sound. Llompart measured the vocabulary size and categorization of English /ɛ/—/æ/ in one intermediate and one advanced L1 German group. The goal was to test whether vocabulary size and perception accuracy were predictors of precise encoding. The results showed that, for the intermediate group, vowel categorization was the only significant predictor of accurate LDT responses. For the advanced group, who overall had more accurate categorization than the intermediate group, vocabulary size was the only significant predictor of LDT accuracy. Llompart explained that with increasingly accurate categorization, learners capture and feed more crucial phonetic details to lexical representations that would have otherwise gone unnoticed. Once perception has improved past a critical threshold (which advanced participants were more likely to have reached), vocabulary size becomes important. As a learner's lexicon grows, more words containing /ɛ/—/æ/ are acquired which, because the contrast can now be perceived clearly, are able to act as evidence that the vowel contrast exists and is important. This may trigger a dynamic relexification process where phonolexical representations *update* to become more precise.

The development of phonolexical representations is still under-researched, but potential factors that lead to precise encoding are

[1] We use the term "perception" when referring generally to segmental discrimination/categorization.

[2] Although some studies have shown that /ɾ/—/d/ can be perceptually confusable (e.g., Daidone & Darcy, 2014; Rose, 2010).

[3] Functional load is a way of measuring how much "work" a segmental contrast does. The most basic calculation (Catford, 1987) counts the number of minimal pairs that exist for all possibly confusable segmental contrasts. The contrasts are then ranked so that those with many minimal pairs have "high" functional load and those with few have "low" functional load.

https://doi.org/10.1017/S1366728924000865 Published online by Cambridge University Press

becoming clearer. First, acquiring a larger vocabulary correlates with better segmental perception, but this correlation wanes after learners' lexicons become large enough that they can function adequately in the L2 (Bundgaard-Nielsen et al., 2011). Second, measures of segmental perception are not predictive of encoding if a contrast is confusable at the phonolexical level but not very confusable at the perceptual level (Daidone & Darcy, 2021). Finally, for perceptually confusable contrasts, perception is important up to a certain threshold, but after that, vocabulary size becomes an important predictor (Llompart, 2021a). What is still unclear is what exactly these thresholds are and what happens after they have been met. Do all words in the lexicon update simultaneously? Do some words update before others? If vocabulary size is a missing piece of the puzzle, then looking more closely at the structure of the lexicon and lexical characteristics is of great interest.

One of the first approaches to examining the role of lexical characteristics was Darcy and Holliday (2019)'s. They examined the encoding of a confusable vowel contrast (/o/—/ʌ/) for L1 Mandarin-L2 Korean learners and hypothesized that words learned earlier (i.e., at the "beginner level") should be less precise than words learned more recently when a more developed L2 phonological system was in place. This prediction was based on what they named the *age of words* hypothesis, which asserted that phonolexical "…updates enter the lexicon through new words, and then gradually permeate the system retroactively to update older forms…" (p. 13). A competing hypothesis was the *phonological update hypothesis*, which predicted that after learners acquire a new perceptual dimension (e.g., a vowel contrast), lexical representations update wholesale (i.e., all at the same time). The results showed a trend towards more recently learned words being more accurate than words learned earlier; however, the results were not robust enough to support either of the hypotheses, and further investigation is needed to shed additional light on this issue (but see Rothgerber, 2020, who found support for the age of words effect).

One way to extend Darcy and Holliday (2019) is to draw on Bundgaard-Nielsen et al. (2011, 2012), who highlight the potential role minimal pairs play in updating the phonological system. Bungaard-Nielsen et al. (2012) suggest that acquiring a larger vocabulary leads to improvements in segmental perception during immersion "…as the need to decipher and comprehend L2 speech rapidly guides the learner to tune in to the phonological system of that particular language, rather than continue to perceive L2 speech on the basis of its superficial phonetic similarities (and dissimilarities) to the L1" (p. 646). If meaningful differences in word forms are indeed what drive improvements in the phonological system, then this may also occur for the phonolexical system, as suggested by Daidone and Darcy (2021). This means that words that have a minimal pair (i.e., are contrastive) should be represented more precisely than words that have no minimal pairs.

1.2 The role of minimal pairs and PND in phonolexical encoding

Minimal pairs may help learners notice that there is a meaningful difference between the sounds in a confusable contrast. One study that provides evidence of this is Llompart and Reinisch (2020). In this study, L1 German-L2 English participants learned novel words in a word-picture association task in one of three conditions: (1) /ɛ/—/æ/ minimal pairs (*tenzer* versus *tanzer*) presented simultaneously (i.e., in direct contrast) in one-third of training trials; (2) minimal pairs, although present in the experiment so that listeners may notice them, are presented with filler items but never presented in direct contrast with the other member of the minimal pair; or (3) no minimal pairs – words also differed by the second syllable (*tenzer* versus *tandek*). The participants from condition (1) showed evidence of more accurate phonolexical encoding. Presenting minimal pairs together helped learners notice a difference between /ɛ/—/æ/ and create a phonological distinction between them in their phonolexical representations. Llompart and Reinisch argue that their work provides evidence that adult L2 learners can only create distinct lexical representations for difficult L2 contrasts in novel word learning when "the word learning situation favours that learners' attention is focused on those particular sounds" (p. 1604; see also Llompart & Reinisch, 2020). Therefore, it may not simply be that a larger lexicon offers more exemplars of /ɛ/—/æ/ words that help learners realize that contrast exists and is important. It may be that, as Daidone and Darcy (2021) suggest, having a larger lexicon means that more minimal pairs have been acquired, which helps learners notice (Doughty, 2001; Schmidt, 2001) a difference between what they have encoded and what is in the input.

One metric used to count the number of minimal pairs for a word is phonological neighborhood density (PND). Phonological neighbors are words that differ by adding, subtracting, or replacing one phoneme (Vitevitch & Luce, 2016). For example, the word "cat" has high PND (or a *dense* phonological neighborhood) because it has many minimal pairs (e.g., cot, kit, bat, vat, gnat, coat, scat at, etc.) while the word "strap" has low PND (or a *sparse* phonological neighborhood) because it has few minimal pairs (only 5: scrap, strep, strip, stripe, trap). Many studies have examined the effects of PND on word learning, word recognition, and speech production (see Vitevitch & Luce, 2016 for an overview) but little is known about how PND affects L2 phonolexical encoding.

Llompart (2021b) provides preliminary evidence for the idea that L2 speakers create more precise representations for words with high PND by re-examining LDT data from several studies focused on the encoding of English /ɛ/—/æ/ by L1 German speakers. The re-analysis examined whether the PND and lexical frequency of the stimuli predicted nonword rejection. The results showed that /æ/ nonwords (match → *m[ɛ]tch) were responded to more accurately if their real word form had high PND and low lexical frequency, but these lexical characteristics did not affect /ɛ/ nonword (desk → *d[æ]sk) rejection accuracy. This provides evidence that the representations of words which underlyingly contain the non-dominant category (/æ/ nonwords in this study) are affected by lexical characteristics. In explaining why these representations are affected by PND, Llompart suggested that this may be because a high PND word has more similar-sounding words containing the same vowel. These neighbors could help strengthen the connection between the nonnative phonetic category and the lexical items in the neighborhood. As the author acknowledges, this was a post-hoc analysis of data from a task in which PND had not been systematically manipulated. Therefore, it is crucial to replicate and extend these findings by testing PND effects under tighter experimental control conditions.

The research reviewed so far suggests that minimal pairs and high PND are important catalysts in helping learners create more precise phonolexical representations. This leads us to propose the *lexicon-driven update hypothesis*: after a new segmental contrast is acquired, the phonolexical representations of words in dense phonological neighborhoods will update more quickly and more robustly than words in sparse neighborhoods because there is stronger evidence of a phonological contrast. Updating one word in the neighborhood may cause an internal comparison to other words in the neighborhood, which triggers these representations

https://doi.org/10.1017/S1366728924000865 Published online by Cambridge University Press

to update as well as part of the relexification process. Conversely, words in sparse neighborhoods will update later or may not update at all if there is not enough evidence of contrast (i.e., little/no competition). Using a similar design to Llompart (2021a) while manipulating PND allows us to investigate this hypothesis.

2. The current study

We set out to test the lexicon-driven update hypothesis and extend Llompart (2021b) by directly examining the effect of PND on the encoding of confusable L2 segmental contrasts. We compare words with high and low PND while controlling for lexical frequency. Furthermore, we examine how potential PND effects interact with learners' perceptual abilities and L2 vocabulary size.

One group of L1 Korean-L2 English participants and one group of L1 English participants were recruited. Two English contrasts were identified as targets: /ɛ/—/æ/, which has been found to be perceptually confusable for L1 Koreans (Barrios and Hayes-Harb, 2021; Lee and Cho, 2018), and /f/—/p/, reported to be less confusable for L1 Koreans (Park and de Jong, 2008).[4] Therefore, English /f/—/p/ is interesting in that it allows a comparison between a segmental contrast that may be perceptually confusable to an extent but is easier to learn (/f/—/p/) to one that is harder (/ɛ/—/æ/). Henceforth, we will refer to /f/—/p/ as the "easy" contrast and /ɛ/—/æ/ as the "difficult" contrast.

By testing an easy and difficult contrast, the current study extends Llompart (2021a) and Daidone and Darcy (2021). It is possible that if a contrast is perceptually not "confusable enough", the primary predictor of accuracy will be vocabulary size and not perception (Daidone & Darcy, 2021).

2.1 Research questions

1. When words contain perceptually confusable segmental contrasts, do advanced learners of English create more precise phonolexical representations if those words have high PND or low PND?
2. Is the effect of PND modulated by vocabulary size?
3. Is the effect of PND modulated by perception accuracy (i.e., phonetic categorization)?

2.2 Hypotheses

RQ1) When words contain perceptually confusable contrasts, we hypothesize that learners create more precise phonolexical representations for words with high PND regardless of whether it is an easy or difficult contrast.

RQ2) We hypothesize that phonolexical encoding could be modulated by vocabulary size in two opposite ways:

(A) Learners with accurate perception benefit from having a larger vocabulary because they have more evidence of phonological contrasts, pushing phonolexical representations to update more quickly and robustly. It is possible that the effect of PND is stronger the larger a participant's vocabulary is.

(B) Because some studies (see Brysbaert et al., 2018) show that L2 frequency effects tend to be stronger for lower proficiency participants, one could also predict the same for PND in principle: that the smaller a participant's vocabulary size is, the greater the effect a dense neighborhood has. However, a smaller vocabulary is also likely to correlate with lower proficiency and worse perception, so PND may not have an effect on these participants.

RQ3) We hypothesize that learners with accurate perception benefit more from words with high PND because they can take advantage of the evidence presented by contrastive forms. When miscommunication occurs and a learner receives some kind of negative feedback, learners with better perception can notice a mismatch between the target sound and what they have stored.

3. Method

3.1 Participants

Two groups of participants ages 18–40 were recruited for this online experiment. The test group consisted of 35 L1 Korean-L2 English speakers (henceforth the Korean group). To qualify for this study, participants had to have taken at least one course at a university or college that used English as the main language of communication (ESL/EFL courses at a university did not qualify). Participants were recruited on campus, on Korean student groups' social media pages, and through snowball recruiting. All participants indicated that they were most familiar with American English except for three (one was most familiar with Canadian English, one with British English, and one reported both British and American English). At the time of testing, 27 participants were living in the USA. The control group consisted of 15 L1 American English speakers (henceforth the English group). Table 1 provides an overview of the participants' demographics.

This study was run online so that participants were able to take part from anywhere as long as they wore headphones, used a keyboard, and were in a quiet space. They met via video conference with the first author to complete three experimental tasks. They were not paid for their participation in the study. As a form of compensation, they were offered feedback on their perception of the test contrasts based on their results. No participants needed to be excluded for speech or hearing disorders.

3.2 Materials and procedures

The three tasks in the study were programmed in PsychoPy 3 (Pierce et al., 2019) and run via Pavlovia, PsychoPy's online platform, and were administered in the following order: (1) an auditory LDT, (2) a phonetic categorization task, and (3) a vocabulary test. Participants returned to the video conference twice to debrief: once after the LDT and once after they completed the vocabulary test (see Appendix A for debrief questions). After completing the experimental tasks, participants completed a background questionnaire and a word familiarity survey. The experiment lasted 50–70 minutes.

3.2.1 Auditory lexical decision task

This task assessed the phonolexical encoding of /ɛ/—/æ/ and /f/—/p/ in words with dense and sparse phonological neighborhoods. Participants heard a string of sounds and had to decide whether it

[4] Anecdotal evidence from teacher trainers identifies English /f/—/p/ as a confusable contrast for Koreans (Lee, 2001), but Park and de Jong (2008) show that learners can acquire relatively high accuracy in a mapping and goodness of fit task. Low goodness of fit ratings indicated that /f/ is a bad exemplar of both Korean /p^h/ and /p'/. Having a lower goodness of fit may make it easier to create a new category for English /f/ because learners are more likely to notice that English /f/ differs from English /p/.

https://doi.org/10.1017/S1366728924000865 Published online by Cambridge University Press

Table 1. Participant demographics

	Average (SD)	Median
English participants		
Age	28 (4.2)	29
Experience with L2 speech (1 = none; 5 = extensive)	3.9 (1.3)	5
Korean participants		
Age	29.3 (5.4)	29
Age of English onset (speaking)	8.7 (5.0)	7
Years in English-speaking country	5.3 (4.8)	5
English use (1 = no English, 6 = only English):		
at home	2.3 (1.7)	1
at university/work	4.7 (1.6)	5
in emails/chats/texts/social media	4.3 (1.1)	4
in conversation with American English speakers	5.1 (1.6)	6
in conversation with L2 English speakers	2.4 (1.5)	2
Self-reported proficiency: (1 = No ability, 5 = Perfect)		
English comprehension	4.0 (0.8)	4
Spoken English	3.9 (0.8)	4
English reading	4.3 (0.8)	5
Overall English proficiency	3.9 (0.7)	4
Accent in English (higher = weaker accent)	3.2 (0.9)	3

Note: "English usage" reflects current usage, meaning how often English was used within the past 5 weeks.

corresponded to a real English word or a nonword. Each word had a real word and a nonword version where the target segment was switched with the other sound in contrast. For consonants, the switch always occurred word-initially. For vowels, the switch always occurs in the first syllable. For example, for the /f/—/p/ contrast, there is the real word *push*, and the nonword is created by switching [p] for [f] to create *[f]ush. All nonwords conformed to English phonotactics.[5]

3.2.1.1 Design The task contained three conditions: test, control, and distractor. Lexical frequency was held constant for the test and control stimuli while PND was manipulated so that there was an equal number of dense and sparse items in each condition. Lexical characteristics were not manipulated in the distractor condition.

The test condition contained the contrasts /ɛ/—/æ/ and /f/—/p/, which were predicted to be confusable for Korean participants, making it more likely that they would accept nonwords in this condition. There were 32 real words and 32 nonwords per contrast for a total of 128 test items. Half of the items were monosyllabic and half disyllabic. We chose words in the /ɛ/—/æ/ condition such that no target vowel was followed by a nasal consonant. This was done because American English speakers often raise /æ/ when it is followed by a nasal, and this could have presented a potential confound in the experiment (see Appendix B for a list of test and control stimuli).

The control condition contained the segmental contrasts /eɪ/—/oʊ/ and /s/—/t/. Vowel identification scores in Lee and Cho (2018) show that intermediate Korean learners of English identify /eɪ/ and /oʊ/ with comparable accuracy (79% and 84%, respectively), and they never misidentify /eɪ/ for /oʊ/ or vice versa. Consequently, nonwords created by substituting these two vowels were predicted to be salient for Korean listeners and hence easy to reject. Similarly, /s/ and /t/ differ in the manner of articulation in the same way as /f/—/p/, but /s/ and /t/ exist in Korean, so nonwords created by substituting these two sounds were predicted to be rejected with high accuracy. For the L1 English participants, accuracy was predicted to be high in all conditions. There were 32 real words and 32 nonwords per contrast for a total of 128 control items.

In the test and control conditions, PND was manipulated so that half of the items had high PND (10–39 neighbors) and half had low PND (1–9 neighbors; see Appendix C for comparisons of PND across conditions). All PND figures were taken from the CLEARPOND database (Marian et al., 2012). The lexical frequency of items was 0.35–181 words per million (SUBTLEX-US corpus; Brysbaert & New, 2009), but the mean frequency was balanced across conditions (see Appendix D).

No words in the test or control condition were in minimal pairs with one another. This was done to avoid testing words from the same neighborhood. This could either over- or underestimate the reality of phonolexical updates because words in different neighborhoods may update at different times/rates. Thus, testing different neighborhoods provided a fuller examination of participants' mental lexicons.

The distractor condition contained three segmental contrasts which exist in Korean and were expected to be salient when switched to create nonwords: /k/—/m/, /i/—/u/, and /dʒ/—/w/. For each contrast, nine words (nouns, verbs, adjectives, or adverbs; three mono-, three di-, and three trisyllabic) were selected. One nonword was created for each real word, resulting in 18 items per contrast.

Because participants hear both the real word and nonword version of each word, we repeated some distractor items to prevent participants from developing a strategy of noticing that similar sounding items are presented twice – which could lead them to react differently to any item that appears like a (near) repetition. A second version of six items was recorded for each distractor contrast to encourage participants to think that any item can occur multiple times during the experiment (e.g., in the /k/—/m/ contrast, participants heard *campus-realword-1*, *campus-realword-2*, **[m]ampus-nonword*). In total, the distractor condition was comprised of 72 items (18 items per contrast +6 repeated items = 24 items per contrast).

Words were recorded in Praat by a 30-year-old female speaker of American English. The speaker was asked to read all the words aloud clearly, and care was taken so that she produced the key contrasts as distinctly as possible. Stimuli acoustics were measured in Praat (Boersma & Weenink, 2020) and matched so that there were no significant differences across PND conditions: the vowels were matched using formant measurements, the fricatives were matched using duration and center of gravity, and the stops were matched using voice onset time (see Appendices E-H).

3.2.1.2 Procedure Participants were seated at their personal computers wearing headphones. They were instructed to decide after

[5] One nonword item in the /s/—/t/ control contrast does not conform to English phonotactics: traitor → [s]raitor. Though, in some word onsets in reduced or connected speech, [sɹ] is a possible onset. For example, that's right → [sɹaɪt], or serrated knife → [sɹeɪ.ɾɪɾ naɪf].

every item they heard whether it was a real English word or not. They indicated their response by pressing 1 or 0 on their keyboard. All participants rejected words with their dominant hand and accepted them with their non-dominant hand.

The task began with eight practice items. Automatic feedback of correct/incorrect and the amount of time it took to respond was given after each key press. No feedback was provided on the main task. There was no time limit on responses, and the next item was presented 0.8 s after the previous key press. The 328 experimental trials were fully randomized, and participants received three short breaks.

3.2.2 Phonetic categorization task

This was a two-alternative forced-choice identification task that determined how well-defined participants' perceptual categories were for /ɛ/–/æ/ and /f/–/p/. The minimal pairs of *bet/bat* and *fan/pan* were recorded by the same speaker who recorded the LDT stimuli. A 21-step continuum was created for each minimal pair in MATLAB (R2021B) using the STRAIGHT morphing algorithm (Kawahara et al., 1999). This program takes a production of *bet* and *bat* (or *fan/pan*) and incrementally morphs them into one another in 4.7% increments (from 100% bet-0% bat to 0% bet-100% bat).

Categorization of /ɛ/–/æ/ was presented first. The task was then repeated with the words *fan/pan* to assess /f/–/p/ categorization. On their computer screen, participants saw a picture representing "bet" (and the orthography) on the left and a picture representing "bat" on the right. In each trial, participants heard one word randomly selected from the 21 step-continuum and pressed 1 or 0 to indicate whether they heard "bet" or "bat." All 21 items from the continuum were presented 10 times, resulting in a total of 210 trials, divided into three blocks. The task was not speeded, and the next item was presented 0.8 s after the previous key press. Responses were used to calculate a categorization slope for each participant to determine how clear-cut their category boundaries were for each contrast.

3.2.3 Vocabulary test

Immediately after finishing the categorization task, participants completed the vocabulary component of the Shipley-2 (Shipley et al., 2009). This is a multiple-choice test with 40 items of differing lexical frequencies. The directions read: "You will see a WORD in all capital letters. Press the key (number) of the word that best matches the meaning of the WORD in all capital letters." For example, TALK: 1. draw, 2. eat, 3. speak, 4. sleep. The items were presented in the fixed order provided by Shipley et al. After responding to an item, participants were not allowed to go back. This is the same task used in Llompart (2021a) to assess participants' vocabulary size. The number of correct responses was used as an estimate of participants' vocabulary size.

3.2.4 Word familiarity survey

Following the vocabulary test, participants completed a Qualtrics survey composed of a background questionnaire and a word familiarity survey. In the familiarity survey, participants rated how well they knew words from the LDT test condition on a four-point scale: (1) *Never seen this word*; (2) *I know it's a word, but I'm not sure what it means*; (3) *I recognize this word and know more or less what it means*; or (4) *Very familiar. I know how to use this word.*

4. Results

4.1 Auditory LDT

Fifty participants completed the LDT but one Korean participant's data did not save on Pavlovia. Several steps were taken to ensure the quality of the remaining data (49 participants × 328 items = 16,072 data points). First, any items that were responded to within 250 ms from the beginning of the trial or after 10,000 ms were excluded as being an accidental button press (too fast) or a participant losing focus (too slow). This led to an exclusion of 15 trials (0.1% of LDT data). Second, we removed any unknown items based on word familiarity survey ratings. To ensure that participants had lexical representations for items, participants had to rate each item as "Very familiar. I know how to use this word" for the trial to remain in the data set. Notably, if one word was marked as unfamiliar, this excluded two items – a real word and its nonword version. This led to the elimination of 23 words (46 items; 0.3% of total LDT data). Third, the LDT items were screened based on the English group's responses. Any item with a mean accuracy rate of 2.5 SD below the mean was excluded. These exclusions were done separately for each contrast as well as by real word and nonwords. Examining all items together would have led to an elimination of more items that are perceptually confusable rather than items that are potentially problematic. Using this exclusion criteria led to discarding 8 items (389 data points, which is 2.4% of total LDT data). Three of these items were nonwords from the test condition (based on the items "draft", "jacket", and "flame"). Finally, any participant within their language group whose accuracy in the distractor and control conditions was 2.5 SD from the mean of the group was considered an outlier and their scores were excluded. This excluded one participant from the English group and one from the Korean group, leaving a total of 14 English participants and 33 Korean participants for the LDT analysis.

Figure 1 shows the mean LDT accuracy for the test vowels (/ɛ/–/æ/; see Appendices I–J for accuracy on individual nonword items). From Figure 1, it appears that the Korean group's (in white) real-word acceptance accuracy is comparatively higher than nonword rejection accuracy. Focusing on nonword rejection, there appears to be an advantage where /ɛ/ nonword items (e.g. t[æ]st) are rejected more accurately than /æ/ nonwords (e.g. m[ɛ]tch), and it seems like dense nonwords may be rejected more accurately than sparse nonwords. For the English group (in gold), accuracy is near the ceiling for both real words and nonwords. However, accuracy is lowest in responses to sparse /æ/ nonwords, possibly indicating that participants are more tolerant of mispronunciations of these words where /æ/ has been switched with /ɛ/.

Figure 2 shows the accuracy results for the test consonant contrast (/f/–/p/). For English participants, accuracy is high for both real word and nonword conditions. For the Korean group, accuracy is overall substantially higher for this contrast than /ɛ/–/æ/, and there does not appear to be any visible difference as a function of PND.

For the control contrasts (/eɪ/–/oʊ/ and /s/–/t/), the mean results suggest that accuracy was high for both the English and Korean speakers, and there does not seem to be any effect of PND (see Appendix K for details and plots).

4.2 Lexical decision: group differences and test versus control contrasts

All analyses were preregistered on the Open Science Framework website unless otherwise stated. To begin answering RQ1, we

https://doi.org/10.1017/S1366728924000865 Published online by Cambridge University Press

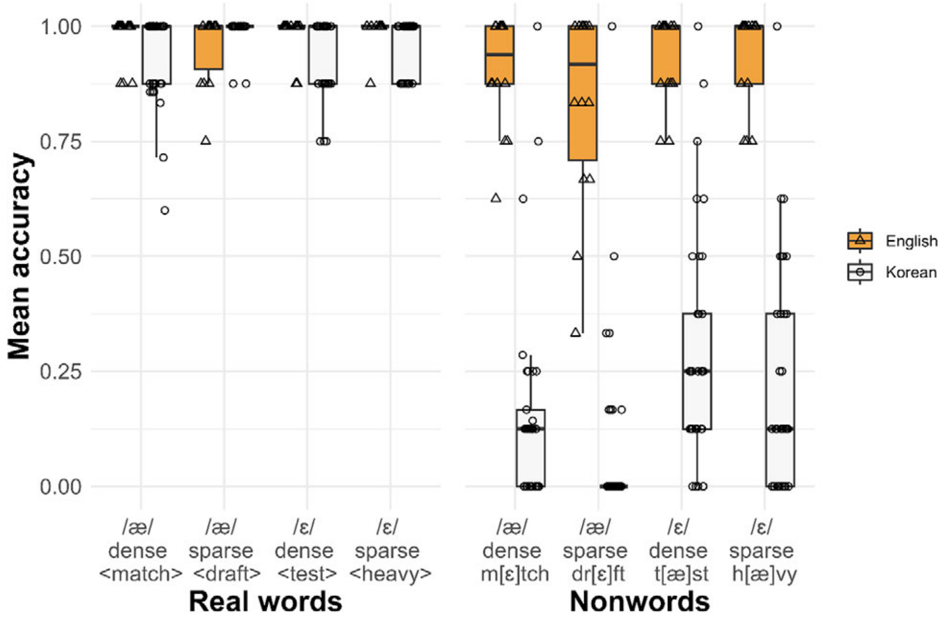

Figure 1. Mean lexical decision accuracy (proportion correct) on vowel test items.
Note: Whiskers represent 1.5 times the inter-quartile range. Real words are shown in the left panel and nonwords in the right panel.

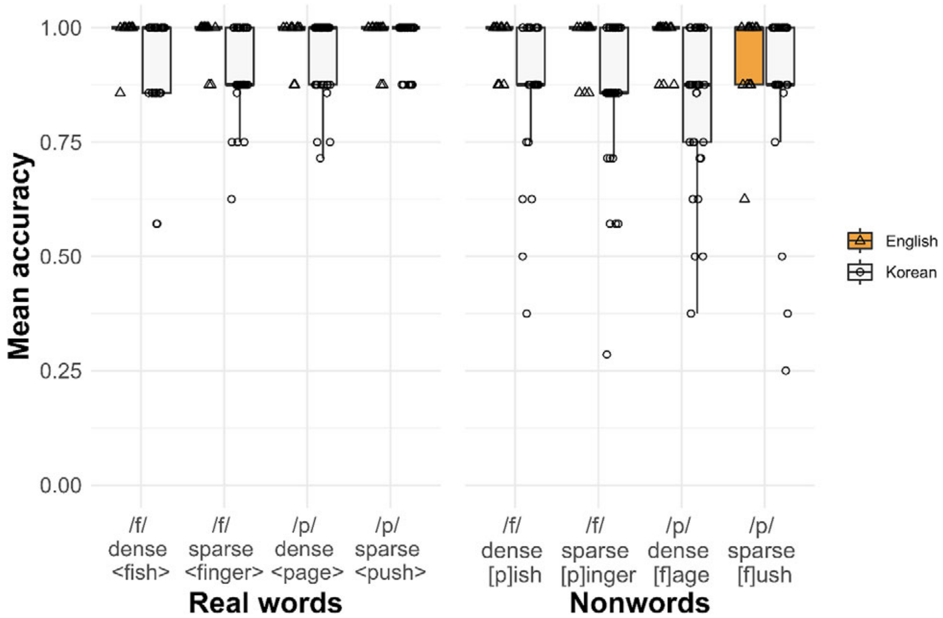

Figure 2. Mean lexical decision accuracy (proportion correct) on consonant test items.
Note: Whiskers represent 1.5 times the inter-quartile range.

conducted a basic analysis to check that the English group outperformed the Korean group and that participants responded more accurately on control items than test items to ensure that the task worked properly. Our pre-registration planned on only including nonwords in this analysis because we predicted that real word acceptance rates would be at ceiling, which they are (see Figures 1–2). All data for nonword trials for test and control items were submitted to a generalized linear mixed-effects regression model (GLMM) in R (version 4.2.1; R Core Team, 2022) with a logitistic linking function (lme4 package; Bates et al., 2015). The categorical dependent variable was Response (0 = incorrect, 1 = correct), and the independent variables were Contrast type (Test/Control), Segment type (Vowel/Consonant), Language group (Korean/English), and their interactions. The independent variables were contrast coded: Contrast type with Test as −0.5 and Control as 0.5, Segment type with Vowel as −0.5 and Consonant as 0.5, and Language group with Korean −0.5 and English as 0.5. Random-effects structures for all analyses in this study were chosen by a model fitting procedure using log-likelihood ratio tests, and random slopes were only included if they improved the model's fit. The random-effects structure for this model included random intercepts for Participants and Items, a random slope for Contrast

Table 2. GLMM results for nonword rejection accuracy in the LDT for Korean and English participants

Predictor	*b*	Std Error	*z*	*p*
Intercept	2.65	0.19	14.26	<.001***
Group	2.37	0.35	6.69	<.001***
Contrast type	1.63	0.25	6.58	<.001***
Segment type	1.20	0.25	4.87	<.001***
Group × Contrast type	−1.58	0.44	−3.56	<.001***
Group × Segment type	−2.41	0.44	−5.46	<.001***
Contrast type × Segment type	−3.05	0.42	−7.27	<.001***
Group × Contrast type × Segment type	1.89	0.74	2.56	0.010*

Note: Language Group = Korean/English, Contrast Type = test/control, Segment Type = Vowel/Consonant. Marginal R2 = 0.44, adjusted R2 = 0.63.

type and Segment type over Participants, and a random slope for Language group over Items, as these slopes improved the model's fit compared to models that did not include them (see Appendix L for model comparisons).

The model reveals a significant effect of Language group, Contrast type, and Segment type, as well as several significant interactions (see Table 2). To help understand these results, Figure 3 was created using the GLMM from Table 2 to predict the probability of correctly rejecting a nonword from each contrast. Overall, the figure shows that the English group was more accurate than the Korean group, and it shows that both groups had lower accuracy in the /ɛ/—/æ/ contrast compared to the /eɪ/—/oʊ/ contrast. Confidence intervals do not overlap between the English and Korean groups, and they are much further apart for the /ɛ/—/æ/ contrast than /f/—/p/.

4.3 Effects of PND and segment type on Korean Participants nonword rejection accuracy

A second step in answering RQ1 requires looking only at the Korean group's nonword rejection accuracy on test contrasts to assess whether PND affects accuracy. This data was submitted to a GLMM with the dependent variable Response and the independent variables PND (Sparse/Dense), Segment type (Vowel/Consonant), and their interaction. The independent variables were contrast coded: PND with Sparse as −0.5 and Dense as 0.5, and Segment type with Vowel as −0.5 and Consonant as 0.5. The random-effects structure included random intercepts for Participants and Items and a random slope for Segment type over Participant. The model reveals a main effect of Segment type (b = 4.58, SE = 0.409, z = 11.207, p < 0.001), showing that participants are more accurate on consonant items, but there is no main effect of PND (b = 0.31, SE = 0.292, z = 1.091, p = .275) nor an interaction between PND and segment type (b = −0.55., SE = 0.584, z = −0.951, p = .342).

While there is no significant interaction between PND and Segment type, an exploratory analysis of the effect of PND on response accuracy for the /ɛ/—/æ/ contrast is warranted for two reasons. First, the Korean group showed unexpectedly high nonword rejection accuracy for /f/—/p/ items, which suggests the contrast is not as difficult to encode lexically as expected (see categorization results). Therefore, this may obscure any difference in accuracy due to PND. For the /ɛ/—/æ/ contrast, the LDT nonword rejection scores are much lower (see Figure 3), indicating that this is indeed a difficult contrast and warrants a more focused analysis into whether PND is truly driving the differences in mean dense and sparse nonword rejection accuracy (see Figure 1). Second, the only previous study reporting a PND effect in an LDT focused on the /ɛ/—/æ/ contrast (Llompart, 2021b). This exploratory analysis can thus establish a parallel using a new L2 population while presenting a more controlled approach to PND effects. Additionally, because Llompart only finds an effect of PND for

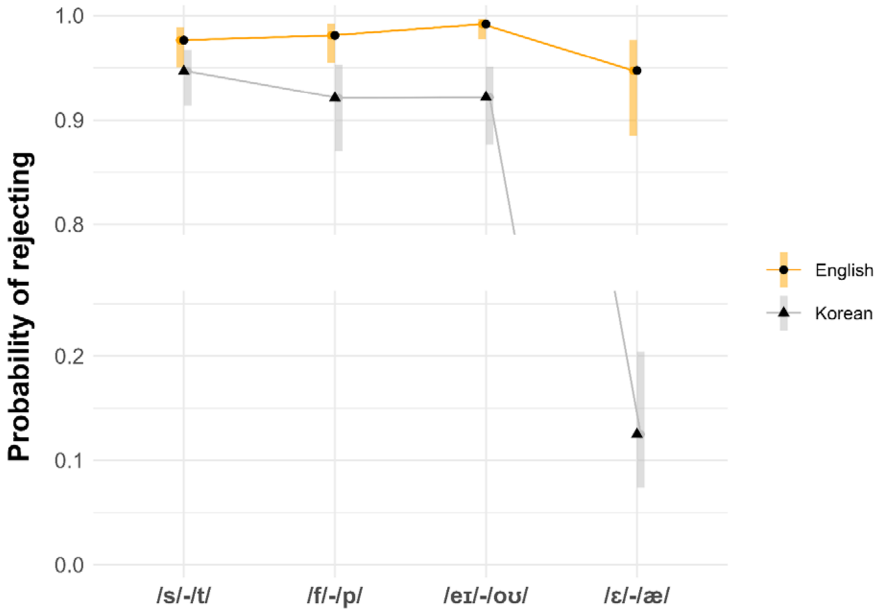

Figure 3. Predicted probability of nonword rejection by group and contrast.
Note: This plot was created based on the GLMM in Table 2 using the emmeans (Lenth, 2022) and ggbreak (Xu et al., 2021) packages in R. Vertical bars represent confidence intervals. Because /ɛ/—/æ/ accuracy was much lower than the other contrasts, there is a gap in the y-axis between 25% and 80% to make the data easier to interpret.

https://doi.org/10.1017/S1366728924000865 Published online by Cambridge University Press

/æ/ items, we include Vowel as a predictor to see whether PND and Vowel interact in a similar way.

For this exploratory analysis, the Korean group's data was submitted to a GLMM with the dependent variable Response and the independent variables PND (Sparse/Dense) and Vowel (æ/ɛ), as well as their interaction. The independent variables were contrast coded as in the model above and the random-effects structure included random intercepts for Participants and Items, but no random slopes because they did not improve the model's fit. This model resulted in three findings. Firstly, responses to dense nonwords were significantly more accurate than sparse nonwords (b = 0.74, SE = 0.366, z = 2.032, p = .042). Secondly, responses to /ɛ/ nonwords (e.g., *t[æ]st, *h[æ]vy) – which is the dominant category – were significantly more accurate than responses to /æ/ nonwords (e.g., *m[ɛ]tch, *sh[ɛ]dow; b = 1.31, SE = 0.368, z = 3.546, p < .001). Thirdly, there is no significant interaction of PND × Vowel, indicating that PND does not differentially affect the dominant and nondominant vowel (b = −0.58, SE = 0.730, z = −0.788, p = .43).

In summary, the initial model reveals that participants are significantly more accurate at rejecting /f/—/p/ nonwords, and there is no overall main effect of PND nor an interaction. An exploratory analysis examining only /ɛ/—/æ/ nonwords, however, shows that participants are significantly more accurate in rejecting /ɛ/ nonwords than /æ/ nonwords and, crucially, in rejecting dense nonwords than sparse nonwords.

4.4 Obtaining individual scores for perception and vocabulary size

Performance on the phonetic categorization task was used as a measure of perception for the two test contrasts (/ɛ/—/æ/ and /f/—/p/). We measured the steepness (i.e., slope) of the categorization curve for each participant by contrast. The perception data was split by contrast and the following procedure was used. Following Llompart (2021a), individual slopes were calculated by submitting the categorization data to a GLMM in R with a logistic linking function with Response (coded as 0 and 1) as the categorical dependent variable, an intercept term, and a random slope for Continuum step over Participants. The slope coefficient for each participant was extracted from the model. This coefficient quantifies the increase in log-odds of a "bat" response (for /ɛ/—/æ/) or a "fan" response (for /f/—/p/) for each one-unit increase of the continuum step. A coefficient of 0 would indicate no change (i.e., poor perception). Therefore, the higher the slope coefficient, the steeper the slope of the categorization function.

After extracting the slopes, two Korean participants' data sets were removed from the following analyses because they had clearly negative perception scores for the /ɛ/—/æ/ contrast. Although we cannot be absolutely certain, the negative scores likely mean that they reversed the buttons in the task. Another Korean participant's perception data did not save on Pavlovia, leaving a total of 30 L1 Korean data sets for the remaining analyses. For /ɛ/—/æ/, the English group had a mean slope of 1.02 (SD = 0.3) while the Korean group had a mean slope of 0.47 (SD = 0.27). For /f/—/p/, the English group had a mean slope of 1.13 (SD = 0.54) while the Korean group had a mean slope of 1.05 (SD = 0.57). Figure 4 visualizes the results for both contrasts. The English group seems to perform similarly on both contrasts with some steeper slopes on the consonant contrast. The Korean group appears to perform similarly to the English group on the /f/—/p/ contrast but has less steep slopes on the /ɛ/—/æ/ contrast.

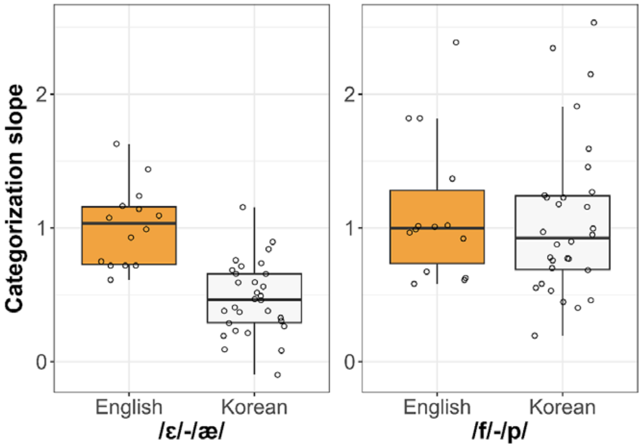

Figure 4. Categorization slope for the /ɛ/—/æ/ (left panel) contrast and /f/—/p/ (right panel) for English and Korean participants.

Note: Whiskers represent 1.5 times the inter-quartile range. The dots are slightly jittered to better visualize the data.

Individual results for the vocabulary test were used as a measure of vocabulary size. Results were calculated by taking the total number of items correct out of 40. The English group had a mean score of 86% (SD = 5%) while the Korean group had a mean score of 73% (SD = 12%).

4.4 Lexical decision: Effects of perception, vocabulary size, and PND

To answer RQs 2–3, the Korean group's responses to nonword trials for test items were submitted to a GLMM with Response as the categorical dependent variable. The independent variables were PND, Segment type, Vocabulary size, and Perception (one categorization slope for each contrast), as well as the interactions between PND x Segment type, PND × Perception, PND × Vocabulary size, Perception × Segment type, Perception × Vocabulary size, Vocabulary size × Segment type, and a three-way interaction between PND × Perception × Vocabulary size. Vocabulary size and perception

Table 3. Results of the GLMM on LDT nonword rejection accuracy for Korean participants

Predictor	*b*	Std Error	*z*	*p*
Intercept	0.20	0.23	0.89	0.376
PND	0.21	0.30	0.70	0.487
Segment type	4.62	0.33	13.93	<.001***
Perception	0.36	0.14	2.56	0.010*
Vocab	0.54	0.20	2.74	0.006**
PND × Segment type	−0.55	0.60	−0.91	0.366
PND × Perception	−0.40	0.18	−2.17	0.030*
PND × Vocab	0.03	0.17	0.15	0.880
Perception × Segment type	0.80	0.26	3.15	0.002**
Vocab × Segment type	−0.86	0.18	−4.93	<.001***
Perception × Vocab	0.54	0.16	3.31	0.001***
PND × Perception × Vocab	0.30	0.17	1.83	0.067

Note: *p < 0.05, **p < .01, ***p < .001. Vocab = Vocabulary size. Marginal R2 = 0.54, adjusted R2 = 0.70.

https://doi.org/10.1017/S1366728924000865 Published online by Cambridge University Press

Table 4. Results of the follow-up GLMM on /f/—/p/ LDT nonword rejection accuracy by Korean participants

Predictor	*b*	Std Error	*z*	*p*
Intercept	2.89	0.39	7.33	<.001***
PND	−0.13	0.58	−0.22	0.823
Perception	1.13	0.38	2.98	0.003**
Vocab	−0.02	0.22	−0.10	0.917
PND × Perception	−0.32	0.46	−0.69	0.491
PND × Vocab	0.12	0.25	0.46	0.646
Perception × Vocab	0.32	0.23	1.43	0.152
PND × Perception × Vocab	0.32	0.26	1.24	0.216

Notes: *p < 0.05, **p < .01, ***p < .001. Vocab = Vocabulary size. Marginal R2 = 0.18, adjusted R2 = 0.60.

were scaled using the *scale()* function in R (the phonetic categorization slopes for vowels were scaled separately from consonants). The random-effects structure for this model included random intercepts for Participants and Items as well as random slopes for Perception over Items. The results of this GLMM can be found in Table 3. Firstly, the model reveals a significant main effect of Perception, Vocabulary size, and Segment type. This means that better perception and a larger vocabulary led to higher overall accuracy, and that accuracy was higher on Consonant items than on Vowels. Secondly, there are four significant interactions that require us to split the data for further analysis.

To follow up on the significant interactions, the data was split by Segment type and submitted to two GLMMs: one for consonant data and one for vowel data. For both models, the dependent variable was Response, and the independent variables were PND, Perception, Vocabulary Size, as well as their interaction. The random-effects structure for these models included random intercepts for Participants and Items, and the Consonant model also included random slopes for Perception over Item. The results of the Consonant and Vowel models are shown in Table 4 and Table 5, respectively.

For the Consonant GLMM, there is a significant main effect of Perception, meaning participants with better perception reject nonwords more accurately. Converting the beta value from Table 4 into an odds ratio, the model indicates that for each increase in the scaled perception variable, participants are 3.1 times more likely to correctly reject an /f/—/p/ nonword.

Table 5. Results of the follow-up GLMM on Vowel LDT nonword rejection accuracy by Korean participants

Predictor	*b*	std error	*z*	*p*
Intercept	−2.12	0.30	−7.10	<.001***
PND	0.43	0.44	0.97	0.335
Perception	0.09	0.29	0.30	0.765
Vocab	0.97	0.24	4.08	<.001***
PND × Perception	−0.69	0.32	−2.13	0.033*
PND × Vocab	0.10	0.25	0.41	0.682
Perception × Vocab	0.50	0.25	1.98	0.048*
PND × Perception × Vocab	0.55	0.28	2.01	0.045*

Note: *p < 0.05, **p < .01, ***p < .001. Vocab = Vocabulary size. Marginal R2 = 0.21, adjusted R2 = 0.50.

For the Vowel GLMM, there is a significant main effect of Vocabulary size, which means that participants with a larger vocabulary have higher nonword rejection accuracy. Converting the beta value from Table 5 into an odds ratio, the model indicates that for each increase in the scaled Vocabulary size variable (which roughly corresponds to 4–5 points on the Shipley-2 vocabulary test), participants are 2.6 times more likely to correctly reject an /ɛ/—/æ/ nonword. Table 5 also shows several significant interactions. To help interpret these, Figure 5 was created using the GLMM output shown in Table 5 to plot the predicted probability of rejecting dense and sparse /ɛ/—/æ/ nonwords based on participants' vocabulary and perception scores. Firstly, there is almost always a higher predicted accuracy for dense nonwords (dotted lines) than sparse nonwords. Secondly, all learners seem to benefit from acquiring a large vocabulary, as this leads to a higher probability of correctly rejecting a nonword even for learners with low perception scores. Thirdly, the plot shows that as perception abilities improve, the predicted accuracy slope becomes steeper so that the benefits that come with having a larger vocabulary are magnified with higher perception scores.

5. Discussion

The present study examined the extent to which PND, vocabulary size, and perception contribute to the phonolexical encoding of one confusable but relatively easy (/f/—/p/) and one difficult (/ɛ/—/æ/) segmental contrast. Individual performances of L1 English and L1 Korean-L2 English participants were examined in an auditory LDT. For this task, English nonwords were created using systematic mispronunciations of words containing /f/—/p/ and /ɛ/—/æ/, and lexical items were chosen so that they either had dense or sparse PND. As predicted, the English group outperformed the Korean group, and both groups performed more accurately on control items than on test items. Further analyses showed that both groups performed more accurately when responding to consonant items than vowel items.

The findings from this study inform the role of PND in phonolexical encoding. Analyzing the Korean group's data, LDT nonword rejection accuracy for the /f/—/p/ contrast was unexpectedly high, and therefore PND-driven differences were not found. Accuracy for the /ɛ/—/æ/ contrast was comparatively much lower, and an effect of PND was found: words with high PND were represented more precisely than words with low PND. Furthermore, PND was found to have an effect across the board, similarly affecting both the dominant (/ɛ/) and nondominant (/æ/) vowels. The effect of PND partially replicates Llompart (2021b); however, Llompart found that dense PND only affected the nondominant vowel in the contrast (/æ/).

The findings from this study also inform the role of perception and vocabulary size in phonolexical encoding. Individual performance was examined in a phonetic categorization task (one task for each contrast) and a vocabulary test. These scores were then used to test whether they could predict the Korean participants' LDT nonword rejection accuracy. For both the easy (/f/—/p/) and difficult contrast (/ɛ/—/æ/), acquiring accurate perception is needed to reliably encode these contrasts in words. However, it is important to note that perception was the only significant predictor of encoding for /f/—/p/. For /ɛ/—/æ/, we found a significant effect of vocabulary size that interacted with perception, indicating that the better perception is, the stronger the effect of vocabulary size (as argued in Llompart, 2021a). In addition, the finding that

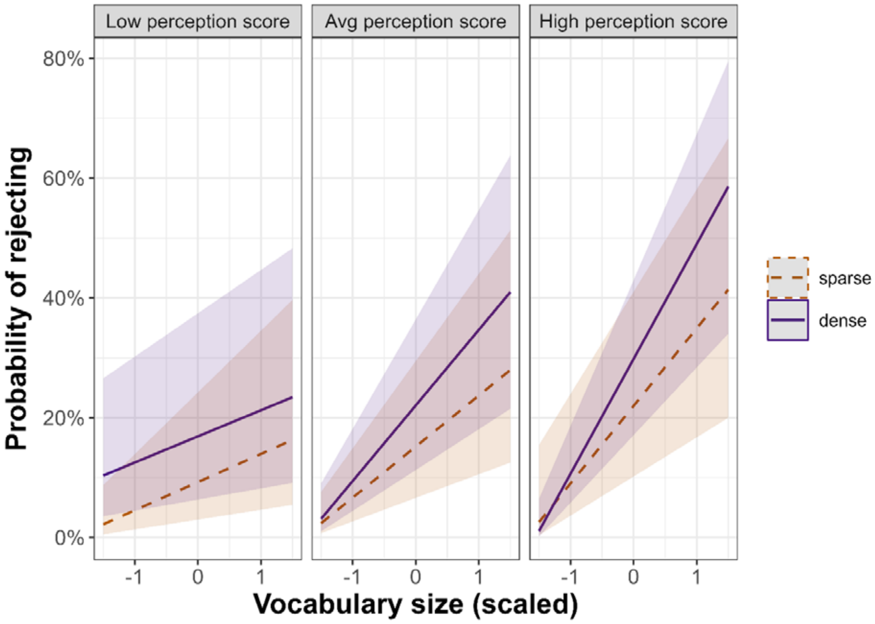

Figure 5. Predicted probability of rejecting /ɛ/—/æ/ nonwords for Korean participants.
Note: In terms of the slopes in phonetic categorization task (see Figure 4), the slices can be interpreted as: low perception score = ~0.25, average perception score = ~0.5, high perception score = ~0.7. This plot was created in R using the sjPlot (Lüdecke, 2022). Ribbons represent standard errors. See the online version to clearly distinguish ribbons.

vocabulary size is a significant predictor of /ɛ/—/æ/ encoding replicates a similar finding from Daidone and Darcy (2021) and Llompart (2021a) but extends it to a different population with a larger L1–L2 typological distance (see Llompart et al., 2023).

Overall, the analysis of the /ɛ/—/æ/ data aligns with the lexicon-driven update hypothesis which says that words with dense phonological neighborhoods will update more quickly and more robustly than words with sparse neighborhoods (which may not update if there is not enough evidence of a contrast). However, the /f/—/p/ data does not align with our hypothesis as there was no effect of PND, although this may have to do with the learners' very high /f/—/p/ perception accuracy. It is possible that, at an earlier proficiency level, PND effects were also present for /f/—/p/, but this remains speculation at this point. Future studies could replicate this with an intermediate and advanced group to test whether there is a difference between proficiency levels, or learners could be tested at multiple time points to assess accuracy over time.

The lexicon-driven update hypothesis cannot be accepted in its current form because it states that the lexicon only plays a role after a perceptual dimension is acquired while the current study provides evidence that both vocabulary size and perception are important concurrent predictors of accuracy even while perception is still developing (/ɛ/—/æ/). This study also shows that if a segmental contrast is perceptually confusable in the initial learning stages but relatively easy to acquire, learners can develop both accurate perception and precise encoding for words containing those segments. This leads to a reformulation of the lexicon-driven update hypothesis: *the lexical friction hypothesis.*

The lexical friction hypothesis says that phonolexical encoding is influenced by both the structure of the lexicon (e.g., PND, minimal pairs, lexical frequency, orthography) and how perceptually confusable a segmental contrast is. Encoding often relies on perception, but accurate perception does not mean that encoding will be precise (e.g., Daidone & Darcy, 2021; Darcy et al., 2013). *Friction* is anything that makes a contrast more salient for learners and leads to noticing. The more confusable a contrast is, the more friction will be needed to encode precisely. The structure of the lexicon can modulate the amount of friction and influence encoding in different ways. As this study suggests, high PND may lead to more friction because perception and production have to be more precise or a different word will be produced/interpreted. This would lead to miscommunication (i.e., friction). While miscommunication can be frustrating for learners, it can also be beneficial because it can lead them to notice the phonetic difference between two similar words. Over time, this miscommunication will carve out more precise representations. If a word has low PND, there are few or no similar-sounding words, so mispronunciation/misperception is less likely to lead to miscommunication. If there is no friction, then there is no push to update that representation. Depending on how confusable a contrast is, friction could facilitate encoding in the following ways:

- If a contrast is perceptually difficult to the extent that few learners acquire it (i.e., very confusable), both perception and vocabulary size will be predictors of phonolexical encoding because perception and a learner's lexicon are co-evolving (e.g., /ɛ/—/æ/ in this study). Because accurate perception is so difficult to acquire, learners need a larger vocabulary (and hence more minimal pairs) to increase the salience of the contrast. A larger vocabulary is also a proxy for more exposure to the language (Kuperman & Van Dyke, 2013). This means learners likely have more fine-grained knowledge of the words they have acquired, and learners have likely had more time communicating with others and therefore benefitted from the feedback that comes from miscommunication.
- If a contrast is difficult but learners are still able to acquire it, perception will be important until a certain threshold, and then vocabulary size will become more important (/ɛ/—/æ/ in Llompart, 2021a). Of course, vocabulary size and PND may still help sharpen perception (Best & Tyler, 2007; Bundgaard-Nielsen et al., 2011; 2012); however, acquiring accurate perception should act as more of a springboard so that once learners can

perceive a difference, they then can take advantage of the /ɛ/—/æ/ input as they acquire more and more words.

- If a contrast is relatively easy to acquire and has a high functional load but is still somewhat perceptually confusable, perception will be the main factor (Koreans' encoding of /f/—/p/ in the current study). Learners' attention will be drawn to the contrast early in acquisition due to miscommunication, helping to facilitate learning.
- If, however, discrimination accuracy is high from the onset of learning but there is not enough evidence in the lexicon that the contrast matters (i.e., low functional load – no friction), then vocabulary size will be the only significant predictor of accuracy because there is no friction if a learner's vocabulary is too small (Daidone & Darcy, 2021). The lexicon has to become large enough that it begins putting pressure on the phonolexical system to update.

6. Conclusion

Overall, this study adds to the literature on phonolexical encoding by examining factors beyond perception that lead to precise encoding. We find that perception is key for encoding both a perceptually easy and difficult contrast in words, and we find that acquiring a larger vocabulary is a significant predictor of precise encoding for a difficult contrast. Additionally, this effect interacts with perception so that the effect of vocabulary size is stronger for those with better perception. This replicates previous findings that learners need to be able to perceive a contrast in order to fully benefit from lexical knowledge. Importantly, while controlling for lexical frequency, we find an effect of PND for the perceptually difficult contrast where words with high PND are encoded more precisely than words with low PND. Additionally, the effects of PND interacted with vocabulary size and perception abilities.

Further work is needed to better understand the role of PND in phonolexical encoding in different contrasts (perceptually easy versus difficult) and at what point in the acquisition process PND may play a role. Relatedly, some L1 English speakers in this study were more accepting of /ɛ/—/æ/ mispronunciations than others, but these "over acceptances" tended to occur for /æ/ sparse nonwords. Future research could explore this further to examine whether listeners are more tolerant of mispronunciations depending on a word's lexical characteristics. Both L1 and L2 speakers might allow more variation in low PND words because there is less competition from similar sounding words. More variation might also be allowed for higher-frequency words. These words have lower activation thresholds (i.e., are activated more easily; Dufour et al., 2013) so variation might just be treated as noise in the signal that does not disrupt lexical retrieval. In other words, listeners might recognize a "mispronunciation" as a dialectal/accent variation or as an L2-like production that deviates from what is expected but is still accepted as a possible production. This might be an even more likely scenario for L2 listeners, as their representations tend to be weaker and their perception of some L2 contrasts is less reliable. Furthermore, it may be warranted to study the effects of various lexical characteristics in the same study to determine whether these variables modulate one another. For example, Karimi and Diaz (2020) found that the role of PND in picture naming changes from facilitatory to inhibitory depending on a word's age of acquisition and name agreement.

Concerning how PND is measured, future work could compare the effects of different PND variables. For example, is the total number of neighbors a better predictor of accuracy and reaction time than how many words in the neighborhood contain the contrast of interest? Or, perhaps, research should also explore other ways to compute phonological similarity besides adding/subtracting/replacing one phoneme (see Karimi and Diaz, 2020 for an example). Furthermore, when using the add/subtract/replace one phoneme metric for PND, a related issue is which PND database researchers should use. The current study uses the CLEARPOND database (Marian et al., 2012) to estimate the PND of words, but this database is built to simulate a native speaker's mental lexicon. For L2 learners, the PND figures could be overestimated. Hence, creating and using an L2-focused database to select stimuli could help push research into the effects of PND on L2 representations forward (for example, see Luef, 2022; Rocca et al., 2024). Be that as it may, the present study presents further evidence in favor of a modulating role for phonological neighborhood density in the phonological encoding of challenging L2 contrasts, and we hope that it will pave the way for further work on this exciting topic in the future.

Supplementary material. To view supplementary material for this article, please visit http://doi.org/10.1017/S1366728924000865.

Data availability. The data that support the findings of this study are available in the OSF at https://osf.io/5mswh/.

Acknowledgements. The authors wish to thank Sadi Phillips, Ken de Jong, and Rex Sprouse for their feedback on earlier versions of this article. Furthermore, we are indebted to Sadi Phillips for her help recording stimuli. Finally, we thank the IU Second Language Psycholinguistics lab for feedback on this project throughout its lifespan, and we would also like to acknowledge several precious suggestions from two anonymous BLC reviewers.

Competing interest. None declared.

References

Amengual, M. (2016). The perception of language-specific phonetic categories does not guarantee accurate phonological representations in the lexicon of early bilinguals. *Applied Psycholinguistics*, **37**(5), 1221–1251. https://doi.org/10.1017/S0142716415000557

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

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

Best, C. T., & **Tyler, M. D.** (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In Bohn OS & Munro MJ (Eds.), *Language experience in second language speech learning: Inhonor of James Emil Flege* (pp. 13–34). John Benjamins. https://doi.org/10.1075/lllt.17.07bes

Boersma, P. & **Weenink, D.** (2020). *Praat: doing phonetics by computer* [Computer software]. http://www.praat.org/

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** (2), 131–158. https://doi.org/10.1017/S0272263100010792

Brysbaert, M. & **New, B.** (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, **41**(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., **Mandera, P.**, & **Keuleers, E.** (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, **27**(1), 45–50. https://doi.org/10.1177/0963721417727521

Bundgaard-Nielsen, R. L., **Best, C. T.**, **Kroos, C.**, & **Tyler, M. D.** (2012). Second language learners' vocabulary expansion is associated with improved second language vowel intelligibility. *Applied Psycholinguistics*, **33**, 643–664. https://doi.org/10.1017/S0142716411000518

Bundgaard-Nielsen, R. L., **Best, C. T.**, & **Tyler, M. D.** (2011). Vocabulary size is associated with second-language vowel perception performance in adult learners. *Studies in Second Language Acquisition*, **33**(3), 433–461. https://doi.org/10.1017/S0272263111000040

Catford, J. C., (1987). Phonetics and the teaching of pronunciation: a systemic description of English phonology. In: Morley, J. (Ed.), *Current perspectives on pronunciation: Practices anchored in theory* (pp. 87–100). TESOL.

Daidone, D., & **Darcy, I.** (2021). Vocabulary size is a key factor in predicting second-language lexical encoding accuracy. *Frontiers in Psychology*, **12**, 688356. https://doi.org/10.3389/fpsyg.2021.688356

Daidone, D., & **Darcy, I.** (2014). Quierro comprar una guitara: Lexical encoding of the tap and trill by L2 learners of Spanish. In R. T. Miller, K. I. Martin, C. M. Eddington, A. Henery, N. Marcos Miguel, A. M. Tseng, A. Tuninetti, D. Walter (Eds.), *Selected Proceedings of the 2012 Second Language Research Forum: Building Bridges between Disciplines* (pp. 39–50). Cascadilla Proceedings Project.

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

Darcy, I., **Dekydtspotter, L.**, **Sprouse, R. A.**, **Glover, J.**, **Kaden, C.**, **McGuire, M.**, & **Scott, J. H.** (2012). Direct mapping of acoustics to phonology: On the lexical encoding of front rounded vowels in L1 English–L2 French acquisition. *Second Language Research*, **28**(1), 5–40. https://doi.org/10.1177/0267658311423455

Darcy, I. & **Holliday, J. J.** (2019). Teaching an old work new tricks: Phonological updates in the L2 mental lexicon. In J. Levis, C. Nagle, & E. Todey (Eds.), *Proceedings of the 10th Pronunciation in second language learning and teaching*. Conference, ISSN 2380-9566, September 2018 (pp. 10–26). Iowa State University.

Doughty, C. (2001). Cognitive underpinnings of focus on form. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 206–257). Cambridge University Press. https://doi.org/10.1017/CBO9781139524780

Dufour, S., **Brunellière, A.**, & **Frauenfelder, U. H.** (2013). Tracking the time course of word-frequency effects in auditory word recognition with event-related potentials. *Cognitive Science*, **37**(3), 489–507. https://doi.org/10.1111/cogs.12015

Flege, J. E., **Bohn, O. S.**, & **Jang, S.** (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, **25**(4), 437–470. https://doi.org/10.1006/jpho.1997.0052

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

Hulstijn, J. (2001). Intentional and incidental second-language vocabulary learning. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258–286). Cambridge University Press. https://doi.org/10.1017/CBO9781139524780

Karimi, H., & **Diaz, M.** (2020). When phonological neighborhood density both facilitates and impedes: Age of acquisition and name agreement interact with phonological neighborhood during word production. *Memory & Cognition*, **48**, 1061–1072. https://doi.org/10.3758/s13421-020-01042-4

Kawahara, H., **Masuda-Katsuse, I.**, & **De Cheveigne, A.** (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, **27**(3–4), 187–207. https://doi.org/10.1016/S0167-6393(98)00085-5

Kuperman, V., & **Van Dyke, J. A.** (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance*, **39**(3), 802. https://doi.org/10.1037/a0030859

Lee, J. (2001). Korean speakers. In S. Swan & B. Smith (Eds.) *Learner English: A teacher's guide to interference and other problems* (2nd ed., pp. 325–342). Ernst Klett Sprachen. https://doi.org/10.1017/cbo9780511667121

Lee, S., & **Cho, M. H.** (2018). Predicting L2 vowel identification accuracy from cross-language mappings between L2 English and L1 Korean. *Language Sciences*, **66**, 183–198. https://doi.org/10.1016/j.langsci.2017.09.006

Lenth, R. (2022). *emmeans: Estimated marginal means, aka least-squares means.* R package version 1.8.3. https://CRAN.R-project.org/package=emmeans.

Llompart, M. (2021a). Phonetic categorization ability and vocabulary size contribute to the encoding of difficult second-language phonological contrasts into the lexicon. *Bilingualism: Language and Cognition*, **24**(3), 481–496. https://doi.org/10.1017/S1366728920000656

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

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), 1590–1610. https://doi.org/10.1177/0023830918803978

Llompart, M., **Rocca, B.**, & **Darcy, I.** (2023). Is the effect of vocabulary size on lexical encoding modulated by L1-L2 similarity? In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th international congress of phonetic sciences* (pp. 2447–2451). Guarant International.

Lüdecke, D. (2022). *sjPlot: Data visualization for statistics in social science.* R package version 2.8.12, https://CRAN.R-project.org/package=sjPlot.

Luef, E. M. (2022). Growth algorithms in the phonological networks of second language learners: A replication of Siew and Vitevitch (2020a). *Journal of Experimental Psychology: General*, **151**(12), e26–e44. https://doi.org/10.1037/xge0001248

Marian, V., **Bartolotti, J.**, **Chabal, S.**, & **Shook, A.** (2012). CLEARPOND: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PLoS One*, **7**(8), e43230. https://doi.org/10.1371/journal.pone.0043230

Ota, M., **Hartsuiker, R. J.**, & **Haywood, S. L.** (2009). The KEY to the ROCK: Near-homophony in nonnative visual word recognition. *Cognition*, **111**(2), 263–269. https://doi.org/10.1016/j.cognition.2008.12.007

Pallier, C., **Colomé, A.**, & **Sebastián-Gallés, N.** (2001). The influence of native-language phonology on lexical access: Exemplar-based versus abstract lexical entries. *Psychological Science*, **12**(6), 445–449. https://doi.org/10.1111/1467-9280.0038

Park, H., & **de Jong, K. J.** (2008). Perceptual category mapping between English and Korean prevocalic obstruents: Evidence from mapping effects in second language identification skills. *Journal of Phonetics*, **36**(4), 704–723. https://doi.org/10.1016/j.wocn.2008.06.002

Peirce, J. W., **Gray, J. R.**, **Simpson, S.**, **MacAskill, M. R.**, **Höchenberger, R.**, **Sogo, H.**, **Kastman, E.**, **Lindeløv, J.** (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-01193-y

R Core Team (2022). *R: A language and environment for statistical computing* (version 4.2.1) [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rocca, B., **Martino, F.**, & **Darcy, I.** (2024). How misperception affects the structure of the L2 mental lexicon: A conceptual replication of Cutler (2005). *Proceedings of 14th Pronunciation in Second Language Learning and Teaching Conference*, (pp. 1–12). Purdue University, September 2023. https://doi.org/10.31274/psllt.17573.

Rose, M. (2010). "Differences in discriminating L2 consonants: a comparison of Spanish taps and trills." In M. T. Prior, Y. Watanabe, & S. Lee (Eds.) *Selected proceedings of the 2008 second language research forum: Exploring SLA perspectives, positions, and practices* (pp. 181–196). Cascadilla Proceedings Project.

Rothgerber, J. R. (2020). *The influence of native language phonotactics on second language lexical representation in Japanese learners of English* [Doctoral dissertation, Indiana University]. ProQuest.

Schmidt, R. W. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge University Press. https://doi.org/10.1017/CBO9781139524780

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). Iowa State University.

Shipley, W. C., **Gruber, C. P.**, **Martin, T. A.**, & **Klein, A. M.** (2009). *Shipley-2 manual*. Western Psychological Services.

Vitevitch, M. S., & **Luce, P. A.** (2016). Phonological neighborhood effects in spoken word perception and production. *Annual Review of Linguistics*, **2**, 75–94. https://doi.org/10.1146/annurev-linguistics-030514-124832

Weber, A., & **Cutler, A.** (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, **50**(1), 1–25. https://doi.org/10.1016/S0749596X(03)00105-0

White, K. S., **Yee, E.**, **Blumstein, S. E.**, & **Morgan, J. L.** (2013). Adults show less sensitivity to phonetic detail in unfamiliar words, too. *Journal of Memory and Language*, **68**(4), 362–378. https://doi.org/10.1016/j.jml.2013.01.003

Willis, E. W., & **Bradley, T. G.** (2008). Contrast maintenance of taps and trills in Dominican Spanish: Data and analysis. In L. Colantoni & J. Steele (Eds.), *Selected proceedings of the 3rd Conference on Laboratory Approaches to Spanish Phonology* (pp. 87–100). Cascadilla Proceedings Project.

Xu, S., **Chen, M.**, **Feng, T.**, **Zhan, L.**, **Zhou, L.**, & **Yu, G.** (2021). Use ggbreak to effectively utilize plotting space to deal with large datasets and outliers. *Frontiers in Genetics*, **12**, 774846. https://doi.org/10.3389/fgene.2021.774846

https://doi.org/10.1017/S1366728924000865 Published online by Cambridge University Press