Malleability of previously-established and newly-established phono-lexical representations: a design

Miren Adrian Vinagre

Supervisor: Dr. Joan Carles Mora
MA Thesis

Applied Linguistics and Language Acquisition in Multilingual Contexts
M.A. Program

Department of Modern Languages and Literatures and
English Studies

Academic year: 2019-2020

Màster Oficial en Lingüística Aplicada
i Adquisició de Llengües en Contextos Multilingües
LAALCM

Joan Carles Mora com a supervisor/a del treball (Tesina de
(nom i cognoms)

Màster) presentat com a requeriment per a l'avaluació de l'assignatura **Projecte de**

Recerca en Lingüística Aplicada

presentat per l'alumne/a: Miren Adrian Vinagre
(nom i cognoms)

amb el títol de: Malleability of newly-established and previously-established phono-lexical representations: a design

certifico que he llegit el treball i l'aprovo perquè pugui ser presentat per a la seva defensa pública.

I perquè consti i tingui els efectes oportuns signo aquest certificat en

Barcelona, a 7 de Setembre de 2020

Dr/a. Joan Carles Mora

Official MA programme in
Applied Linguistics and Language Acquisition in Multilingual Contexts (LAALCM)

Universitat de Barcelona

Non-Plagiarism Statement

This form must be completed, dated and signed and must be included at the beginning of every copy of the MA Thesis you submit for assessment.

Name and surnames:	Miren Adrian Vinagre
MA Thesis title:	Malleability of newly-established and previously-established phono-lexical representations: a design
Supervisor:	Dr. Joan Carles Mora

I HEREBY DECLARE THAT:

- This MA Thesis that I am submitting for assessment is entirely my own work and I have written it completely by myself.
- I have not previously submitted this work or any version of it for assessment in any other programme or institution.
- I have not used any other sources or resources than the ones mentioned.
- I have identified and included the source of all facts, ideas, opinions and viewpoints of others through in-text referencing and the relevant sources are all included in the list of references at the end of my work. Direct quotations from books, journal articles, internet sources or any other source whatsoever are acknowledged and the sources cited are identified in the list ofreferences.

I understand that plagiarism and copying are serious offences. In case of proof that this MA Thesis fails to comply with this declaration, either as negligence or as a deliberate act, I understand that the examiner has the right to exclude me from the assessment act and consequently all research activities conducted for this course will be declared null and the MA Thesis will not be presented for public defense, thus obtaining the lowest qualification.

Date: **1/09/20** **Signature:**

ACKNOWLEDGEMENTS

First and foremost, I am thankful to my supervisor, not only for his guidance but also for his kindness when I approached him at the very beginning of the course and shared my interest on Phonetics with him.

Secondly, to the professors of the MA course for their feedback and sharing their knowledge in the different courses.

And finally, I am thankful to everyone who encouraged me to pursue my goals and enroll this MA.

Initially, this thesis was meant to include a study, however, due to the sanitary condition (Covid-19) we were not able to collect any data. As a consequence, this paper only presents the research design that we created.

ABSTRACT

Previous phonetic training research has yet to provide evidence of its effectiveness in modifying L2 learners' inaccurate phonological representations in perception and production. In addition, no study to date has explored whether all the words that constitute learners' L2 lexicon are equally malleable and susceptible to improvement. The study we propose in this paper sets out to test, within a high-variability phonetic training (HVPT) paradigm involving AX discrimination, an identification, and immediate repetition tasks, i) the malleability of the phonological representation of L2 words targeting the difficult vowel contrast /æ/-/ʌ/. And ii) potential differential effects of training participants with lexical (words) or non-lexical (nonwords) materials. To compare the malleability of previously-established and newly-established phonological representations participants are asked to complete a series of word-learning tasks for the latter. Afterwards, they are trained on the target vowel contrast and tested on its perception and production in both word types (i.e.: *old* and *new*). To test the differential effects of training materials participants were randomly assigned to lexical or non-lexical training groups. Hypothesized results will be provided based on the evidence that the literature has provided.

Keywords: high variability phonetic training, phono-lexical representations, malleability, L2 perception and production.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Listening plays an important role in human communication, and for the most part, it is effortless when we communicate in our native languages (L1s). The situation is completely different when we listen to someone in a language that is not our own. Foreign language (FL) learners often struggle with similar sounding sounds/words prompting the activation of unwanted words during a conversation. For instance, L1 Spanish learners of English may wrongly activate *cut* (/kʌt/) instead of *cat* (/kæt/) because for these learners, /æ/ and /ʌ/ are acoustically similar, and they are likely to map both sounds onto their closest sounding vowel (Spanish /a/). Indeed, predictions on which phonemic contrasts could be troublesome to acquire by different L1 speakers can be explained by two influential models -the speech learning model (Flege, 1995) and the perceptual assimilation model (Best & Tyler, 2007)-.

From very early on, we learn considerable amounts of words, and our vocabulary in the FL keeps growing as we progress in its acquisition. The way we learn our first words may be crucial since they represent the first encounters with FL sounds. Initially, if we are not taught to perceive/produce differences in *cat-cut*, it is very unlikely that we will be able to do so at more advanced levels when we learn *gash-gush*. That is, if the contrast /æ/-/ʌ/ is not accurately encoded in the mental lexicon from the very beginning, phono-lexical representations of words containing that contrast are likely to be inaccurate. Recent evidence (Darcy & Thomas, 2009; Darcy & Holliday, 2019; John & Cardoso, 2020) suggests that many FL Learners function with inaccurate phono-lexical representations. For example, some L1 Korean learners perceive the English word *blue* as [bʊˈluː]. This is a big concern because if phono-lexical representations are inaccurate, we will not be able to activate the correct words when listening to someone speak in the FL, and our pronunciation will never reach native-like standards. It has been shown that learners can be trained to differentiate /æ/ from /ʌ/ acoustically, yet, are inaccurate phono-lexical representations "trainable"? Or do they work differently? If so, are all the words that constitute FL learners' mental lexicon equally "trainable"? And, finally, what is the cause for developing inaccurate representations? The purpose of this paper is to discuss these questions and present the design of a study that would advance our knowledge about the malleability of phono-lexical representations. To be more precise, it would aim to shine new light into the malleability[1] of *previously-established*[2] and *newly-established*[3] phono-lexical representations. The paper is structured into seven parts. The literature review will examine previous findings related to phono-lexical representations and gaps found in the literature. Parts three and four present the research design of the study. After presenting the research design, in part five, we will conduct some preliminary analysis of previously collected data[4]. Afterwards, some expected results will be advanced, and finally, part seven presents the conclusion and pedagogical implications.

2. LITERATURE REVIEW

2.1 ACQUISITION OF FL PHONOLOGY

Acquiring FL phonology entails the acquisition of the segmental inventory[5], phonotactic grammar[6] and phonological processes[7] of the FL. In addition, FL learners need to develop accurate phono-lexical representations, which are stored in the mental lexicon. The mental lexicon is a long-term memory store which contains information about the morphological, syntactic, semantic, phonological, and orthographic properties of words, that is, their lexical representation (Hayes-Harb & Masuda, 2008). For the purpose of his paper, we are mainly concerned with phono-lexical representations, which include information such as the meaning of a word and what it sounds like (Darcy & Holliday, 2019). We will also use the term 'FL categories" to refer to "the discrete elements that make up a [FL] phonological representation" (Boersma, 2010, p. 1). It is necessary to mention that this paper mainly examines studies which analysed learners' speech perception and production in a FL context, however, there are exceptions that include (second language) L2 learners. For this reason, we will use the abbreviation "FL", but "NNL"/"NN" will be introduced when referring to L2 learners. We shall start with the factors that influence the acquisition of FL sounds, and how

[1] "Capable of being changed, molded, trained, etc." (Collins Dictionary, n.d.).
[2] Representations of words that learners are likely to have learnt a long time ago. More detailed information will be provided in section 5.3.2.
[3] Representations of words that learners will be taught for the purpose of the study. More detailed information will be provided in section 5.3.2.
[4] These data could not be used to answer the research questions of the thesis, but section 6 will explain in detail the reasons to analyse it.
[5] Set of vowel and consonant sounds.
[6] Phonotactic grammar: "constraints that govern the arrangement of phonemes" (Darcy & Thomas, 2019, p. 1441)
[7] Phonological processes: variations in speech for easiness of articulation. E.g.: pronouncing *listen as* / ˈlɪsn̩/ instead of /ˈlɪsən/ (referred as syllabicity).

initial encounters with the FL sounds may or may not be related to the development of phono-lexical representations.

2.1.1 Factors influencing the acquisition of L2 categories

A large and growing body of literature has investigated the acquisition of non-native (NN) sounds in a FL context (Gomez Lacabex, García Lecumberri & Cooke, 2008; Aliaga-García & Mora, 2009; Goble, 2013; Cebrian & Carlet, 2014; Mora-Plaza, Mora & Gilabert, 2017; Cooke & Garcia Lecumberri, 2018). In this context, the opportunities for acquiring FL sounds are more limited than in a L2 context (Tyler, 2019). Among the challenges learners face, we find scarcity of input (Carlet & Kivistö de Souza, 2018) which is restricted to teachers' instruction and partially delivered in the learners' native language (NL/L1) (Muñoz, 2008). Some learners may be exposed to out-of-classroom activities such as watching movies, playing video games or travelling abroad, which may be carried out in the FL (Tyler, 2019). Therefore, differences in quality of exposure range from being exposed to a single-source NN teacher who may not pronounce correctly, or to a wider range of native speech (Garcia Lecumberri & Garcia Mayo, 2003). In line with this argument, in a FL context, it is very likely that learners will be exposed mostly to foreign-accented speakers (if not their teachers it could be their classmates). Having a foreign accent would not necessarily be a problem if teachers maintained the phonological distinctions of the FL, but this is not always the case (Tyler, 2019). Consequently, for those FL learners exposed only to teachers' speech which does not maintain the phonological distinctions, acquisition of FL categories would be more troublesome than for those exposed to native speakers. Another obstacle to FL phonological acquisition is the orthography of both the L1 and FL. It has been suggested that lack of grapheme-phoneme correspondence in the FL is partially responsible for making pronunciation one of the most difficult areas of language to acquire (Cenoz & Garcia Lecumberri, 1999). Indeed, Tyler (2019) adds that while written forms are an attractive method to learn a large number of words, it is usually at the expense of learning new words through perception.

Scarcity of input, foreign-accented speech and lack of grapheme-phoneme correspondence are some of the factors that make the acquisition of FL sounds difficult at the acoustic level. However, as mentioned earlier, learners do not only need to establish two separate phonetic categories corresponding to /æ/ and /ʌ/ at the acoustic level, they also have to accurately encode them into representations of words like *gash-gush* at the lexical level (see *Figure 1* for a better understanding of each level). *Figure 1* presents a model of speech perception and production processing proposed by Ramus *et al.*, (2010). According to this model, when we are exposed to auditory input, we analyse its acoustic patterns. From these acoustic representations, speech sounds are mapped onto phonetic categories. This process takes place at different stages referred to as "sublexical phonological representations". We advance from this stage to the lexical level, through lexical activation, selection, and retrieval, that is, through word recognition (Ramus et al., 2010). Considering that the acoustic and the lexical levels of representation are two aspects of phonology FL learners have to master, the questions that emerge are: what is the relationship between them? and, does one of them precede the other?

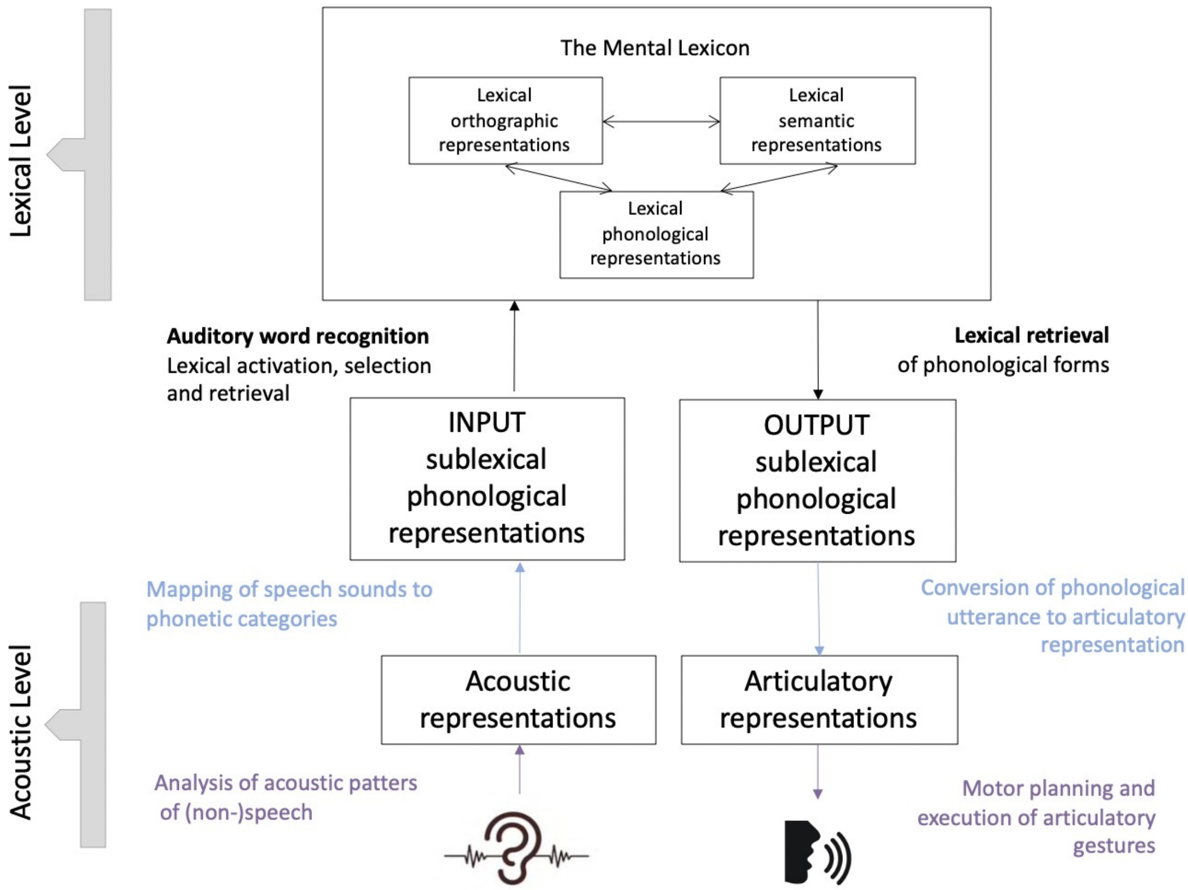

Figure 1. An information processing model of speech perception and production (adapted from Ramus et al., 2010, p. 313).

2.1.2 Relationship between the acoustic and the lexical level

The previous section opens the discussion to the relationship between the acoustic and the lexical level. There are two approaches currently being adopted: the "category first" and the "lexicon first". As for the former one (Pallier, Bosch & Sebastian-Gallés, 1997; Pallier, Colomé & Sebastian-Gallés, 2001; Dufour, Nguyen & Hans Frauenfelder, 2007), poor discrimination abilities at the acoustic level do not allow accurate mapping of speech sounds into phonetic categories, and this may be the cause for developing a single phonological representation[8] at the lexical level. Let's clarify what this means with an example of Spanish-dominant Spanish-Catalan bilinguals, being this the population studied by the first two groups of authors aforementioned. Spanish phonology only contains one mid front vowel (/e/), on the contrary, Catalan has a close-mid front vowel (/e/) and an open-mid front vowel (/ɛ/), therefore, those speakers learning Catalan whose L1 is Spanish would need to add this latter sound to their phonetic repertoire. Presumably, if a new category (Catalan [ɛ]) has been acquired at the acoustic level, learners would be able to distinguish minimal pair words containing the /ɛ/ - /e/ contrast. As a consequence, they may not develop homophonic[9] representations at the lexical level. However, this is not what Pallier *et al.* (1997) and Pallier *et al.* (2001) found. These authors proposed that since Spanish dominant-Catalan bilinguals could not distinguish /ɛ/ from /e/ acoustically, a single phonological representation (e.g.: [bens]) may have been stored in their mental lexicon for two Catalan lexical units (e.g.: [bɛns] – [bens] *to sell-to come*). Interestingly, these authors concluded that despite early exposure and being highly proficient in Catalan participants were not able to distinguish /ɛ/ from /e/ (Pallier *et al.*, 2001). In sum, for the "category first" authors, it seems that mastery of discrimination at the acoustic level precedes accurate performance at the lexical level.

According to the "lexicon first" approach, (Weber & Cutler, 2004; Cutler, Weber & Otake, 2006; Escudero, Hayes-Harb & Mitterer, 2008; Darcy *et al.*, 2012; Llompart & Reinisch, 2018) learners can establish a phonological contrast at the lexical level despite exhibiting constant perceptual errors. In other words, at the acoustic level, learners perceptually neutralize the contrast, yet, they manage to maintain a distinction at the lexical level. This finding seems to contradict that from the "category first" and suggests that "the establishment of a lexical contrast is independent of the previous acquisition of phonetic categories" (Darcy *et al.*, 2012, p. 28). So, how did these authors come to such a conclusion? In the following paragraph, we will explore this question with an example from the studies by Weber and Cutler (2004) and Escudero *et al.* (2008) who tested L1 Dutch learners of English.

L1 Dutch speakers experience difficulty discriminating the English /æ/-/ɛ/ contrast and are likely to map both L2 sounds to Dutch category /ɛ/. In order to examine speech perception

[8] Phonological representation: information about how a word sounds.
[9] Homophony: "the linguistic phenomenon whereby words of different origins become identical in pronunciation" (Collins Dictionary, n.d.).

and lexical processing of this contrast, Weber and Cutler (2004) and Escudero *et al.* (2008) employed eye-tracking technology. In these experiments, participants were presented with four images on the screen: two distractors (e.g. *strawberry* and *dress*); one target (e.g.: *panda*); and a competitor (e.g.: *pencil*). Using auditory input, subjects were instructed in the following way: "Click on the panda. Now put it on top of the circle.". Results from Weber and Cutler (2004) and Escudero *et al.* (2008) revealed that when participants listened to the onset of a target word like ***pan**da,* they activated its competitor word ***pen**cil*, this implied looking at the picture of the pencil until the second syllable was heard. When presented with the word *pencil*, the scenario was different, *pen-* did not activate *pan-*. This asymmetric fixation pattern indicates that the contrast /æ/-/ɛ/ has been encoded lexically but in a non-target like manner. The asymmetry is driven by the dominant category [ɛ] which is acoustically similar to Dutch [ɛ], prompting both members of the contrast (/æ/-/ɛ/) to be mapped onto /ɛ/ (Weber & Cutler, 2004; Cutler *et al.*, 2006).

Both the category-first and the lexicon-first approaches have failed to directly disentangle whether accurate perception of NN contrasts is a pre-requisite so that it is accurately encoded in the mental lexicon. From a pedagogical perspective, understanding how learners can encode a contrast accurately would be crucial for the design of classroom-tasks, and/or phonetic training. The evidence that has been presented of learners encoding a phonological contrast in the mental lexicon showed it was non-target like, hence, we might be interested in exploring the following questions: how are these non-target like FL contrasts encoded in the mental lexicon? And, how is this related to the development of phono-lexical representations? The section that follows is concerned with these questions.

2.2 FL PHONO-LEXICAL REPRESENTATIONS IN THE MENTAL LEXICON

Examining the mental lexicon is a challenging task given that it appears to be rather abstract. This section is divided into three subsections: i) how sounds are encoded in the mental lexicon; ii) how phono-lexical representations change over time; and iii) phonetic training and encoding of phonological contrasts in the mental lexicon.

2.2.1 Encoding of FL phonetic contrast in the mental lexicon

Let us first consider how contrasting pairs of sounds may be encoded in the mental lexicon based on Hayes-Harb and Masuda's (2008) paper. These authors tested L1-English L2-Japanese learners on geminate (e.g.: /tt/) and singleton consonants (e.g.: /t/). After analysing learners' speech, researchers suggested that participants may have encoded the geminate consonant as /t*/. The asterisk indicates that learners still did not know what that sound (/tt/) was, but it was not confused with the singleton /t/. If we were to apply this conclusion to the case of L1 Dutch learners of English, the contrast /æ/-/ɛ/ may have been inaccurately encoded as /ɛ*/-/ɛ/. We cannot confirm that this was the case because testing methods only allow us to observe whether a contrast is accurately or inaccurately encoded. With respect to the form in which a contrast is represented in the mental lexicon, research to date cannot analyse such aspects, but Hayes-Harb and Masuda's (2008) notion of the * may be suitable. But, how is inaccurate encoding related to the establishment of phono-lexical representations? According to Darcy, Daidone, and Kojima (2013) inaccurate encoding of a contrast gives way to *fuzzy phono-lexical representations*. A fuzzy representation could be defined as follows:

> Mental representation of phonolexical form that does not represent the word as a fixed phonological sequence. Such a representation may leave some phonemes underspecified (e.g., either a final /d/ or /t/) or contain some uncertainty (and ensuing optionality) regarding the exact phonemes and their sequence (Cook *et al.*, 2016, p. 3).

Given that inaccurate encoding of the phonological contrast might be the cause for developing fuzzy phono-lexical representations, we should consider what is interfering with its accurate encoding.

Up till now, learners who encoded a contrast lexically in a non-target like manner experienced difficulties in discriminating one sound from another, as in ***pan**da-**pen**cil* (Weber & Cutler, 2004; Cutler et al, 2006; Escudero *et al.*, 2008; Darcy *et al.*, 2012; Llompart & Reinisch, 2017). Therefore, one would hypothesize from these results that accurate discrimination of a given contrast is a pre-requisite to encode it accurately. This assumption is partially challenged by the work of Hayes and Masuda (2008), Darcy et al. (2013), Amengual (2016), Simonchyk and Darcy (2017) and Darcy and Holliday (2019) which suggests that accurate discrimination of a contrast in isolation will not guarantee accurate encoding. Hayes-Harb and Masuda (2008), for instance, propose that learners may need to have additional information such as words spelling to encode a contrast accurately.

It seems like there may be many factors interfering with the encoding of a phonological contrast, and, in turn, with the development of accurate phono-lexical representations. All in all, most literature seems to suggest that learners do not manage to encode NN contrasts accurately and as a consequence, they operate with fuzzy phono-lexical representations. After reviewing these findings, one wonders if fuzzy phono-lexical representations can become more target-like over time, and if that is the case, how this "change" happens. These concerns will be discussed in the following section.

2.2.2 Malleability of phono-lexical representations

Recent evidence (Darcy & Thomas, 2019; Darcy & Holliday, 2020; John & Cardoso, 2020) suggests that accurate phono-lexical representations are achievable for words which have been previously misrepresented. It is important to emphasize that this evidence is limited to perception, and there is still no study which has looked at the form of phono-lexical representations through production. In Darcy and Thomas (2019), for instance, L1 Korean learners perceived both [bʊˈluː] and [bluː] to be the English word *blue*. The phonotactic grammar of Korean does not allow word-initial obstruent-liquid clusters[10], for this reason they perceptually repair the input by adding an epenthetic vowel[11] ([ʊ]) (Darcy & Thomas, 2019). This example appears to be consistent with the claim that accurate representations are attainable for words which were misrepresented in the mental lexicon. For the moment, the studies by Darcy and Thomas (2019), Darcy and Holliday (2019), and John and Cardoso (2020) have demonstrated that accurate phono-lexical representations are achievable, yet, how does this "update"[12] happen? The next paragraph will discuss Darcy and Holliday (2019)'s view on how representations seem to get "updated".

Darcy and Holliday's (2019) pioneering study is of great significance as it marks the first attempt to investigate if the "update" of phono-lexical representations depends on the age a word has been learnt -age of words hypothesis-. According to this hypothesis, the latest acquired words will be the most accurately represented. A second possibility they propose is that all the representations are "updated" once a new contrast is acquired -phonological update hypothesis-. These authors tested L1 Mandarin learners of Korean and compared recent[13] and old words in perception. It was the absence of a statistically significant effect that supported the phonological update hypothesis, but descriptively, the researchers found phono-lexical representations of recently-learnt words to be easier to "update". One must bear in mind that participants were tested after nearly two years of residence in Korea. This is important to note

[10] Cluster: group of consonants which have no vowel in between (e.g.: /**bl**uː/, ***bl**ue*).
[11] Epenthetic vowel: vowel that is inserted to repair phonological sequences (Darcy & Thomas, 2019)
[12] The term *update* will be used with inverted commas throughout the paper because inaccurate phono-lexical representations will not disappear and become accurate. Accurate representations will coexist together with inaccurate ones in the mental lexicon.
[13] The classification of *old* and *recent* words was based on a Likert-scale questionnaire. The authors do not specify how much is "recent".

because exploring how "update" takes place requires testing participants before and after a phonetic training, or in the case of these L2 Korean learners, before and after a period of living in the L2 community. Gains (T2-T1) observed for each type of word may provide an answer to how "update" happens when being exposed to native input.

L2 learners in Darcy and Holliday (2019) and Darcy & Thomas (2019) achieved accurate phono-lexical representation over time for some words. This finding invites to look for a phonetic training that will help learners encode a contrast accurately, and that will "update" FL learners' inaccurate phono-lexical representations.

2.2.3 Phonetic training beyond the acoustic level

Considerable work has accumulated on the role of phonetic training for the perception and/or production of FL phonological contrasts (Gomez Lacabex, *et al.*, 2009; Goble, 2013; Cooke & Garcia Lecumberri, 2018). Far less studied is the effect of phonetic training on phono-lexical representations. None of the studies mentioned in this literature review carried out phonetic training. To the best of our knowledge, only the papers by Escudero *et al.* (2008) and Llompart and Reinisch (2017) provide some guidance to design a training that may help learners encode a FL contrast accurately in the mental lexicon. These authors included word-learning tasks through a visual-word eye-tracking paradigm during one session of no longer than an hour. Participants in these tasks received instructions through headphones (e.g.: "Click on the [tɛnzə] and then on the triangle"). On the one hand, in Escudero *et al.* (2008), learners obtained feedback in two different ways: one group was exposed to the orthographic, visual, and phonological form of the words (e.g.: this is a *tenzer*), while the other group did not have the orthographic form. By comparing both groups, Escudero *et al.* (2008) concluded that adding orthographic forms of words along with auditory input resulted in a more accurate encoding of the phonological contrast. On the other hand, Llompart and Reinisch (2017) exposed learners to visual and articulatory information. In the former condition, a group of learners was exposed to videos of NSs producing the target words. For the latter condition, another group was asked to repeat the utterances produced by the NSs. This procedure was repeated twice and allow learners to compare their own production to that of the NS. According to Llompart & Reinisch (2017), these practices helped learners acquire more articulatory knowledge about the L2 contrast, which in turn, led to better encoding of the contrast.

A commonly used phonetic training for perception and production of FL sounds is High Variability Phonetic Training (HVPT), and it has demonstrated to be effective (Aliaga-García & Mora, 2009; Cebrian & Carlet, 2014; Carlet & Cebrian, 2019; Ortega, Mora & Mora-Plaza, submitted). By means of this training, learners encounter the target sounds in a variety of phonetic environments (i.e., preceded and followed by a variety of sounds) and produced by different native voices (Barriuso & Hayes-Harb, 2018). However, it remains unexplored whether HVPT is sufficient to tap into the mental lexicon and, therefore, phono-lexical representations. It would be interesting to observe if positive results with this training can be transferred to the "update" of phono-lexical representations. Indeed, this is one of the questions that the study that we present in the following section aims to address.

3. THE STUDY

The study builds on previous findings from the literature review and aims to address whether *previously-established* phono-lexical representations are as malleable as *newly-established* representations. The phonological contrast that we examine in this study is /æ/-/ʌ/, which is a challenging contrast for L1 Spanish-Catalan bilinguals because it is often assimilated to Spanish /a/ (Cebrian, 2019). The study also sets out to assess whether accurate discrimination of the /æ/-/ʌ/ is needed to develop accurate phono-lexical representations. To explore this, a cohort of L1 Catalan-Spanish learners of English will undergo a HVPT bracketed by an identical pre-and post-test. Previous research (Darcy & Holliday, 2019; Darcy & Thomas, 2019; John & Cardoso, 2020) has shown that learners operate with inaccurate phono-lexical representations, therefore, we decided to include two different types of stimuli in the training: non-word vs. word. This comparison would allow us to observe whether learners can take more advantage from a training in which the stimuli will not prompt the activation of previous representations that are likely to be inaccurate. To our knowledge, there is no study comparing the malleability of phono-lexical-representations after a phonetic training in both *old* and *new* word forms neither at a perceptual nor productive level. Apart from this, no previous study has tested the effectiveness of HVPT for inaccurate phono-lexical representations and/or if this training can make phono-lexical representations more accurate. The design of the study will allow us to examine whether the benefits of HVPT can be extended to the development of accurate phono-lexical representations and how this "update" happens. It does not consider the possibility of removing completely inaccurate phono-lexical representations due to a limited HVPT of 2 hours and 20 minutes.

As a consequence, the study would seek to address the following questions:

RQ1: Can HVPT "update" the phono-lexical representations of *old* and *new* word forms containing the /æ/-/ʌ/ contrast in EFL learners?

RQ1.1: Are there differential training gains as a function of the lexical status of the training stimuli (*words* vs. *nonwords*)?

RQ2: Are the phono-lexical representations of *old* and *new* word forms equally malleable?

RQ 3: Do EFL learners need to be able to discriminate /æ/ from /ʌ/ to develop accurate phono-lexical representations for words that contain /æ/ and /ʌ/?

Data collection for this study was not possible due to the sanitary condition. For practical purposes, no hypotheses are advanced in this part, but they will be presented further below in a different section.

4. METHODOLOGY

4.1. Participants

Participants for the study would be L1-Catalan-Spanish bilinguals studying the first year of English Studies at the University of Barcelona. Their level of proficiency is expected to be B1-B2 up to C1. A language background questionnaire will be distributed like the one shown in *Appendix A*. They will be randomly assigned to three groups: two experimental and one control.

4.2. Research Design

L1-Catalan-Spanish learners of English will be tested on their ability to perceive and produce English near-open front unrounded vowel (/æ/) and open-mid back unrounded vowel (/ʌ/) in *old* and *new* words before and after a HVPT. Participants are expected to perform the testing and training tasks in seven sessions, separated by at least a day in between. In sessions 2 and 7, participants will be assessed in their discrimination of the /æ/-/ʌ/ contrast (ABX), their ability to produce words in isolation (delayed word repetition (DWR)), and the malleability of their phono-lexical representation (through a lexical decision (LD) and a delayed sentence repetition (DSR)). Four sessions (3,4,5,6) will be devoted to the training tasks which include AX discrimination and identification (ID) (for perception) and immediate repetition (for production). Our study compares the phono-lexical representations of *old* and *new* word forms. Thus, we needed to include a word-knowledge test and word-learning activities for participants to learn the *new* words before the phonetic pre-test. In session 1 and 2, participants will perform a set of activities (for 30 minutes). Both sessions include the same *new* words, however, on the first session activities are designed for learners to memorize the *new* words, while on the second session, *new* words have to be used in meaningful sentences. In session 3,4 and 5, participants will also be tested on their knowledge of the *new* words to validate that these new words have been stored in long-term memory and are comparable to the *old* words. Measures of overall L2 proficiency (Elicited Imitation (EI): Ortega *et al.,* 2002) and the LLAMA battery of tests will be obtained in session 1, along with the consent form. EI was chosen because different studies have reported EI to be an effective method to test overall L2 proficiency (Yan, Maeda, LV & Ginther, 2016). The EI allows us to control for differences in proficiency. The LLAMA subtests allow accounting for individual differences in sound recognition (LLAMA_E), and phonemic coding ability (LLAMA_D) (Roger, Meara, Barnett-Leght & Curry, 2017).

Figure 2. Research design

4.3 Instruments

4.3.1 Speech Materials

Training and part of the testing stimuli will be drawn from the PTC Project (information about this project will be provided further below in section 5). The target sounds that would be analysed are the near-open front unrounded vowel (/æ/) and the open-mid back unrounded vowel (/ʌ/). L1 Spanish speakers have considerable difficulties in teasing these vowels apart since they are perceived as Spanish /a/, as shown by Cebrian (2019), so it is a contrast worth investigating. This contrast would appear embedded in monosyllabic CVC real words and non-words produced by 6 British English speakers (3 males + 3 females) with the target vowels appearing in a variety of phonetic environments.

4.3.2 Vocabulary: word- learning and testing

This study looks at two types of words, 16 *old* words, which presumably participants would know before the training, and 16 *new* words, which they will learn for our study (see *Appendix B*). This categorization is based on: i) the frequency[14] of the selected words (i.e., *high* or *low* frequency) (see *Appendix C,* Table 2) and ii) the rating scores obtained from a 7-point-Likert-scale word familiarity questionnaire filled out by B2-C1 EFL learners in the PTC Project (see *Appendix C,* Table 3).

To teach participants the set of *new* words, two days of thirty minutes have been included in the design based on previous studies which have reported that short periods (30-60 minutes) of word-learning are enough to learn about 30 words (Escudero *et al.* 2008; Shatzman & McQueen, 2006). The activities were piloted with four EFL learners and the time set was enough for learners to learn all the words. Both days will be devoted to memorization and practical activities through *Quizlet* (Sutherland, 2007) (see *Figure 3*).

For participants to memorize the *new* words, we will be using PowerPoint slides that present each word with its corresponding picture, definition, translation into Spanish and pronunciation. This would allow learners to associate the orthographic, phonological, visual, and semantic forms of the words. As for the activities, *Quizlet* was considered the appropriate platform because it provides individual feedback. Researchers like Beyer and Lynch (n.d.), Bar (2016), and Chaikovska and Zbaravska (2020) have provided evidence that learners using Quizlet learn a larger amount of words than students who only use flashcards or pen and paper traditional methods. In particular, Bar (2016) found that students who used a wider range of activities offered in Quizlet performed better than those who only used flashcards and matching games. As a consequence, we decided to include the following list of activities: multiple-choice, dictation, fill-in-the-gaps, and matching games.

[14] Measures of frequency have been taken from SubtlexUS database. The information was taken from SUBTLWF, since "it is a standardized measure of word frequency independent of the corpus size" (Brysbaert & New, 2009, p. 988).

Learners will start with a multiple-choice. They will be presented with a picture and four different word options, they have to select which of the options corresponds to the image. This would help them associate the orthographic form with its visual representation. Afterward, they will perform a dictation activity. Participants will first listen to a word, and then the screen will display its corresponding picture, their task is to write down the word correctly. This activity aims at connecting the phonological representation of a word with its corresponding visual and orthographic form. Subsequently, they will complete a different type of activity which requires to write down the word shown as a picture on the screen, we will refer to this activity as a "fill in the gaps". This activity would help to link orthographic and visual information. Finally, the last activity consists of matching the picture with its corresponding word. In this case, participants will have the 16 *new* words at the same time. This activity also aims at associating orthographic and visual information (see *Appendix E.1* for examples of each activity). On the second day, they will perform the same activities, however, there will be no visual information and learners will have to use the words in sentences (see *Appendix E.2* for examples). The following paragraph describes how participants will be tested on their knowledge of the *new* words.

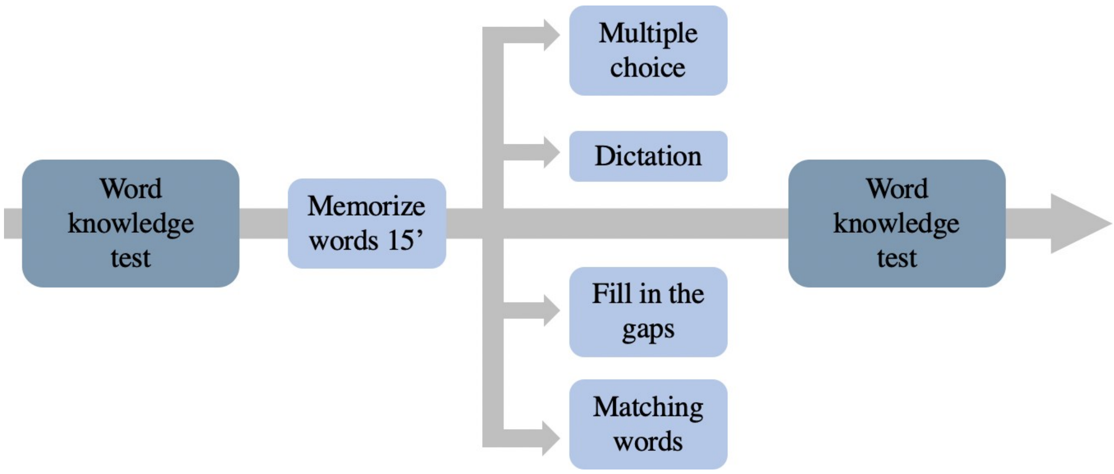

Figure 3. Example of the word-learning task design.

As I mentioned previously, words have been pre-categorized as *old* and *new,* however, in the first session, participants will be tested on both word types to make sure that the classification is accurate. Vocabulary knowledge testing consists of presenting (in a randomized order) the set of words and requires participants to write down the translation into their L1 (see *Appendix D*). Participants will be tested on their knowledge of the *new* words with the same test, before and after the word learning activities to track their learning progress. We decided to include two vocabulary-testing sessions because there is evidence that a period of sleep may be necessary for lexical competition to occur (Dumay & Gaskell, 2007). Besides, every session includes word-knowledge testing because we want to test if these *new* words are stored in the long-term memory so that they are comparable to the *old* words.

4.3.3 Testing procedure

4.3.3.1 Perception and production (pre- and post-tests)

During the pre- and post-tests, which are identical in all respects, participants will be tested in a set of tests to measure their discrimination abilities as well as the form of their phono-lexical representations. To test their perceptual discrimination abilities, they will perform an ABX test. In this test, participants have to decide in an interval of 2500ms whether the third word (X) matches the first (A) or the second (B) by pressing the designated key on the computer keyboard (e.g.: A cat B cut X cat). This will be presented in four counterbalanced orderings (ABA, ABB, BAA, BAB). The ABX test consists of a total of 136 trials (30 test trials x 4 orders) produced by two voices (see *Appendix F,* Table 4). Apart from this, there are 16 control items (e.g.: A shab B sheeb X shab) to ensure subjects' accurate performance in the test (see *Appendix F,* Table 5). A LD test will be used as a measure of sensitivity to the lexical encoding of the contrast, to compare how well the phonological contrast has been encoded at both pre- and post-test. Apart from this, it also allows to compare malleability of phono-lexical representations at a perceptual level. In the LD test, participants are presented auditorily with a single word produced by a female voice, and they have to determine whether it is a real English word or not by pressing the corresponding labelled key on the computer keyboard. They will have an interval of 2500ms to respond, and there is a total of 56 trials (28 control trials + 28 test trials) (see *Appendix F,* Table 6).

As regards production, participants will complete a Delayed Word Repetition (DWR) and a Delayed Sentence Repetition (DSR) test. On the one hand, the DWR test looks at the accuracy in producing the words in isolation. On the other hand, for the DSR test learners have to focus on meaning, process the sentence and retrieve it from memory. This test involves a meaningful context. The DSR test serves as a measure for the accuracy of phono-lexical representations in production. In the DWR, participants are presented auditorily with a word and they have to repeat it as accurately as possible after a *beep* sound. This test consists of a total of 76 trials (60 test trials + 16 control trials) (see *Appendix F,* Table 7). In the DSR test, participants will be presented with a sentence on the screen, they will have to read before it disappears from the screen. Afterwards, they will listen to that sentence through the headphones and will have to repeat it as accurately as possible after a *beep* sound. The DSR contains 32 trials (16 sentences) (see *Appendix F,* Table 8). Both the DWR and the DSR include a female and a male voice.

All these tests will be run through the software *DmDx* (Forster & Forster, 2003) in the GRAL lab, and will have a short practice with feedback for accuracy prior to the testing. ABX, DWR and DSR contain the same *old* and *new* words, whereas the LD contains a different set of *new* and *old* words. This was necessary because stimuli in the ABX, DWR and DSR tests are minimal pairs (e.g.: cat-cut), while for the LD, it is required that a word is only possible with one member of the contrast (e.g.: black-*bluck) (see *Appendix F,* Table 9 for an easier understanding of the differences).

4.3.3.2 Proficiency

Participants' proficiency will be measured through an elicited imitation (EI) test (Ortega *et al.*, 2002). In this test, learners are presented auditorily with a recorded stimulus and they have to repeat it as accurately as possible (Ortega *et al.*, 2002) (see *Appendix H*). The test assumes that those learners who are capable of repeating the stimuli with easiness had previously acquired the grammatical features that the EI test contains. EI has not only shown to be an effective method to test overall L2 proficiency but it also distinguishes learners across different proficiency levels (Yan *et al.*, 2016[15]).

4.3.3.3 **Aptitude tests: *LLAMA_E and LLAMA_D***

The LLAMA test battery (LLAMA-E, and LLAMA-D) has been included in the design to account for learners' individual differences in language learning aptitude. This would allow to group participants according to their aptitude profile. LLAMA-F has not been included because we are not concerned about learners' ability to infer grammatical rules. And LLAMA LLAMA-B was excluded as well because this subtest examines associative memory between unknown names and unknown objects (Roger *et al.*, 2017).

Aptitude for repeated sound-recognition and sound-symbol correspondence will be measured employing LLAMA-D and LLAMA-E respectively. Those learners scoring high in LLAMA-D are considered to be capable of noticing small variations in speech since they are

[15] This is a reference from a meta-analysis based on 1089 participants from 21 different studies.

able to "recognize repeated stretches of sounds" (p. 51). This test consists of two phases, a first one in which test-takers are presented with sound clips in a non-existent language, and a second one in which they have to indicate if the utterance presented has been heard (smiley emoji) or not (sad emoji) in Phase 1 (see *Figure 4*). The program will display both repeated and new sound clips (Roger *et al.*, 2017). In Saito, Suzukida & Sun (2019) they found the subtest LLAMA-D to be related to accuracy in L2 segmental attainment.

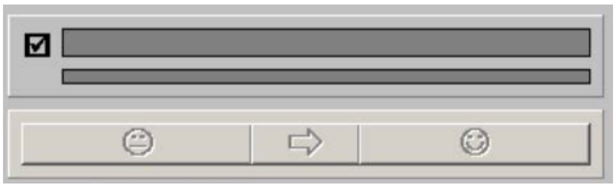

Figure 4. Part of the interface of LLAMA_D.

As regards LLAMA-E, it consists of 24 labelled Roman alphabetic buttons which correspond to syllabic sounds. Learners in Phase 1 will need to get acquainted with how each labelled button corresponds to a particular syllable structure and how it sounds. In Phase 2, they are presented auditorily with "words" made up from these syllables, and they will have to decide which of the two spellings they are presented with matches the "word" they have heard (Roger *et al.*, 2017) (see *Figure 5*). As for LLAMA-E, it has been reported to be related to L2 pronunciation accuracy, however, it must be acknowledged that this relationship was medium to small (Saito, 2017).

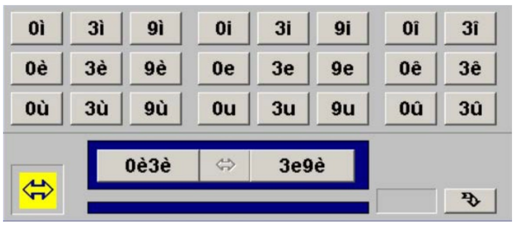

Figure 5. Example of LLAMA_E.

4.3.4 Phonetic Training

The phonetic training chosen for the study is HVPT under two conditions: a lexically-oriented training (real word stimuli (WDT)) and a phonetically-oriented training (nonword stimuli (NWDT)). The procedure is the same for both types of training. The training tasks are different from those of the testing in order to avoid familiarity with the testing methodology. Apart from this, this type of training has been used before and it proved to be effective. Participants will perform an AX discrimination task (96 trials with feedback for each group in each session (96x4384) produced by 2 females and 2 males) and an identification (ID) task (32 trials with feedback for each group in each session (a total of 128) produced by 2 females and 2 males) (see *Appendix G,* Table 10). In the AX, trainees are exposed to auditory input and they have to decide whether the words they listen are the same or different. In the ID task, trainees listen to a single word and have to identify to which picture on the screen it corresponds to. As regards production, they will perform an Immediate Repetition task in which they had to repeat the words/non-words as accurately as possible (a total of 128 practice trials, 32 trials x 4 sessions, for each group produced by 2 females and 2 males) (see *Appendix G,* Table 11). Trainees are exposed auditorily to an item they have to repeat, they will be presented with this item a second time, and would have to repeat again the same utterance. This allows self-monitoring on accuracy of imitation since they are able to compare the production heard auditory and their concurrent repetition. This type of task has been shown to be beneficial for the encoding of a contrast in Llompart and Reinisch (2017).

4.3.5 Data Analysis

Training effectiveness and differential gains as a function of word type (*old* or *new*) will be assessed for both training groups (NWDT, WDT). Gains in ABX discrimination will be assessed through a Generalized Linear Mixed Model (GLMM) in which *Time* (T1, T2), *Training Type*, and *Vowel* (/æ/, /ʌ/) are set as fixed factors ad *Subject* and *Item* as random factors. For the LD test, accuracy scores will be fitted to a GLMM with the same fixed and random factors as in the ABX. Apart from GLMM, a *d*-prime[16] measure avoiding response bias will be calculated. This measure includes both word and non-word items and takes into account the following measures: hit (words judged as words), false alarm (non-words judged as words), miss (word judge as nonword) and correct rejection (nonword judged as nonword). For the production tests (DWR, DSR), vowel frequency measurements will be extracted. Only values from frequencies *f*0, F1, F2 will be taken into consideration, *f*0 refers to the physical property of a sound, F1 to height and F2 to frontness. Vowel frequency values are given in Hertz (Hz), however, to minimize any vocal track difference effects, Hz will be converted into Bark[17] (B0, B1, and B2). Afterwards, Bark-normalized distance metrics will be

[16] *d*-prime formula: d' = z(H) - z(F) (H indicates the number of Hits and F the number of false-alarms)
[17] Formula to transform Hz into Barks: Zi = 26.81/(1+1960/ Fi)-0.53 (Syrdal & Gopal, 1986).

obtained through a normalization procedure[18] (B1B0 (height) and B2B1 (frontness)). Spectral distance scores (SDS) from both DWR and DSR will be fitted to a Linear Mixed effects Model (LMM)[19] with *Time* (T1, T2), *Training Type*, *Word Type* (old, new), and *Vowel* (/æ/, /ʌ/) as fixed effects and *Subject* and *Item* as random factors.

Finally, chi-square tests will be conducted between performance in ABX and LD at T2 to determine whether EFL learners need to be able to discriminate /æ/ from /ʌ/ to start developing accurate phono-lexical representations for words that contain /æ/ and /ʌ/. We will set two thresholds: performance above 85% in the ABX and above 70% in the LD. For the ABX, above 85% is considered to be a good score even though it does not reach native-like standards. For the LD, we interpret performing over 70% to indicate that learners are starting to develop accurate phono-lexical representations. Apart from this, correlation analyses will be run to observe if learners who are highly skilled in their production of words in isolation also perform at nearly native-like levels when words are embedded in a meaningful context.

As noted earlier, data collection was not possible, however, we will now be taking a look at some other available data from a large project (Phonetic Training Condition (PTC) Project). These data have been included because i) we can assess the effectiveness of HVPT in "updating" phono-lexical representations and ii) its methodology served as a standpoint illustrating which changes needed to be made to the word-learning task. The following section will explain briefly which groups from the PTC Project were taken into consideration and why. A comparison of the PTC methodology and our design for the word-learning part will be presented as well. Criticism of the methodology only applies when analysing it from the perspective of the study proposed in sections 3 and 4. The data presented in the following section have not been analysed in the way it will be done in this paper.

5. PHONETIC TRAINING CONDITIONS (PTC)[20] PROJECT

Aims from this project were very varied, however, researchers collecting data for the PTC Project were also interested in the comparison of a phonetically- and a lexically-oriented training and comparing the effects of a HVPT in *old* and *new* words (see *Appendix I* for more

[18] Euclidean distance formula: $\sqrt{(VaB2 - B1) - (VbB2 - B1))2 + ((VaB1 - B0) - (VbB1 - B0))2}$
Va and Vb correspond to the vowels under analysis.

[19] GLMM does not allow quadruple interactions, so for analysis to be possible, we will need to split the file by one of the variables (e.g.: *training type*). We want to avoid this for production, and since SDS is a continuous variable, we are able to carry out LMM analyses.

[20] Data collected by: Cristina Aliaga, Pace Bailey, Eva Cerviño, Joan C. Mora, Ingrid Mora-Plaza and Mireia Ortega (alphabetically)

information on PTC project). Of the initial cohort of 120 participants that were recruited, only 26 will be analysed. These participants underwent a non-word (NWD, 14, EI: 80.9%) and a word training (WD, 12, EI: 74.6%), which are the same conditions as those of the study outlined above. *Figure 7* presents the research design from the PTC Project, including only the word-learning part, the testing methodology, and the training.

Figure 6. Research design from the PTC project.

Before the experiment took place, learners filled out online a word familiarity questionnaire from which researchers were able to categorize words as *known* or *unknown* (I will refer to them as *old* and *new* respectively). To test participants' knowledge of these words, they were asked to translate orally the meaning of each word, which was presented through a test in *DmDx*. As illustrated in *Figure 7*, learners performed the same word-learning task in every session. It consisted of identifying the word they heard auditorily with its corresponding image on the screen. The word-learning task was 100% based on listening, which may have interfered with the phonetic training. In our design, we included a variety of activities that were not only based on listening, and most importantly, that were not included after the phonetic pre-test. Last but foremost, we aim to compare the malleability of *old* and *new* word forms in the mental lexicon, for this to be possible the LD and DSR must include the set of *old* and *new* words. This was not the case in the PTC project, only the ABX and the DWR contain a set of *old* and *new* words (see *Appendix J*, Table J.1 for this list). However, we observed that the DSR contained a number of highly familiar and less familiar words which we decided to analyse separately. In the following pages the reader will indeed encounter a comparison of *old* and *new* words of data coming from the DSR. Yet, this categorization[21] is based on the following criteria: if the word was highly familiar for the learners, it was classified as *old*, on the contrary, if it was not so familiar to the learners it was categorized as *new*. The phono-lexical representation of a highly familiar word would presumably be more robust than for a less familiarized word, this is why they have been classified as *old* and *new* respectively. The majority of the words fall within the category of *old*, and those classified as *new* were only three (*fuss*, *numb,* and *van*). One could argue this cannot be taken as a reliable comparison, but this is just a preliminary analysis of previously collected data that was worth examining. Our design has proposed an even comparison of *old* and *new* words, the latter being taught for the experiment.

Effectiveness of the HVPT for discrimination (measured through ABX) and production of words in isolation (measured through DWR) has been analysed in *Appendix J.* This was considered necessary because we hypothesize that learners' accuracy to distinguish /æ/ from /ʌ/ and their ability to produce words in isolation precedes accurate performance in lexical contexts (measured through LD and DSR). Results from the ABX and DWR tests showed that the HVPT was beneficial in improving trainees' i) discrimination of the contrast and ii) pronunciation of the words in isolation. Consequently, the next section examines the effectiveness of the HVPT for the "update" of phono-lexical representations.

5.1 HVPT and FL learners' phono-lexical representations

Phonetic training efficiency was assessed through a Generalized Linear Mixed Model (GLMM) for the LD test. Nonword rejection accuracy scores were fitted to GLMM with *Training Type* (nonwords, words), *Time* (T1, T2), and *Vowel* (/æ/, /ʌ/) as fixed effects, and *Subject* and *Item* as random factors. GLMM revealed a main effect of *Time* ($F(1, 790) = 8.65$, $p=.003$). None of the other fixed effects or interactions reached significance, which suggests that there was an improvement for both groups and both vowels (see *Appendix J.3*, Table J.3.1). Bonferroni-adjusted pairwise comparisons revealed that the main effect of *Time* was mostly driven by EFL learners' improvement in the NWDT condition and for both vowels (/ae/: $t(790)=2$, $p=.046$; /ʌ/: $t(790)=2.27$, $p=.024$) (see *Appendix J.3*, Table J.3.2). These results seem to indicate that only the nonword training (NWDT) helped trainees "update" the form of their phono-lexical representations. This may be due to the fact that the stimuli from the word

[21] This classification has been done by the author of the paper for its convenience use, not by the researchers from the PTC Project. Words categorized as *old* obtained a 6.82/7 rate of familiarity while *new* words obtained 4.1/7.

training (WDT) were highly-familiar words. It is very likely that these words were misrepresented and have minimized the effectiveness of the lexically-oriented HVPT.

A closer examination to the LD results demonstrated that on average test words showed higher accuracy rates than nonwords at T2, for both groups and both contrasts indicating that learners have not yet encoded the contrast accurately (NWDT: test words, /æ/-/ʌ/ = 77% > test nonwords, /æ/-/ʌ/ = 44%; WDT: test words, /æ/-/ʌ/ = 81% > test nonwords /æ/-/ʌ/ = 42%) (see *Figure 7* and *Appendix J.3*, Table J.3.3). In the LD, we will focus on the performance of nonword rejection since we take successful rejection of a non-word to mean that the phono-lexical representations are accurate. On the contrary, failure to reject a non-word would suggest that representations are still imprecise. On average, for test nonwords, the NWDT group improved to a larger extent than the WDT group (M *diff,* 13%, and 6% respectively, see *Figure 8*) as we observed in the GLMM analysis. Response latencies (RTs) for correct responses were also taking into consideration and the pattern is similar, both groups took longer to answer at T2, but it was the NWDT that responded faster (M *diff,* NWDAT= 50ms; WDAT= 115ms) (see *Figure 9* and *Appendix J.3*, Table J.3.3). Altogether, these results suggest the training was more effective for the NWDT group both for test words and nonwords, yet, inaccurate phono-lexical representations are still perceivable, as reflected by the percentage of accuracy in nonword rejection (<50%). No comparisons of *old* and *new* words can be made for the LD test since all the words from the stimuli are considered to be *old*. *Figure 7* displays accuracy scores for nonwords and words in the LD test separately.

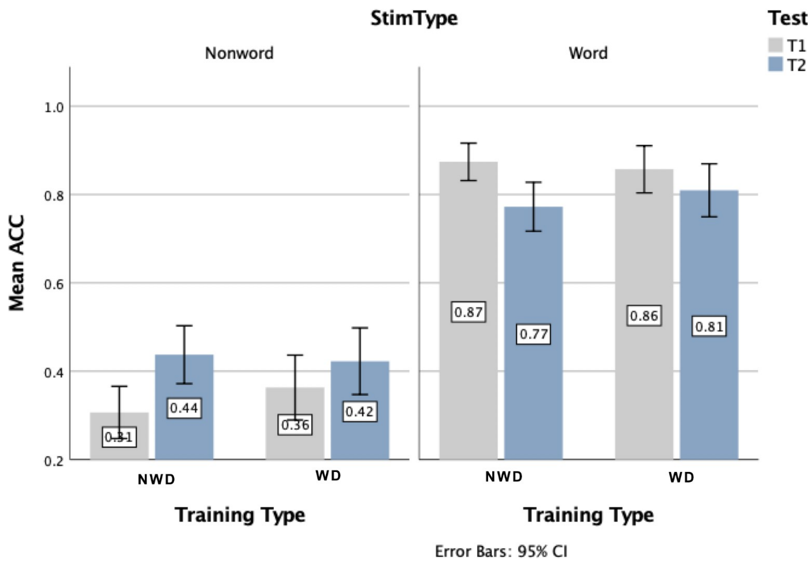

Figure 7. Bar graphs for accuracy scores for /æ/-/ʌ/ in nonwords and words separately (T1, T2).

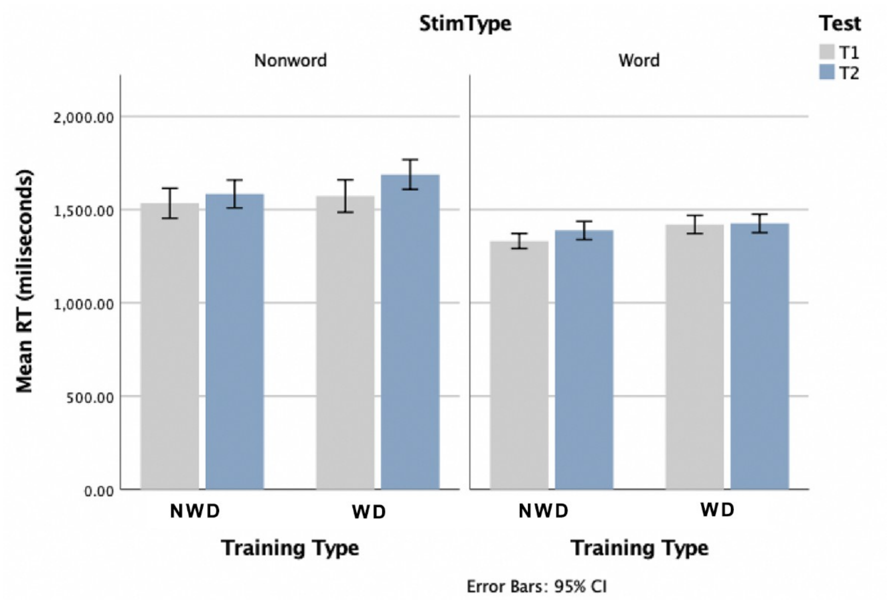

Figure 8. Bar graphs for RT for /æ/-/ʌ/ in nonwords and words separately (T1, T2).

As regards the efficiency of the HVPT for phono-lexical representations in *old* and *new* word forms in production, analysis of spectral distance scores (SDS)[22] were fitted to a Linear Mixed effects Model (LMM) analysis. *Train Type* (NWDT, WDT), *Time* (T1, T2), *Word Type* (old, new), and *Vowel* (/æ/, /ʌ/) were set as fixed effects, and *Subject* and *Item* as random factors. Type III tests of fixed effects revealed a significant main effect of *Time,* ($F(1, 749) = 9.94$, $p = .002$), and *Training Type,* ($F(1, 37.46) = 8.39$, $p = .006$), and a non-significant main effect of *WordType* ($p = .635$) and *Vowel* ($p = .464$). The interactions *TrainType**Time* and *TrainingType**Known**Vowel* were statistically significant (*TrainType**Time:* $F(1, 749) = 30.07$, $p<.001$; *Training Type**Test*: $F(1, 749) = 6.43$, $p = .011$) (see *Appendix J.3*, Table J.4.1). Bonferroni-adjusted pairwise comparisons indicated that trainees' enhanced production of the target items was only significant for the NWDT group, both for *new* and *old* words and for both vowels (*old*, /æ/ $p = <.001$, /ʌ/: $p = <.001$; *new*, /æ/: $p = .006$, /ʌ/: $p = .005$) (see *Figure* 9 and Appendix *J.3*, Table J.4.2). These findings seem to suggest that the malleability of *newly-established* and *previously established* phono-lexical representations is similar for people who were trained on nonwords. Despite the absence of statistical support, descriptively, in the NWDT condition, *new* words improved on average to a larger extent than *old* words (M *diff,* new: -1.137; old: -.896). For the WDT group, trainees performed worse at T2 (see *Figure 10* and *11*, and *Appendix 3*, Table J.4.3). *Figure 9* presents SDS scores by training type and illustrates that improvement took place only for the NWD group. *Figure 10* presents a vowel

[22] Measure used to compare the distance between the production of a native speaker and a non-native speaker. The shorter the distance the better.

plot graph indicating frontness and height values for /æ/ and /ʌ/ for both *old* and *new* words. *Figure 11* displays the SDS mean for *old* and *new* words.

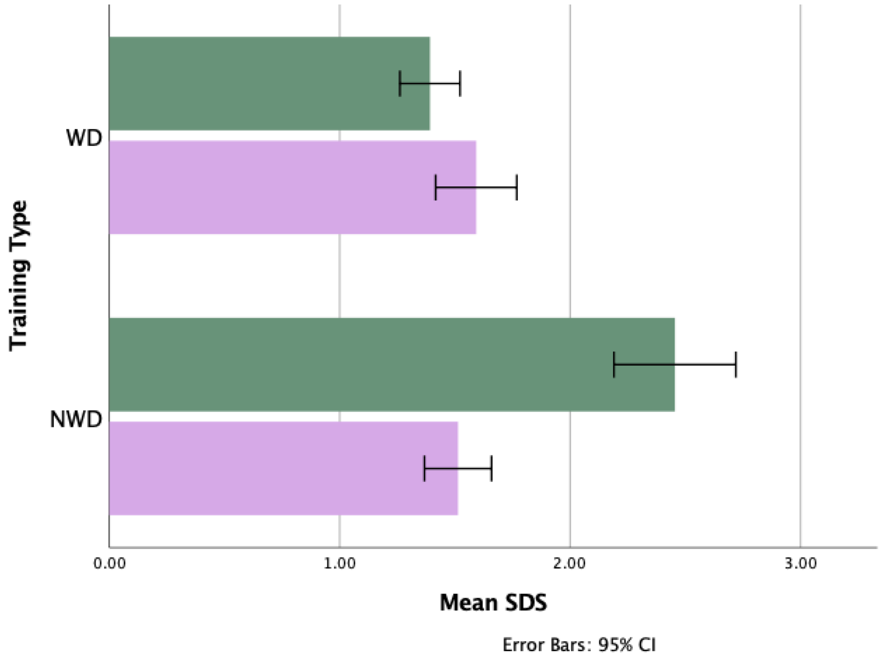

Figure 9. Spectral distance score mean for the /æ/-/ʌ/ contrast by training type (T1, T2).

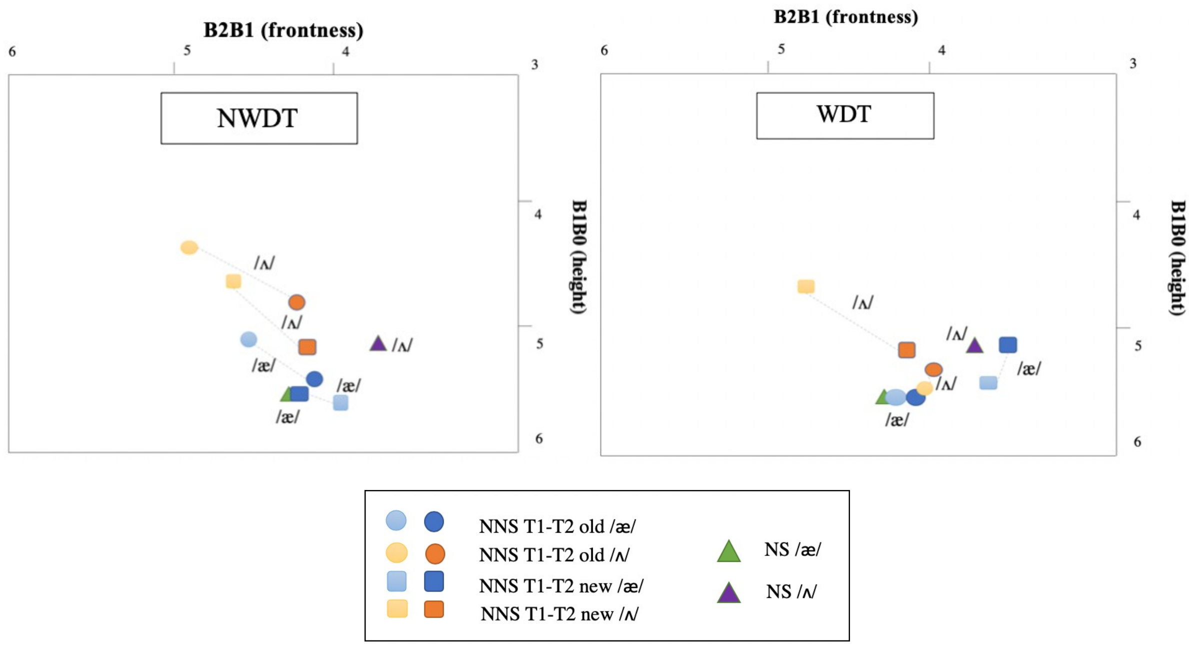

Figure 10. Vowel plot for *old* and *new* words and for both vowels (T1, T2).

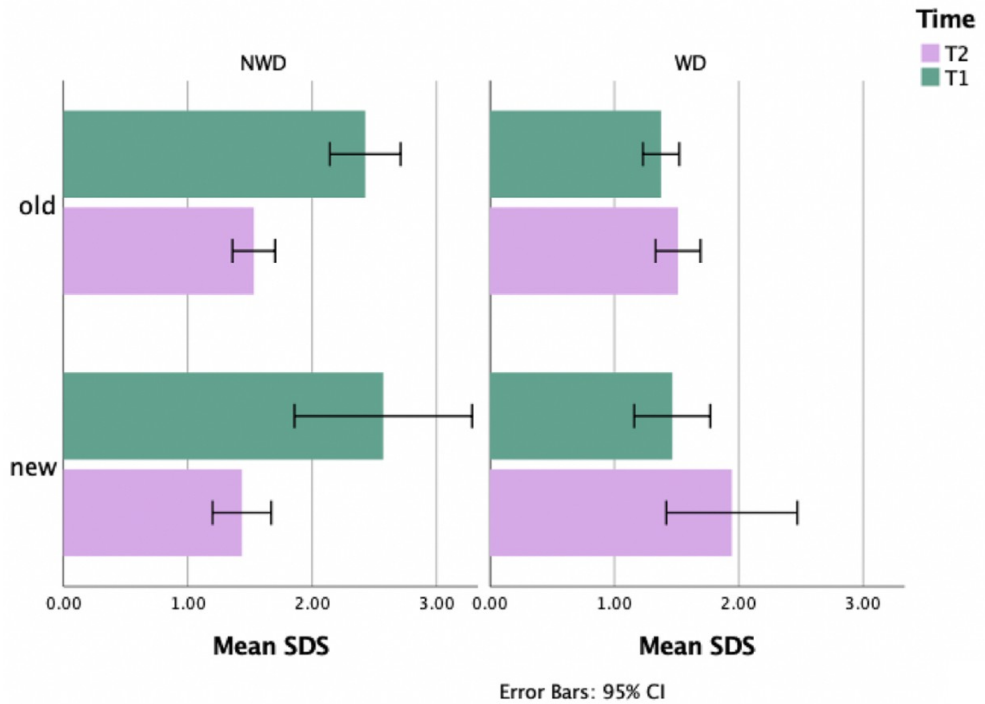

Figure 11. Spectral distance score mean for *old* and *new* words by training type (T1, T2).

Before concluding the analyses of this data, we shall take a look at the relationship between discrimination of the /æ/-/ʌ/ contrast and the form of learners' phono-lexical representations. To observe this relationship, we run chi-square tests, and we set two thresholds: learners performing over 85% in the ABX and over 70% in the LD. The criteria are the same as the ones explained in section 4.3.5. Fisher's Exact test indicated that there was a significant association (p = .003) between being able to distinguish /æ/ from /ʌ/ and starting to develop accurate phono-lexical representations (see *Appendix J* Table J.5.2). No participant scored below 85% in the ABX and over 70% in the LD (see *Figure 12* red rectangle and *Appendix J* Table J.5.1). However, we also found some participants (3/24) who scored over 85% in the ABX and below 70% in the LD (see *Figure 12* green rectangle). These results are difficult to interpret, therefore, a larger sample size may be able to provide some conclusive answers. *Figure 12* presents accuracy scores for each participant in the ABX and LD tests at T2.

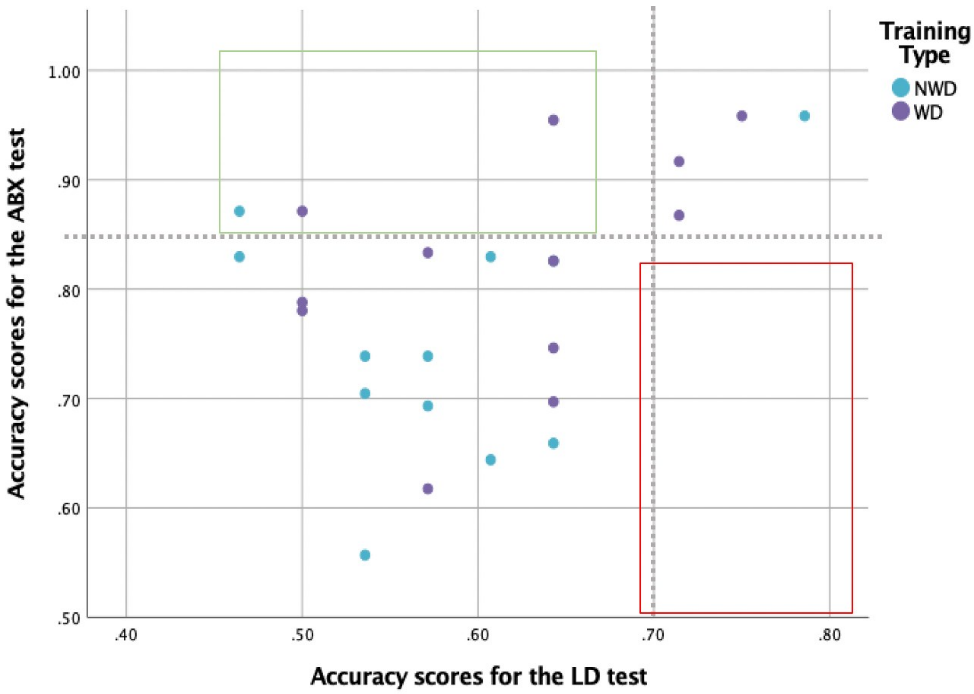

Figure 12. Accuracy scores for the ABX and LD test at T2 by participant and training type.

Based on these results, we are optimistic about the effectiveness of HVPT for inaccurate phono-lexical representations, and while these preliminary analyses are informative, it is yet to be determined what results we will find after applying our methodology. The section that follows advances some hypotheses on what to expect in the study we outlined in sections 3 and 4. We will present what we originally predicted when designing the study, but will also discuss whether data from the PTC project supported/disconfirmed our hypotheses.

6. EXPECTED RESULTS

First, we would investigate if the advantages of using HVPT can be extended to include the "updating" of inaccurate phono-lexical representations both in perception and production. Considering that previous studies (Aliaga-García & Mora, 2009; Cebrian & Carlet, 2014; Carlet & Cebrian, 2019) have shown improvement after a HVPT for discrimination and production of words in isolation, we initially hypothesized that the HVPT would also be beneficial for inaccurate phono-lexical representations. It has been shown that changes in production require more time (Sakai & Moorman, 2017), consequently, we expect HVPT to have a stronger effect in perception than in production. Descriptively, for perception, results from the PTC Project support our prediction as both groups improved over time. However, for production results do not fully support our hypothesis. In spite of this, we remain hopeful that

HVPT will be helpful for the “update” of inaccurate phono-lexical representations both for perception and production.

The second research question we propose would explore if two different HVPT paradigms (*nonwords* vs. *words*) produce differential gains in the “update” of phono-lexical representations. In 2016, Thomson and Derwing tested a similar comparison to ours with regard to the effectiveness of a phonetically-oriented. (NWD) and a lexically-oriented (WD) training for the pronunciation of words in isolation. The phonetically-oriented training these authors used involved syllables and was more beneficial than the lexically-oriented training. This finding leads to expect larger gains for phono-lexical representations when using nonword stimuli. Previous research (Darcy & Thomas, 2019; Darcy & Holliday, 2019) has shown that learners function with inaccurate phono-lexical representation, hence, it is very likely that the stimuli included in the lexically-oriented training contain words which are misrepresented, minimizing the effectiveness of the HVPT. Data from the PTC project supports the hypothesis that NWDT may be more advantageous for the development of accurate phono-lexical representations both in perception and production.

The third research question we present concerns our major objective and asks whether *newly-established* representations are as malleable as *previously-established* representations. Regardless of the lexical status of the training stimuli, we hypothesize that *newly-established* representations will be more malleable, and therefore, more target-like than *previously-established* representations both at the level of perception and production. The study by Darcy and Holliday (2019) despite the absence of statistical support and a phonetic training, calls for an age of words hypothesis in which *old* words will be more reticent to “updates” at a perceptual level. In the same vein, data from the PTC project showed no significant main effect of word type, however, descriptives support the hypothesis that representations of *old* words will be more reticent to “updates” when implementing a NWD training.

Last but not least, we were interested in investigating whether learners need to be able to discriminate /æ/ from /ʌ/ to develop accurate phono-lexical representations. We predict that learners’ ability to discriminate the /æ/-/ʌ/ contrast at a nearly native-like level is a prerequisite for the development of accurate phono-lexical representations. In the PTC project, we found support for this prediction since participants who performed “poorly” (below 85%) in the discrimination test did not seem to start developing accurate phono-lexical representation. However, we observed that a few participants who attained high discrimination abilities for the /æ/-/ʌ/ contrast, exhibited inaccurate phonological encoding and their phono-lexical representations were still imprecise. As proposed by Hayes and Masuda (2008), Darcy et al.

(2013), Amengual (2016), Simonchyk and Darcy (2017), and Darcy and Holliday (2019), accurate discrimination of a given contrast may not be enough for accurate encoding to take place, but it seems like it is one of the preconditions to start developing accurate phono-lexical representations.

This section has advanced some hypothesized results. We hope that once we have the possibility to carry out the study, we can find more conclusive answers.

7. CONCLUSION

The conclusion that can be drawn from this paper is that "updating" FL inaccurate phono-lexical representations is not easy. Preliminary analyses from the PTC Project showed that none of the participants performed within the native-like standards after the phonetic training. In terms of directions for future research, further work could involve testing a more pedagogical phonetic training for the development of phono-lexical representations that can be implemented in the FL classroom. For example, listening activities with a focus on minimal pair words containing a FL contrast, or communicative tasks together with pronunciation tips. Apart from this, including a variety of conditions (with noise or visual information) could help identify which training condition leads to larger gains at the two levels discussed in this paper -the acoustic and the lexical level- and also determine from which type of training *old* words can take more advantage. The phonetic training proposed in this paper is limited to 2 hours and 20 minutes. One wonders how much exposure to the target language is needed to remove FL learners' inaccurate phono-lexical representations, or if this is possible at all. Testing L2 learners who are mostly exposed to native language and differ in the length of residence in the L2 country might help to answer this question. It is hoped that in the upcoming years more studies are conducted on the malleability and development of phono-lexical representations so that little by little we have a better understanding of how it works.

7.1 Pedagogical implications

Our objective with this study is to emphasize the idea that inaccuracies in phono-lexical representations are a common issue in FL learners, and that for "update" to happen, learners need to undertake a phonetic training, which will not guarantee native-like representations. Being the case that we find our main hypothesis to be true and *previously-established* phono-lexical representations are less malleable, we would like to call for the teaching of pronunciation as early as possible to avoid long-term negative effects. If results do not go in line with this hypothesis, we would still like to encourage teachers to devote some time to FL perception and production in their teaching curriculum. How can FL speech perception and production practices be implemented in the classroom?

At early stages, exposing learners to authentic language would be desirable, and this could be achieved by accessing online sources or using audio-books in the classroom. A good practice apart from being exposed to native language is the imitation of texts. This has shown to be beneficial with the platform Golden Speaker Builder. This software grants the possibility to imitate native voices that are modified to approximate learners' voices but with a native pronunciation (Ding *et al.*, 2019). By the time learners can have a conversation, a good practice to raise phonological awareness could be by means of task-based pronunciation teaching (TBPT). Previous studies such as Mora-Plaza, Mora and Gilabert (2017), Solon, Long and Gurzynski-Weiss (2017) or Parlak and Ziegler (2017) have shown its advantages for pronunciation. At university, which is a higher level, especially in the cases of undergraduate courses that prepare students to become teachers (e.g.: English studies), transcription practice may raise phonological awareness. Usually, phonetic courses are offered, therefore, introducing IPA at this stage seems reasonable. This may seem like a time-consuming activity but webpages like Web Transcription Tool (WTT) provide automatic individual feedback.

10255 words + 184 abstract.

LIST OF REFERENCES

Aliaga-García, C. and Mora, J. C. (2009) Assessing the effects of phonetic training on L2 sound perception and production. In M. A. Watkins, A. S. Rauber, & B. O. Baptista (Eds.) *Recent Research in Second Language Phonetics/Phonology: Perception and Production* (pp. 2-31). Newcastle upon Tyne, UK: Cambridge Scholars Publishing.

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*, 1221–1251.

Barr, B. W. B. (2016). Checking the effectiveness of quizlet as a tool for vocabulary learning. *The Center of EFL Journal, 1(2),* 36–48.

Barriuso, A & Hayes-Harb, C. (2018). High Variability Phonetic Training as a Bridge From Research to Practice. *The CATESOL Journal, 30*, 177-194.

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.

Beyer, L. M., & Lynch, K. N. (n.d.). Flashcards versus quizlet: Achieving maximal verbosity.

Boersma, P. (2010) Modelling phonological category learning. In A. Cohn, C. Fougeron & M. Huffman (Eds.) *The Oxford handbook of laboratory phonology* (pp 207-218). Oxford: Oxford University Press.

Brysbaert, M. & New, B. (2009). Moving beyond Kuera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behaviour Research Methods, 41(4),* 977 – 990.

Carlet, A. & Kivistö de Souza, H. (2018). Improving L2 pronunciation inside and outside the classroom. *Ilha do Desterro, 71(3)* 99-123.

Carlet, A. & Cebrian, J. (2019) Assessing the effect of perceptual training on L2 vowel identification, generalization and long-term effects. In Anne Mette Nyvad, Michaela Hejná, Anders Højen, Anna Bothe Jespersen & Mette Hjortshøj Sørensen (eds.): *A Sound Approach to Language Matters – In Honor of Ocke-Schwen Bohn*. Dept. of English, School of Communication & Culture, Aarhus University, pp. 91-119.

Cebrian, J. & Carlet, A. (2014). Second language learners' identification of target language phonemes: A short-term phonetic training study. *Canadian Modern Language Review 70*(4), 474-499.

Cenoz, J. & García Lecumberri, ML. (1999). The acquisition of English pronunciation: learners' views. *International Journal of Applied Linguistics, 9,* 3-17.

Cook, S. V., Pandža, N. B., Lancaster, A. K., & Gor, K. (2016). Fuzzy Nonnative Phonolexical Representations Lead to Fuzzy Form-to-Meaning Mappings. *Frontiers in Psychology, 7,* 1345.

Cooke, M. & García Lecumberri, M. L. (2002). *Web Transcription Tool (WTT).* [23]www.wtt.org.uk.

Cooke, M. & García Lecumberri, M. L. (2018). Effects of exposure to noise during perceptual training of non-native language sounds. *Acoustical Society of America, 143*, 2602-2610.

Cutler, A., Weber, A., & Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics, 34,* 269 –284.

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

Darcy, I., Daidone, D., & Kojima, C. (2013). Asymmetric lexical access and fuzzy lexical representations in second language learners. *The Mental Lexicon, 8*, 372-420.

Darcy, I. & Holliday, J. (2019). Teaching an old word new tricks: phonological updates in the L2 lexicon. In J. Levis, C. Nagle, & E. Todey (Eds.), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference*, ISSN 2380-9566, Ames, IA, September 2018 (pp. 10-26). Ames, IA: Iowa State University.

Darcy, I., & Thomas, T. (2019). When blue is a disyllabic word: Perceptual epenthesis in the mental lexicon of second language learners. Bilingualism: Language and Cognition, 1-19.

Ding, S. *et al.* (2019). Golden speaker builder -An interactive tool for pronunciation training. *Speech Communication, 115*, 51-66.

Dufour, S., Nguyen, N. & Hans Frauenfelder, U. (2007). The perception of phonemic contrasts in a non-native dialect. *Acoustical Society of America, 121(4)*, 131-136.

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

[23] This webpage is not accessible at the moment since it will become a software which is predicted to be available from September 2020 onwards.

Flege, J. E. (1995). Second-language speech learning: theory, findings and problems. In W. Strange (ed.), *Speech Perception and Linguistic Experience. Issues in Cross- Language Research* (pp. 233-277). Timonium, MD: York Press.

Forster, K. I. and Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods*, 35(1), 116-124.

García Lecumberri, ML. & García Mayo, MP. (2003). English FL Sounds in School Learners of Different Ages. In M.L. García Lecumberri & M.P. García Mayo (Eds.), *Age and the Acquisition of English As a Foreign Language* (pp. 115-135). UK: Multilingual Matters Ltd**.**

Goble, B. (2013). Pronunciation training for Chinese English es foreign language (EFL) learners: an evaluation of two approaches. *Int J Innovation Engl Lang Teach Res*, *2*, 47–63.

Gómez Lacabex, E., García Lecumberri, ML. & Cooke, M. (2009). Identification of the contrast full vowel-schwa: training effects and generalization to a new perceptual context. *Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies, 55*, 173-196.

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

Homophony. (n.d.). In *collinsdictionary.com*. Retrieved from: https://www.collinsdictionary.com/es/diccionario/ingles/homophony.

John, P. & Cardoso, W. (2020). Are word-final consonants codas? Evidence from Brazilian Portuguese ESL/EFL learners. In J. Volin & R. Skarnitzl (eds.). *Pronunciation of English by speakers of other languages* (pp. 117-138). Newcastle upon Tyne: Cambridge Scholars Publishing.

Llompart, M., & Reinisch, E. (2017). Articulatory information helps encode lexical contrasts in a second language. *Journal of Experimental Psychology: Human Perception and Performance, 43*(*5*), 1040–1056.

Llompart, M., & Reinisch, E. (2018). Robustness of phonolexical representations relates to phonetic flexibility for difficult second language sound contrasts. *Bilingualism: Language and Cognition*, 1-16.

Malleable. (n.d.). In *collinsdictionary.com*. Retrieved from: https://www.collinsdictionary.com/es/diccionario/ingles/malleable.

Meara, P. M. (2005). Llama language aptitude tests. Swansea: Lognostics.

Mora-Plaza, I. Mora, J. C., & Gilabert, R. (2018). Learning L2 pronunciation through communicative tasks. In J. Levis (Ed.), *Proceedings of the 9th Pronunciation in Second Language Learning and Teaching conference*, ISSN 2380-9566, University of Utah, September, 2017 (pp. 174-184). Ames, IA: Iowa State University.

Muñoz, C. (2008). Symmetries and asymmetries of age effects in naturalistic and instructed L2 learning. *Applied Linguistics*, *24(4)*, 578-596.

Ortega, L., Iwashita, N., Norris, J. M. and Rabie, S. (2002). An investigation of elicited imitation tasks in crosslinguistic SLA research. Paper presented at *Second Language Research Forum*, Toronto.

Ortega, M., Mora-Plaza, I. & Mora, J.C. (submitted). Differential effects of lexical and non-lexical HVPT on the production of L2 vowels.

Pallier, C., Bosch, L., & Sebastian-Gallés, N. (1997). A limit on behavioural plasticity in speech perception. *Cognition, 64*, B9-B17.

Pallier, C., Colomé, A., & Sebastian-Gallés, N. (2001). The influence of native-language phonology on lexical access: Concrete exemplar-based vs. abstract lexical entries. *Psychological Science 12*, 445–449.

Parlak, O. & Zeigler, N. (2017). The impact of recasts on the acquisition of primary stress in a computer- mediated environment. *Studies in Second Language Acquisition, 39(2)*.

Rogers, V., Maera, P., Barnett-Leght, T., Curry, C. (2017). Examining the LLAMA aptitude test. *Journal of the European Second Language Association, 1(1),* 49–60.

Saito, K., *Suzukida, Y., & *Sun, H. (2019). Aptitude, experience and second language pronunciation proficiency development in classroom settings: A longitudinal study. *Studies in Second Language Acquisition, 41*, 201-225.

Saito, K. (2017). Effects of sound, vocabulary and grammar learning aptitude on adult second language speech attainment in foreign language classrooms. *Language Learning, 67*, 665-693.

Setter, J. and Jenkins, J. (2005). State-of-the-art review article. *Language Teaching*, *38(1)*, 1-17.

Shatzman, K. B., & McQueen, J. M. (2006). Prosodic knowledge affects the recognition of newly acquired words. *Psychological Science, 17(5),* 372–377.

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.

Solon, M., Long, A. Y., & Gurzynski-Weiss, L. (2016). Task complexity, language-related episodes, and production of L2 Spanish vowels. *Studies in Second Language Acquisition, 39(2)*.

Sutherland, A. (2007). *Quizlet*. https://quizlet.com/es

Syrdal, A. K. and Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America, 79*, 1086-100.

Thomson, R. I. and Derwing, T. M. (2016). Is phonemic training using nonsense or real words more effective? In J. Levis, H. Le., I. Lucic, E. Simpson and S. Vo (Eds.). *Proceedings of the 7thPronunciation in Second Language Learning and Teaching Conference*. (pp. 88-97). Ames, IA: Iowa State University.

Traunmüller. H. (1990). Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America, 88*, 97–100.

Tyler, M. (2019). PAM-L2 and Phonological Category Acquisition in the Foreign Language Classroom. In Anne Mette Nyvad, Michaela Hejná, Anders Højen, Anna Bothe Jespersen & Mette Hjortshøj Sørensen (Eds.), *A Sound Approach to Language Matters – In Honor of Ocke-Schwen Bohn* (pp. 607-630). Dept. of English, School of Communication & Culture, Aarhus University.

Weber, A., & Cutler, A. (2006). First-language phonotactics in second-language listening. *The Journal of the Acoustical Society of America, 119,* 1-25.

Yan, X., Maeda, Y., LV, J. & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing, 33(4)*, 497-528.

Appendix A. Language Background questionnaire.

Language Background Questionnaire

*Obligatorio

1. Dirección de correo electrónico *

2. Numero de telefono móvil

Personal Data

3. Surnames: *

   Please type your surnames here.

4. Name: *

   Please type your name here.

5. Gender: *

   Marca solo un óvalo.

   - Male
   - Female

6. Date of Birth: *

Ejemplo: 7 de enero del 2019

7. You are: *

Marca solo un óvalo.

- Right-handed
- Left-handed
- Both

Language Background

8. Indicate which language(s) you normally use on a daily basis: *

You can tick more than one box.

Selecciona todos los que correspondan.

- Catalan
- Spanish
- English

Otro:

9. Indicate which is/are your native language(s) (i.e., mother tongue, "r“t language): *

Language(s) you learnt to speak from birth.

Selecciona todos los que correspondan.

- Catalan
- Spanish
- English

Otro:

10. Indicate which language(s) is/are used at home most of the time: *

You can tick more than one box.

Selecciona todos los que correspondan.

- [ ] Catalan
- [ ] Spanish
- [ ] English

Otro: [ ] ____________________

11. List all the languages you use in order of dominance, and include the age at which you stared to learn them: *

For example: Catalan/0.

____

12. *

For example: Spanish/2.

____

13. *

For example: English/8

14. *

____

15.

____

16. ____________________

17. Indicate your estimate percentage of daily use of Catalan: *

Type in a number from 0 to 100. Make sure that the percentages in the next three questions add up to 100%. For example: Catalan 70%, Spanish 20% and English 10% = 100%.

18. Indicate your estimate percentage of daily use of Spanish: *

19. Indicate your estimate percentage of daily use of English: *

20. Indicate your total estimate percentage of daily use of other languages: *

English Learning Experience

21. Indicate at which age you sta!ed learning English: *

22. Indicate for how many hours a week you studied English in primary school: *

For example: 2 hours.

23. Indicate for how many hours a week you studied English in secondary school: *

For example: 2 hours.

24. Indicate for how many hours a week you studied English in high school: *

For example: 2 hours.

25. Indicate for how many hours a week you studied/study English at university: *

For example: 2 hours.

26. Indicate whether you have taken extracurricular English courses or private classes : *

If yes, specify whether it was in primary/secondary/high school/university, where (e.g., British Council) and how lon you took those classes (e.g., 2 years, 4 hours a week).

27. Indicate whether you have any certificate of English level? *

If yes, specify whether it was in primary/secondary/high school/university, where

28. Indicate the estimate number of hours you spend listening to English on a weekly basis: *

For example: 2 hours.

29. Indicate the estimate number of hours you spend speaking in English on a weekly basis: *

For example: 2 hours.

30. Indicate the estimate number of hours you spend reading in English on a weekly basis: *

For example: 2 hours.

31. Indicate the estimate number of hours you spend writing in English on a weekly basis: *

For example: 2 hours.

32. Rate your command of the English language on the scale: 1=VERY POOR and 7=NEAR-
NATIVE *

Marca solo un óvalo por >la.

	1	2	3	4	5	6	7
Reading:	⬭	⬭	⬭	⬭	⬭	⬭	⬭
Listening:	⬭	⬭	⬭	⬭	⬭	⬭	⬭
Speaking:	⬭	⬭	⬭	⬭	⬭	⬭	⬭
Writing:	⬭	⬭	⬭	⬭	⬭	⬭	⬭

33. Indicate whether you have ever lived in an English-speaking country for more than two weeks: *

Marca solo un óvalo.

- Yes
- No

34. If yes, specify how many times, where, when, for how long and the purpose of the stay: *

35. Indicate how often you engage in each of the following activities in English: *

1 - –ever; 2 - – few times a year; 3 - –onthly; 4 - –eekly; 5 - –aily

Marca solo un óvalo por >la.

	1	2	3	4	5
Watching English language television					
Reading newspapers/magazines in English					
Reading books in English					
Listening to songs in English					
Watching movies or videos in English					
Speaking English with native or duent speakers					
Speaking English with nonnative speakers					
Writing e-mails/letters in English					

36. Indicate whether you have any speech-language pathology

(e.g., dyslexia): * *Marca solo un óvalo.*

- Yes
- No

37. If yes, specify: *

If no, type in 'N'':

Appendix B. Classification of words

Table 1. Classification of *old* and *new* words.

***Old* words**		***New* words**	
æ	**ʌ**	**æ**	**ʌ**
cat	cut	gash	gush
mad	mud	hatch	hutch
cap	cup	mat	mutt
bag	bug	rag	rug
black	duck	jab	pug
fat	gun	wham	hub
man	punk	lass	chub
map	sun	chaff	buff

Appendix C. Word knowledge categorization

Table 2. Word knowledge categorization based on frequency of words.[24]

Type of word	*SUBTLWF*	Type of word	*SUBTLWF*
Old Words			
man	1845.75	cup	51.65
cut	229.76	cat	66.33
black	167.94	map	31.82
gun	213.2	duck	24.76
mad	113.41	punk	21.98
bag	94.04	cap	18.75
fat	79.43	bug	20.94
sun	69.67	mud	14.82
New words			
hatch	12.82	gush	0.71
rug	10.41	gash	0.69
rag	4.78	chaff	0.22
mat	3.49	wham	
mutt	3.96	hub	
buff	2.49	chub	
jab	3.35	hutch	
lass	2.67	pug	

[24] Words in red do not appear in SubtlexUS database.

Table 3. Word knowledge categorization based on a 7-point Likert-scale questionnaire from Project PTC.[25]

Type of word			
Old Words	*M*		*M*
cat	7.00	gun	6.98
man	7.00	duck	6.94
bag	6.99	cup	6.90
cut	6.99	mad	6.86
sun	6.99	punk	6.44
map	6.98	bug	6.17
black	6.98	cap	5.95
fat	6.98	mud	5.20
New words			
mat	3.38	jab	
rug	3.25	lass	
rag	2.44	chaff	
hatch	2.34	wham	
gash	2.09	buff	
gush	2.02	hub	
hutch	1.80	chub	
mutt	1.48	pug	
Valid N (listwise)			

[25] Words in red were not included in the questionnaire from Project PTC.

Appendix D. Word knowledge test

WRITE DOWN THE TRANSLATION FOR EACH WORD TO YOUR L1 IN "OTRO" (WHETHER THIS IS IN SPANISH OR CATALAN). IF YOU DO NOT KNOW THE MEANING OF THE WORD PLEASE SELECT A) I DON'T KNOW THIS WORD. THANK YOU FOR YOUR PARTICIPATION!!!
*Obligatorio

Please enter your code here: *

____

1. cat *

   Marca solo un óvalo.

   ( ) A) I don't know this word
   ( ) Otro: _______________

2. cut *

   Marca solo un óvalo.

   ( ) A) I don't know this word
   ( ) Otro: _______________

3. mad *

   Marca solo un óvalo.

   ( ) A) I don't know this word
   ( ) Otro: _______________

4. mud *

Marca solo un óvalo.

A) I don't know this word

Otro:

5. bag *

Marca solo un óvalo.

A) I don't know this word

Otro:

6. bug *

Marca solo un óvalo.

A) I don't know this word

Otro:

7. black *

Marca solo un óvalo.

A) I don't know this word

Otro:

8. duck *

Marca solo un óvalo.

A) I don't know this word

Otro:

9. fat *

Marca solo un óvalo.

A) I don't know this word

Otro:

10. gun *

Marca solo un óvalo.

A) I don't know this word

Otro:

11. gash *

Marca solo un óvalo.

A) I don't know this word

Otro:

12. gush *

Marca solo un óvalo.

A) I don't know this word

Otro:

13. mat *

Marca solo un óvalo.

A) I don't know this word

Otro: ____________________

14. mut *

Marca solo un óvalo.

A) I don't know this word

Otro: ____________________

15. rag *

Marca solo un óvalo.

A) I don't know this word

Otro: ____________________

16. rug *

Marca solo un óvalo.

A) I don't know this word

Otro: ____________________

17. hatch *

Marca solo un óvalo.

A) I don't know this word

Otro: ____________________

18. hutch *

Marca solo un óvalo.

A) I don't know this word

Otro: ______________________

19. jab *

Marca solo un óvalo.

A) I don't know this word

Otro: ______________________

20. pug *

Marca solo un óvalo.

A) I don't know this word

Otro: ______________________

21. wham *

Marca solo un óvalo.

A) I don't know this word

Otro: ______________________

22. hub *

Marca solo un óvalo.

A) I don't know this word

Otro: ______________________

23. lass *

Marca solo un óvalo.

A) I don't know this word

Otro:

24. chub *

Marca solo un óvalo.

A) I don't know this word

Otro:

25. chaff *

Marca solo un óvalo.

A) I don't know this word

Otro:

26. map *

Marca solo un óvalo.

A) I don't know this word

Otro:

27. cap *

Marca solo un óvalo.

A) I don't know this word

Otro:

28. cup *

Marca solo un óvalo.

A) I don't know this word

Otro:

29. man *

Marca solo un óvalo.

A) I don't know this word

Otro:

30. punk *

Marca solo un óvalo.

A) I don't know this word

Otro:

31. sun *

Marca solo un óvalo.

A) I don't know this word

Otro:

32. buff *

Marca solo un óvalo.

A) I don't know this word

Otro:

Appendix E. Quizlet activities examples

E.1 Quizlet activities examples (session 1)

Figure 13. Example of words to memorize (S1)

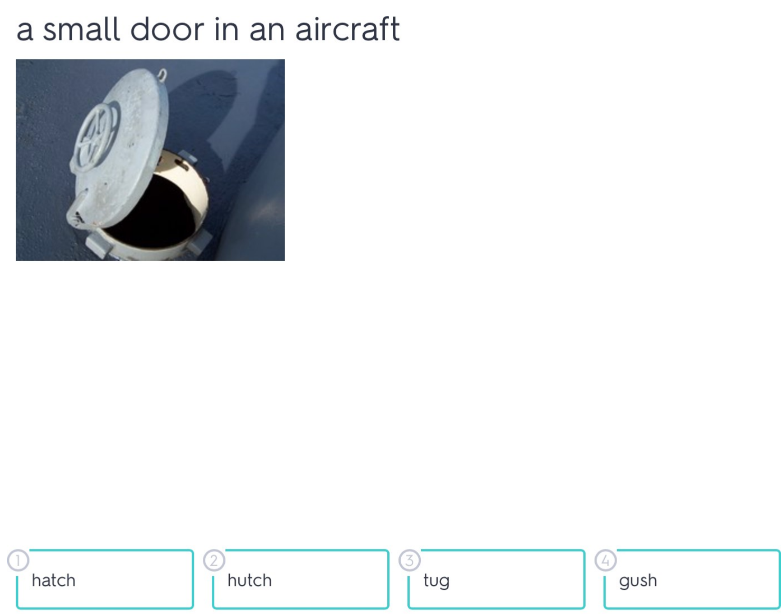

Figure 14. Example of multiple-choice activity (S1)

Figure 15. Example of fill-in-the-gaps activity (S1)

Figure 16. Example of dictation activity (S1)

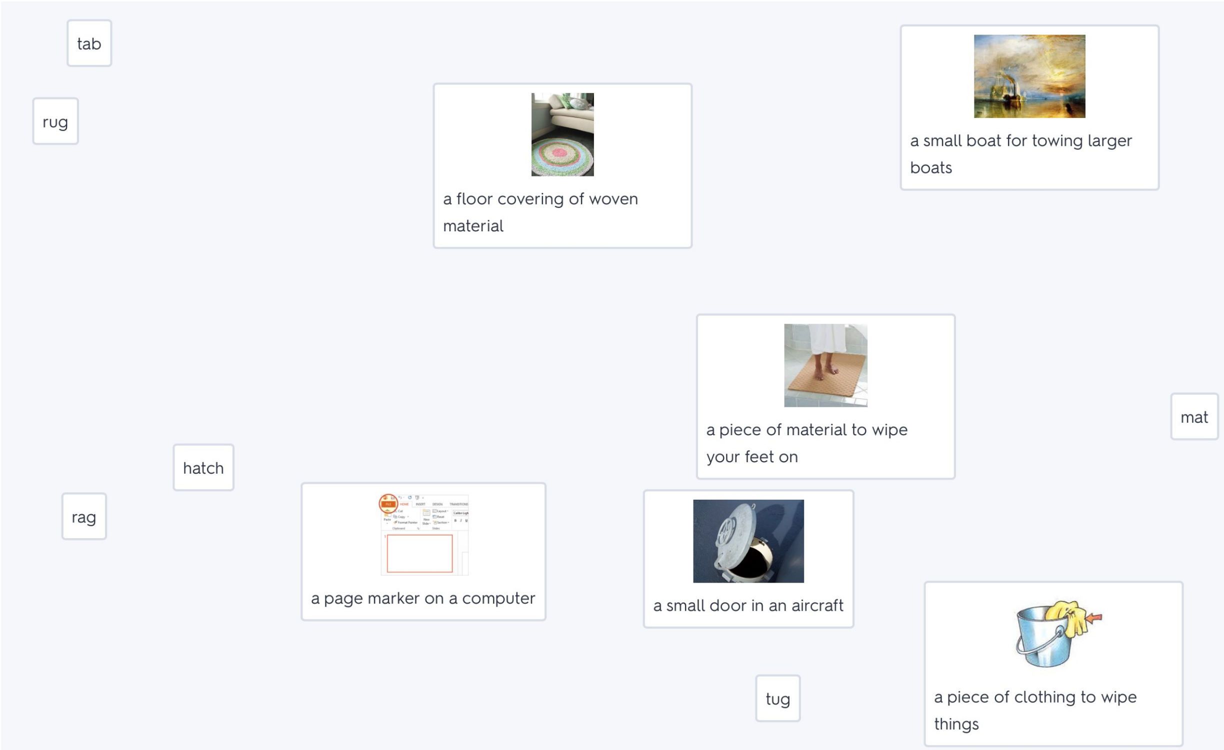

Figure 17. Example of matching words activity (S1)

E.2 Quizlet activities examples (session 2)

Click on the word to hear it!

hatch

a small door in an aircraft (escotilla)

He opened the **HATCH** in the cockpit ceiling.

Figure 18. Example of words to memorize (S2)

I had to close the ……… when the storm started.

1. gash
2. mutt
3. hutch
4. hatch

Figure 19. Example of multiple-choice activity (S2)

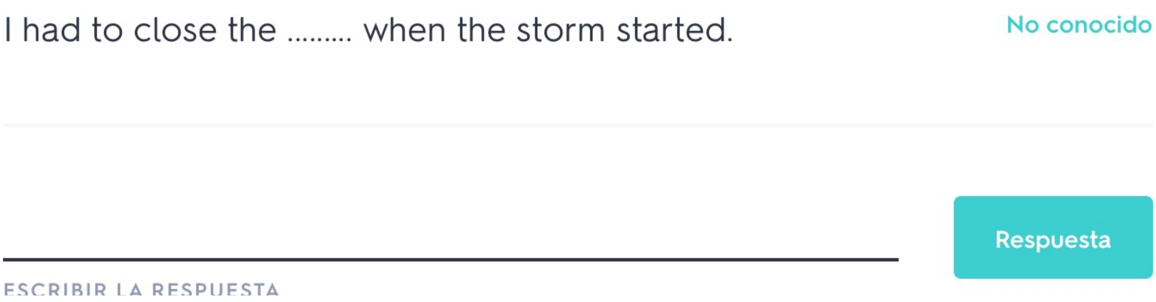

Figure 20. Example of fill-in-the-gaps activity (S2)

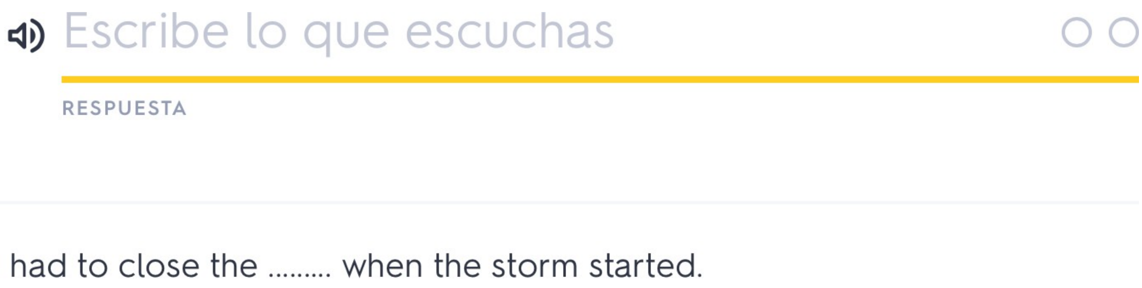

Figure 21. Example of dictation activity (S2)

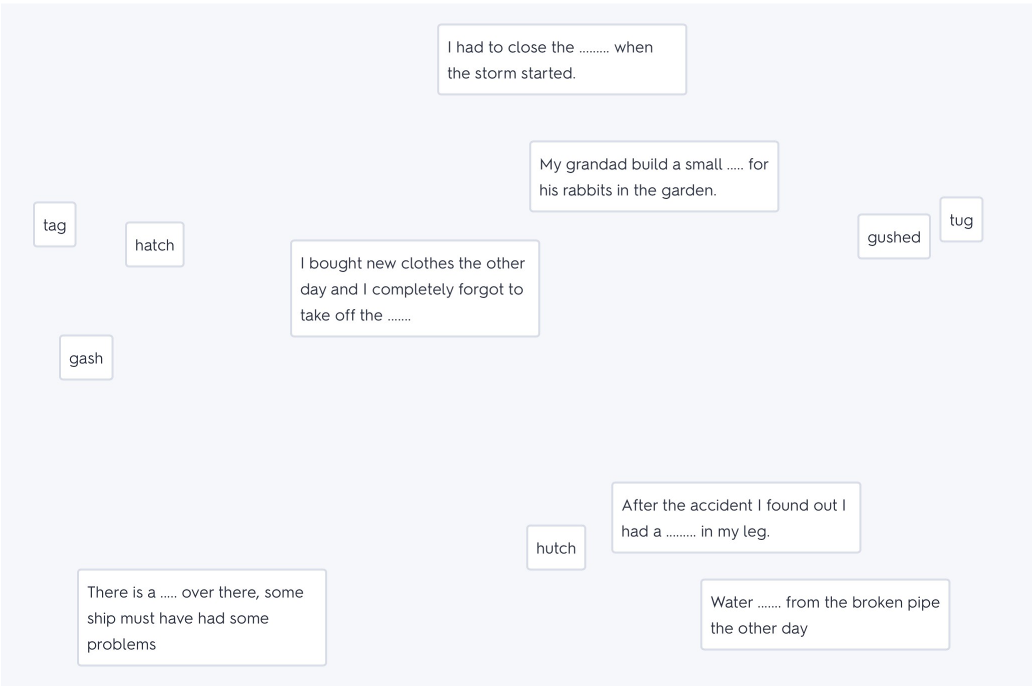

Figure 22. Example of matching words **activity (S2)**

Appendix F. Testing stimuli

Table 4. *List of stimuli used for the ABX test*

Test (æ-ʌ) items			
Words		**Nonwords**	
æ	**ʌ**	**æ**	**ʌ**
cat	cut	fash	fush
mad	mud	thatt	thutt
cap	cup	mab	mub
bag	bug	chang	chung
match	much	tam	tum
sack	suck	thack	thuck
hat	hut	datt	dutt
back	buck	tazz	tuzz
fan	fun	shad	shudd
lack	luck	tadge	tudge
bad	bud	thapp	thupp
pan	pun	gab	gub
gash	gush	sazz	suzz
hatch	hutch	vack	vuck
mat	mutt		
rag	rug		
tab	tub		
tag	tug		

Table 5. List of stimuli for the LD test

Test (æ-ʌ) items			
Word (æ)	**Nonword (competitor)**	**Word (ʌ)**	**Nonword (competitor)**
black	bluck	duck	dack
fat	fut	gun	gan
lap	lup	plum	plam
man	mun	punk	pank
map	mup	sun	san
jab	jub	pug	pag
wham	whum	hub	hab
lass	luss	chub	chab
chaff	chuff	buff	baff
		tough	taff

Table 6. *List of stimuli used for the delayed word repetition (DWR) test*

Test æ-ʌ			
Words		**Nonwords**	
æ	**ʌ**	**æ**	**ʌ**
cat	cut	fash	fush
mad	mud	thatt	thutt
cap	cup	mab	mub
bag	bug	chang	chung
match	much	tam	tum
sack	suck	thack	thuck
hat	hut	datt	dutt
back	buck	tazz	tuzz
fan	fun	shad	shudd
lack	luck	tadge	tudge
bad	bud	thapp	thupp
pan	pun	gab	gub
gash	gush	sazz	suzz
hatch	hutch	vack	vuck
mat	mutt		
rag	rug		
tab	tub		
tag	tug		

Table 7. *List of stimuli used in the delayed word repetition (DWR) test*

Control items			
Words			
map	fat	can	dad
cash	jam	bun	sad
pub	gun	bus	sun
fuss	dull	numb	nut

Table 8. *List of stimuli used in the delayed sentence repetition (DSR) test*

Test æ-ʌ	
Words	
æ	**ʌ**
My friend has a cat with grey eyes	The doctor cut the patient to start the surgery
He got mad after receiving bad news	You still have mud on your shoes
She brought me a cap from Barcelona	Your cup fell down to the floor
I forgot my bag at the station	There is a bug in that window
He got a gash on his face	There was a gush of water coming from there
We have a mat at the door entrance	That mutt is always barking
Helicopters have a small hatch for safety	She bought a huge hutch for the rabbits
Bring a rag to clean the window	We bought a new rug for the entrance

Table 9. Classification of *old* and *new* words in each test.

***Old* words**		***New* words**	
æ	**Test**	**æ**	**ʌ**
cat	ABX, DWR, DSR	gash	ABX, DWR, DSR
mad	ABX, DWR, DSR	hatch	ABX, DWR, DSR
cap	ABX, DWR, DSR	mat	ABX, DWR, DSR
bag	ABX, DWR, DSR	rag	ABX, DWR, DSR
cut	ABX, DWR, DSR	gush	ABX, DWR, DSR
mud	ABX, DWR, DSR	hutch	ABX, DWR, DSR
cup	ABX, DWR, DSR	mutt	ABX, DWR, DSR
bug	ABX, DWR, DSR	rug	ABX, DWR, DSR
black	LD	jab	LD
fat	LD	wham	LD
man	LD	lass	LD
map	LD	chaff	LD
duck	LD	pug	LD
gun	LD	hub	LD
punk	LD	chub	LD
sun	LD	buff	LD

Appendix G. Training stimuli

Table 10. *List of stimuli used in the AX and identification (ID) test*

Test æ-ʌ			
Word training		**Nonword training**	
æ	**ʌ**	**æ**	**ʌ**
cat	cut	datt	dutt
match	much	fash	fush
mad	mud	thatt	thutt
ban	bun	tazz	tuzz
cap	cup	mab	mub
sack	suck	tam	tum
bag	bug	thack	thuck
mag	mug	chang	chung

Table 11. *List of stimuli used in the immediate repetition (IR) test*

Test æ-ʌ			
Word training		**Nonword training**	
æ	**ʌ**	**æ**	**ʌ**
cat	cut	datt	dutt
match	much	fash	fush
mad	mud	thatt	thutt
ban	bun	tazz	tuzz
cap	cup	mab	mub
sack	suck	tam	tum
bag	bug	thack	thuck
mag	mug	chang	chung

Appendix H. Elicited Imitation Sentences[26]

1. I have to get a haircut (7)
2. The red book is on the table (8)
3. The streets in this city are wide (8)
4. He takes a shower every morning (9)
5. What did you say you were doing today? (10)
6. I doubt that he knows how to drive that well (10)
7. After dinner I had a ling, peaceful nap (11)
8. It is possible that it will rain tomorrow (12)
9. I enjoy movies which have a happy ending (12)
10. The houses are very nice but too expensive (12)
11. The little boy whose kitten died yesterday is sad (13)
12. That restaurant is supposed to have very good food (13)
13. I want a nice, big house in which my animals can live (14)
14. You really enjoy listening to country music, don't you (14)
15. She just finished painting the inside of her apartment (14)
16. Cross the street at the light and then just continue straight ahead (15)
17. The person I'm dating has a wonderful sense of humor (15)
18. She only orders meat dishes and never eats vegetables (15/16)
19. I wish the price of town houses would become affordable (15)
20. I hope it will get warmer sooner this year than it did last year (16)
21. A good friend of mine always takes care of my neighbour’s three children (16)
22. The black cat that you fed yesterday was the one chased by the dog(16)

[26] Test developed by Ortega, L., Iwashita, N., Norris, J. M., & Rabie, S. (2002, October 3-6). *An investigation of elicited imitation tasks in crosslinguistic SLA research.* Paper presented at the Second Language Research Forum, Toronto.

23. Before he can go outside, he has to finish cleaning his room (16)

24. The most fun I've ever had was when we went to the opera (16)

25. The terrible thief whom the police caught was very tall and thin (17)

26. Would you be so kind as to hand me the book which is on the table? (17)

27. The number of people who smoke cigars is increasing every year (17/18)

28. I don't know if the 11:30 train has left the station yet (18)

29. The exam wasn't nearly as difficult as you told me it would be (18)

30. There are a lot of people who don’t eat anything at all in the morning (19)

Appendix I. Extra information on the PTC Project

Data from Project PTC was collected in 2018 by Cristina Aliaga, Pace Bailey, Eva Cerviño, Joan C. Mora, Ingrid Mora-Plaza and Mireia Ortega (alphabetically). There was a total of eight experimental groups and one control group. First, group 4 (NWDA) and 8 (WDA) were trained with audio (A) under a non-word (NWD) and a word (WD) training respectively. Learners in group 1 (NWDVN) and 5 (WDVN), were trained with nonwords/words, they had visual information (V) about their own articulation, and were exposed to noise (N) during the training. Then, we have group 2 (NWDV) and 6 (WDV), who were also trained with nonwords or words, and trainees in this condition were exposed to visual information. Finally, we find group 3 (NWDN) and 7 (WDN), in these cases, trainees were trained in noise, with nonwords and words respectively. Participants were randomly assigned to each group depending on the stimulus type, monitoring, and listening (See Figure 12). To determine their level of proficiency I used information from the elicited imitation task. Overall participants in the experimental group scored on average 78.52 (SD = 12.94). Proficiency for each individual group is displayed in Table 4 below.

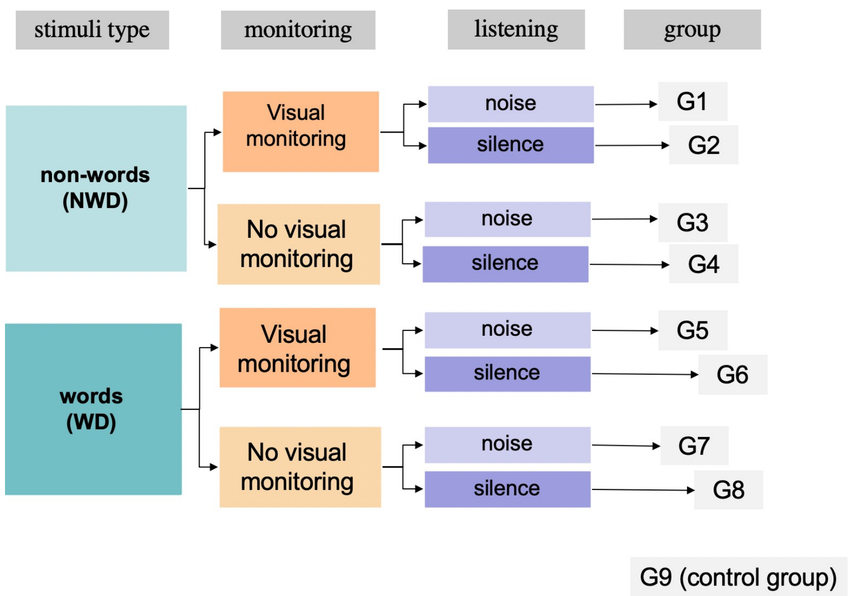

Figure I 1 Project PTC experimental group conditions.

Descriptive Statistics[a]

Group		N	Mean	SD
NWDN	**EIT%**	13	79.1026	10.49946
NWDA	**EIT%**	14	80.9524	9.49503
NWDVS	**EIT%**	13	77.7564	10.08864
NWDV	**EIT%**	11	74.3939	18.68262
WDN	**EIT%**	15	82.0556	13.82956
WDA	**EIT%**	12	74.6528	7.88537
WDVS	**EIT%**	13	78.0128	16.40961
WDV	**EIT%**	15	79.4444	14.95916

Table I 1 *Mean* and *SD* proficiency scores for each group.

The categorization of *old* and *new* words presented below in Figure I.2 was used for analyses in the DSR. This classification has been done by the author of the paper for the preliminary analysis presented in Section 5.

***Old* words (6.82/7)**		***New* words (4.1/7)**	
æ	**ʌ**	**æ**	**ʌ**
map	pub	gash	gush
cash	gun		numb
can	bus		
fat	sun		
dad	dull		
jam	nut		
sad			

Table I. 2 Familiarity average score and classification of *old* and *new* words for the DSR test for Project PTC.

The tables below (I.3 and I.4) present the list of stimuli used in the LD and DSR. This testing stimuli from the PTC Project differs from the study we have presented in this paper.

Table I 3 *List of stimuli used in the delayed word repetition (LD) test*

æ-ʌ test items			
Word		**Nonword**	
æ	**ʌ**	**æ**	**ʌ**
black	duck	dack	bluck
crack	dull	dall	cruck
fat	gun	gan	fut
lap	plum	plam	lup
man	punk	pank	mun
map	sun	san	mup
tap	tough	taff	tup

Table I 4 *List of stimuli used in the delayed sentence repetition (DSR) test*

Words	
æ	**ʌ**
He looked at the map to find his way	We go to the pub on Saturdays
I have no cash in my pocket	Nobody saw the gun in his pocket
I don't want to open up that can of worms	I`ll get off the bus at the next stop
She bought a van to travel around the world	He caught the sun at the beach
He became so fat he could hardly walk	Don’t make such a fuss about it
She loves her dad more than anyone else	She finds it dull living in the country
She loves jam on her toast	My fingers go numb in the cold
It's very sad that she didn't get the job	To eat the nut first crack the shell

Appendix J. Tables and figures from the PTC Project

This section presents the analyses run for data from the PTC Project. It examines all the tests that researchers used in the PTC Project (i.e., ABX, LD, DWR and DSR). Only group 4 (NWDT) and 8 (WDT) will be analysed because they share the same training conditions as the study we proposed in this paper. ABX and DWR were analyzed in order to assess the effectiveness of HVPT. We hypothesize that learners first need to discriminate /æ/ from /ʌ/ accurately and produce words containing the /æ/-/ʌ/ contrast accurately without a lexical context before being able to perform well in tasks that involve a lexical context (i.e., in the LD for perception and in the DSR for production). Table J1 below presents the classification of *old* and *new* words used for the ABX and DWR. Classification of *old* and *new* words in the ABX and DWR differs from the classification of *old-new* in the DSR. Classification is different because the LD contained mostly familiar words. Section 5 has already explain the criterion followed for the classification of *old* and *new* in the DSR. Classification for ABX and DWR was based on the 7-point-likerscale familiarity questionnaire that learners filled out at the beginning of the experiment.

***Old* words (6.82/7)**		***New* words (4.1/7)**	
æ	**ʌ**	**æ**	**ʌ**
map	pub	van	fuss
cash	gun		numb
can	bus		
fat	sun		
dad	dull		
jam	nut		
sad			

Table J 1 Classification of *old* and *new* words for the ABX and DWR tests.

J.1 ABX results

Phonetic training efficiency for discrimination of the /æ/-/ʌ/ contrast was assessed though a Generalized Linear Mixed Model (LMM) in which *Time* (T1, T2), *Training Type* (NWD, WD), *Word Type* (old, new) and *Vowel* (/æ/, /ʌ/) are set as fixed factors ad *Subject* and *Item* as random factors. The GLMM analysis indicated that there was a significant main effect of *Testing Time* ($F(1, 2768) = 32.26$, p=.000). None of the other fixed effects or interactions reached significance (Table J.1.1). GLMM does not allow to observe quadruple interactions, for this reason, in order to run Bonferroni-pairwise comparisons the data file had to be split by one of the variables. Since *Training Type* did not have a main effect it was decided to split the file by training type.

For the NWD training, Bonferroni pairwise-comparisons determined that the main effect of *Testing Time* was mainly driven by improvement for /æ/ in both *old* and *new* words (*old:* $t(1285)=2.151$ p=.032; *new*: $t(373)= 2.807$ p=.005) and by improvement for /ʌ/ in *new* words ($t(970)= 2.449$ p=.014) (Table J.1.2). For the WD training condition, participants did not improve significantly from T1 to T2 (Table J.1.3).

It was interesting to observe that there was no main effect of *Word Type* ($F(1, 46)$ = .973, p=.329) suggesting that training people to discriminate æ/ from /ʌ/ has the same effect in *old* and *new* words. Descriptively, for the NWD and the WD training, we found *new* words to improve to a larger extent than *old* words (T2-T1, NWD: *old:* 8% *new:*11%; WD: *old:* 4% *new:*6%) (Table J.1.4 and *Figure J.1.1*). As for reaction time (RT), people performed at a faster rate at T2, and for the *new* words, trainees seemed to be faster than for the *old* words (Table J.1.4 and *Figure J.1.2*).

Fixed Effects[a]

Source	F	df1	df2	p
Corrected Model	3.222	15	343	.000
Training Type	.422	1	25	.522
Testing Time	32.263	1	2768	.000
Vowel	3.411	1	46	.071
Word Type	.973	1	46	.329
Training Type * Time	.592	1	2768	.442
Training Type * Vowel	.231	1	2768	.631
Training Type * Word Type	.089	1	2768	.765
Time * Vowel	.685	1	2768	.408

Time* Word Type	3.631	1	2768	.057
Vowel * Word Type	.172	1	46	.680
Training Type * Time* Vowel	3.064	1	2768	.080
Training Type * Time * Word Type	.390	1	2768	.532
Training Type * Vowel * Word Type	.735	1	2768	.391
Time * Vowel * Word Type	.054	1	2768	.816
Training Type * Time* Vowel * Word Type	.014	1	2768	.905

Probability distribution: Binomial
Link function: Logit[a]
a. Target: ACC

Table J.1. 1 GLMM: Fixed Effects results for ABX.

Pairwise Contrasts

Vowel	Known	Test Pairwise Contrasts	Contrast Estimate	Std. Error	t	df	Adj. Sig.	95% Confidence Interval	
								Lower	Upper
æ	**Old**	**T1 - T2**	-.088	.041	-2.151	1285	.032	-.169	-.008
		T2 - T1	.088	.041	2.151	1285	.032	.008	.169
	New	**T1 - T2**	-.112	.040	-2.807	373	.005	-.191	-.034
		T2 - T1	.112	.040	2.807	373	.005	.034	.191
ʌ	**Old**	**T1 - T2**	-.084	.047	-1.778	1624	.076	-.176	.009
		T2 - T1	.084	.047	1.778	1624	.076	-.009	.176
	New	**T1 - T2**	-.113	.046	-2.449	970	.014	-.203	-.022
		T2 - T1	.113	.046	2.449	970	.014	.022	.203

The sequential Bonferroni adjusted significance level is .05.
Confidence interval bounds are approximate.

Table J.1. 2 Bonferroni-adjusted pairwise contrasts (*Time, Vowel*, ***Word Type***) **for the NWD group.**

Pairwise Contrasts

Vowel	Word Type	Time Pairwise Contrasts	Contrast Estimate	Std. Error	t	df	Adj. Sig.	95% Confidence Interval Lower	95% Confidence Interval Upper
æ	Old	T1 - T2	.007	.040	.167	1144	.868	-.072	.085
		T2 - T1	-.007	.040	-.167	1144	.868	-.085	.072
	New	T1 - T2	-.060	.039	-1.557	1144	.120	-.136	.016
		T2 - T1	.060	.039	1.557	1144	.120	-.016	.136
ʌ	Old	T1 - T2	-.100	.055	-1.800	1144	.072	-.209	.009
		T2 - T1	.100	.055	1.800	1144	.072	-.009	.209
	New	T1 - T2	-.143	.053	-2.679	479	.008	-.247	-.038
		T2 - T1	.143	.053	2.679	479	.008	.038	.247

The sequential Bonferroni adjusted significance level is .05.
Confidence interval bounds are approximate.

Table J.1. 3 Bonferroni-adjusted pairwise contrasts (*Time, Vowel, Word Type*) for the WD group.

			T1		T2		T2-T1	
NWD			**ACC (%)**	**RT (ms.)**	**ACC (%)**	**RT (ms.)**	**ACC (%)**	**RT (ms.)**
	***Old* words**	**æ**	75	969.02	84	966.63	9	-2.39
		ʌ	68	991.70	75	939.44	7	-52.26
		Total	72	980.36	80	953.03	8	-27.33
	***New* nonwords**	**æ**	76	992.01	88	960.44	12	-31.57
		ʌ	72	1021.96	83	1020.92	11	-1.04
		Total	74	1006.99	86	990.68	12	-16.31
WD								
	***Old* words**	**æ**	84	1078.02	83	963.37	-1	-114.65
		ʌ	67	1076.84	76	972.94	9	-103.9
		Total	76	1077.43	80	968.15	4	-109.28
	***New* nonwords**	**æ**	81	1087.64	88	968.17	7	-119.47
		ʌ	73	1067.54	88	986.18	15	-81.36
		Total	82	1077.59	88	977.17	6	-100.42

Table J.1. 4 Percentages of accurate scores for the NWDN and WDN group (T1,T2, M *diff*).

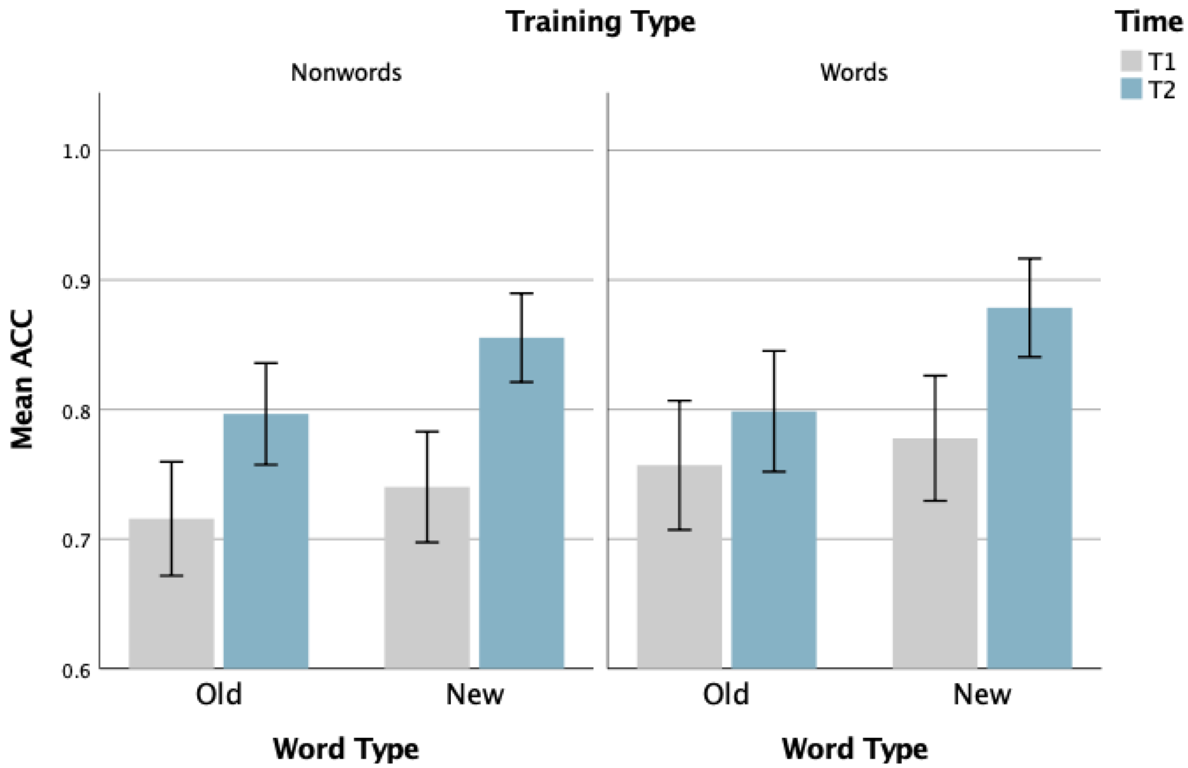

Figure J.1. 1 Bar graph for accuracy scores in the ABX by training type and word type (T1, T2).

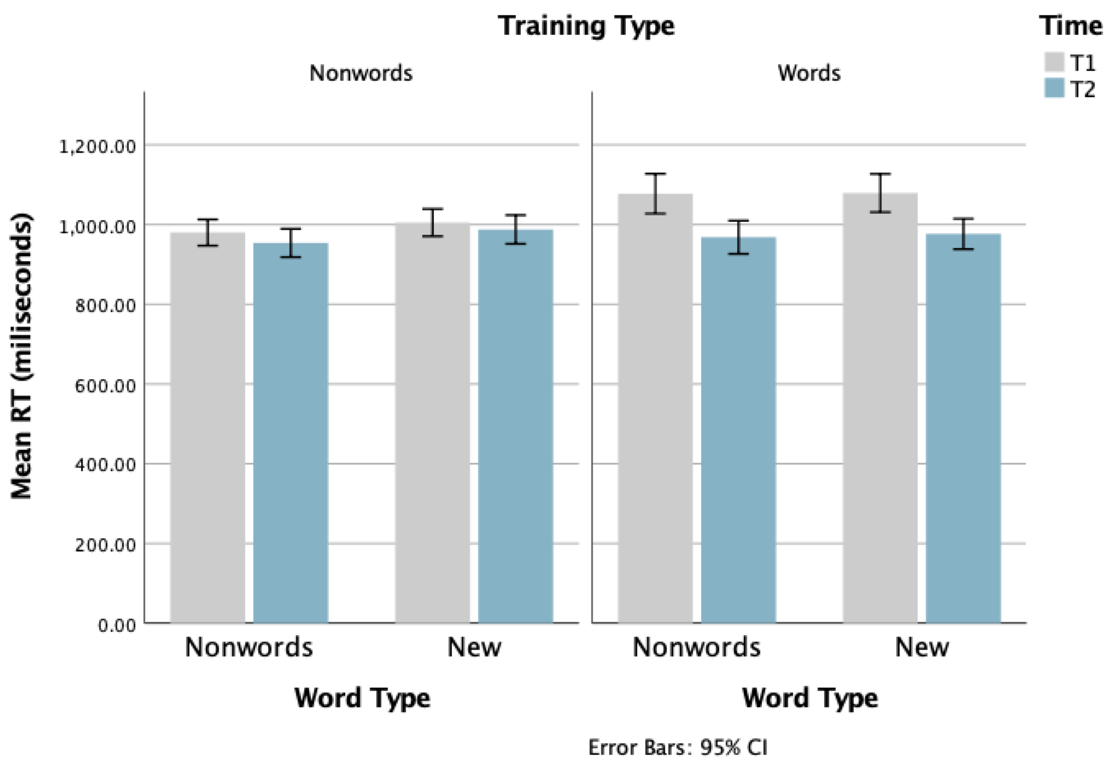

Figure J.1. 2 Bar graph for RT in the ABX by training type and word type (T1, T2).

J.2. DWR Results

For the DWR, spectral distance scores were fitted to a Linear Mixed effects Model (LMM) with *Time* (T1, T2), *Training Type*, *Word Type* (old, new), and *Vowel* (/æ/, /ʌ/) as fixed effects and *Subject* and *Item* as random factors. Type III of fixed effects showed that there was a main effect of *Testing Time* ($F(1, 1187.99) = 7.99$, $p = .005$) and *Vowel* ($F(1, 20.56) = 4.91$, $p = .038$). The *Training Type * Time* ($F(1, 1187.99) = 4.42$, $p = .036$) interaction and *Training Type * Time * Word Knowledge* ($F(1, 1191.26) = 8.82$, $p = .003$) reached significance. None of the other fixed effects or interactions reached significance (Table J.2.1). Bonferroni-adjusted pairwise comparisons indicate that the main effect of *Training Type* was driven by improvement in the NWD condition for *old* words for /æ/ and /ʌ/ (/æ/ $p = .001$; /ʌ/: $p = .044$) (Table J.2.2).

Results from the DWR do not show any significant differences between word type (*old* vs. *new*). These results are similar to those found in the ABX test where there was absence of statistical support for *new* words to be perceived more accurately. For the accuracy of production of words in isolation, descriptively, for the NWD group, *old* words improved to a larger extent after the phontic training (T2-T1, *old:* -0.41; *new:* 0.135)[27], while for the NW group, *new* words took more advantage of the training (T2-T1, *old:* 0.151; *new:* -0.231) (*Figure J.2.1*).

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	*p*
Intercept	1	29.816	162.453	.000
Training Type	1	23.995	1.765	.196
Testing Time	1	1187.995	7.991	.005
Word Type	1	29.722	.113	.739
Vowel	1	20.566	4.917	.038
Training Type * Time	1	1187.995	4.426	.036
Training Type * Word Type	1	1193.204	.096	.757

[27] These measures have been taken form SDS scores, the shorter the distance the better the vowel has been produced. In this case, -0.41 for the *old* words indicates that learners' pronunciation approximated more to native speakers' production at T2, while for the *new* words, the positive value .135 indicates that at T2, the distance between learners' and native speakers' pronunciation was larger.

*Training Type * Vowel*	1	1187.993	3.232	.072
*Time * Word Type*	1	1191.262	.233	.629
*Time * Vowel*	1	1187.993	.097	.755
*Word Type * Vowel*	1	29.568	.142	.709
*Training Type * Time * Word Type*	1	1191.262	8.828	.003
*Training Type * Time * Vowel*	1	1187.993	1.695	.193
*Training Type * Word Type * Vowel*	1	1193.084	2.627	.105
*Time * Word Type * Vowel*	1	1191.425	.095	.758
*Training Type * Time * Word Type * Vowel*	1	1191.425	.433	.511

a. Dependent Variable: SDS1.

Table J.2. 1 LMM: Type II Tests of Fixed Effects results

Pairwise Comparisons[a]

Training Type	Word Type	Vowel	(I) Time	(J) Time	Mean Diff. (I-J)	Std. Error	df	Sig.[c]	95% Confidence Interval for Difference[c] Lower Bound	Upper Bound
NWDT	**Old**	**æ**	**T1**	**T2**	.518*	.151	1189.014	.001	.222	.815
			T2	**T1**	-.518*	.151	1189.014	.001	-.815	-.222
		ʌ	**T1**	**T2**	.301*	.149	1191.435	.044	.008	.594
			T2	**T1**	-.301*	.149	1191.435	.044	-.594	-.008
	New	**æ**	**T1**	**T2**	.205	.149	1191.578	.171	-.088	.498
			T2	**T1**	-.205	.149	1191.578	.171	-.498	.088
		ʌ	**T1**	**T2**	.065	.151	1194.935	.667	-.232	.362
			T2	**T1**	-.065	.151	1194.935	.667	-.362	.232

WDT	**Old**	**æ**	**T1**	**T2**	-.259	.162	1187.800	.110	-.577	.059
			T2	**T1**	.259	.162	1187.800	.110	-.059	.577
		ʌ	**T1**	**T2**	-.043	.162	1187.800	.792	-.361	.275
			T2	**T1**	.043	.162	1187.800	.792	-.275	.361
	New	**æ**	**T1**	**T2**	.230	.162	1187.800	.157	-.088	.548
			T2	**T1**	-.230	.162	1187.800	.157	-.548	.088
		ʌ	**T1**	**T2**	.232	.162	1187.800	.152	-.086	.550
			T2	**T1**	-.232	.162	1187.800	.152	-.550	.086

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

a. Dependent Variable: SDS1.

c. Adjustment for multiple comparisons: Bonferroni.

Table J.2. 2 Bonferroni-adjusted pairwise comparisons (*Training type, Time, Word Type, Vowel*) for the DWR.

Group NWDN			**T1**	**T2**	**M *diff***
	Old	**æ**	1.721	1.203	-0.518
		ʌ	2.06	1.758	-0.302
		Total	1.89	1.48	-0.41
	New	**æ**	1.641	1.436	-0.205
		ʌ	1.767	1.702	-0.065
		Total	1.704	1.569	-0.135
Group WDN					
	Old	**æ**	1.209	1.468	0.259
		ʌ	1.386	1.429	0.043
		Total	1.297	1.448	0.151
	New	**æ**	1.385	1.155	-0.23
		ʌ	1.563	1.331	-0.232
		Total	1.474	1.243	-0.231

Table J.2. 3 Spectral distance score mean for group NWDN and WDN (T1, T2, M *diff*) for the DWR.

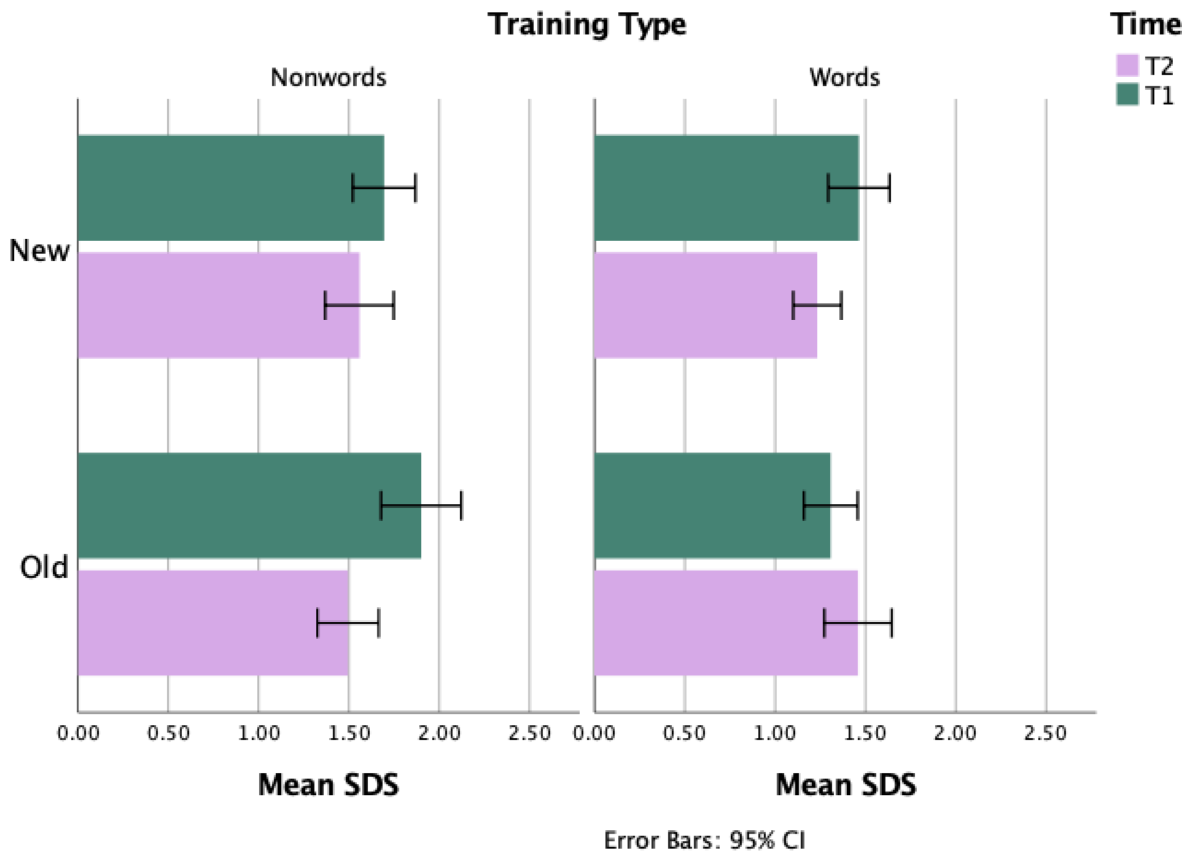

Figure J.2. 1 Bar graph for spectral distance scores in the ABX by training type and word type (T1, T2).

J.3 Lexical Decision results

The GLMM analysis revealed that there was a significant main effect of *Time*, F(1, 790) = 8.65, *p*=.003. No other fixed effect or interaction reached significance (Table J.1.1). Further Bonferroni-adjusted pairwise comparisons showed that the main effect of *Time* was mainly driven by learners who improved in the NWDA condition. Trainees improved significantly for both vowels (/æ/: *t*(790)=2, *p*=.046; /ʌ/: *t*(790)=2.27, *p*=.024) (Table J.1.2). Descriptively, we observed improvement in nonword-rejection for both groups, nonetheless, the NWDA group improved to a larger extent than the WDA group and was faster to respond (M *diff*, NWD, Acc.: 15%, RT: 50ms; WDT, Acc.: 6%, RT: 115ms) (Table J.1.3 and *Figure J.1.1*). Thus, the training period was effective at the perceptual level in improving how well the contrast was encoded at T2. Apart from this, despite trainees exhibiting inaccurate phono-lexical representations, these were at a fewer rate than at T1.

Fixed Effects[a]

Source	F	df1	df2	*p*
Corrected Model	1.916	7	95	.075
Time	8.653	1	790	.003
Training Type	.184	1	25	.672
Vowel	.909	1	12	.360
Time * TrainingType	1.166	1	790	.281
Time * Vowel	.752	1	790	.386
TrainingType * Vowel	.074	1	790	.786
Time*TrainingType *Vowel	1.008	1	790	.316

Probability distribution: Binomial
Link function: Logit
a. Target: ACC

Table J.3. 1 GLMM: Fixed Effects results.

Pairwise Contrasts

Training Type	Vowel	Time Pairwise Contrasts	Contrast Estimate	SE	t	df	*p*	95% Confidence Interval	
								Lower	Upper
NWDAT	**æ**	**T1 - T2**	-.131	.066	-2.000	790	.046	-.261	-.002
		T2 - T1	.131	.066	2.000	790	.046	.002	.261
	ʌ	**T1 - T2**	-.160	.071	-2.269	790	.024	-.299	-.022
		T2 - T1	.160	.071	2.269	790	.024	.022	.299
WDAT	**æ**	**T1 - T2**	-.130	.076	-1.699	790	.090	-.279	.020
		T2 - T1	.130	.076	1.699	790	.090	-.020	.279
	ʌ	**T1 - T2**	.000	.084	.000	790	1.000	-.165	.165
		T2 - T1	.000	.084	.000	790	1.000	-.165	.165

The least significant difference adjusted significance level is .05.

Table J.3. 2 Bonferroni-adjusted pairwise contrasts (*Training type, Time, Vowel*).

			T1		T2		M *diff*	
NWDAT			**ACC (%)**	**RT (ms.)**	**ACC (%)**	**RT (ms.)**	**ACC (%)**	**RT (ms.)**
	Test words	**æ**	84	1358	70	1403	-14	45
		ʌ	91	1306	85	1375	-6	69
		Total	87	1331	77	1388	-10	57
	Test nonwords	**æ**	26	1452	38	1507	12	55
		ʌ	35	1597	49	1644	14	47
		Total	31	1534	44	1584	13	50
WDAT								
	Test words	**æ**	81	1430	80	1402	-1	-28
		ʌ	90	1412	82	1449	-8	37
		Total	86	1420	81	1426	-5	6
	Test nonwords	**æ**	27	1540	39	1715	12	175
		ʌ	45	1591	45	1663	0	72
		Total	36	1573	42	1688	6	115

Table J.3. 3 Percentages of correctly identified test words and non-words for the WD group.

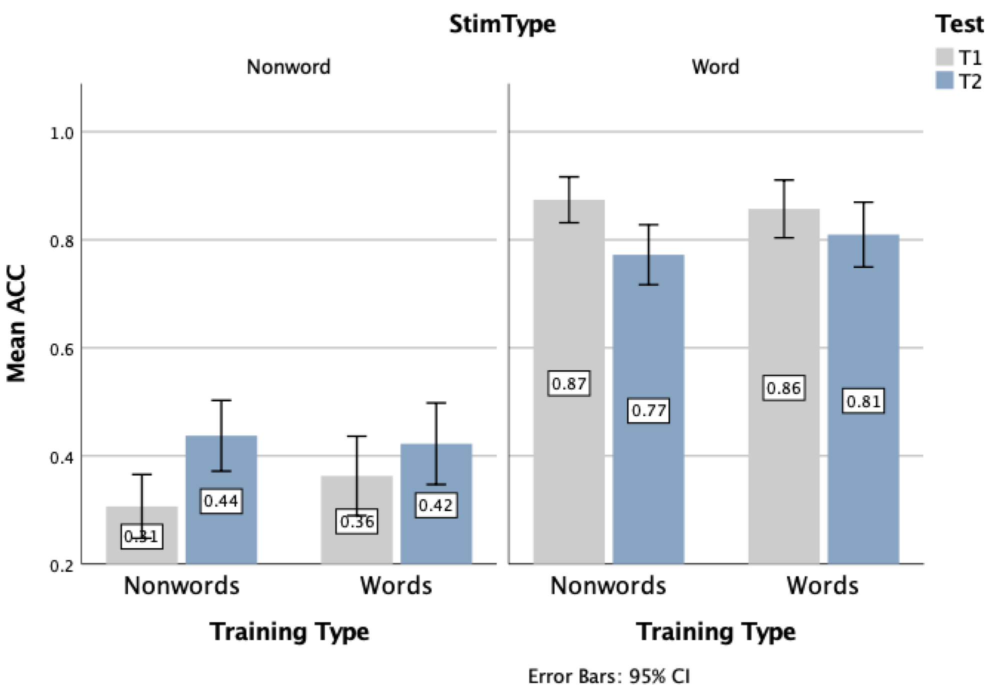

Figure J.3. 1 Non-word rejection accuracy means for NWDA and WDA at T1 and T2 by vowel.

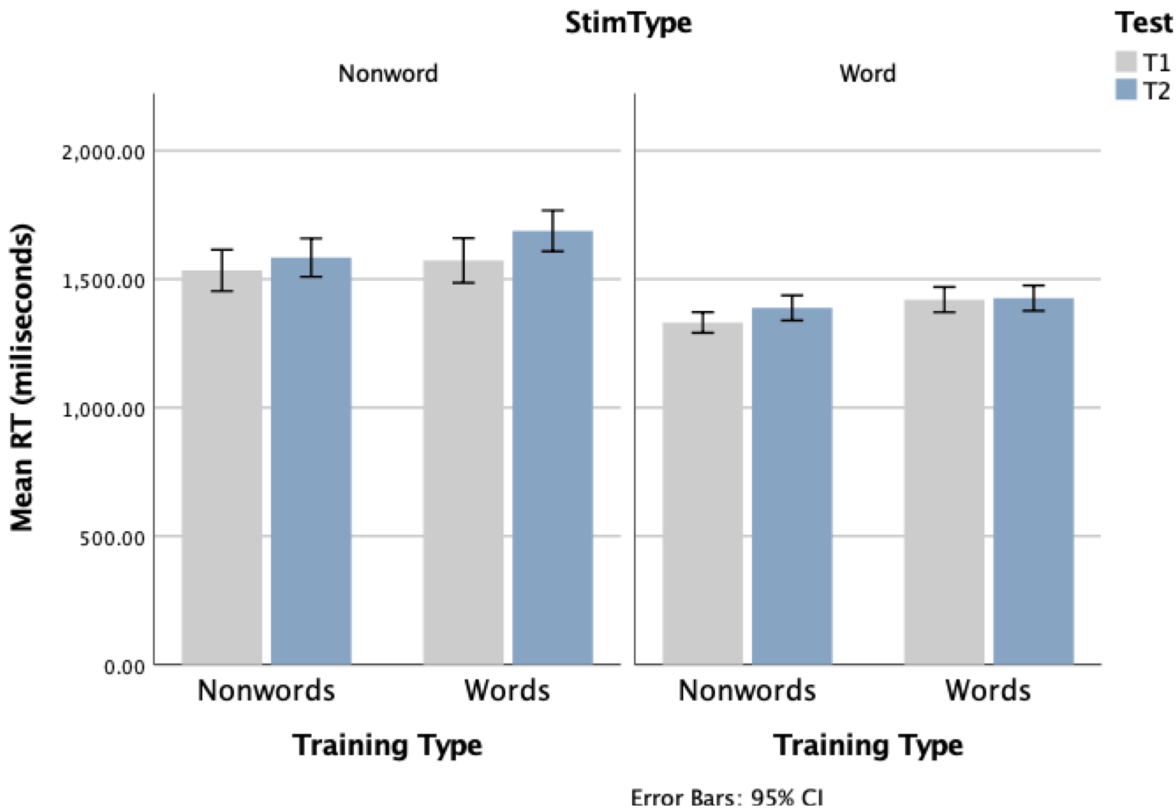

Figure J.3. 2 Bar graphs for RT for /æ/-/ʌ/ in nonwords and words separately (T1, T2).

J.4 Delayed Sentence Repetition results

Regarding production, in the LMM analysis, Type III of fixed effects indicated that there was a significant main effect of *Training Type,* $F(1, 749) = 9.94$, $p = .002$, and *Time,* $F(1, 37.46) = 8.39$, $p = .006$. No main effects were found neither for *Word Type* ($p = .635$) nor for *Vowel* ($p = .464$) (Table J.1.4). The only interaction that reached significance was the *TrainingType*Known*Vowel* interaction. Bonferroni-adjusted pairwise comparisons revealed that performance differed significantly only for the NDWA condition for both vowels and for *old* (/æ/: $F(1,749) = 19.40$, $p < .001$; /ʌ/: $F(1,749) = 22.31$, $p < .001$) and *new* (/æ/: $F(1,749) = 7.74$, $p = .006$; /ʌ/: $F(1,749) = 8.07$, $p = .005$) words (Table J.1.5). Descriptively, however, spectral distance scores indicate that trainees' phono-lexical representations of *new* words might be more malleable than for *old* words (M *diff*, NWDA, old: -.896, new: -1.137; WDA, old: .136, new: .479) (Table J.1.6 and *Figure J.1.*2).

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	*p*
Intercept	1	17.221	156.900	.000
Training Type	1	37.463	8.395	.006
Time	1	749	9.945	.002
Word Type	1	12.009	.237	.635
Vowel	1	12.009	.571	.464
TrainingType * Time	1	749	30.072	.000
TrainingType * WordType	1	749	.106	.745
TrainingType * Vowel	1	749	2.449	.118
Time * WordType	1	749	.000	.990
Time * Vowel	1	749.000	.130	.718
WordType * Vowel	1	12.009	.450	.515
TrainingType * Time * WordType	1	749	1.554	.213
TrainingType * Time * Vowel	1	749	.027	.870
TrainingType * WordType * Vowel	1	74	6.428	.011
Time * WordType * Vowel	1	749	.910	.340
TrainType * Time * WordType * Vowel	1	749	.013	.910

a. Dependent Variable: SDS.

Table J.4. 1 LMM: Type II Tests of Fixed Effects results for Study 1.

Pairwise Comparisons[a]

Train Type	Word Type	Vowel	(I) Time	(J) Time	Mean *diff* (I-J)	SE	df	*p*[c]	95% Confidence Interval for Difference[c] Lower Bound	Upper Bound
NWDA	**old**	**æ**	**T1**	**T2**	.835*	.190	749	.000	.463	1.207
			T2	**T1**	-.835*	.190	749	.000	-1.207	-.463
		ʌ	**T1**	**T2**	.967*	.205	749	.000	.565	1.369
			T2	**T1**	-.967*	.205	749	.000	-1.369	-.565
	new	**æ**	**T1**	**T2**	1.395*	.501	749	.006	.411	2.380
			T2	**T1**	-1.395*	.501	749	.006	-2.380	-.411
		ʌ	**T1**	**T2**	1.008*	.355	749	.005	.312	1.704

			T2	**T1**	-1.008*	.355	749	.005	-1.704	-.312
WDA	**old**	**æ**	**T1**	**T2**	-.209	.197	749	.290	-.596	.178
			T2	**T1**	.209	.197	749	.290	-.178	.596
		ʌ	**T1**	**T2**	-.052	.213	749	.807	-.470	.366
			T2	**T1**	.052	.213	749	.807	-.366	.470
	new	**æ**	**T1**	**T2**	-.310	.522	749	.552	-1.335	.714
			T2	**T1**	.310	.522	749	.552	-.714	1.335
		ʌ	**T1**	**T2**	-.563	.369	749	.128	-1.287	.161
			T2	**T1**	.563	.369	749	.128	-.161	1.287

Based on estimated marginal means
*. The mean difference is significant at the .05 level.
a. Dependent Variable: SDS.
c. Adjustment for multiple comparisons: Bonferroni.

Table J.4. 2 Bonferroni-adjusted pairwise comparisons (*Training type, Time, Word Type, Vowel*).

NWDA			**T1**	**T2**	**M *diff***
	Old	**æ**	2.127	1.338	-0.789
		ʌ	2.723	1.756	-0.967
		Total	2.427	1.531	-0.896
	New	**æ**	3.019	1.569	-1.45
		ʌ	2.348	2.197	-0.151
		Total	2.572	1.435	-1.137
WDA					
	Old	**æ**	1.224	1.433	0.209
		ʌ	1.550	1.602	0.052
		Total	1.375	1.511	0.136
	New	**æ**	1.201	1.512	0.311
		ʌ	1.592	2.159	0.567
		Total	1.464	1.943	0.479

Table J.4. 3 Spectral distance score mean for group NWDN and WDN (T1, T2, M *diff*).

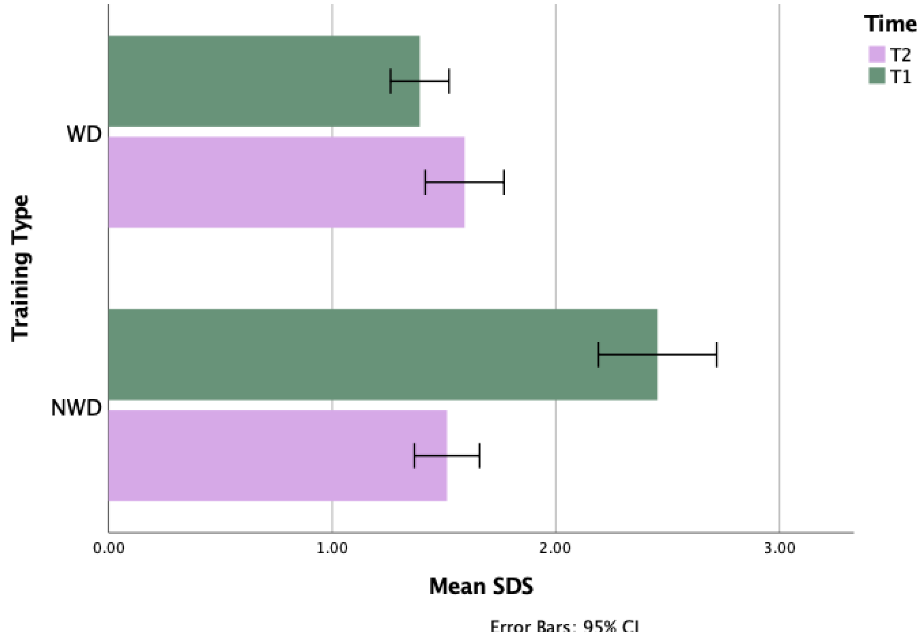

Figure J.4. 1 Spectral distance score mean for the /æ/-/ʌ/ contrast by training type (T1, T2).

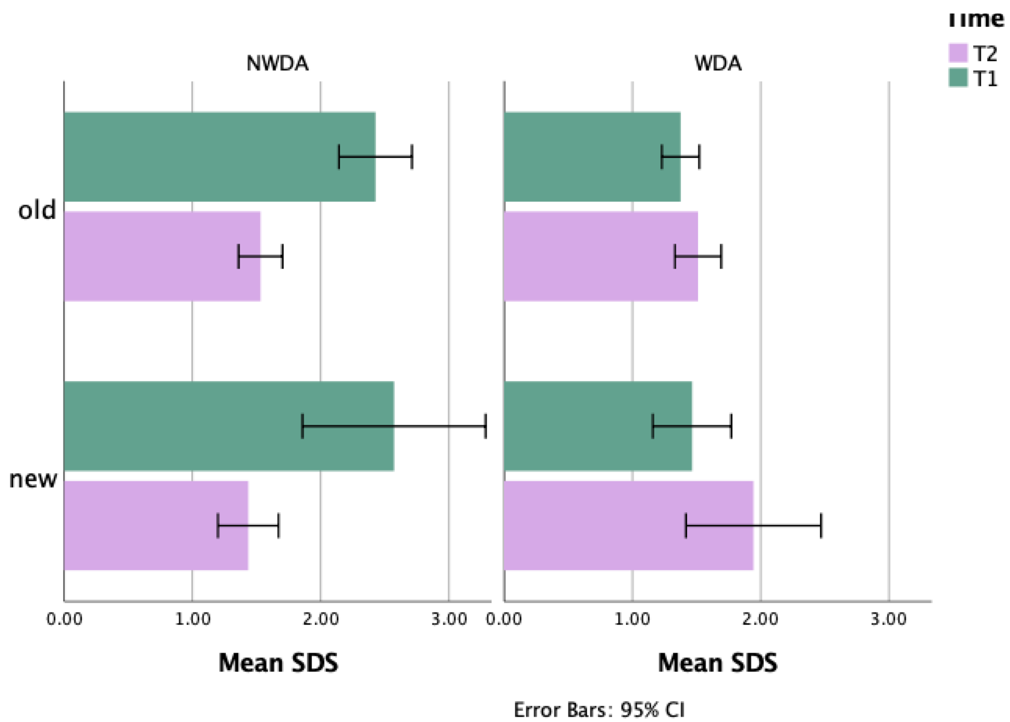

Figure J.4. 2 Bar graph for spectral distance scores for *old* and *new* words.

J.5 Relationship between ABX and LD at T2

ABX* LD Crosstabulation

			LD		Total
			Above 70%	Below 70%	
ABX	Above 85%	Count	4	3	7
		Expected Count	1.2	5.8	7.0
	Below 85%	Count	0	17	17
		Expected Count	2.8	14.2	17.0
Total		Count	4	20	24
		Expected Count	4.0	20.0	24.0

Table J.5. 1 ABX*LD Crosstabulation

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	11.657[a]	1	.001		
Continuity Correction[b]	7.906	1	.005		
Likelihood Ratio	12.066	1	.001		
Fisher's Exact Test				.003	.003
Linear-by-Linear Association	11.171	1	.001		
N of Valid Cases	24				

a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 1.17.

b. Computed only for a 2x2 table

Table J.5. 2 Chi-square tests results

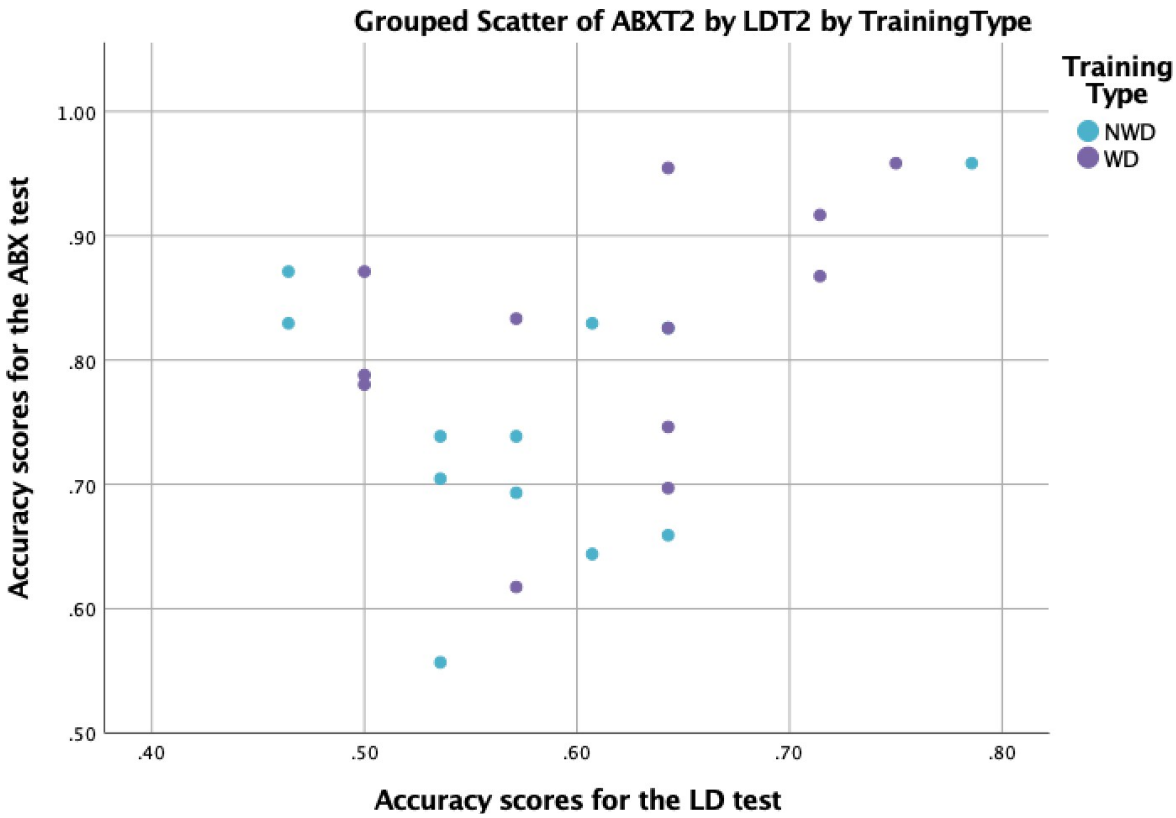

Figure J.5. 1 Accuracy scores for the ABX and LD test by participant and training type at T2.