

# Phonetic categorization ability and vocabulary size contribute to the encoding of difficult second-language phonological contrasts into the lexicon

## Research Article

Miquel Llompart

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

Received: 11 October 2019  
Revised: 5 October 2020  
Accepted: 5 October 2020  
First published online: 17 November 2020

**Keywords:**

second language acquisition; phonological representations; L2 phonology; L2 lexicon; speech perception

**Address for correspondence:**

Miquel Llompart,  
E-mail: [llompart.garcia@fau.de](mailto:llompart.garcia@fau.de)

Friedrich Alexander University Erlangen-Nuremberg

**Abstract**

This study investigated the contribution of second-language (L2) phonetic categorization abilities and vocabulary size to the phonolexical encoding of challenging non-native phonological contrasts into the L2 lexicon. Two groups of German learners of English differing in L2 proficiency (advanced vs. intermediate) participated in an English lexical decision task including words and nonwords with /ɛ/ and /æ/ (/æ/ does not exist in German), an /ɛ/-/æ/ phonetic categorization task and an English vocabulary test. Results showed that the effects of phonetic categorization and vocabulary size on lexical decision performance were modulated by proficiency: categorization predicted /ɛ/-/æ/ nonword rejection accuracy for intermediate learners, whereas vocabulary did so for advanced learners. This suggests that sufficient phonetic identification ability is key for an accurate phonological representation of difficult L2 phones, but, for learners for whom robust phonetic identification is already in place, their ultimate success is tightly linked to their vocabulary size in the L2.

**Introduction**

Learning a second language (L2) involves acquiring phones that are not part of the native phonological inventory. For late L2 learners, the perceptual system is already attuned to the phonological contrasts that are relevant in the native language (L1), which largely determines how L2 phones are perceived (Kuhl, Williams, Lacerda, Stevens & Lindblom, 1992; Polka & Werker, 1994; Werker & Tees, 1984). Therefore, as detailed in models of L2 phonology learning (Best & Tyler, 2007; Escudero, 2005; Flege, 1995), the relationship between the phonological systems of L1 and L2 is highly predictive of the difficulties that specific phones and phonological contrasts of the non-native language may cause. Especially problematic for late L2 learners are cases in which two L2 phones are mapped onto the same L1 phonological category (i.e., single-category assimilations; Best & Tyler, 2007). In such situations, difficulties to perceptually distinguish the two L2 contrastive sounds and to produce the corresponding acoustic differences between them have been repeatedly observed (Casillas & Simonet, 2015; Escudero & Boersma, 2004; Escudero & Wanrooij, 2010; Flege, Bohn & Jang, 1997; Goto, 1971, to name but a few). A well-known example of a “difficult L2 phonological contrast” is the English distinction between the open-mid front vowel /ɛ/ and the near open front vowel /æ/ for native speakers of German. This is because German speakers tend to associate both English /ɛ/ and /æ/ with their native vowel /ɛ/, which complicates the task of teasing the two apart (Bohn & Flege, 1990, 1992; Eger & Reinisch, 2019a, 2019b; Flege et al., 1997; Llompart & Reinisch, 2017, 2019a, 2019b, 2020).

Crucially, in order for native speakers of German to master a phonological contrast like /ɛ/-/æ/ and use it to their advantage in L2 communication, they need to reach two interrelated goals. First, they have to improve perceptual discrimination between the two vowels to the point that they can be accurately identified as two different non-native phonetic categories. Secondly, learners need to assign the L2 phones in question to the L2 words that contain them. In other words, they have to encode the phonetic categories corresponding to said phones into the higher-level phonological representations of L2 lexical items. This means that /ɛ/ has to be encoded as one of the phonological building blocks in words like *bet*, *chess* and *lemon*, and /æ/ has to be encoded into the phonological representations of *bat*, *flag* and *dragon*, and not the other way around. This process, which is tightly related to efficiency in lexical access and spoken word recognition<sup>1</sup>, is known as PHONOLEXICAL ENCODING

© The Author(s), 2020. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

<sup>1</sup>A thorough discussion of the architecture of the speech perception system is outside of the scope of this study. For this reason, its focus lies on tasks probing the auditory-phonetic and (phono)lexical levels of speech perception and I refrain from making claims on whether pre-lexical abstraction takes place in between (Goldinger, 1998; McQueen, Cutler & Norris, 2003), and if so, on the units that could be therein involved (Mitterer, Reinisch & McQueen, 2018; Llompart & Reinisch, 2018).

(Cook & Gor, 2015; Cook, Pandza, Lancaster & Gor, 2016; Darcy & Thomas, 2019; Hayes-Harb & Barrios, 2019; Llopart & Reinisch, 2019b, 2020).

The present study focuses on the phonolexical encoding of the English /ε/-/æ/ contrast by German learners of English with three main objectives. The first one is to reach a better understanding of the relationship between phonetic identification ability for the critical phonological contrast and its encoding into the representations of L2 words. Secondly, building on the fact that phonolexical encoding cannot be understood without the involvement of lexical knowledge, this study also attempts to establish to what extent a robust assignment of the vowels to English words (e.g., /ε/mon, dr/æ/gon) may be related to the learner's English vocabulary size. Finally, the third objective is to shed light on whether – and if so, how – L2 proficiency modulates the aforementioned relationships. For this purpose, a group of university students who had English instruction until secondary school is compared to an “expert” group of English instructors and English and American Studies students.

A number of studies have concentrated on comparing L2 learners' performance in tasks targeting the phonetic categorization of difficult L2 phones and tasks involving the retrieval of L2 phonological representations containing these phonetic categories (Amengual, 2016; Darcy, Daidone & Kojima, 2013; Darcy & Holliday, 2019; Díaz, Mitterer, Broersma & Sebastián-Gallés, 2012; Llopart & Reinisch, 2019a, 2019b; Pajak, Creel & Levy, 2016; Simonchyk & Darcy, 2017, 2018). These studies point towards the existence of a large gap between the establishment of phonetic distinctions for confusable L2 phones and their accurate phonological encoding into the lexicon. This is evidenced by the fact that most learners struggle with the auditory recognition of words containing phones in difficult L2 contrasts, even when they exhibit patterns of phonetic categorization that are close to native in segmental speech perception tasks (e.g., Amengual, 2016; Darcy et al., 2013). In the light of this mismatch, Amengual (2016) argues that accurate segmental judgments in explicit categorization and discrimination tasks do not always entail appropriate representations at the lexical level. Alternatively, such a gap could also indicate that the higher cognitive demands of lexical tasks prevent learners from accurately decoding fine phonetic detail in the input (Pajak et al., 2016) regardless of the state of their phonological representations. However, note that the same poorer performances in lexical tasks vs. phonetic tasks have been documented in production even when the latter involved higher cognitive demands during testing (Llopart & Reinisch, 2019a).

In spite of the highly consistent results outlined above for L2 learners as a group, the relationship between phonetic categorization and phonolexical encoding of difficult-to-distinguish L2 phones at the level of the individual learner is far from well understood. In principle, it is sensible to expect a relationship between the two because some degree of differentiation of the phonetic properties of the phones should be a prerequisite for a distinction to be encoded into the lexicon (Darcy et al., 2013; Darcy & Holliday, 2019; Simonchyk & Darcy, 2017, but see Cutler, 2015). However, the few studies that have indeed looked into relationships between individual patterns of performance for tasks tapping into phonetic categorization and phonolexical encoding have provided rather mixed results (Darcy & Holliday, 2019; Llopart & Reinisch, 2019a; Silbert et al., 2015; Simonchyk & Darcy, 2017). For example, Darcy and Holliday (2019) report a strong and significant correlation, for Chinese

learners of Korean, between their accuracy in a /o/-/Λ/ vowel identification task and a lexical decision task containing /o/-/Λ/ words and nonwords. In contrast, Simonchyk and Darcy (2017), having American learners of Russian as population of interest, failed to find a significant correlation between their individual scores in an ABX task designed to assess the perceptual distinction of plain and palatalized Russian consonants and a word-picture matching task probing their encoding of the plain/palatalized contrasts into Russian words.

A first possible explanation for these mixed results is that the strength of the relationship between phonetic categorization and phonolexical encoding may vary as a function of L2 proficiency, only being robust in the stages of L2 acquisition in which there are meaningful between-subject differences in identification ability. This idea refers to a scenario in which some learners may already be able to distinguish between the two phones relatively well in perceptual categorization tasks while others may still struggle to do so. This is usually the case for intermediate L2 learners (Darcy, Park & Yang, 2015; Kim, Clayards & Goad, 2017; Simonchyk & Darcy, 2017) but not for very inexperienced L2 learners, who generally fail to differentiate between the phones in difficult L2 contrasts even in the simplest categorization tasks (e.g., Bohn & Flege, 1990), or highly-proficient L2 learners, who are often found to perform in these tasks with close to native-like accuracy (Amengual, 2016; Darcy et al., 2013).

A second and non-exclusive potential explanation for the frequent mismatch between phonetic categorization and phonolexical encoding is that the latter intertwines with a level of knowledge that is not necessarily involved in the former: lexical knowledge. While the robustness of L2 phonetic categories can be assessed on the spot in a perceptual categorization task, any paradigm assessing their phonolexical encoding relies on learners activating and retrieving the phonological representations of non-native lexical items, and L2 learners may differ in terms of lexical knowledge in important ways. Firstly, learners differ in vocabulary size (or breadth) – that is, in how many words are part of their L2 lexicon (Cameron, 2002; Nation, 2006). Secondly, learners may also differ in how and how often they have been exposed to the words they know. For example, some words may have been encountered very often but mostly in the visual modality, while others may have come across in very limited instances, but always auditorily. These differences can determine how well learners know the words that they know (i.e., vocabulary depth; Wesche & Paribakht, 1996). Furthermore, now specifically with regard to the phonological representations of known words, the type of input in the oral modality can also have critical consequences on the native-likeness of such representations. In a non-immersion classroom setting, L2 vocabulary may have been taught by non-native speakers sharing an L1 with the learners. This means that, even if certain words are well known and even frequently used by learners, phonological representations may contain inaccuracies regarding particular L2 phones due to their exposure to accented input (Bohn & Bundgaard-Nielsen, 2009; Bundgaard-Nielsen, Best, Kroos & Tyler, 2012; Eger & Reinisch, 2019b).

In the present study, I assess the potential role of vocabulary size as a predictor of the phonolexical encoding of difficult L2 phonological contrasts. It is known that, in early L1 acquisition, vocabulary size relates to one's phonolexical encoding of native phonetic categories. Vocabulary size correlates with young children's ability to use native phonological contrasts in novel word learning paradigms (Mani & Plunkett, 2010; Werker, Fennell,

Corcoran & Stager, 2002; Yoshida, Fennell, Swingle & Werker, 2009) and to recognize familiar words produced in unfamiliar L1 accents (Best, Tyler, Gooding, Orlando & Quann, 2009; Mulak, Best, Tyler, Kitamura & Irwin, 2013). Because of this, it has been argued that vocabulary expansion in L1 acquisition drives the attunement to the phonological structures of the native language (Kemp et al., 2017; Metsala, 1999). Similarly, larger vocabulary sizes also seem to be beneficial in adult L2 learning, as they relate to better perceptual discrimination and production of L2 phones (Bundgaard-Nielsen, Best & Tyler, 2011a, 2011b) and enhanced L2 word segmentation skills (Darcy et al., 2015). However, it remains to be seen whether this same connection is also observed particularly in the acquisition of difficult non-native phonological contrasts by adult L2 learners.

In the L2, it is possible that, at least in the initial stages of learning, large vocabularies may not only not be beneficial, but even be detrimental for the establishment of challenging L2 phonological contrasts. As reviewed above, when L2 learning starts, perception of L2 phones is based on their similarity to L1 phones and, consequently, some are perceived inaccurately (Best & Tyler, 2007; Escudero, 2005; Flege, 1995). Additionally, perception is also conditioned by the orthography-to-phonology mappings of the L1 (e.g., Escudero, Simon & Mulak, 2014). Because of this, Best and Tyler (2007) suggested that rapidly acquiring a large vocabulary in a non-native language may lead to the fossilization of L2 phonological representations reflecting L1 properties (i.e., accented phonological representations). In contrast to this view, Bundgaard-Nielsen et al. (2011a, 2011b) proposed that vocabulary learning should actually facilitate learners' perceptual reattunement to non-native phones and their rephonologization as L2-specific phonological categories making up the representations of L2 words. This also fits well with the hypothesis outlined in Llompart (2019) that extensive L2 input and repeatedly encountering L2 words containing the confusable phones are essential to bolster the phonolexical encoding of these phones. Partial support for this claim comes from spoken word recognition tasks with bilingual speakers showing that the phonolexical encoding of challenging phonological contrasts in the non-dominant language is positively related to how much learners use that language (Amengual, 2016; Sebastián-Gallés, Echeverría & Bosch, 2005).

In the present study, individual measures of /ɛ/-/æ/ phonetic categorization and English vocabulary were obtained and used to predict the phonolexical encoding of /ɛ/ and /æ/ into English words by two groups of German learners of English differing in their L2 experience and proficiency (i.e., advanced vs. intermediate learners). Robustness of /ɛ/-/æ/ phonolexical encoding was assessed using a lexical decision task containing real words and nonwords with systematic mispronunciations (e.g., \*[æ]mon, \*dr[ɛ]gn). Phonetic categorization for /ɛ/-/æ/ was quantified by means of a two-alternative forced-choice perceptual categorization task on the steps of a *bet-bat* speech continuum. Finally, English vocabulary size was assessed through the vocabulary component of the revised and restandardized version of the Shipley Institute of Living Scale (Shipley-2; Shipley, Gruber, Martin & Klein, 2009; see also Kaya, Delen & Bulut, 2012).

In terms of specific predictions, both phonetic categorization ability and vocabulary size were expected to relate to learners' phonolexical encoding of /ɛ/ and /æ/ into English words. However, these relationships were predicted to be modulated by the learners' L2 proficiency. In particular, learners in the advanced group were expected to be rather homogeneous

regarding their phonetic categorization of the critical contrast, for most of them should already exhibit a clear-cut perceptual differentiation of /ɛ/ and /æ/. Because of this, a strong relationship between phonetic categorization of /ɛ/ and /æ/ and its phonolexical encoding into L2 words (in the lexical decision task) was predicted to be less likely for this group than for the intermediate learner group, for whom more variation in categorization was expected. For the relationship between L2 vocabulary size and phonolexical encoding, by contrast, two different outcomes could in principle be predicted. On the one hand, following Best and Tyler (2007), one would expect large English vocabularies to relate to a less robust encoding of /ɛ/- and /æ/- into the phonological representations of known L2 words. On the other hand, in accordance with Bundgaard-Nielsen et al. (2011a, 2011b), learners with larger vocabulary sizes would be expected to show a more robust and native-like phonolexical encoding of /ɛ/ and /æ/. If so, this relationship should be apparent for advanced learners of English but might not surface so clearly for intermediate learners due to the potential hindering effects of inaccurate phonetic categorization abilities on their phonolexical encoding.

## Methods

### Participants

Seventy-nine native speakers of German participated in this study in exchange for course credit or a small payment. All participants grew up in monolingual German households and had not learned any language other than German before they started to learn English at school. Participants were divided into two groups based on their education and the role of English in their academic careers. First, an advanced group was established encompassing English professionals and students aiming to become English professionals. A total of 30 participants (17 females) were assigned to this group. They were either students currently enrolled in the BA and MA programs at the Department of English and American Studies of the University of Erlangen-Nuremberg (N = 25) or English instructors at the Center for Languages of the same university (N = 5). Secondly, a group of intermediate learners (N = 49, 35 females) was established by recruiting students enrolled in degree programs at the University of Erlangen-Nuremberg but in departments other than English and American Studies. Grouping participants according to their education, which is an objective criterion and is expected to have had consequences on their English proficiency, was considered preferable to dividing them according to subjective self-reported measures of English proficiency or similar. This is because the reliability of these measures has recently been put into question even when, as in the present study, all participants come from similar learner populations (Tomoschuk, Ferreira & Gollan, 2019). Nevertheless, participants were indeed asked to provide several self-reported measures of English proficiency and use together with information on their language learning history in a background questionnaire. This allowed for an examination of how well these measures captured critical differences between participants in the advanced and intermediate groups. Participant's age, age of English acquisition, length of residence in an English-speaking country plus self-reported measures of English use and English proficiency for the two groups of participants are shown in Table 1.

Comparisons via Welch two sample t-tests, in order to account for the samples' unequal variances (Welch, 1947), indicated that

**Table 1.** Mean age, age of acquisition, years in an English-speaking country, and self-reported English use and proficiency measures for the advanced and intermediate groups. Standard deviations are in parentheses. Use and contact with native speakers' ratings were given on a 1 to 6 scale and proficiency and accent ratings on a 1 to 5 scale, with 6 and 5, respectively, being the highest scores and 1 the lowest score. Statistical differences were calculated by means of Welch two sample t-tests.

Measure	Advanced	Intermediate	Difference
Age	28.5 (12.32)	24.22 (4.26)	$t(33.1) = 1.84, p = .08$
Age of acquisition	8.7 (2.69)	9.11 (1.89)	$t(47.18) = -0.72, p = .48$
Years in English-speaking country	1.45 (3.42)	0.42 (0.86)	$t(31.4) = 1.62, p = .12$
English use at home	2.4 (0.97)	2.12 (0.95)	$t(60.51) = 1.25, p = .22$
English use at university	3.3 (0.95)	1.98 (0.8)	$t(53.63) = 6.34, p < .001$
Current contact with native speakers	3.07 (1.51)	2.65 (1.59)	$t(63.68) = 1.16, p = .25$
English comprehension	4.53 (0.57)	3.9 (0.92)	$t(76.96) = 3.79, p < .001$
Spoken English	4.03 (0.67)	3.53 (0.87)	$t(72.87) = 2.89, p < .01$
Overall English proficiency	4.23 (0.57)	3.78 (0.71)	$t(71.73) = 3.14, p < .01$
Accent in English (higher = weaker accent)	3.73 (0.69)	3.20 (0.79)	$t(67.74) = 3.13, p < .01$

the two groups did not differ in age, age of acquisition and length of residence in an English-speaking country. Importantly, they did not differ in self-estimated use of English at home and contact with native speakers in their everyday lives at the moment either. In contrast, advanced and intermediate learners did differ in how much they used English at the university and, critically, in their self-reported measures for English comprehension, spoken English, overall English proficiency and strength of German accent when speaking English. Hence, Table 1 highlights the similarity between the two groups in that neither of them is currently exposed to English in an immersion setting, but also indicates that self-reported proficiency scores aligned well with the distinction based on educational background that was implemented.

### Materials and procedure

The present study consisted of three tasks: a lexical decision task, a phonetic categorization task and a vocabulary test. Materials and procedure for each particular task are described in detail in the following subsections. Participants were tested individually or in groups of two in a quiet room at the Friedrich Alexander University Erlangen-Nuremberg. The tasks were completed over two experimental sessions on two consecutive days. In the first session, participants completed the lexical decision task, whereas in the second session they were tested on the phonetic categorization task and the vocabulary test. Splitting the tasks into two sessions had the advantage that the two speech perception tasks were run on different days, therefore minimizing the influence they could have had on each other. All three tasks were implemented in Psychopy3 (v. 3.0.2; Peirce, Gray, Simpson, MacAskill, Höchenberger, Sogo, Kastman & Lindeløv, 2019) and were run on Lenovo Thinkpad L580 (15.6-inch screen) laptops. Auditory stimuli were presented binaurally at a comfortable listening level by means of Beyerdynamic DT770 PRO headphones and responses were recorded using Cedrus RB-740 response boxes.

### Lexical decision task

The lexical decision task used to assess the phonolexical encoding of /ɛ/-/æ/ was the same as the one described in Llopart and Reinisch (2019b). Learners were presented with real words and nonwords with systematic mispronunciations and had to decide

whether the stimuli were real English words. Critical items were words containing /ɛ/ and /æ/ (e.g., *desert*) and nonwords created by swapping the two vowels (e.g., \*[ɛ]mon). This type of lexical decision has been recurrently used to probe the robustness of the phonological representations of lexical items in monolingual and bilingual speakers (Amengual, 2016; Sebastián-Gallés et al., 2005) as well as late L2 learners (Darcy & Thomas, 2019; Darcy et al., 2013; Díaz et al., 2012; Llopart & Reinisch, 2019b). While it should be acknowledged that this task only allows for binary responses (word/nonword) to key acoustic modifications that are in many cases better characterized as gradient (e.g., changes in vowel quality), lexical decision tasks with this kind of phonological substitutions are a widely used paradigm to prompt explicit comparisons between a determined acoustic input (e.g., \*[ɛ]mon) and stored phonological representations (e.g., /'lɛmən/).

Materials for the lexical decision task were 304 English words including mono-, di- and tri-syllabic nouns, adjectives and verbs. 52 of these words contained the critical /ɛ/-/æ/ contrast, 26 for each vowel. The remaining 252 words were used as fillers and they involved five L2 contrasts (i.e., 2 vowel and 3 consonant contrasts) that were expected to be easy for our participant population because they also exist in German: /i/-/ɪ/, /ɔ/-/ʊ/, /p/-/t/, /k/-/m/ and /b/-/v/. Learners were predicted to be quite accurate with both words and nonwords containing the filler contrasts. Filler items thus provided an ideal baseline for performance in the lexical decision task in the absence of confusable L2 phones, which allowed for a controlled observation of the effects that the difficult L2 phonological contrast (i.e., /ɛ/-/æ/) had on word recognition and nonword rejection.

Half of the words in the set were selected to be used as nonwords created by exchanging the two phones in each contrast. For /ɛ/-/æ/, this means that 13 /ɛ/-words and 13 /æ/-words appeared as containing a single-sound mispronunciation in the first stressed vowel (e.g., \*[ɛ]mon for *lemon* and \*dr[ɛ]gon for *dragon*), while the remaining 13, in each case, appeared as correctly-pronounced real words. The same procedure applied to the 5 filler contrasts, even though for them the position of the mispronunciation in the word could vary. Lexical frequency for /ɛ/-/æ/ items was controlled using the Zipf-scale frequency measures provided by Subtlex-UK (van Heuven, Mandera, Keuleers & Brysbaert, 2014). Mean frequencies were 4.66 for

/ɛ/-real words, 4.66 for /ɛ/-mispronounced nonwords, 4.82 for /æ/-real words and 4.41 for /æ/-mispronounced nonwords. Additionally, phonological neighborhood density for each set of items was assessed using the neighborhood size measures provided by CLEARPOND (Marian, Bartolotti, Chabal & Shook, 2012), respectively. Mean neighborhood sizes were 9.38 for /ɛ/-real words, 7.92 for /ɛ/-mispronounced nonwords, 8.23 for /æ/-real words and 6.15 for /æ/-mispronounced nonwords. Frequency, phonological neighborhood size and cognate status for all /ɛ/-/æ/ items are provided in Appendix B.

Words were recorded by a 26-year-old male speaker of Standard Southern British English who lived in London until moving to Germany at the age of 22. All words were recorded in their correct forms and items selected to work as nonwords were recorded with the suitable substitutions (see Llopart & Reinisch, 2019b, for more detail and formant values for /ɛ/-/æ/ real words and mispronounced nonwords). The sets of items presented as real words and nonwords with mispronunciations were the same for all participants. Because of the study's focus on individual differences and between-group comparisons, this was necessary to minimize the risk that performance differences between participants (and groups) were caused by the lexical properties of specific items. This would have been a likely outcome had items been divided in various lists and only a subset of participants had responded to each list. With the present design, however, it is unlikely that variability in the lexical properties of target items had a major influence on the results reported. The order of presentation of the items was fully randomized.

Participants were told that they would be hearing real English words and invented words that in some cases could sound similar to English words and that their task was to decide whether what they heard was an existing English word. Two small boxes were shown on the screen on every trial: a green box with "word" written on it on the left-hand side, and a red one with "not a word" written on it on the right-hand side. Participants were instructed to press the leftmost key of the response box (with a green key top) if they considered the presented item to be a real word, and the rightmost key (with a red key top) if they considered it not to be a word. There was no time limit on responses and the next item was presented 0.8 s after the previous key press. Ten practice trials were presented before the start of the experiment. The task took between 15 and 20 minutes to complete.

### Phonetic categorization task

A two-alternative forced-choice task (2AFC) assessed the learners' phonetic categorization of the /ɛ/-/æ/ contrast. The task used here was the same as the one described in Llopart and Reinisch (2019a). A 21-step *bet-bat* continuum was created in Praat (Boersma & Weenink, 2010) from natural productions from the same speaker who recorded the words for the lexical decision task. Vowel duration and first and second formant values (F1 and F2; in Hertz) for the endpoints of the continuum were taken from naturally-produced tokens of *bet* and *bat*, and continuum steps were set to change linearly in all three dimensions (see Llopart & Reinisch, 2019a, for further details). Participants were told that they would hear English words and see two pictures on the screen, and that their task would be to decide which of the pictures matched the word just heard better. Note here that, even though two real words are used as carrier contexts for the critical L2 phones and as response options, the

experimental demands associated to this task direct participants' attention towards the acoustic properties of the critical vowel in the stimuli. For this reason, its outcome is widely considered a reflection of phonetic (i.e., perceptual) identification abilities rather than fully-fledged (phono)lexical identification (e.g., Amengual, 2016; Broersma, 2005; Pallier, Bosch & Sebastián-Gallés, 1997).

To enter their responses, participants were instructed to press the leftmost key of the response box when the auditory stimulus matched the picture on the left (i.e., *bet*) and the rightmost key when it matched the picture on the right (i.e., *bat*). Orthographic representations of the words (BET-BAT) were provided while participants received written instructions prior to the task, but not during the task itself. The 21 steps of the continuum were presented 10 times. All 21 steps had to be presented before any of them was repeated, but, within each set of 21, steps were presented in random order. No time limit was imposed on responses and presentation of the next trial started 0.8 s after the previous button press. The categorization task took approximately 5 minutes to complete.

### Vocabulary test

A computerized implementation of the vocabulary component of Shipley-2 (Shipley et al., 2009) was used as English vocabulary size test. This is a written, normed, multiple-choice receptive-vocabulary test consisting of 40 English words of diverging lexical frequencies. A written vocabulary test was chosen to keep task demands low and to minimize the influence of L2 oral comprehension abilities on vocabulary scores quantifying learners' lexical-semantic knowledge. In each trial, one English word was presented on the screen in uppercase letters (e.g., TALK) and four possible responses were provided in lowercase preceded by numbers 1 to 4 (e.g., 1. Draw, 2. Eat, 3. Speak, 4. Sleep, "speak" being the correct answer). Participants were instructed to press the key of the response box (labelled from 1 to 4) corresponding to the word whose meaning approximated that of the uppercase word the most. No time limit was imposed on responses and presentation of the next trial started 0.5 s after the previous button press. The vocabulary test took between 5 and 8 minutes to complete.

## Results

The main goal of the present study was to assess the relative influence of phonetic categorization ability and vocabulary size on the phonological encoding of a difficult L2 contrast at two different levels of L2 proficiency. To do this, it was necessary to obtain individual scores from the /ɛ/-/æ/ phonetic categorization task and the vocabulary test in order to examine their effectiveness as predictors of performance with /ɛ/-/æ/ in the lexical decision task. Hence, the individual measures derived from the two former tasks are described in detail in the first subsection, whereas analyses on lexical decision data are presented in the second and third subsections. In the second subsection, effects of L2 proficiency (i.e., advanced vs. intermediate) and contrast type (i.e., /ɛ/-/æ/ vs. filler contrasts) on accuracy and reaction times (RT) in the lexical decision task are assessed. In the third subsection, analyses on the effects of phonetic categorization and vocabulary, and their interactions with L2 proficiency, on lexical decision scores and RTs for /ɛ/- and /æ/-items only are reported.

### Categorization and vocabulary: obtaining individual measures

Performance in the phonetic categorization task was measured by calculating the steepness (i.e., slope) of the / $\epsilon$ /-/ $\ae$ / categorization curve. Following Llompart and Reinisch (2019a), individual slopes were calculated by submitting the categorization data to a generalized linear mixed-effects regression model with a logistic linking function (lme4 package 1.1–20, Bates, Mächler, Bolker & Walker, 2015) in R (Version 3.5.2, R Core Team, 2017) with Response (button box responses coded as 0 and 1) as the categorical dependent variable, an intercept term and a random slope for Continuum step over Participants. From the model, the slope coefficients by participant were extracted. The coefficients quantify the increase in log-odds of a “bat” response as a function of an increase of one unit in terms of continuum step. A coefficient of 0 indicates no change and therefore a flat categorization function. The higher the slope coefficient, the more dramatic the increase in log-odds, and hence the steeper the slope of the categorization function. For illustration, individual categorization curves for all participants are provided in Figure A1 (Appendix A). For the vocabulary test, individual measures were calculated as the proportion of correct responses each participant obtained out of the 40 test items. Welch two sample t-tests comparing individual values between groups revealed that these measures behaved accordingly to the established participant grouping and to learners’ self-reported proficiency estimations. Advanced learners had steeper categorization slopes ( $t(61.94) = 3.03$ ,  $p < .01$ ; intermediate:  $M = 0.56$ ,  $SD = 0.24$ ; advanced:  $M = 0.73$ ,  $SD = 0.23$ ) and higher vocabulary scores ( $t(63.30) = 4.22$ ,  $p < .001$ ; intermediate:  $M = 0.64$ ,  $SD = 0.11$ ; advanced:  $M = 0.75$ ,  $SD = 0.11$ ) than intermediate learners.

### Lexical decision: / $\epsilon$ /-/ $\ae$ / vs. filler contrasts

The initial lexical decision dataset included 24016 trials (304 trials  $\times$  79 participants). From this, all trials that contained words with which participants indicated to be unfamiliar (as assessed at the end of the second experimental session) were removed, which resulted in a loss of 179 trials (0.75% of the whole dataset). Of these, only 20 involved the / $\epsilon$ /-/ $\ae$ / contrast (0.49% of / $\epsilon$ /-/ $\ae$ / trials). Accuracy with real words was similarly close to ceiling for the two groups for both filler (intermediate:  $M = 94.49\%$  correct,  $SD = 22.82$ ; advanced:  $M = 97.27\%$  correct,  $SD = 16.29$ ) and critical / $\epsilon$ /-/ $\ae$ / items (intermediate:  $M = 96.61\%$  correct,  $SD = 18.1$ ; advanced:  $M = 97.44\%$  correct,  $SD = 15.82$ ). Motivated by these ceiling effects, and in accordance with previous studies (Amengual, 2016; Darcy & Thomas, 2019), only responses to nonwords were included in all forthcoming analyses. Figure 1 shows the proportion of correct responses to nonword items (left panel) and reaction times for correct responses to nonword items (right) by item type (filler items vs. / $\epsilon$ /-/ $\ae$ / items) for the two groups of L2 learners.

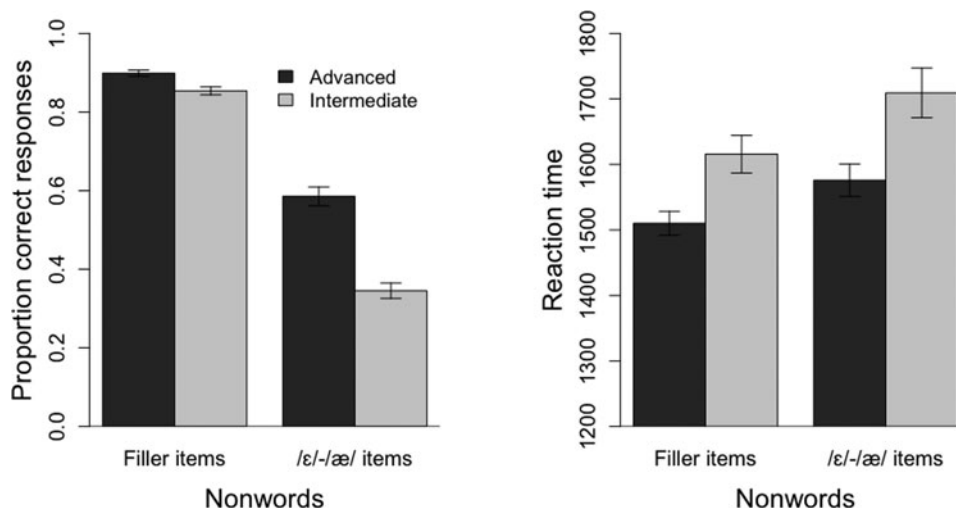
Before directly assessing the effects of phonetic categorization ability and vocabulary size on performance with / $\epsilon$ /-/ $\ae$ / nonwords in the lexical decision task, a preliminary analysis was conducted in order to ascertain i) that the L2 learners in the present study indeed exhibited enhanced difficulties with nonwords containing the a priori difficult / $\epsilon$ /-/ $\ae$ / contrast in comparison to the remaining filler nonwords containing easy L2 contrasts, and ii) that accuracy was higher for the advanced group than for the intermediate group. These were highly expected findings, given that even early consecutive bilinguals have difficulties with

nonwords containing challenging L2 contrasts in this type of task, and proficiency and usage have previously been found to modulate learners’ performance (Amengual, 2016; Sebastián-Gallés et al., 2005). However, both were essential premises for the experimental paradigm to be able to shed light on this study’s main research questions and needed to be confirmed.

Hence, data for all nonword trials (i.e., responses to filler and critical / $\epsilon$ /-/ $\ae$ / nonwords) were first submitted to a generalized linear mixed-effects model (henceforth GLMM) with a logistic linking function with Response (0 = incorrect, 1 = correct) as the categorical dependent variable and Contrast Type (Filler / Critical) and Proficiency Group (Advanced / Intermediate), as well as their interaction, as predictors. Contrast Type was contrast coded with Critical as -0.5 and Filler as 0.5. Proficiency Group was coded with Intermediate as -0.5 and Advanced as 0.5. Random-effects structures for all analyses in this study were chosen by a model fitting procedure using log-likelihood ratio tests and random slopes were only included if they improved the model’s fit. The random-effects structure for this model included random intercepts for Participants and Items, a random slope for Contrast Type over Participants, and a random slope for Proficiency Group over Items, as both slopes improved the model’s fit in comparison to models in which they were not included (Contrast Type over Participants:  $\chi^2(2) = 76.16$ ,  $p < .001$ ; Proficiency Group over Items:  $\chi^2(2) = 16.32$ ,  $p < .001$ ). Parameter estimates and significance values for all variables in GLMMs are provided as rendered by the lme4 package (Bates et al., 2015), which makes use of maximum likelihood estimation based on the Laplace approximation (Raudenbush, Yang & Yosef, 2000). Results are provided in Table 2.

The model revealed significant effects of Contrast Type and Proficiency Group, showing that learners were less accurate with nonword items containing / $\epsilon$ / and / $\ae$ / than with filler items, and that learners in the advanced group were overall more accurate than those in the intermediate group. In addition, the interaction between Contrast Type and Proficiency Group was also significant. To follow up on this interaction, data were split by Proficiency Group and separate GLMMs were run for each group with Response as dependent variable and Contrast Type as fixed factor. Both models included random intercepts for Participants and Items and a random slope for Contrast Type over Participants. The random slope improved the fit of both models (Intermediate:  $\chi^2(2) = 37.03$ ,  $p < .001$ ; Advanced:  $\chi^2(2) = 39.75$ ,  $p < .001$ ). The model on data from the intermediate learner group revealed a significant effect of Contrast Type ( $b = 3.42$ ;  $z = 10.79$ ;  $p < .001$ ) and so did the model on data from the advanced group ( $b = 2.51$ ;  $z = 6.42$ ;  $p < .001$ ). These additional analyses confirmed that both groups of L2 learners showed evidence of a more unreliable phonolexical encoding of the difficult non-native contrast versus the other, easier, L2 phonological distinctions. Furthermore, as suggested by the estimates and by Figure 1 (left panel), accuracy differences between the two contrast types appear to be larger for the intermediate learners than the advanced learners<sup>2</sup>.

<sup>2</sup>Within the set of nonwords with / $\epsilon$ / and / $\ae$ /, both groups showed numerically lower accuracies for items in which / $\ae$ / was mispronounced as [ $\epsilon$ ] (e.g., \*dr[ $\epsilon$ ]gon; Advanced: 51.03% correct (SD = 50.05); Intermediate: 25.44% (SD = 43.59)) than when the substitution pattern was the opposite (e.g., \*[ $\ae$ ]mon; Advanced: 66.15% correct (SD = 47.37); Intermediate: 47.74% (SD = 49.65)). This is in line with the patterns outlined in previous studies using similar tasks with Dutch (Díaz et al., 2012; Simon, Sjerps & Fikkert, 2014)



**Fig. 1.** Proportion of correct responses (left panel) and reaction times (in milliseconds) for correct responses (right panel) for nonword trials in the lexical decision task as a function of item type (filler items - /ε/-/æ/ items) and learner group (advanced - in black; intermediate - in grey).

Parallel analyses were conducted on reaction times (RT) only for correct nonword trials (87.77% of total nonword trials). Trials with RTs 2 standard deviations shorter or longer than the mean for each participant for each contrast type (filler/critical) were excluded. This resulted in the exclusion of 947 trials (4.53% of correct trials). A linear mixed-effects model (henceforth LMM) was fit with RT as dependent variable and Contrast Type, Proficiency Group and their interaction as predictors. Variables were contrast coded as in the model on accuracy reported above. Random intercepts by Participants and Items and a random slope for Contrast Type over Participants ( $\chi^2(2) = 28.18$ ,  $p < .001$ ) were included as random-effects structure. A random slope for Proficiency Group over Items was not included because it did not improve the model's fit ( $\chi^2(2) = 0.87$ ,  $p = .65$ ). Significance of variables was assessed by means of Satterthwaite's approximation for degrees of freedom using the lmerTest package (Kuznetsova, Brockhoff & Christensen, 2017). As reported in Table 3, the model revealed significant effects of Contrast Type and Proficiency Group, but no significant interaction between the two. Thus, this analysis showed that participants in the advanced group provided faster responses than those in the intermediate group and that RTs were shorter for filler nonwords than for the critical nonwords containing /ε/ and /æ/.

#### Lexical decision: effects of phonetic categorization and vocabulary size

The main analysis of interest was the one assessing the contribution of phonetic categorization ability with /ε/-/æ/ and L2

and German learners of English (Llompert & Reinisch, 2019b). In order to assess the effects of mispronunciation direction, a GLMM was run with Response as dependent variable, Mispronunciation Direction (/ε/ → [æ] vs. /æ/ → [ε]), Proficiency Group and their interaction as predictors, and the most complex random-effects structure justified by log-likelihood comparisons. Results revealed a marginally significant effect of Mispronunciation Direction ( $b = -1.05$ ;  $z = -1.77$ ;  $p = .08$ ) and no interaction between Mispronunciation Direction and Proficiency Group ( $b = 0.13$ ;  $z = 0.36$ ;  $p = .72$ ). This indicates that the effect of mispronunciation direction on nonword rejection was not particularly robust for the present sample and, importantly, that the two groups did not differ in how much their accuracies were affected by the direction of the mispronunciation.

vocabulary size to performance with mispronounced nonwords containing /ε/ and /æ/ in the lexical decision task. Data for lexical decision trials in which /ε/-/æ/ nonwords were presented were submitted to a GLMM with a logistic linking function with Response (0 = incorrect, 1 = correct) as the categorical dependent variable. Phonetic Categorization (i.e., individual categorization slopes) and Vocabulary Size (i.e., vocabulary scores), the interaction between the two, as well as the interactions between Phonetic Categorization and Proficiency Group (Advanced / Intermediate) and between Vocabulary Size and Proficiency Group, were entered as predictors. Phonetic Categorization and Vocabulary Size were entered as continuous variables centered around the mean for all participants. Proficiency Group was contrast coded with Intermediate as -0.5 and Advanced as 0.5. The random-effects structure included random intercepts for Participants and Items, as well as a random slope for Phonetic Categorization over participants. A log-likelihood ratio test confirmed that adding the random slope improved the model's fit ( $\chi^2(2) = 7.06$ ,  $p < .05$ ). A random slope for Proficiency Group over Items was not included because it did not improve the model's fit ( $\chi^2(2) = 0.36$ ,  $p = .83$ ). Finally, a random slope for Vocabulary Size over participants was not included because its inclusion resulted in severe non-convergence issues. Results are shown in Table 4. The model revealed a significant effect of Vocabulary Size and, most importantly, significant interactions between Phonetic Categorization and Proficiency Group, and between Vocabulary Size and Proficiency Group. These interactions indicate that L2 learners' ability to reject nonwords containing /ε/ > [æ] and /æ/ > [ε] mispronunciations related to vocabulary test scores and phonetic categorization slopes differently for the two groups of learners.

In order to follow up on the interactions, data were split by Proficiency Group and two GLMMs were run, one for each group. The models had Response as dependent variable and Phonetic Categorization and Vocabulary Size as predictors. Their random-effects structures included random intercepts for Participants and Items and random slopes for Vocabulary Size over Participants because they improved the fit of the models (Advanced:  $\chi^2(2) = 7.03$ ,  $p < .05$ ; Intermediate:  $\chi^2(2) = 11.55$ ,  $p < .01$ ). In addition, a random slope for Phonetic Categorization

**Table 2.** Results of the GLMM on the effects of Contrast Type (Filler / Critical), Proficiency Group (Advanced / Intermediate) and their interaction on nonword rejection accuracy in the lexical decision task.

Predictor	<i>b</i>	<i>z</i>	<i>p</i>
Intercept	1.25	6.99	< .001
Contrast Type	2.96	9.40	< .001
Proficiency Group	1.12	4.98	< .001
Contrast Type x Proficiency Group	-0.92	-3.17	< .01

**Table 3.** Results of the LMM on the effects of Contrast Type (Filler / Critical), Proficiency Group (Advanced / Intermediate) and their interaction on reaction times (RT) of correct responses in the lexical decision task.

Predictor	<i>b</i>	<i>z</i>	<i>p</i>
Intercept	1619.29	54.20	< .001
Contrast Type	-83.96	-2.57	< .05
Proficiency Group	-118.82	-2.20	< .05
Contrast Type x Proficiency Group	22.59	0.55	.59

over Participants was included for the intermediate participant data because it marginally improved the model's fit ( $\chi^2(2) = 5.36, p = .07$ ). These follow-up analyses showed that, for the advanced group, the effect of Vocabulary Size was significant ( $b = 4.97; z = 2.69; p < .01$ ), whereas that of Phonetic Categorization failed to reach significance ( $b = -1.09; z = -0.99; p = .32$ ). For the intermediate group, the opposite pattern was found: Phonetic Categorization had a significant effect on / $\epsilon$ /-/ $\ae$ / nonword recognition accuracy ( $b = 1.88; z = 2.82; p < .01$ ), while Vocabulary Size did not ( $b = -0.29; z = -0.18; p = .86$ ). These results indicate that, for the intermediate group, participants with steeper categorization slopes were better able to reject the / $\epsilon$ /-/ $\ae$ / nonwords, whereas, for the advanced group, the learners who were more accurate were those with higher vocabulary scores. For a better visualization of how these relationships are instantiated at the level of the individual learner, scatterplots are provided in Figure 2, with the two groups differentiated by color (advanced – in black; intermediate – in grey).

As in the previous section, parallel analyses were conducted on RTs. This time the dataset only included correct responses to / $\epsilon$ /-/ $\ae$ / nonwords. A LMM was fit with RT as dependent variable and Phonetic Categorization, Vocabulary Size, the interaction between the two and the interactions between Phonetic Categorization and Proficiency Group and Vocabulary Size and Proficiency Group as predictors. The random effects structure only included random intercepts for Participants and Items, as random slopes over Participants for Phonetic Categorization ( $\chi^2(5) = 6.29, p = .28$ ) and Vocabulary Size ( $\chi^2(2) = 2.04, p = .36$ ) and over Items for Proficiency Group ( $\chi^2(2) = 0.30, p = .86$ ) all failed to improve the model's fit. The model, whose results are shown in Table 5, rendered a marginally significant effect of Vocabulary Size, while the effect of Phonetic Categorization was not significant and neither were any of the interactions. These results suggest that, overall, participants with higher vocabulary scores were faster at rejecting the nonwords in the trials in which their responses were correct.

**Table 4.** Results of the GLMM on the effects of Phonetic Categorization, Vocabulary Size, their interaction and their respective interactions with Proficiency Group (Advanced / Intermediate) on / $\epsilon$ /-/ $\ae$ / nonword rejection accuracy in the lexical decision task.

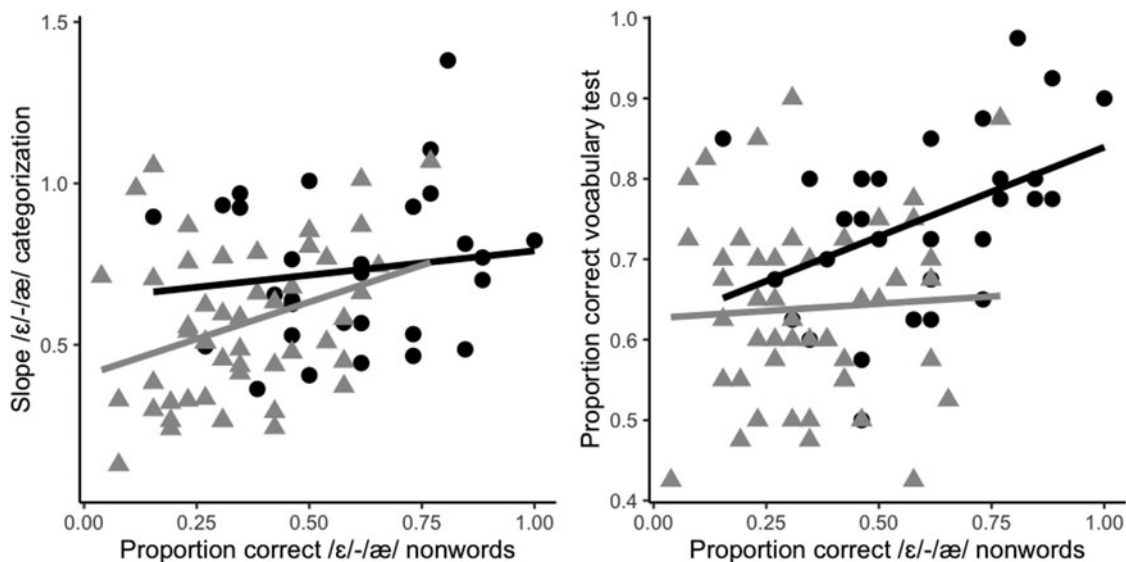
Predictor	<i>b</i>	<i>z</i>	<i>p</i>
Intercept	-0.42	-1.25	.25
Phonetic Categorization	0.98	1.26	.21
Vocabulary Size	4.90	3.60	< .001
Phonetic Categorization x Proficiency Group	-3.34	-2.21	< .05
Vocabulary Size x Proficiency Group	7.42	2.26	< .05
Phonetic Categorization x Vocabulary Size	0.97	0.16	.88

## Discussion

The present study aimed to provide a better understanding of the relationship between the accurate phonolexical encoding of difficult L2 phonological contrasts and the phonetic categorization of the same contrasts, on the one hand, and phonolexical encoding and vocabulary size in the non-native language, on the other hand. Additionally, it was investigated whether these relationships are modulated by L2 proficiency. Individual performances of two groups of German learners of English (advanced vs. intermediate) were examined in a lexical decision task in English including words and mispronounced nonwords with / $\epsilon$ / and / $\ae$ /, a phonetic categorization task on the same vowel contrast and an English vocabulary test. Individual measures were obtained for phonetic categorization and vocabulary size and it was subsequently tested whether these would predict learners' performance in the lexical decision task that assessed the phonolexical encoding of / $\epsilon$ / and / $\ae$ . While, as expected, learners in the advanced group obtained higher scores in all three tasks than those in the intermediate group, results of the lexical decision task confirmed that even the former exhibited considerable difficulties when judging the status of nonwords in which / $\epsilon$ / and / $\ae$ / had been swapped (e.g., \*l[ $\ae$ ]mon, \*dr[ $\epsilon$ ]gon). The most crucial finding was, however, that / $\epsilon$ /-/ $\ae$ / phonetic categorization and vocabulary size related to accuracy in lexical decision differently for the two learner groups. / $\epsilon$ /-/ $\ae$ / nonword rejection was significantly related to how clear-cut the distinction between the two contrastive L2 sounds was in the phonetic categorization task for the intermediate learners but not for the advanced learners, while scores in the vocabulary test related to / $\epsilon$ /-/ $\ae$ / nonword rejection for the advanced but not the intermediate learners.

Similar analyses were conducted on RTs for correct responses to nonword stimuli and, while intermediate learners were also slower to reject / $\epsilon$ /-/ $\ae$ / nonwords than advanced learners, the critical interactions between phonetic categorization and proficiency group and vocabulary size and proficiency group did not arise for RTs. Instead, participants with larger vocabulary sizes were found to be marginally faster to reject / $\epsilon$ /-/ $\ae$ / nonwords across the board. This indicates that participants with larger vocabulary sizes needed less time to access lexical entries and make decisions on their lexical status (see Fernald, Perfors & Marchman, 2006 and Marchman, Fernald & Hurtado, 2010). However, RT analyses only considered correct responses and did not provide information on how accurate the phonological representations of the accessed lexical entries were. For this





**Fig. 2.** Scatterplots showcasing individual values for /ε/-/æ/ phonetic categorization slopes and accuracy in /ε/-/æ/ lexical decision (left panel) and accuracy in vocabulary test and accuracy in /ε/-/æ/ lexical decision (right panel). The black circles correspond to values for the advanced learner group and the grey triangles to values for the intermediate learner group. Regression lines for each group are also provided for illustration purposes.

**Table 5.** Results of the LMM on the effects of Phonetic Categorization, Vocabulary Size, their interaction and their respective interactions with Proficiency Group (Advanced / Intermediate) on reaction times (RT) of correct responses to /ε/- and /æ/-nonwords in the lexical decision task.

Predictor	<i>b</i>	<i>z</i>	<i>p</i>
Intercept	1649.99	41.79	< .001
Phonetic Categorization	211.84	1.30	.20
Vocabulary Size	-557.57	-1.85	.07
Phonetic Categorization x Proficiency Group	82.36	0.25	.81
Vocabulary Size x Proficiency Group	264.13	0.40	.69
Phonetic Categorization x Vocabulary Size	-195.71	-0.17	.87

reason, the remainder of the general discussion is mainly concerned with accuracy in lexical decision, which is also the measure that most previous research on the topic has taken into account (e.g., Amengual, 2016; Díaz et al., 2012; Llompарт & Reinisch, 2019b; Sebastián-Gallés, Vera-Constán, Larsson, Costa & Deco, 2009).

Accuracy results reaffirmed, in the first place, that initial difficulties at the phonetic level with specific L2 phonological contrasts translate into long-lasting lexical deficits that are extremely difficult to overcome (Darcy et al., 2013; Díaz et al., 2012; Hayes-Harb & Masuda, 2008; Llompарт, 2019; Llompарт & Reinisch, 2019a, 2019b). This is supported by the fact that even the highly-proficient late L2 learners in the advanced group made numerous mistakes when judging the lexicality of /ε/- and /æ/-nonwords. This was the case even when, based on their performance on the categorization task, many of them do not appear to have major phonetic difficulties with these phones anymore. This fits well with previous research showing that even early sequential bilinguals, whose L2 proficiency is virtually

native-like in many other domains, are still far from perfect in tasks assessing their phonolexical encoding of difficult L2 contrasts (Amengual, 2016; Sebastián-Gallés & Bosch, 2003; Sebastián-Gallés et al., 2005).

Results for the intermediate learner group also point towards the idea that becoming able to accurately identify the two difficult L2 phones in perception contributes greatly to the improvement of the phonolexical encoding of these phones. For this group, steeper slopes in the phonetic categorization task predicted learners' /ε/-/æ/ nonword rejection accuracy: the steeper the categorization slope, the higher nonword rejection scores. This was the case even though the latter were predominantly rather low (see Figure 2, left panel). By contrast, the same relationship was not found for advanced learners, who, in general, exhibited a more robust phonetic categorization and phonolexical encoding of the critical contrast. Hence, it appears that learners need to reach a certain state where they can already differentiate the two L2 phones with some degree of success. Past that threshold, individual differences in phonetic categorization do not seem to bear a significant impact on learners' success at encoding these L2 phones into phonological representations. These results therefore provide further evidence that phonolexical encoding is not a straightforward reflection of phonetic categorization abilities (Darcy et al., 2013; Llompарт, 2019; Llompарт & Reinisch, 2019a) and suggest that the diverging results of previous studies comparing the two may have been related to differences in L2 proficiency between their learner populations (Darcy & Holliday, 2019; Llompарт & Reinisch, 2019a, Silbert et al., 2015; Simonchyk & Darcy, 2017, 2018).

In the introduction, it was hypothesized that another potential predictor of the encoding of difficult non-native phonological categories into L2 words could be vocabulary size in the non-native language. This hypothesis was based on the premises that i) vocabulary size strongly relates to phonological development in early L1 learning (e.g., Best et al., 2009; Mani & Plunkett, 2010), and ii) diverging predictions have been made regarding this link in the L2 (Best & Tyler, 2007; Bundgaard-Nielsen

et al., 2011a; Llopart, 2019). With regard to L2 learning, Best and Tyler (2007) suggested that vocabulary expansion in the early stages of L2 acquisition might be detrimental for the establishment of native-like phonological representations and may lead to long-lasting fossilization. By contrast, Bundgaard-Nielsen et al. (2011a) posited that vocabulary acquisition could potentially help bootstrap phonological learning in the L2 in a similar way as it does in early L1 learning. Two important findings of the present study align with the latter proposal. First, advanced learners of English, who on average obtained higher vocabulary scores than intermediate learners, showed a more reliable phonolexical encoding of the critical phonological contrast. Secondly and most importantly, for the advanced learner group, /ɛ/-/æ/ nonword rejection results could be predicted by their scores in the vocabulary test. This suggests that vocabulary expansion in the L2 contributes to approximating native phonolexical encoding norms for difficult non-native phonological distinctions (Broersma, 2005; Sebastián-Gallés & Bosch, 2003; Sebastián-Gallés et al., 2005). Results for the advanced learner group mirror thus findings of early L1 learning, where 14-month-olds with larger vocabularies were found to be more successful at learning to distinguish similar-sounding words than children with smaller vocabularies (Mani & Plunkett, 2010; Yoshida et al., 2009; Werker et al., 2002). As hinted at in the introduction, the lack of a relationship between phonolexical encoding and vocabulary size for intermediate learners may be attributed to the fact that, in general, their phonolexical encoding is rather weak and strongly conditioned by individual differences in phonetic categorization.

Crucially, even if limited to one learner population and one L2 phonological contrast, the insights gained here allow for a tentative characterization of the development of the phonetics-to-lexicon mapping over the L2 learning process. In L2 acquisition, phonetic and (phono)lexical learning take place in parallel. Learners need to learn to perceive non-native phonetic distinctions while simultaneously learning L2 words, some of which actually contain the phones of these distinctions. The present findings suggest that these two constantly evolving processes interrelate in a very specific way: while both phonetic abilities and lexical knowledge improve with L2 experience and instruction, they both constrain the process that is responsible to bridge the divide between the two: phonolexical encoding.

Regarding phonetic abilities, it can be proposed that poor phonetic perception prevents feeding lexical processing with the necessary detail to capture challenging phonetic distinctions accurately and encode them into the representations of L2 words (e.g., Amengual, 2016; Darcy et al., 2013). This results in learners being uncertain of which non-native phone is to be assigned to each L2 word that they encounter (Cook & Gor, 2015; Cook et al., 2016; Darcy & Thomas, 2019; Kleinschmidt, Weatherholtz & Jaeger, 2018; Llopart & Reinisch, 2019b). As long as phonetic (i.e., perceptual) difficulties remain, this seems to be the case irrespective of the state of one's lexical knowledge. These difficulties therefore need to be overcome for phonolexical encoding to be in a position to improve substantially.

Nonetheless, mastery of the contrast at the phonetic level is not an automatic guarantee of robust phonolexical encoding, since, once the required phonetic identification abilities are in place, the quality of the phonological encoding into the lexicon appears to be further shaped by lexical knowledge, and, in particular, by one's vocabulary size. An explanation for this is that the aforementioned uncertainty in non-native lexical representations can

only be reduced by accumulating evidence from the input received and learners with larger vocabulary sizes are likely to have learned more words containing, and contrasting, the critical L2 phones. Hence, a reasonable possibility is that the link between the specific phonetic properties of a given phone (e.g., [æ]) and known words including it as part of their phonological representation (e.g., *dr/æ/gon*) strengthens as a function of the number of words in the L2 lexicon that provide evidence of a phonological contrast. Vocabulary expansion may thus provide a beneficial boost to the updating of phonological representations, facilitating that non-native-like phonological forms (e.g., *dr/?/gon*) are substituted by phonological forms closer to native-like (e.g., *dr/æ/gon*) in a dynamic relexification process. By contrast, with smaller vocabularies, the phonetics-to-lexicon mapping is likely to remain weak and unstable because, with a smaller set of lexical entries, evidence pointing towards a phonological contrast between the challenging L2 phones will be more limited. This idea is in line with proposals from early L1 acquisition (Kemp et al., 2017; Metsala, 1999; Thiessen, 2007) and with Bundgaard-Nielsen et al.'s (2011a, 2011b) view of L2 phonological acquisition.

Critically, the relationship between vocabulary size and phonolexical encoding discussed above gives rise to two interesting questions for future research. The first one is to what extent a larger vocabulary size is beneficial simply because it involves having obtained a critical mass of lexical entries with the difficult-to-distinguish L2 phonological categories, and to what extent its contribution is dependent on the lexical properties of the words in the L2 lexicon. This includes frequency (e.g., Gollan, Montoya, Cera & Sandoval, 2008), phonological neighborhood density (Imai, Walley & Flege, 2005; Storkel, 2002) and, relatedly, the number of minimal pairs contrasting the relevant phonological categories (Llopart & Reinisch, 2020). A second question, related to the fact that German and English have relatively similar phonological inventories and share a large part of their vocabulary, is to what extent the relationship between vocabulary size and phonolexical encoding is modulated by the relatedness of the phonological systems and vocabularies of L1 and L2. In principle, relatedness could strengthen as well as loosen the link between the two. On the one hand, the similarities across languages in both phonological restrictions and lexical items could favor the appearance of an observable relationship between the two. On the other hand, as the high similarity between German and English forces learners to adjust their phonology to the L2 while learning words that are in many cases very similar to those in the L1, L1-interference may be larger in this case and make the connection between phonology and vocabulary less straightforward.

To conclude, this study highlights the need for more research on the encoding and representation of challenging L2 phonological distinctions into non-native words. To date, this is an area that remains understudied, especially in comparison with the abundant literature on the evolution of phonetic-level perception and production abilities for this type of contrasts (e.g., Kartushina, Hervais-Adelman, Frauenfelder & Golestani, 2015, 2016; Thorin, Sadakata, Desain & McQueen, 2018). While it is undeniable that improvements in phonetic perception and production are essential for ultimate success in L2 learning, the present results show that the struggles of L2 learners with specific non-native phonological contrasts cannot be fully understood if only phonetic aptitudes are considered. Communication involves words and this is why our focus should also lie on how L2 contrastive sounds are incorporated into these larger, meaningful

speech units. Hence, further research on this topic will be key in order for our field to move closer to drawing a clear picture of how L2 learners effectively master challenging L2 phonological distinctions.

**Acknowledgements.** This project was funded by an Alexander von Humboldt Professorship (ID-1195918) awarded to Ewa Dąbrowska, Chair of Language and Cognition at the Department of English and American Studies of Friedrich Alexander University Erlangen-Nuremberg. I would like to thank Armine Garibyan for her help with data collection and Eva Reinisch and three anonymous reviewers for their thoughtful comments on this work.

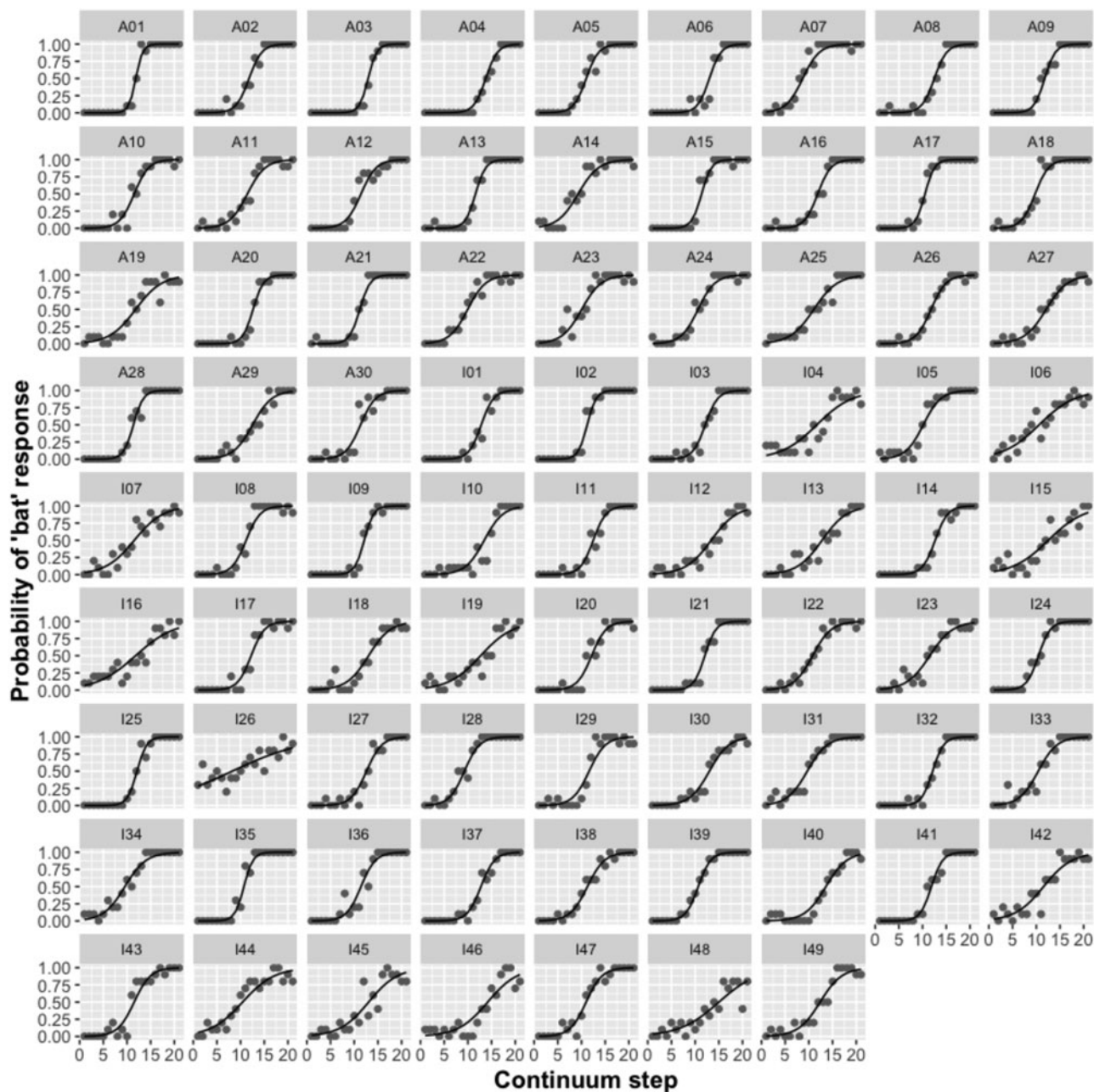
## References

- Amengual M** (2016) The perception of language-specific phonetic categories does not guarantee accurate phonological representations in the lexicon of early bilinguals. *Applied Psycholinguistics* 37, 1221–1251. <https://doi.org/10.1017/s0142716415000557>
- Bates D, Mächler M, Bolker B and Walker S** (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67, 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Best CT and Tyler MD** (2007) Nonnative and second-language speech perception: Commonalities and complementarities. In Bohn OS & Munro MJ (eds), *Language experience in second language speech learning: In honor of James Emil Flege*. Amsterdam: John Benjamins, pp. 13–34. <https://doi.org/10.1075/llt.17.07bes>
- Best CT, Tyler MD, Gooding TN, Orlando CB and Quann CA** (2009) Development of phonological constancy: Toddlers' perception of native and Jamaican-accented words. *Psychological Science* 20, 539–542. <https://doi.org/10.1111/j.1467-9280.2009.02327.x>
- Boersma P and Weenink D** (2010) Praat: doing phonetics by computer. Version 5.4.22, <http://www.praat.org/>.
- Bohn O-S and Bundgaard-Nielsen RL** (2009) Second language speech learning with diverse inputs. In Piske T and Young-Scholten M (eds), *Input Matters in SLA*. Clevedon, UK: Multilingual Matters, pp. 207–218.
- Bohn OS and Flege JE** (1990) Interlingual identification and the role of foreign language experience in L2 vowel perception. *Applied Psycholinguistics* 11, 303–328. <https://doi.org/10.1017/s0142716400008912>
- Bohn OS and Flege JE** (1992) The production of new and similar vowels by adult German learners of English. *Studies in Second Language Acquisition* 14, 131–158. <https://doi.org/10.1017/s0272263100010792>
- Broersma M** (2005) *Phonetic and lexical processing in a second language*. Ph.D. dissertation, Radboud University.
- Bundgaard-Nielsen RL, Best CT, Kroos C and Tyler MD** (2012) Second language learners' vocabulary expansion is associated with improved second language vowel intelligibility. *Applied Psycholinguistics* 33, 643–664. <https://doi.org/10.1017/s0142716411000518>
- Bundgaard-Nielsen RL, Best CT and Tyler MD** (2011a) Vocabulary size is associated with second-language vowel perception performance in adult learners. *Studies in Second Language Acquisition* 33, 433–461. <https://doi.org/10.1017/S0272263111000040>
- Bundgaard-Nielsen RL, Best CT and Tyler MD** (2011b) Vocabulary size matters: The assimilation of second-language Australian English vowels to first-language Japanese vowel categories. *Applied Psycholinguistics* 32, 51–67. <https://doi.org/10.1017/s0142716410000287>
- Cameron L** (2002) Measuring vocabulary size in English as an additional language. *Language Teaching Research* 6, 145–173. <https://doi.org/10.1191/1362168802lr1030a>
- Casillas JV and Simonet M** (2015) Production and perception of the English /æ/–/a/ contrast in switched-dominance speakers. *Second Language Research* 32, 171–195. <https://doi.org/10.1177/0267658315608912>
- Cook SV and Gor K** (2015) Lexical access in L2: Representational deficit or processing constraint? *The Mental Lexicon* 10, 247–270. <https://doi.org/10.1075/ml.10.2.04coo>
- Cook SV, Pandza NB, Lancaster AK and Gor K** (2016) Fuzzy nonnative phonological representations lead to fuzzy form-to-meaning mappings. *Frontiers in Psychology* 7, 1345. <https://doi.org/10.3389/fpsyg.2016.01345>
- Cutler A** (2015) Representation of second language phonology. *Applied Psycholinguistics* 36, 115–128. <https://doi.org/10.1017/s0142716414000459>
- Darcy I, Daidone D and Kojima C** (2013) Asymmetric lexical access and fuzzy lexical representations in second language learners. *The Mental Lexicon* 8, 372–420. <https://doi.org/10.1075/ml.8.3.06dar>
- Darcy I and Holliday JJ** (2019) Teaching an old word new tricks: Phonological updates in the L2 mental lexicon. In Levis J, Nagle C and Todey E (eds), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference*. Ames, IA: Iowa State University, pp. 10–26.
- Darcy I, Park H and Yang C-L** (2015) Individual differences in L2 acquisition of English phonology: The relation between cognitive abilities and phonological processing. *Learning and Individual Differences* 40, 63–72. <https://doi.org/10.1016/j.lindif.2015.04.005>
- Darcy I and Thomas T** (2019) When blue is a disyllabic word: Perceptual epenthesis in the mental lexicon of second language learners. *Bilingualism: Language and Cognition* 22, 1141–1159. <https://doi.org/10.1017/s1366728918001050>
- Díaz B, Mitterer H, Broersma M and Sebastián-Gallés N** (2012) Individual differences in late bilinguals' L2 phonological processes: From acoustic-phonetic analysis to lexical access. *Learning and Individual Differences* 22, 680–689. <https://doi.org/10.1016/j.lindif.2012.05.005>
- Eger NA and Reinisch E** (2019a) The impact of one's own voice and production skills on word recognition in a second language. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 45, 552–571. <https://doi.org/10.1037/xlm0000599>
- Eger NA and Reinisch E** (2019b) The role of acoustic cues and listener proficiency in the perception of accent in nonnative sounds. *Studies in Second Language Acquisition* 41, 179–200. <https://doi.org/10.1017/s0272263117000377>
- Escudero P** (2005) *Linguistic Perception and Second Language Acquisition: Explaining the Attainment of Optimal Phonological Categorization*. Ph.D. dissertation, Utrecht University.
- Escudero P and Boersma P** (2004) Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition* 26, 551–585. <https://doi.org/10.1017/s0272263104040021>
- Escudero P, Simon E and Mulak KE** (2014) Learning words in a new language: Orthography doesn't always help. *Bilingualism: Language and Cognition* 17, 384–395. <https://doi.org/10.1017/S1366728913000436>
- Escudero P and Wanrooij K** (2010) The effect of L1 orthography on non-native vowel perception. *Language and Speech* 53, 343–365. <https://doi.org/10.1177/0023830910371447>
- Fernald A, Perfors A and Marchman VA** (2006) Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Developmental Psychology* 42, 98–116. <https://doi.org/10.1037/0012-1649.42.1.98>
- Flege JE** (1995) Second language speech learning: Theory, findings, and problems. In Strange W (ed), *Speech perception and linguistic experience: Issues in cross-language research*. Timonium, MD: York Press, pp. 233–277.
- Flege JE, Bohn OS and Jang S** (1997) Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics* 25, 437–470. <https://doi.org/10.1006/jpho.1997.0052>
- Goldinger SD** (1998) Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105, 251–279. <https://doi.org/10.1037/0033-295X.105.2.251>
- Gollan TH, Montoya RI, Cera C and Sandoval TC** (2008) More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language* 58, 787–814. <https://doi.org/10.1016/j.jml.2007.07.001>
- Goto H** (1971) Auditory perception by normal Japanese adults of the sounds “L” and “R.” *Neuropsychologia* 9, 317–323. [https://doi.org/10.1016/0028-3932\(71\)90027-3](https://doi.org/10.1016/0028-3932(71)90027-3)
- Hayes-Harb R and Barrios S** (2019) Investigating the phonological content of learners' “fuzzy” lexical representation for new L2 words. In Levis J, Nagle C and Todey E (eds), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching conference*. Ames, IA: Iowa State University, pp. 55–69.
- Hayes-Harb R and Masuda K** (2008) Development of the ability to lexically encode novel second language phonemic contrasts. *Second Language Research* 24, 5–33. <https://doi.org/10.1177/0267658307082980>

- Imai S, Walley AC and Flege JE (2005) Lexical frequency and neighborhood density effects on the recognition of native and Spanish-accented words by native English and Spanish listeners. *The Journal of the Acoustical Society of America* 117, 896–907. doi:10.1121/1.1823291
- Kartushina N, Hervais-Adelman A, Frauenfelder UH and Golestani N (2015) The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *The Journal of the Acoustical Society of America* 138, 817–832. <https://doi.org/10.1121/1.4926561>
- Kartushina N, Hervais-Adelman A, Frauenfelder UH and Golestani N (2016) Mutual influences between native and non-native vowels in production: Evidence from short-term visual articulatory feedback training. *Journal of Phonetics* 57, 21–39. <https://doi.org/10.1016/j.wocn.2016.05.001>
- Kaya F, Delen E and Bulut O (2012) Test review: Shipley-2 manual. *Journal of Psychoeducational Assessment* 30, 593–597. <https://doi.org/10.1177/0734282912440852>
- Kemp N, Scott J, Bernhardt BM, Johnson CE, Siegel LS and Werker JF (2017) Minimal pair word learning and vocabulary size: Links with later language skills. *Applied Psycholinguistics* 38, 289–314. <https://doi.org/10.1017/S0142716416000199>
- Kim D, Clayards M and Goad H (2017) Individual differences in second language speech perception across tasks and contrasts: The case of English vowel contrasts by Korean learners. *Linguistics Vanguard* 3. <https://doi.org/10.1515/lingvan-2016-0025>
- Kleinschmidt DF, Weatherholtz K and Jaeger TF (2018) Sociolinguistic perception as inference under uncertainty. *Topics in Cognitive Science* 10, 818–834. <https://doi.org/10.1111/tops.12331>
- Kuhl P, Williams K, Lacerda F, Stevens K and Lindblom B (1992) Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* 255, 606–608. <https://doi.org/10.1126/science.1736364>
- Kuznetsova A, Brockhoff PB and Christensen RHB (2017) lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software* 82, 1–26. <https://doi.org/10.18637/jss.v082.i13>
- Llompart M (2019) *Bridging the gap between phonetic abilities and the lexicon in second language learning*. Ph.D. dissertation, Ludwig Maximilian University Munich.
- Llompart M and Reinisch E (2017) Articulatory information helps encode lexical contrasts in a second language. *Journal of Experimental Psychology: Human Perception and Performance* 43, 1040–1056. <https://doi.org/10.1037/xhp0000383>
- Llompart M and Reinisch E (2018) Acoustic cues, not phonological features, drive vowel perception: Evidence from height, position and tenseness contrasts in German vowels. *Journal of Phonetics* 67, 34–48. <https://doi.org/10.1016/j.wocn.2017.12.001>
- Llompart M and Reinisch E (2019a) Imitation in a second language relies on phonological categories but does not reflect the productive usage of difficult sound contrasts. *Language and Speech* 62, 594–622. <https://doi.org/10.1177/0023830918803978>
- Llompart M and Reinisch E (2019b) Robustness of phonolexical representations relates to phonetic flexibility for difficult second language sound contrasts. *Bilingualism: Language and Cognition* 22, 1085–1100. <https://doi.org/10.1017/s1366728918000925>
- Llompart M and Reinisch E (2020) The phonological form of lexical items modulates the encoding of challenging second-language sound contrasts. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 46, 1590–1610. <https://doi.org/10.1037/xlm0000832>
- Mani N and Plunkett K (2010) Twelve-month-olds know their *cups* from their *keps* and *tups*. *Infancy* 15, 445–470. <https://doi.org/10.1111/j.1532-7078.2009.00027.x>
- Marchman VA, Fernald A and Hurtado N (2010) How vocabulary size in two languages relates to efficiency in spoken word recognition by young Spanish–English bilinguals. *Journal of Child Language* 37, 817–840. <https://doi.org/10.1017/s0305000909990055>
- Marian V, Bartolotti J, Chabal S and Shook A (2012) CLEARPOND: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PLoS ONE* 7, e43230. <https://doi.org/10.1371/journal.pone.0043230>
- McQueen JM, Cutler A and Norris D (2003) Flow of information in the spoken word recognition system. *Speech Communication* 41, 257–270. [https://doi.org/10.1016/s0167-6393\(02\)00108-5](https://doi.org/10.1016/s0167-6393(02)00108-5)
- Metsala JL (1999) Young children’s phonological awareness and nonword repetition as a function of vocabulary development. *Journal of Educational Psychology* 91, 3–19. <https://doi.org/10.1037/0022-0663.91.1.3>
- Mitterer H, Reinisch E and McQueen JM (2018) Allophones, not phonemes in spoken-word recognition. *Journal of Memory and Language* 98, 77–92. <https://doi.org/10.1016/j.jml.2017.09.005>
- Mulak KE, Best CT, Tyler MD, Kitamura C and Irwin JR (2013) Development of phonological constancy: 19-month-olds, but not 15-month-olds, identify words in a non-native regional accent. *Child Development* 84, 2064–2078. <https://doi.org/10.1111/cdev.12087>
- Nation ISP (2006) How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review* 63, 59–82. <https://doi.org/10.3138/cmlr.63.1.59>
- Pajak B, Creel SC and Levy R (2016) Difficulty in learning similar-sounding words: A developmental stage or a general property of learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 42, 1377–1399. <https://doi.org/10.1037/xlm0000247>
- Pallier C, Bosch L and Sebastián-Gallés N (1997) A limit on behavioral plasticity in speech perception. *Cognition* 64, B9–B17. [https://doi.org/10.1016/s0010-0277\(97\)00030-9](https://doi.org/10.1016/s0010-0277(97)00030-9)
- Peirce JW, Gray JR, Simpson S, MacAskill MR, Höchenberger R, Sogo H, Kastman E and Lindeløv J (2019) Psychopy2: experiments in behavior made easy. *Behavior Research Methods*. 51, 195–203. <https://doi.org/10.3758/s13428-018-01193-y>
- Polka L and Werker JF (1994) Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance* 20, 421–435. <https://doi.org/10.1037/0096-1523.20.2.421>
- R Core Team (2017) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.r-project.org/>.
- Raudenbush SW, Yang M-L and Yosef M (2000) Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics* 9, 141–157. <https://doi.org/10.1080/10618600.2000.10474870>
- Sebastián-Gallés N and Bosch L (2003) The representation of native and non-native lexical items in early bilinguals. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 2885–2888).
- Sebastián-Gallés N, Echeverría S and Bosch L (2005) The influence of initial exposure on lexical representation: Comparing early and simultaneous bilinguals. *Journal of Memory and Language* 52, 240–255. <https://doi.org/10.1016/j.jml.2004.11.001>
- Sebastián-Gallés N, Vera-Constán F, Larsson JP, Costa A and Deco G (2009) Lexical plasticity in early bilinguals does not alter phoneme categories: II. Experimental evidence. *Journal of Cognitive Neuroscience* 21, 2343–2357. <https://doi.org/10.1162/jocn.2008.21152>
- Shipley WC, Gruber CP, Martin TA and Klein AM (2009) *Shipley-2 manual*. Los Angeles, CA: Western Psychological Services.
- Silbert NH, Smith BK, Jackson SR, Campbell SG, Hughes MM and Tare M (2015) Non-native phonemic discrimination, phonological short term memory, and word learning. *Journal of Phonetics* 50, 99–119. <https://doi.org/10.1016/j.wocn.2015.03.001>
- Simon E, Sjerps M and Fikkert P (2014) Phonological representations in children’s native and non-native lexicon. *Bilingualism: Language and Cognition* 17, 3–21. <https://doi.org/10.1017/S1366728912000764>
- Simonchik A and Darcy I (2017) Lexical encoding and perception of palatalized consonants in L2 Russian. In O’Brien M & Levis J (eds), *Proceedings of the 8th Pronunciation in Second Language Learning and Teaching Conference* (pp. 121–132). Ames, IA: Iowa State University.
- Simonchik A and Darcy I (2018) The effect of orthography on the lexical encoding of palatalized consonants in L2 Russian. *Language and Speech* 61, 522–546. <https://doi.org/10.1177/0023830918761490>
- Storkel HL (2002) Restructuring of similarity neighbourhoods in the developing mental lexicon. *Journal of Child Language* 29, 251–274. doi:10.1017/s0305000902005032

- Thiessen ED** (2007) The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language* **56**, 16–34. <https://doi.org/10.1016/j.jml.2006.07.002>
- Thorin J, Sadakata M, Desain P and McQueen JM** (2018) Perception and production in interaction during non-native speech category learning. *The Journal of the Acoustical Society of America* **144**, 92–103. <https://doi.org/10.1121/1.5044415>
- Tomoschuk B, Ferreira VS and Gollan TH** (2019) When a seven is not a seven: Self-ratings of bilingual language proficiency differ between and within language populations. *Bilingualism: Language and Cognition* **22**, 516–536. <https://doi.org/10.1017/s1366728918000421>
- van Heuven WJB, Mandera P, Keuleers E and Brysbaert M** (2014) SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology* **67**, 1176–1190. <https://doi.org/10.1080/17470218.2013.850521>
- Welch BL** (1947) The generalization of 'student's' problem when several different population variances are involved. *Biometrika* **34**, 28–35. <https://doi.org/10.1093/biomet/34.1-2.28>
- Werker JF, Fennell CT, Corcoran KM and Stager CL** (2002) Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy* **3**, 1–30. [https://doi.org/10.1207/S15327078IN0301\\_1](https://doi.org/10.1207/S15327078IN0301_1)
- Werker JF and Tees RC** (1984) Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development* **7**, 49–63. [https://doi.org/10.1016/s0163-6383\(84\)80022-3](https://doi.org/10.1016/s0163-6383(84)80022-3)
- Wesche M and Paribakht TS** (1996) Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review* **53**, 13–40. <https://doi.org/10.3138/cmlr.53.1.13>
- Yoshida KA, Fennell CT, Swingley D and Werker JF** (2009) Fourteen-month-old infants learn similar-sounding words. *Developmental Science* **12**, 412–418. <https://doi.org/10.1111/j.1467-7687.2008.00789.x>

## Appendix A



**Fig. A1.** Individual response patterns for the phonetic categorization task (A = Advanced group; I = Intermediate group). The grey dots indicate the predicted probability of a 'bat' response for each continuum step (0 = most *bet*-like step; 20 = most *bat*-like step). Black lines depict the fitted categorization curve for each participant.

## Appendix B

**Appendix B** List of /ɛ/- and /æ/-items including lexical status (word / nonword), lexical frequency, neighborhood size and cognate status.

Word	Vowel	Status	Frequency	NS	Cognate
Centre	ɛ	word	5.19	15	Yes
Check	ɛ	word	5.1	20	No
Chest	ɛ	word	4.48	20	No
Death	ɛ	word	5.14	15	No
Dentist	ɛ	word	3.76	1	No
Desert	ɛ	word	4.3	0	No
Helmet	ɛ	word	4.08	1	Yes
Helpful	ɛ	word	4.21	0	No
Lecture	ɛ	word	3.74	2	No
Record	ɛ	word	5.11	3	Yes
Rest	ɛ	word	5.28	29	Yes
Smell	ɛ	word	4.84	12	No
Special	ɛ	word	5.32	4	Yes
Cherry	ɛ	nonword	4.23	11	No
Chess	ɛ	nonword	3.91	18	No
Desk	ɛ	nonword	4.34	6	No
Dress	ɛ	nonword	4.85	9	No
Fresh	ɛ	nonword	4.89	4	Yes
Health	ɛ	nonword	5.14	9	No
Help	ɛ	nonword	5.75	12	Yes
Legend	ɛ	nonword	4.34	1	Yes
Lemon	ɛ	nonword	4.46	5	No
Lesson	ɛ	nonword	4.46	5	Yes
Sentence	ɛ	nonword	4.34	1	No
Weather	ɛ	nonword	5.12	12	Yes
Yellow	ɛ	nonword	4.84	10	No
Champion	æ	word	4.81	1	No
Damage	æ	word	4.81	1	No
Fact	æ	word	5.45	21	No
Fashion	æ	word	4.63	7	No
Hammer	æ	word	4.66	15	Yes
Hand	æ	word	5.44	24	Yes
Happy	æ	word	5.56	10	No
Language	æ	word	4.82	0	No
National	æ	word	5.2	2	Yes
Saturday	æ	word	4.83	1	No
Shadow	æ	word	4.49	6	Yes
Smash	æ	word	4.08	6	No
Snack	æ	word	3.82	13	No
Channel	æ	nonword	4.51	6	Yes
Dragon	æ	nonword	4.28	1	Yes

(Continued)

**Appendix B** (Continued.)

Word	Vowel	Status	Frequency	NS	Cognate
Factor	æ	nonword	4.5	7	Yes
Factory	æ	nonword	4.59	0	No
Flag	æ	nonword	4.42	11	Yes
Gallery	æ	nonword	4.28	2	Yes
Habit	æ	nonword	4.04	4	No
Lamp	æ	nonword	4.09	14	Yes
Ramp	æ	nonword	3.67	16	Yes
Scandal	æ	nonword	4.22	3	Yes
Splash	æ	nonword	4.13	3	No
Standard	æ	nonword	4.72	2	Yes
Thank	æ	nonword	5.87	11	Yes