Author accepted manuscript

Asymmetric lexical access and fuzzy lexical representations in second language learners

Isabelle Darcy, Danielle Daidone, and Chisato Kojima
Indiana University

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

Darcy, I., Daidone, D., & Kojima, C. (2015). Asymmetric lexical access and fuzzy lexical representations in second language learners. In G. Jarema & G. Libben (Eds.), *Phonological and phonetic considerations of lexical processing* (pp. 119-167). Philadelphia, PA: John Benjamins. http://dx.doi.org/10.1075/bct.80.06dar

Abstract

For L2-learners, confusable phonemic categories lead to ambiguous lexical representations. Yet, learners can establish separate lexical representations for confusable categories, as shown by asymmetric patterns of lexical access, but the source of this asymmetry is not clear (Cutler et al., 2006). Two hypotheses compete, situating its source either at the lexical coding level or at the phonetic categorization level. The lexical coding hypothesis suggests that learners' encoding of an unfamiliar category is not target-like but makes reference to a familiar L1 category (encoded as a poor exemplar of that L1 category).

Four experiments examined how learners lexically encode confusable phonemic categories. American English learners of Japanese and of German were tested on phonetic categorization and lexical decision for geminate/singleton contrasts and front/back rounded vowel contrasts.

Results showed the same asymmetrical patterns as Cutler et al.' s (2006), indicating that learners encode a lexical *distinction* between difficult categories. Results also clarify that the source of the asymmetry is located at the lexical coding level and does not emerge during input categorization: the distinction is not target-like, and makes reference to L1 categories. We further provide new evidence that asymmetries can be resolved over time: advanced learners are establishing more native-like lexical representations.

Keywords

lexical representations in a second language; Japanese; German; phonetic categorization; lexical encoding;

1. Introduction

In recent years, there has been a growing interest in understanding the way spoken words are stored and accessed in the mental lexicon of bilingual language users (e.g. Broersma, 2012; Cutler, Weber & Otake, 2006; Darcy, Dekydtspotter, Sprouse, et al., 2012; Dupoux, Sebastián-Gallés, Navarrete & Peperkamp, 2008; Ju & Luce, 2004; Marian & Spivey, 2003; Ota, Hartsuiker, & Haywood, 2009; Pallier, Colomé, & Sebastián-Gallés, 2001; Sebastián-Gallés, Rodríguez-Fornells, de Diego-Balaguer, & Díaz, 2006; Spivey & Marian, 1999; Weber & Cutler, 2004).

When recognizing words spoken in their native language, adult listeners activate multiple word candidates (i.e. *a cohort*) that match at least part of the acoustic input. As the input unfolds, the candidate that offers the best match wins the competition (Marslen-Wilson, 1987) and is selected. This process is usually fast and largely error-free. Under normal circumstances, native listeners' perception of the acoustic input is rather faithful, and activation of candidates is straightforward because lexical representations are accurate. Native listeners are also able to exploit phonetic detail bottom-up during on-line word recognition to reduce the number of activated competitors, and make the competition process more selective (Ju & Luce, 2006; McMurray, Tanenhaus & Aslin, 2002; Sumner & Samuel, 2009). Conversely, ambiguous input will lead to a larger number of activated competitors (e.g. Gaskell & Marslen-Wilson, 2001), and slow down recognition.

Recognizing words in a second language (L2) is a much more complicated affair because lexical items from both languages can enter lexical competition under certain conditions. There is strong evidence that bilinguals experience parallel competition from all their languages even in a monolingual mode, both in the auditory modality (Ju & Luce, 2004; Marian & Spivey, 2003; Spivey & Marian, 1999; Weber & Cutler, 2004) and in the visual modality (Dijkstra & van Heuven, 1998; Jared & Szucs, 2002). For instance, in an eye-tracking paradigm, when instructed to click on the picture of a marker, Russian-English bilinguals often looked briefly at the picture of a stamp as well, because its Russian name 'marka' phonologically overlaps with the English word *marker* (Spivey & Marian, 1999; Marian & Spivey, 2003). This effect indicates that the Russian word was at least temporarily part of the cohort. To complicate matters further, not only is the number of competitors to choose from larger, but L2 listeners' perception of the spoken input is often unreliable (see Strange and Shafer, 2008, and Sebastián-Gallés, 2005, for reviews). If the input is less reliable, it is also less efficient in constraining the number of competitors activated (Broersma, 2012; Broersma & Cutler, 2011; Dupoux et al., 2008; Weber & Cutler, 2004). For example,

Broersma and Cutler (2011) found in a cross-modal priming study that Dutch listeners' difficulty with the English /ɛ-æ/ contrast led them to activate real words like *deaf* [dɛf] and *lamp* [læmp] more often than native English listeners upon hearing near-word fragments such as [dæf] and [lɛmp] extracted from the words DAFfodil and eviL EMPire.

A further problem is that non-native listeners cannot always rely on fully accurate lexical representations in the first place. The goal of this study is to evaluate whether the less efficient word recognition often observed in L2 learners is due to inaccurate input perception or to fuzzy lexical representations. We now turn to the previous studies more specifically examining the form of lexical representations as the basis for our study.

1.1. Background

One of the potential reasons for L2 learners' less efficient word recognition is that L2 learners do not always encode new, L2-specific contrasts accurately for lexical processing (Darcy, Dekydtspotter, Sprouse, et al., 2012; Dupoux et al., 2008; Ota, Hartsuiker, & Haywood, 2009; Pallier, Colomé, & Sebastián-Gallés, 2001; Sebastián-Gallés, Rodríguez-Fornells, de Diego-Balaguer, & Díaz, 2006). One possible explanation for less accurate encoding is that confusable L2 phonemic categories lead to imprecise, or fuzzy, lexical representations, and/or less effective mismatch from competitors during on-line lexical access (Broersma, 2012; Ota et al., 2009). For example, it is not rare for Japanese learners of English to confuse words such as *rock* and *lock*, an observation explained by the difficulty that Japanese native speakers have in distinguishing these two English sounds /l/ vs. /r/. As a result, these two words might be either perceived as one another, or encoded in long-term representations as homophones, or both.

Previous studies have shown that L2 learners exhibit repetition priming for minimal pairs in lexical decision, where no such priming was found for native listeners (Darcy et al., 2012; Pallier et al., 2001). Similarly, L2 learners experienced false-alarm recognition of non-words as words (Broersma & Cutler, 2008; Sebastian-Gallés, Echeverría & Bosch, 2005). In online tasks, abstract encoding of certain non-native dimensions "can remain unavailable even to advanced learners of a second language" (Dupoux et al. 2008, p. 699). These data (see also Trofimovitch & John, 2011) suggest that learners might have conflated, fragmentary and imprecise lexical representations for words of their second language. Taken together, these findings suggest that L2 learners' lexical encoding and/or lexical access does not always make use of the necessary linguistic dimensions, and is therefore less efficient than for native listeners.

For instance, Pallier, Bosch and Sebastián-Gallés (1997), observing unreliable discrimination of Catalan words containing /ɛ/ from minimally different words containing /e/ in Spanish-dominant early bilinguals, argue that these bilinguals did not generally establish a new category for Catalan /ɛ/, despite exposure from an early age. Pallier et al. (2001) then observed repetition priming for Catalan /e/-/ɛ/ minimal pairs, which they interpreted as evidence that Spanish-dominant bilinguals treated such word pairs as homophones (that is, lexical entries are merged), unlike Catalan native speakers. Even though the lexical homophony interpretation is not the only one possible (the same results could be due to the listeners' inability to auditorily *distinguish* the minimal pairs in the first place), these authors attribute the lack of lexical distinction to the bilinguals not having established a distinct phonological category for the Catalan-specific vowel.

However, other researchers have also found cases where L2 learners seem to have separate lexical entries for minimal pairs despite "inadequacies in phonetic perception" (Cutler, Weber & Otake, 2006, p. 280; see also Escudero, Hayes-Harb & Mitterer, 2008; Hayes-Harb & Masuda, 2008; Weber & Cutler 2004). Weber and Cutler (2004) and Cutler et al. (2006) examined pairs of L2 sounds that are both mapped onto one L1 category (and are therefore confusable). They suggest that perceptual mappings, based on acoustic phonetic similarity between the L2 sounds and the corresponding L1 category, establish the most similar one as the dominant category. In the case of English /ɛ/ and /æ/, both mapping onto Dutch /ɛ/, the dominant category is English /ɛ/ because it exhibits a greater acoustic phonetic similarity to Dutch /ɛ/ than English /æ/. In Weber and Cutler's (2004) eye-tracking study, they asked native English listeners and Dutch-English bilinguals to click on pictures of words that they heard, while their eye-movements were recorded (visual world paradigm). For example, participants were instructed to click on the picture of a target word (e.g. *panda*, /pænda/), while a phonetically confusable competitor (e.g. *pencil*, /pɛnsɪl/) was also displayed on the screen. For Dutch listeners, /ɛ/ and /æ/ being confusable categories, the first syllable of the auditory target was ambiguous with the first syllable of the competitor (both perceived as /pɛn…/). For native listeners, looks to the competitor were suppressed very early upon hearing the first vowel as the auditory input unfolded. By contrast, Dutch listeners looked longer at the "pencil" during the first syllable, before selecting "panda" upon hearing the second syllable. This finding indicates that for the Dutch, the initial syllable was ambiguous and not as efficient in constraining lexical activation as it was for the native listeners. However, this ambiguity was not visible in Dutch listeners when the target was "pencil" and the competitor "panda". Hence, fixations for targets containing the dominant (e.g. more

similar) L2 category /ɛ/ (/pɛnsɪl/) were more selective (earlier rejection[1] of a competitor syllable with the new L2 category /æ/, [pæn]) than fixations for targets containing the new category (/pænda/, which induced later rejection of a competitor syllable with the dominant category, [pɛn]).

In a study using a task similar to Weber & Cutler's, Cutler et al. (2006) examined the categories /l-r/, which are famously difficult for Japanese-English bilinguals. Participants were instructed auditorily to click on the picture of a target word (e.g. a "rocket"), while a phonetically confusable competitor (e.g. "locker") was also displayed on the screen. Just as in Weber & Cutler's findings, results showed that a target with a new, non-dominant category /r/ ("rocket") induced looks to a picture of a "locker" in the bilingual group, but not in the native listeners. However, when the target contained the dominant category /l/ ("locker"), it did not induce Japanese-English bilinguals to look at the competitor picture of a "rocket".

This effect parallels the findings for Dutch-English bilinguals (for whom /ɛ/ is dominant). When the target was "panda" (non-dominant category /æ/), they looked at the "pencil", but not vice versa.

So in sum, hearing a non-dominant category produces a less selective lexical activation pattern than hearing a dominant category. This finding is important in showing that the two categories in question are not fully merged in lexical representations, and that the contrast is somehow preserved. If the contrasts had been fully merged in lexical representations of their participants, fixations would have been symmetrically (non)selective.

An additional question arises from these reports that L2 learners maintain a distinction at the lexical level between lexical representations containing difficult contrasts. While having separate lexical entries and demonstrating lexical competence is crucial for learners to avoid confusion of minimally different word pairs such as *rock* and *lock*, this does not automatically imply that the coding of this difference in lexical representations is accurate and target-like. To take an example, Darcy et al. (2012) report that L2 French learners did not show repetition priming for minimal pairs differing in vowels that are difficult to discriminate (/ɔ - œ/). The finding has been interpreted as evidence that L2 French learners have achieved separate lexical representations for words containing these categories. Yet, these findings do not tease apart whether L2 learners have established a target-like phonological encoding in the lexical representation for each, or whether the phonological forms are lexically separated but not (yet) target-like. Repetition priming does not allow us to distinguish between these hypotheses. Findings

[1] The use of the word "rejection" here only refers to the fact that participants look away from the competitor earlier when the competitor contains the new category – while the ambiguity persists longer when the competitor contains a similar category.

of asymmetries in lexical processing are very important because they suggest that lexical representations might be separated and yet still not target-like. Such a possibility is explicitly mentioned by Cutler et al., (2006) and Hayes-Harb and Masuda (2008). In fact, native speakers of English, for whom representations are expected to be target-like, did not show the asymmetry observed in the L2 learners in Cutler et al. or in Weber and Cutler (2004).

Cutler et al. leave open the question of how the phonological distinction is coded in the lexical representation. While the dominant category is accurately represented at the lexical level, the representation of the non-dominant category is uncertain and compatible with two possible scenarios.

According to the first, the L2 distinction may be accurately represented lexically, that is, in some distinct form for /l/ and /r/ (perhaps due to explicit instruction), but the phonetic input is always misinterpreted as the dominant category (/l/). That this scenario is possible has been shown many times: L2 learners do not always perceive the input faithfully (Sebastián-Gallés, 2005). In this case, even though learners' lexical representations correctly contain the non-dominant category (e.g. /r/), these would not receive positive activation from that stretch of spoken input because the percept would always be [l] (hence the asymmetry, due to temporary ambiguity of the spoken input). Even though listeners misperceive the input and temporarily activate inappropriate lexical representations, this is usually resolved as soon as more input becomes available (e.g. Frauenfelder, Scholten & Content, 2001) – it would only be a problem in the case of minimal pairs such as *lock* and *rock*, where both pairs might in fact be simultaneously activated (see Gaskell & Marslen-Wilson, 2001, for supporting evidence that this is the case even in native speakers, in case of acoustic/phonetic ambiguity).

According to the second scenario, phonetic perception of the spoken input is accurate, but it is the coding of the phonological distinction in the lexical representation itself that is not target-like and makes reference to the L1 category (for instance /l/ vs. /poor l/). The asymmetry then could arise from this difference in coding: although dominant and non-dominant categories are encoded separately (hence, lexical separation is achieved), the non-dominant category is perhaps represented as a poor match of the dominant category. Accordingly, hearing the non-dominant category in the input would activate both /l/ and /poor l/, whereas hearing a familiar (dominant) category /l/ would only specifically activate /l/-containing representations. Just as we explained above, hearing a non-dominant category like /poor l/ produces a less selective lexical activation pattern (activating both) than hearing the dominant category.

It thus appears that the distinction between these two possibilities hinges on the level of processing at which L2 learners are able to successfully

represent the contrast. We developed the following predictions which form the basis of our approach. If the contrast is discriminated correctly during phonetic processing, the asymmetry would arise from a difficulty that is therefore likely located at the lexical coding level (which we term the *lexical coding deficiency* hypothesis). If however phonetic processing indicates that the input containing both categories is interpreted as the dominant category (which we term the *phonetic coding deficiency* hypothesis), then it is possible that lexical representations are accurate, but the ones containing the non-dominant category do not receive (sufficient) activation from auditory input. We develop more specific predictions in the next sections.

1.2. Preliminary experimental considerations

The two hypotheses described above (the *phonetic coding deficiency* vs. the *lexical coding deficiency*) allow mapping out the patterns of lexical decision (see Figures 1a and 1b). The basis for auditory lexical decision is to compare the incoming input (stimulus) to stored phonological representations for words. To take a concrete example from German, the only way to correctly reject a non-word which is a potential word (*König* [kønɪç] 'king' is a word, whereas **Hönig* [hønɪç] is not, it is derived from *Honig* [honɪç] 'honey') is to have an accurate phonological representation of existing words at the lexical level. Therefore, lexical decision patterns will reveal the accuracy of lexical representations. We use the example of two German words *Honig* /honɪç/ 'honey' and *König* /kønɪç/ 'king', and only consider the case where a lexical contrast is present[2], through explicit instruction or other ways. This is likely to be the case for our learners, who all have been exposed to classroom instruction and orthographic forms (see Escudero et al., 2008).

In parallel to the reaction times advantage for words over non-words commonly observed in native language lexical decision performance (e.g. Forster & Chambers, 1973), it is expected that accuracy will generally be higher for words than for non-words, given that it is easier to accept a form as a word than to reject it as a non-word (all non-words used in this study are well-formed and possible words). Participants will tend to respond "yes" more easily than to respond "no".

[2] The full range of possibilities would be four combinations of perceptual categorization and lexical contrast: *absence* of lexical contrast, with or without accurate perceptual distinction, vs. *presence* of lexical contrast, with or without accurate perceptual distinction. However, we only consider the case when a lexical contrast is present because the two hypotheses do not apply to the case when a lexical contrast is absent. In addition, the stimuli used in the lexical decision experiments reported here do not use lexical minimal pairs with which the absence of lexical contrast could reliably be tested, but only pairs of word/non-word.

According to the *phonetic coding deficiency hypothesis* (see Fig. 1a), we assume that listeners have difficulty correctly perceiving the input. Therefore, in the case of German /o/ and /ø/ or /u/ and /y/, both back and front rounded vowels are perceived as back vowels (/o/, /u/). The lexical representations, by contrast, accurately encode the distinction, even though it is not necessarily the case that the phonological form is target-like. It can simply be different from the dominant category (e.g. “not /o/”), without making reference to it. We represent this in Figure 1a as /X/.

In the case shown in Figure 1a, the word/non-word pairs present in the input will not be perceived as different: both inputs [honɪç] and *[hønɪç] are perceived as [honɪç] (as shown in Figure 1a where the percept sometimes differs from the input). Both members of a pair will thus be compared to existent lexical representations in the same way. Lexical representations containing the dominant (“old”) category will be contacted by both words and non-words equally because the percept matches the lexical representation. In the case of lexical representations containing the non-dominant (“new”) category (e.g. /kXnɪç/), the percepts will in each case be a mismatch, and will not contact the lexical representation that contains the non-dominant category, as suggested by Cutler et al. (2006).

lexical representation	/honɪç/		/kXnɪç/	
	match	match	mismatch	mismatch
percept	[honɪç]	*[honɪç]	[konɪç]	*[konɪç]
input	[honɪç]	*[hønɪç]	[kønɪç]	*[konɪç]
expected response	yes	no	yes	no
prediction	easy to accept	difficult to reject	difficult to accept	easy to reject
ordinal accuracy	1	4	3	2

Note: * indicates a non-word.

Figure 1a: Predictions for lexical decision behavior according to the phonetic coding deficiency hypothesis.

In the context of a lexical decision task, this scenario makes the following predictions. Words containing the old category such as [honɪç] in the input will be easy to accept (ordinal accuracy 1), and non-words containing the old category (*[konɪç], ordinal accuracy 2) will also be easy to reject, because they mismatch the lexical representation and therefore do not contact it. Conversely, the words containing the new category in the input but perceived as containing the old category ([konɪç]) will also mismatch the lexical

representation and be difficult to accept (ordinal accuracy 3). Similarly, non-words containing the new category, but perceived as containing the old category (*[honɪç]), will be very difficult to reject (ordinal accuracy 4), because the percept matches the lexical representation.

This pattern of decisions would result in higher accuracy for both words and non-words containing old categories ([honɪç] and *[konɪç]) than for words and non-words containing new categories (*[hønɪç] and [kønɪç]). This pattern is not expected to yield an interaction between lexical status (word vs. non-word) and category type (old vs. new).

According to the *lexical coding deficiency hypothesis*, shown in Figure 1b, we assume that listeners can correctly perceive the input (so percept and input are the same), and the difficulty is located at the lexical coding level, where lexical representations encode the contrast separately but in a fuzzy way (e.g. /o/ vs. /o?/).

lexical representation	/honɪç/		/ko?nɪç/	
	match	≠ mismatch	no mismatch	no mismatch
percept	[honɪç]	*[hønɪç]	[kønɪç]	*[konɪç]
input	[honɪç]	*[hønɪç]	[kønɪç]	*[konɪç]
expected response	yes	no	yes	no
prediction	easy to accept	easy to reject	less easy to accept	difficult to reject
ordinal accuracy	1	3	2	4

Note: * indicates a non-word.

Figure 1b: Predictions for lexical decision behavior according to the lexical coding deficiency hypothesis.

As discussed above, if the L1 does not make use of a certain phonetic category (or acoustic dimension) such as “front rounded vowels” or “long consonants”, L2 lexical encoding of this category might be inaccurate or fragmentary compared to native speakers (at first). We therefore predict in this case that lexical representation containing the new (non-dominant) category might be inaccurate. However, this does not necessarily mean that the dominant category is used *in place of* the new category. As suggested by Cutler et al. (2006) and Hayes-Harb and Masuda (2008), it is possible that the new category is encoded as a poor match to the dominant L1 category. What is important for this hypothesis is that the new category makes reference to the dominant L1 category. For the L2 front-rounded /ø/ category for instance,

this might be represented as a /poor o/, or /o?/ (henceforth we use /?/ to indicate that a category is represented imprecisely in lexical representations), whereas L2 /o/ itself will be represented clearly as the dominant, similar category in L1 /o/.

This scenario makes the following predictions. First, real words containing the old category will likely be most easily recognized as words (ordinal accuracy 1) since the input exactly matches the lexical representation, e.g. [honɪç] matches /honɪç/ 'honey'. Second, real words containing the new category (e.g. [kønɪç] 'king') might be recognized slightly less accurately if they are encoded as a poor match, perhaps as /ko?nɪç/. The input [kønɪç] does not exactly match the representation /ko?nɪç/, but does not clearly mismatch it either. Rejection patterns for non-words will also be asymmetrical: non-words containing the new category (e.g. *[hønɪç]) have to be compared to their real-word counterparts containing the old category (/honɪç/) in order to be recognized as fake and rejected. It is easy to reject the non-word with the new category (ordinal accuracy 3) since the lexical representation with the old category is clear. The percept *[hønɪç] is clearly a mismatch to /honɪç/. Conversely, non-words with an old category (e.g. *[konɪç]) have to be compared to fuzzy lexical representations that contain the new category (e.g. /ko?nɪç/) before they can be rejected as fake. It becomes clear that in this latter case, the rejection will be more difficult if the real word lexical representation is fuzzy (ordinal accuracy 4). Such an asymmetry in accuracy would indicate that there is lexical separation between the old and the new category, but that this separation is not yet target-like (e.g. /o/ vs. /o?/ instead of /o/ vs. /ø/) and still makes reference to the L1 or dominant category.

This pattern of results would mirror the asymmetry obtained with eye-tracking reported by Cutler et al. (2006) and Weber and Cutler (2004), and should produce an interaction between lexical status (word vs. non-word) and category type (old vs. new). If old and new categories are encoded equally well, no interaction is expected, as in the case of native speakers. Such findings combined with very good categorization at the phonetic level would offer support for the lexical coding deficiency hypothesis according to which learners' encoding of phonological contrasts in lexical representations is not target-like and makes reference to the dominant category.

We turn now to the description of our study, designed to examine both hypotheses.

2. The current study

The goal of this study is to expand on these early findings of asymmetries in lexical access and to explicitly investigate whether lexical representations are target-like or not, even if they are separate. We examine the degree to which a novel contrast is target-like in learners' lexical representations by looking at asymmetries in lexical decision patterns, combined with phonetic categorization tasks. By establishing categorization performance in the same participants, we will be able to tease apart whether the contrast is misperceived during input categorization or whether the difficulty is located at the lexical coding level.

We also add to the current understanding of L2 lexical encoding by examining developmental patterns in groups of L2 learners who differ in L2 proficiency and experience, in order to see if lexical processing difficulties can be resolved over time.

Two sets of phonemic contrasts, a consonantal and a vocalic contrast in two languages were used for this study. American English learners of L2 Japanese have difficulties acquiring the durational contrast between long and short consonants (i.e. geminates: *katta* 'bought' vs. singletons: *kata* 'shoulder'). Both are initially mapped onto L1 categories, which in English have only one default duration equivalent to a singleton (Han, 1992; Tajima, Kato, Rothwell, Akahane-Yamada, & Munhall, 2008). Similarly, the German front rounded vowels /y, ø/ (written as <ü, ö>) are confusable with, and thus assimilated to, back rounded vowels /u, o/ for American English listeners (Strange, Weber, Levy, Shafiro, Hisagi et al., 2007). For each language, two experiments were conducted. We tested learners of German and learners of Japanese, and investigated i) learners' phonetic categorization accuracy by examining their response patterns in an ABX task, and ii) the degree to which a novel contrast was target-like in learners' lexical representations by observing any potential asymmetries in their lexical decision patterns.

The vowel and consonant categories used for the experiments reported here are divided in two types: "old" (that is, the dominant category according to Cutler et al., 2006) vs. "new" (the non-dominant category). For the German stimuli, the front-rounded vowels (/y:/, /ʏ/, /ø:/, /œ/) are "new". For the Japanese stimuli, the geminate phonemes (/p:/, /t:/, /k:/) are "new".

3. Experiments 1 and 2: Japanese

In this section, we report the results of two experiments designed to test participants' ability to categorically discriminate between the Japanese long and short phonemes, and to lexically encode this phonemic difference in their lexical representations for Japanese words.

3.1 Experiment 1: *Categorical discrimination with ABX*

3.1.1 Methods and Procedure

Stimuli were pairs of disyllabic non-words in Japanese and English. Test item pairs had the structure CVQV or CVCV, where Q represents a geminate consonant, and C a singleton consonant. An example test pair is [mette] – [mete], where the letter doublet represents a geminate [t:]. Types of consonants used were bilabial, coronal or velar obstruents, as well as the coronal fricative [s]. Control items only contained singleton consonants (CVCV), and differed in the quality of the last V. An example pair is [moke] – [moki]. There were 12 pairs of test items and 4 pairs of control items. Each pair was then arranged into a triplet (A – B – X) where X is either similar to A or to B. (e.g. A-[mette] B-[mete] X-[mete] (X = B); A-[moke] B-[moki] X-[moke] (X = A)). Four counterbalanced orderings for the triplets were used (ABA, ABB, BAA, BAB), which resulted in a total of 64 test triplets.

These 64 trials were presented in four randomized blocks separated by short breaks through the experimental software DMDX (Forster & Forster, 2003). Participants were instructed to decide whether the 3rd non-word (X) matched the first (A) or the second (B) non-word, and indicate their response as fast as possible. Inter-stimulus interval was set to 500 m. Participants had 2000 ms to make their response, before the next trial was initiated. Reaction time (RT) was measured from the onset of the third (X) item.

Stimuli were recorded several times by one female Japanese native speaker of the Tokyo dialect. Two renditions were obtained for each non-word, so that the audio stimulus for A, for example in a triplet ABA, would actually be instantiated by two acoustically different tokens. The speaker in this experiment also recorded the lexical decision items (see Experiment 2). The same speaker was used because we were interested in ascertaining that learners were able to perceptually discriminate short from long consonants in the speaker who also produced the lexical decision tokens.

Both experiments were conducted in one single testing session. A demographic and language background questionnaire was given at the beginning of the testing session. Participants were first tested on the discrimination task and then on the lexical decision task.

3.1.2 Participants

Three groups of participants were tested: two groups of late English learners of Japanese (advanced learners; n = 14, 7 males and 7 females, mean age = 21, intermediate learners; n = 9; 3 males and 6 females, mean

age = 21), and one native speaker group (n = 11, 4 males and 7 females, mean age = 31).

Advanced learners were native speakers of American English, and enrolled either in the second semester of a third-year or fourth-year Japanese class at a large university in the United States, or were teaching Japanese as associate instructors at the same university at the time of recruitment. Average length of time spent living in Japan was 17.8 months. Intermediate learners were native speakers of American English and enrolled in the second semester of a first-year Japanese class. None of them had lived in Japan. Native speakers were all enrolled students at the same university in the United States at the time of testing. They all spoke English as a second language with high proficiency, and most of them were also instructors of Japanese at that university. No participant reported any hearing or speech impairment. Participants received a small payment for their participation in the study.

3.1.3 Results

One native speaker participant and one advanced learner were considered outliers given low accuracy on the control condition (below 2 *SD* from the group mean), and their data excluded from analysis. For each subject, the proportion of accurate answers (%) and mean RT were computed across the four trials for each item, and each condition (test vs. control), resulting in 16 aggregate measures per subject (for 12 test items and four control items). Table 1 presents the mean accuracy and RT in each condition for each group.

Table 1: Mean accuracy (%), mean RT (ms) and standard error (SE) for each group and each condition

Condition	Test				Control			
Group	Accuracy	*SE*	RT	*SE*	Accuracy	*SE*	RT	*SE*
Intermediate	93	1.3	1013	36.9	94	1.7	970	38.5
Advanced	94	1.1	951	29.6	98	1.4	925	30.9
NS	96	1.3	958	35.0	99	1.6	970	36.5

Note: mean RT is computed over correct responses only

Preliminary analysis indicated that the accuracy proportions were not normally distributed, and there was important compression towards the upper accuracy range, unsurprisingly given the high and similar scores of all groups. We therefore used an arcsine transform to expand the top of the scale and increase differences in high accuracy ranges.

A linear mixed effects model was conducted in SPSS 20 on the arcsine-transformed accuracy means. It declared the factors *condition* (test, control) and *group* (intermediate, advanced, NS) as fixed effects, and the factors *condition* and *item* as repeated effects within subjects[3]. The significance threshold was set at $p = 0.05$ for this and all following analyses. The visual analysis of residuals confirmed that the model was a satisfactory fit for the data structure. The parameter estimates are presented in Table 2.

Table 2: Parameter estimate, standard error, t-value, p-value, and 95% Confidence Interval of the predictors for the ABX accuracy (arcsine transformed)

Fixed Effects	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						lower	upper
Intercept	1.557706	.035689	180.512	43.647	.000	1.487285	1.628128
[Condition = test]	-.067632	.037787	477.000	-1.790	.074	-.141880	.006617
[Group = Advanced]	-.027187	.047471	180.512	-.573	.568	-.120856	.066482
[Group = Interm.]	-.110538	.051855	180.512	-2.132	.034	-.212857	-.008218
[Condition = test] * [Group = Advanced]	.000503	.050261	477.000	.010	.992	-.098257	.099264
[Condition = test] * [Group = Interm.]	.062783	.054903	477.000	1.144	.253	-.045098	.170665
Covariance Parameters	**Estimate**	**Std. Error**					
CS diagonal offset (Residual)	.042835	.002774					
CS covariance (Subject)	.002028	.001248					

Note: Interm. = intermediate

When looking at the Type III tests of fixed effects, the F-tests showed a main effect of *condition* ($F(1, 477) = 4.74$, $p < 0.05$), a marginal effect of *group* ($F(2, 40.9) = 2.73$, $p = 0.08$), and no significant interaction between the two factors ($F < 1$). The control condition yielded more accurate performance than the test condition, and the NS group showed the highest accuracy followed by the advanced learners.

Analysis of RTs was performed similarly; the data were normally distributed. Mean RTs were computed over correct responses and entered into a linear mixed model declaring the factors *condition* (test, control) and *group* (intermediate, advanced, NS) as fixed effects, and the factors condition and item as repeated effects within subjects. The parameter estimates are presented in Table 3.

[3] In SPSS 20, this model uses repeated effects within each subject (with compound symmetry correlation structure within subject) in a way that is equivalent to declaring subjects as random effects.

Table 3: Parameter estimate, standard error, t-value, p-value, and 95% Confidence Interval of the predictors for the ABX response times

Fixed Effects	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						lower	upper
Intercept	968.046604	38.168844	39.255	25.362	.000	890.858875	1045.234333
[Condition = test]	-9.584806	19.123175	477.000	-.501	.616	-47.160882	27.991271
[Group = Advanced]	-44.169601	50.769309	39.255	-.870	.390	-146.838878	58.499676
[Group = Interm.]	6.238002	55.458045	39.255	.112	.911	-105.913167	118.389172
[Condition = test] * [Group = Advanced]	36.599469	25.436200	477.000	1.439	.151	-13.381385	86.580323
[Condition = test] * [Group = Interm.]	49.160569	27.785328	477.000	1.769	.077	-5.436204	103.757343

Covariance Parameters	Estimate	Std. Error
CS diagonal offset (Residual)	10970.87	710.39
CS covariance (Subject)	11825.89	3286.0

Note: Interm. = intermediate

When looking at the Type III tests of fixed effects, the F-tests showed that there was a marginal effect of *condition* ($F(1, 477) = 3.08$, $p = 0.08$), no effect of *group* ($F < 1$), and no significant interaction between the two factors ($F(2, 477) = 1.75$, $p > 0.1$). RT in the control condition were slightly faster overall than in the test condition (955 ms vs. 974 ms).

These data suggest that discrimination of the contrasts between geminate and singleton consonants does not appear to be problematic even in early stages of acquisition. We now turn to Experiment 2, where we examine lexical encoding of this contrast among the same participants, with stimuli produced by the same speaker.

3.2. Experiment 2: Lexical decision

3.2.1 Method and Procedure

Forty-two common Japanese words were selected from the textbook used by the first-year and second-year students called *Genki* I and II (Banno et al., 1999), to increase familiarity for all learners. More singleton words were available than geminate words (26 vs. 16). Furthermore, 84 additional words were selected as fillers. Forty-two corresponding test non-words were created from the test words by exchanging a singleton consonant for its geminate counterpart (*akeru* 'to open' vs. **akkeru*) or vice-versa (*kippu* 'ticket' vs. **kipu*). Filler non-words were created by exchanging a feature other than length or by changing or inserting a segment (*tenki* 'weather' vs. **tengi*).

There were 84 test items (42 singleton and 42 geminate), and 168 fillers, resulting in a total of 252 items, which were randomized.

Stimuli were divided into two blocks, so that participants heard both the word and the non-word of a word/non-word pair, but never in the same block. After a short practice, participants were instructed to decide whether each token they heard was a real word or a fake word of Japanese by pressing buttons labeled “yes” or “no” as fast as possible. Participants had 2200 ms to make their response, before the next trial was initiated. RTs were measured from the onset of the word presentation.

3.2.2 Participants

Three groups participated in this experiment: Japanese native speakers, advanced learners and intermediate learners of Japanese. Participants in each of these groups were the same as in Experiment 1.

3.2.3 Results

Accuracy on all items was screened to determine if some non-words were accepted as words by a majority of native speakers or vice-versa. Those items for which accuracy was below 2 *SD* of the mean for the native speaker group were excluded, separately for words and non-words. Six words and four non-words were excluded. The resulting mean accuracy of native speakers in this task was 91% (*SD* = 7.7). Two participants (1 native speaker and 1 intermediate) were excluded because their accuracy on the control condition was below 2 *SD* from their group mean. The total number of participants analyzed was 10 native speakers, 14 advanced, and 8 intermediate learners.

Table 4: Mean accuracy (%), mean RT (ms) and standard error (SE) in the control vs. test conditions in lexical decision for Japanese words and non-words, for each group

	Lexical status	condition	NS		Advanced		Interm.	
			mean	*SE*	mean	*SE*	mean	*SE*
Accuracy	word	control	98	*4.4*	95	*3.7*	79	*4.9*
	non-word	control	91	*4.4*	81	*3.7*	72	*4.9*
	word	test	98	*3.5*	91	*3.0*	71	*3.9*
	non-word	test	93	*3.5*	53	*3.0*	51	*3.9*
RT	word	control	1101	*47.3*	1214	*40.0*	1340	*52.9*
	non-word	control	1232	*47.3*	1345	*40.0*	1415	*52.9*
	word	test	1177	*42.5*	1301	*35.9*	1408	*47.5*
	non-word	test	1288	*42.5*	1513	*35.9*	1523	*47.5*

Note: mean RT is computed over correct responses only

The proportion of accurate answers (%) and mean RT for each subject were computed across items for each condition (test vs. control), within the combination of lexical status (word vs. non-word), and consonant (old vs. new). Mean accuracy and RT in each condition for each group are displayed in Table 4.

A linear mixed effects model was run in SPSS 20 on the accuracy scores. The factors *group* (native speakers, advanced, intermediates), *lexical status* (word vs. non-word), and *condition* (control vs. test) were declared as fixed effects. The factors *condition*, *lexical status*, as well as *consonant* (old vs. new) were also entered as repeated effects within subjects. Tables 5 and 6 display the parameter estimates for each model (accuracy and RT).

Table 5: Parameter estimate, standard error, t-value, p-value, and 95% Confidence Interval of the predictors for the lexical decision accuracy

Fixed Effects	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						lower	upper
Intercept	.977564	.034970	70.253	27.954	.000	.907823	1.047305
[Condition = control]	.002683	.045693	151.000	.059	.953	-.087597	.092963
[Group = Advanced]	-.068681	.045786	70.253	-1.500	.138	-.159993	.022631
[Group = Interm.]	-.267949	.052455	70.253	-5.108	.000	-.372560	-.163338
[Lexical Status = non-word]	-.051522	.037308	151.000	-1.381	.169	-.125236	.022191
[Condition = control] * [Group = Advanced]	.035447	.059826	151.000	.593	.554	-.082757	.153651
[Condition = control] * [Group = Interm.]	.077536	.068539	151.000	1.131	.260	-.057884	.212956
[Lexical Status = non-word] * [Group = Advanced]	-.325366	.048848	151.000	-6.661	.000	-.421880	-.228853
[Lexical Status = non-word] * [Group = Interm.]	-.146374	.055962	151.000	-2.616	.010	-.256944	-.035804
[Condition = control] * [Lexical Status = non-word]	-.015038	.064619	151.000	-.233	.816	-.142713	.112637
[Condition = control] * [Lexical Status = non-word] * [Group = Advanced]	.252057	.084607	151.000	2.979	.003	.084891	.419222
[Condition = control] * [Lexical Status = non-word] * [Group = Interm.]	.146538	.096929	151.000	1.512	.133	-.044975	.338050

Covariance Parameters	Estimate	Std. Error
CS diagonal offset (Residual)	.013919	.001602
CS covariance (Subject)	.005269	.002011

Note: Interm. = intermediate

When looking at the Type III tests of fixed effects, the F-tests revealed that there was a main effect of group on accuracy (native speakers, 95%, advanced, 80%, intermediates, 68%, $F(2, 31.2) = 20.4$, $p < 0.01$). Performance for words was more accurate than for non-words (*lexical status*: $F(1, 157) = 60.8$, $p < 0.01$). Accuracy was also higher in the control condition compared to the test condition (*condition:* $F(1, 157) = 24.1$, $p < 0.01$). All interactions were significant (all $p < .01$), including the triple interaction between group, lexical status and condition ($F(2, 151) = 4.4$, $p = 0.013$). Condition had no effect on the native speaker performance only ($p > 0.1$), whereas both learner groups were significantly more accurate in the control over the test condition (both $p < 0.001$). Similarly, lexical status only marginally influenced native speakers' performance ($p = 0.07$), but learners were significantly more accurate for words than non-words (both $p < 0.001$).

Of particular interest is the interaction between lexical status and condition ($F(1, 157) = 9.8$, $p < 0.01$), for which pairwise comparisons indicated that accuracy for words was similar in the test (88%) and the control (92%) condition ($p > 0.1$), but for non-words, accuracy was much higher in the control condition (82%) compared to the test (65%) condition ($p < 0.001$), suggesting that the difficulty in rejecting non-words was not generalized to all non-words in the experiment. That this effect is mostly due to the learners is confirmed in the triple interaction: for both learner groups, accuracy for non-words was much higher on the control than on the test condition ($p < 0.001$ for both groups). For words, accuracy was similar in the two conditions ($p > 0.1$ for both groups). For native speakers, neither condition nor lexical status influenced their performance (all $p > 0.1$). The analysis of RT was conducted similarly.

Table 6: Parameter estimate, standard error, t-value, p-value, and 95% Confidence Interval of the predictors for the lexical decision RT

Fixed Effects	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						lower	upper
Intercept	1176.561638	42.523051	40.693	27.669	.000	1090.664940	1262.458336
[Condition = control]	-75.246761	35.933487	151.000	-2.094	.038	-146.244105	-4.249417
[Group = Advanced]	124.349338	55.675743	40.693	2.233	.031	11.884161	236.814515
[Group = Interm.]	231.458838	63.784577	40.693	3.629	.001	102.613791	360.303885
[Lexical Status = non-word]	111.704457	29.339569	151.000	3.807	.000	53.735368	169.673546
[Condition = control] * [Group = Advanced]	-11.627803	47.047978	151.000	-.247	.805	-104.585146	81.329541

[Condition = control] * [Group = Interm.]	7.411275	53.900230	151.000	.137	.891	-99.084741	113.907291
[Lexical Status = non-word] * [Group = Advanced]	100.348312	38.414513	151.000	2.612	.010	24.448958	176.247665
[Lexical Status = non-word] * [Group = Interm.]	3.078778	44.009353	151.000	.070	.944	-83.874855	90.032411
[Condition = control] * [Lexical Status = non-word]	19.272912	50.817624	151.000	.379	.705	-81.132495	119.678319
[Condition = control] * [Lexical Status = non-word] * [Group = Advanced]	-100.183328	66.535888	151.000	-1.506	.134	-231.644864	31.278208
[Condition = control] * [Lexical Status = non-word] * [Group = Interm.]	-59.070236	76.226436	151.000	-.775	.440	-209.678346	91.537874

Covariance Parameters	Estimate	Std. Error
CS diagonal offset (Residual)	8608.103	990.681
CS covariance (Subject)	13778.047	3998.469

Note: Interm. = intermediate

When looking at the Type III tests of fixed effects, the F-tests showed that there was a main effect of *group* (mean RT, native speakers = 1199 ms; advanced = 1343 ms; intermediate = 1421 ms, $F(2, 29.7) = 7.6$, $p < 0.01$), *lexical status* (mean RT, words = 1256 ms; non-words = 1386 ms, $F(1, 151) = 78.6$, $p < 0.01$), and of *condition* (mean RT, control = 1274 ms; test = 1368 ms, $F(1, 151) = 41.2$, $p < 0.01$).

There was a marginal interaction between group and lexical status ($F(2, 151) = 2.6$, $p = 0.08$): For non-words, native speakers were significantly faster than both other groups (both $p < 0.01$), whereas for words, native speakers significantly outperformed only the intermediates ($p < 0.01$), not the advanced ($p > 0.1$). No other interaction reached significance. Globally, participants responded to words faster than to non-words, and this difference was visible across both test and control conditions in all three groups.

We now examine whether asymmetrical patterns are visible in the accuracy rates of each group separately. For the test condition only, the mean accuracy for words vs. non-words for each consonant type (new vs. old) is displayed in Figure 2. Average accuracy was low for both learner groups, a finding that contrasts with their performance in the ABX task.

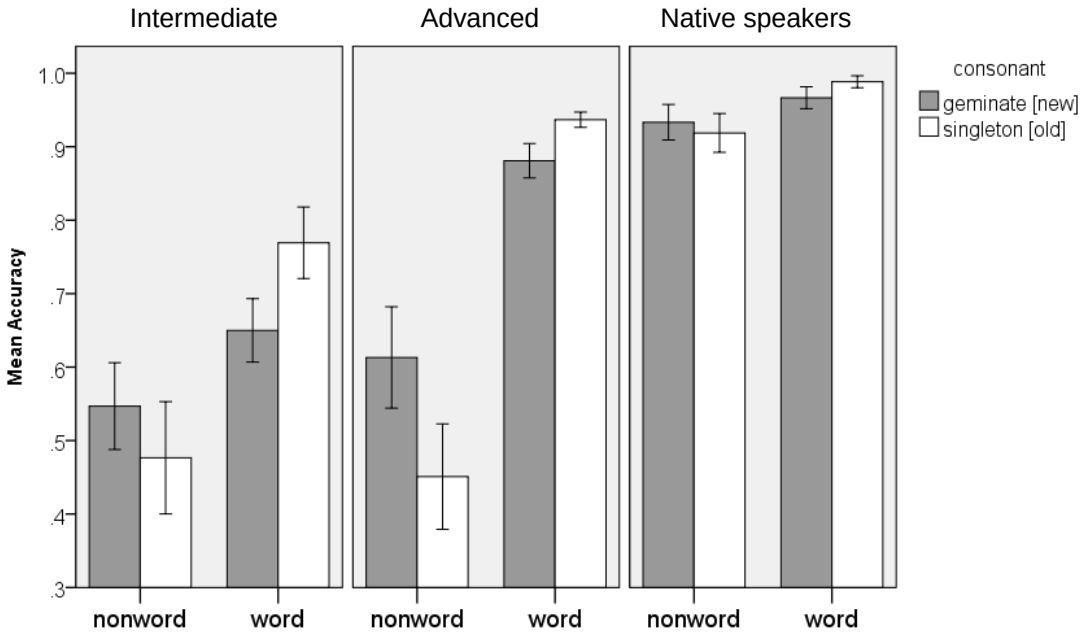

Figure 2: Mean accuracy in lexical decision as a function of consonant type and lexical status for each group. Error bars represent +/- 1 SE.

For each group in turn, a linear mixed effects model declared the mean accuracy as the dependent variable. Fixed factors were *consonant* (old vs. new) and *lexical status* (word vs. non-word). The factors *consonant* and *lexical status* were also entered as repeated effects within subjects. Tables 7, 8, and 9 display the parameter estimates for each group.

Table 7: Parameter estimate, standard error, t-value, p-value, and 95% Confidence Interval of the predictors for the intermediate group

Fixed Effects	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						lower	upper
Intercept	.769231	.058293	16.952	13.196	.000	.646216	.892246
[Lexical Status = non-word]	-.292668	.060237	21	-4.859	.000	-.417938	-.167399
[consonant = geminate]	-.119231	.060237	21	-1.979	.061	-.244500	.006039
[Lexical Status = non-word] * [consonant = geminate]	.189543	.085188	21	2.225	.037	.012386	.366701

Covariance Parameters	Estimate	Std. Error
CS diagonal offset (Residual)	.014514	.004479
CS covariance (Subject)	.012671	.008784

For intermediate learners, accuracy was higher for words than for non-words (71% vs. 51%, *lexical status*: $F(1, 21) = 21.6$, $p < 0.01$), but there was no effect of *consonant* ($F(1, 21) = 0.3$, $p > 0.1$). The interaction was significant ($F(1, 21) = 4.9$, $p < 0.05$): Intermediate learners were marginally more accurate for words containing the old category (77%) than the new (65%) category ($p = 0.06$). Conversely, there was a (non-significant) trend for non-

words containing a new category to be more accurately rejected (55%) than those with an old (48%) category (p = 0.2). This pattern conforms to the predicted ordinal accuracy and suggests that there is an asymmetrical pattern in intermediate learners' lexical representations for these old vs. new consonants.

Advanced learners also show a similar pattern of accuracy even though overall accuracy is higher.

Table 8: Parameter estimate, standard error, t-value, p-value, and 95% Confidence Interval of the predictors for the advanced group

Fixed Effects	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						lower	upper
Intercept	.936813	.051407	41.373	18.223	.000	.833023	1.040604
[Lexical Status = non-word]	-.485920	.061146	39	-7.947	.000	-.609600	-.362240
[consonant = geminate]	-.055861	.061146	39	-.914	.367	-.179541	.067819
[Lexical Status = non-word] * [consonant = geminate]	.218063	.086474	39	2.522	.016	.043153	.392973

Covariance Parameters	Estimate	Std. Error
CS diagonal offset (Residual)	.026172	.005927
CS covariance (Subject)	.010826	.006972

Type III tests of fixed effects revealed that there was a main effect of *lexical status* (F(1, 39) = 75.9, p < 0.01): accuracy was much higher for words (91%) over non-words (53%). Again, the factor *consonant* had no effect (F(1, 39) = 1.5, p > 0.1). The interaction was significant (F(1, 39) = 6.3, p < 0.05). Words containing the old category (94%) were more accurately recognized than those with a new (88%) category, a non-significant trend (p = 0.3). Conversely, non-words containing a new category were more accurately rejected (61%) than those with an old (45%) category (p < 0.01). This suggests that the advanced learners of Japanese resemble the intermediate learners with respect to non target-like lexical representations.

Native speakers show a clearly different pattern.

Table 9: Parameter estimate, standard error, t-value, p-value, and 95% Confidence Interval of the predictors for the native speaker group

Fixed Effects	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						lower	upper
Intercept	.988462	.019842	34.919	49.817	.000	.948177	1.028746
[Lexical Status = non-word]	-.069712	.026598	27.000	-2.621	.014	-.124285	-.015138
[consonant = geminate]	-.021795	.026598	27.000	-.819	.420	-.076368	.032779
[Lexical Status = non-word] * [consonant = geminate]	.036378	.037615	27.000	.967	.342	-.040801	.113557

Covariance Parameters	Estimate	Std. Error
CS diagonal offset (Residual)	.003537	.000963
CS covariance (Subject)	.000400	.000651

Type III tests of fixed effects revealed that there was a main effect of *lexical status* (mean accuracy, words = 98%, non-words = 93%, $F(1, 27) = 7.5$, $p < 0.05$), no effect of *consonant* ($F < 1$) and importantly, there was no interaction ($F < 1$). This pattern suggests that there is no asymmetry in native speakers' lexical representations for each type of consonants, an expected effect.

3.3. Discussion: Experiments 1 and 2

The results of Experiment 1 clearly show that learners can discriminate the geminate and non-geminate contrast, even at the intermediate level, with high accuracy, closely resembling native speakers' performance.

Experiment 2 looked at how learners lexically encode a new contrast that is not in their L1 (i.e., geminate) in comparison with native speakers' performance. Learners, but not native speakers, exhibited a significant interaction of consonant type with lexical status, triggered by the asymmetrical pattern predicted by the lexical coding deficiency hypothesis. In both learner groups, the lowest accuracy was observed for non-words with singletons, the dominant category. This was not the case for native speakers. The order of accuracy and the interactions between consonant type and lexical status observed in the data mirror the asymmetry reported by Cutler et al. (2006) and Weber and Cutler (2004). In addition, we also showed that this asymmetry persists across learner groups that differ in measurable ways in terms of exposure to Japanese. Despite low accuracy overall on the lexical decision task, even intermediate learners were able to maintain a distinction at the lexical level between lexical representations containing the singleton and the geminate categories.

To answer the question whether the lexical encoding of the contrast is target-like however, it is necessary to consider both experiments together. In the ABX task, learners could discriminate geminate from non-geminate very accurately even in early stages of acquisition. This result corroborates previous findings that learners at low-intermediate and advanced levels can discriminate and identify the contrasts in isolated forms (Hardison & Motohashi-Saigo, 2010). For our purpose, the combined results from Experiments 1 and 2 offered evidence that the L2 learners were able to represent the contrast between geminates and singleton at the phonetic level, and that the difficulty was located at the lexical processing level.

Consequently, the observed interaction between category type (old vs. new) and lexical status (word vs. non-word) suggests that the contrast between geminate and singleton was encoded phonologically at the lexical level, but not in a target-like manner. Indeed, it appears that the way this distinction was encoded is dependent on the L1 category, that is, makes reference to the dominant category: the geminate category appears to be encoded as a poor match to the singleton (dominant) category. When lexical representations were target-like, as was the case for native speakers, no interaction has been observed.

It is however possible to object that our categorization task was not sensitive enough to truly establish excellent phonetic categorization. It is worth emphasizing that the cognitive load in this task was low given the limited phonetic variability in stimuli induced by the use of only one voice. The results are comparable to those obtained by Polka (1995) for example. While these findings might not be generalizable to all learners, they served our purpose of examining whether these specific participants were able to discern the singleton-geminate contrast as produced by that particular speaker. However, it would be more compelling to use a more demanding task to establish categorization performance. In Experiment 3, therefore, the design has been slightly modified accordingly. There were more contrasts and two different voices.

4. Experiments 3 and 4: German

In this section, we report the results of two experiments designed to test participants' ability to categorically discriminate between German front-rounded and back-rounded vowel phonemes, and to lexically encode this phonemic difference in their lexical representations for German words.

4.1 Experiment 3: *Categorical discrimination with ABX*

4.1.1 Methods and Procedure

Stimuli were CVC monosyllables which were non-words in German as well as in English. The vowels were surrounded by two consonantal contexts: bilabial (e.g. pVm) or coronal (e.g. sVl). Stimuli included the ten contrasts shown in Table 10.

The front-front contrasts are expected to be easier than front-back according to the Perceptual Assimilation Model for Language Learners (PAM-L2, Best & Tyler 2007), and constitute the control condition. The front-back comparisons are analyzed as the test condition, and the remaining vowel contrasts (/i: - a:/, /i: - o:/) are distracters.

Table 10: contrasts used in the German study

	front-back (test)	front-front (control)	high – non high (fillers)
high	/u: y:/	/i: y:/	/i: a:/, /i: o:/
	/ʏ ʊ/	/ɪ ʏ/	
mid	/o: ø:/	/e: ø:/	
	/ɔ œ/	/ɛ œ/	

There were two pairs of items in each context for each of the 10 contrasts, hence a total of forty pairs of non-words. Each pair was then arranged into a triplet (A – B – X) where X is either similar to A or to B. (e.g. A-[po:m] B-[pø:m] X-[pø:m] (X = B); A-[pe:m] B-[pø:m] X-[pe:m] (X = A)). Four counterbalanced orderings for the triplets were used (ABA, ABB, BAA, BAB), which resulted in a total of 160 triplets. These trials were presented in a block with 3 breaks through the experimental software DMDX (Forster & Forster, 2003). Participants were instructed to decide whether the third non-word (X) matched the first (A) or the second (B) non-word, and indicate their response as fast as possible. Each trial started with a fixation cross at the center of the screen for 250 ms. Interstimulus interval was set at 500 ms. Participants had 2000 ms to make their response, before the next trial was initiated. RT was measured from the onset of the third (X) item in a trial.

To increase task difficulty, contrasts were not blocked: all 10 contrasts could occur in the same block. In addition, two different voices were used. Stimuli were recorded several times by two female German native speakers, and normalized for amplitude. One voice was used for the X token, whereas the other was used for the two different A and B tokens.

Both experiments were conducted in one single testing session. A demographic and language background questionnaire was given at the beginning of the testing session. Participants completed the lexical decision task first, followed by the ABX task. At the end, participants rated their familiarity with the words in the lexical decision task.

4.1.2 Participants

Four groups of participants were tested: two groups of late English learners of German (intermediate, n = 103, vs. advanced, n = 21), one group of American English native speakers without experience with German (“monolingual”, n = 31), and one group of German native speakers (n = 18).

Based on demographic and background information recorded in the questionnaire, participants were excluded from further analysis if one or several of the following criteria applied: Father’s or Mother’s L1 non English; Subject’s L1 non English; Residential history too complex or including parts in

Germany [for intermediate learners and monolingual English native speakers only]; Exposure to German, French or other languages that contain front rounded vowels at home; Exposure to German, French or other languages that contain front rounded vowels in childhood; Speech or hearing disorder; Having had instruction in French/Turkish/Danish/Mandarin or other languages that may contain front rounded vowels under examination; Missing data/subject questionnaire. In total, 39 intermediate participants were excluded, leaving 64 for analysis; six monolingual English native speakers were also excluded, leaving 25 for analysis. No advanced learner or native speaker of German was excluded.

In addition, nine intermediate learners were further excluded because of low familiarity ratings with the words used in Experiment 4 (see below). A total of 55 intermediate participants were retained for the analysis for Experiments 3 and 4.

Intermediate learners (n = 55; 20 females) were enrolled in third-year classes at a large university in the United States, and had taken a maximum of six semesters of German. None of them had spent any time in a German-speaking country. Their mean age was 20.7 years (*SD* = 1.7, range: 19-30).

Advanced learners (n = 21, 7 females) were graduate students and instructors, who had taken at least 8 semesters of German. Except for three of them, they had spent between 4 and 36 months in a German-speaking country (M = 15.1 months, *SD* = 11.1, range: 0-36). Their mean age was 27.2 years (*SD* = 4.73, range: 21-38).

German native speakers (n = 18; 12 females) served as a control group. All were living in the USA at the time of testing, and so had knowledge of English. Their mean age was 26.6 years (*SD* = 5.0, range: 18-33).

Naïve American English native speakers (n = 25; 24 females) with no knowledge of German and no experience with any language containing the target phonemes were also tested as a second control. Their mean age was 20 years (*SD* = 0.85, range 18-22).

None of these participants reported any hearing or speech impairment. They received either course credit or a small monetary compensation for participating.

4.1.3 Results

Two participants were considered outliers (one intermediate, and one German native speaker) given very low accuracy on the control and filler conditions, and were removed from further analysis. Data for one additional intermediate subject were missing, bringing the total number of intermediate learners to 53, and the German native speakers to 17. For each subject,

proportions of accurate answers (%) and RT means were computed across items in each condition, within the combination of *height* (mid vs. high), *tenseness* (tense vs. lax), *context* (bilabial, coronal), and *contrast*. Thus, there were 18 aggregate measures per subject, aggregated over two items (and four trials per item). Table 11 presents the mean accuracy and RT in each condition for each group.

Table 11: Mean accuracy (%), mean RT (ms) and standard error (SE) for each group and each condition

Condition	Test (back-front)				Control (front-front)				Distracter			
Group	Acc.	*SE*	RT	*SE*	Acc.	*SE*	RT	*SE*	Acc.	*SE*	RT	*SE*
Intermediate	87	*0.9*	942	*21.4*	92	*0.9*	896	*21.4*	95	*1.3*	846	*23.2*
Advanced	90	*1.5*	945	*34.0*	94	*1.5*	893	*34.0*	97	*2.1*	857	*36.8*
Native Sp.	96	*1.6*	840	*37.8*	95	*1.6*	845	*37.8*	95	*2.3*	791	*40.9*
Monolingual	83	*1.4*	1021	*31.2*	87	*1.4*	981	*31.2*	93	*1.9*	939	*33.7*

Note: mean RT is computed over correct responses only

An arcsine transform was used for these data for the same reason as in Experiment 1. To compare the performance of the groups, a linear mixed model was conducted in SPSS 20.0 on the arcsine-transformed accuracy means of the four groups. It declared the factors *group* (intermediate, advanced, NS, and monolingual), and *condition* (test, control, and filler) as fixed effects. The variables of *height* (mid vs. high), *tenseness* (tense vs. lax), and *context* (bilabial, coronal) were declared as repeated effects within subjects. Table 12 displays the parameter estimates for this model.

Table 12: Parameter estimate, standard error, t-value, p-value, and 95% Confidence Interval of the predictors for the ABX accuracy (arcsine transformed)

Fixed Effects	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						lower	upper
Intercept	1.284536	.025105	173.381	51.167	.000	1.234986	1.334087
[Condition = test]	-.065048	.021177	1964.000	-3.072	.002	-.106580	-.023516
[Condition = control]	.072852	.033484	1964.000	2.176	.030	.007184	.138520
[Group = Int.]	.102200	.030456	173.381	3.356	.001	.042089	.162312
[Group = Adv.]	.126425	.037156	173.381	3.403	.001	.053089	.199761
[Group = NS]	.149915	.039460	173.381	3.799	.000	.072031	.227799
[Condition = test] * [Group = Interm.]	-.035064	.025691	1964.000	-1.365	.172	-.085448	.015320
[Condition = test] * [Group = Adv.]	-.014067	.031343	1964.000	-.449	.654	-.075535	.047402
[Condition=test] * [Group = NS]	.091619	.033286	1964.000	2.752	.006	.026339	.156900

[Condition = control] * [Group = Interm.]	-.026750	.040621	1964.000	-.659	.510	-.106414	.052915
[Condition = control] * [Group = Adv.]	-.004478	.049557	1964.000	-.090	.928	-.101668	.092712
[Condition = control] * [Group = NS]	-.084453	.052631	1964.000	-1.605	.109	-.187671	.018764

Covariance Parameters	Estimate	Std. Error
CS diagonal offset (Residual)	.044847	.001431
CS covariance (Subject)	.010150	.001691

Note: Interm. = intermediate; Adv. = Advanced; NS = Native speakers

The main effect of condition was significant: accuracy was highest in the distractor condition (95%), followed by the control (92%) and the test (89%) conditions ($F(2, 1964) = 22.5$, $p < 0.01$). Performance also varied by *group* ($F(3, 135.1) = 6.8$, $p < 0.01$). Pairwise comparisons (with Sidak correction) revealed that only the monolingual group performed significantly less accurately than the other three groups (all $p < 0.05$). No other comparison was significant. There was a significant interaction between the two factors ($F(6, 1964) = 4.6$, $p < 0.01$). Univariate tests indicated that performance of all non-native groups was influenced by condition (simple effect of condition significant for all three groups at $p < 0.01$), whereas for the native speaker group, condition had no effect ($F < 1$). More specifically, pairwise comparisons (with Sidak correction) show that in all non-native groups, this significance is driven by performance in the test condition, significantly less accurate than in both other conditions (all $p < 0.01$). Performance in the distractor and the control condition was equally accurate in all non-native groups (intermediates and advanced: $p > 0.1$; monolinguals: $p = 0.09$).

Analysis of RTs was performed similarly; the data were normally distributed. Parameter estimates are displayed in Table 13.

Table 13: Parameter estimate, standard error, t-value, p-value, and 95% Confidence Interval of the predictors for the ABX RT

Fixed Effects	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						lower	upper
Intercept	980.796596	31.151506	119.562	31.485	.000	919.116480	1042.476712
[Condition = test]	40.070043	10.601135	1964.000	3.780	.000	19.279388	60.860699
[Condition = control]	-42.061555	16.761866	1964.000	-2.509	.012	-74.934468	-9.188643
[Group = Interm.]	-85.202285	37.791002	119.562	-2.255	.026	-160.028633	-10.375937
[Group = Adv.]	-87.463574	46.105030	119.562	-1.897	.060	-178.751734	3.824585
[Group = NS]	-135.945198	48.964264	119.562	-2.776	.006	-232.894652	-38.995744

[Condition = test] * [Group = Interm.]	6.080688	12.860615	1964.000	.473	.636	-19.141197	31.302573
[Condition = test] * [Group = Adv.]	11.316010	15.689953	1964.000	.721	.471	-19.454695	42.086715
[Condition = test] * [Group = NS]	-45.063466	16.662975	1964.000	-2.704	.007	-77.742436	-12.384496
[Condition = control] * [Group = Interm.]	-7.689640	20.334417	1964.000	-.378	.705	-47.568942	32.189662
[Condition = control] * [Group = Adv.]	6.021856	24.807993	1964.000	.243	.808	-42.630900	54.674613
[Condition = control] * [Group = NS]	-11.643727	26.346477	1964.000	-.442	.659	-63.313716	40.026262

Covariance Parameters	Estimate	Std. Error
CS diagonal offset (Residual)	11238.406	358.632
CS covariance (Subject)	22855.607	3137.705

Note: Interm. = intermediate; Adv. = Advanced; NS = Native speakers

Type III tests of fixed effects revealed that there was a main effect of *condition* ($F(2, 1964) = 48.3$, $p < 0.01$), and of *group* ($F(3, 115.0) = 3.6$, $p < 0.05$). Pairwise comparisons (with Sidak correction) indicated that RTs were slowest in the test condition (M = 936 ms), faster in the control (M = 903 ms) and fastest in the distractor condition (M = 858 ms, all $p < 0.01$). Latency differences among the groups were less pronounced: pairwise comparisons showed that only the monolinguals were significantly slower than the native speakers ($p = 0.01$). No other comparison reached significance. There was a significant interaction between the two factors ($F(6, 1964) = 2.5$, $p < 0.05$). Univariate tests showed that the simple effect of *group* was significant in each condition (*test*: $F(3, 119.6) = 4.6$, $p < 0.01$; *control*: $F(3, 119.6) = 2.9$, $p < 0.05$; *distracter*: $F(3, 164.3) = 2.9$, $p < 0.05$), and was mainly driven by the significantly slower RT of the monolingual compared to the native speaker group (test: $p < 0.01$; both other conditions: $p < 0.05$). No other pairwise comparison reached significance.

Within each group however, latencies were also influenced by condition: notably, pairwise comparisons indicated that all non-native groups, but not the native speakers, were slower in the test condition compared to both other conditions (all $p < 0.01$). In addition, for both intermediates and monolinguals, all other comparisons were also significant ($p < 0.01$). For advanced, RT on the control and distractor conditions did not differ ($p > 0.1$). The native speakers were faster on the distractor condition compared to both others (both $p < 0.05$), but did not differ in the test vs. control condition ($p > 0.1$).

After excluding the monolingual group, both analyses (arcsine-transformed accuracy, and RT) showed a marginal main effect of *group* (accuracy: $F(2, 105.8) = 2.7$, $p = 0.07$; RT: $F(2, 90.3) = 1.5$, $p > 0.1$). The main effect of *condition* (accuracy: $F(2, 1541) = 14.2$, $p < 0.01$; RT: $F(2, 1541) = 35.0$, $p < 0.01$) and the interaction remained significant in both models (accuracy: $F(4, 1541) = 7.2$, $p < 0.01$; RT: $F(4, 1541) = 3.8$, $p < 0.01$). Pairwise comparisons between the conditions indicated that accuracy on the test condition was significantly lower (91%), and RTs were slower (908 ms), than on both other conditions (control: 93%, 878 ms; distractor: 96%, 831 ms), which did not differ from each other ($p > 0.1$).

The interaction pattern was further explored for both dependent variables. For the arcsine, univariate tests showed that the simple effect of *group* was significant for the test condition ($F(2, 135.3) = 13.0$, $p < 0.01$), but not for the two other conditions ($p > 0.1$). In that condition only, native speakers outperformed both other groups (both $p < 0.01$), who did not differ from each other ($p > 0.1$). For the control and distractor conditions, no comparison reached significance. For the RT, the simple main effect of *group* was marginal for the test condition ($F(2, 93.9) = 3.1$, $p = 0.052$), and not significant in the two other conditions ($F < 1$). Native speakers were marginally faster than the intermediate group ($p = 0.06$) on the test condition only. No other comparison was significant.

The persistent significant interactions in terms of accuracy suggest that the test condition, despite overall high accuracy levels, triggered more errors in the learner groups compared to the native speakers. We now explore this condition in more detail. Mean accuracy rates for these three groups for the test items are presented in Figure 3.

As above, a linear mixed model was conducted on the arcsine-transformed accuracy scores. It declared the factors *group* (intermediate, advanced, NS), *height* (high vs. mid) and *context* (coronal, bilabial) as fixed effects. *Height* and *context* were also entered as repeated effects within subjects. Parameter estimates are shown in Table 14.

Table 14: Parameter estimate, standard error, t-value, p-value, and 95% Confidence Interval of the predictors for ABX with the height and context factors

Fixed Effects	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						lower	upper
Intercept	1.459285	.043574	269.342	33.490	.000	1.373497	1.545074
[height = high]	-.116691	.050822	264.000	-2.296	.022	-.216758	-.016624
[context = bil]	.011620	.050822	264.000	.229	.819	-.088447	.111687
[Group = Interm.]	-.158209	.050077	269.342	-3.159	.002	-.256801	-.059617
[Group = Adv.]	-.057697	.058615	269.342	-.984	.326	-.173099	.057704

[context = bil] * [height = high]	.101058	.071873	264.000	1.406	.161	-.040458	.242574
[height = high] * [Group = Interm.]	-.091473	.058406	264.000	-1.566	.119	-.206474	.023529
[height = high] * [Group = Adv.]	-.144462	.068364	264.000	-2.113	.036	-.279071	-.009853
[context = bil] * [Group = Interm.]	.045559	.058406	264.000	.780	.436	-.069442	.160561
[context = bil] * [Group = Adv.]	-.021971	.068364	264.000	-.321	.748	-.156580	.112638
[context = bil] * [height = high] * [Group = Interm.]	.016990	.082599	264.000	.206	.837	-.145647	.179626
[context = bil] * [height = high] * [Group = Adv.]	.025437	.096682	264.000	.263	.793	-.164929	.215803

Covariance Parameters	Estimate	Std. Error
CS diagonal offset (Residual)	.021954	.001911
CS covariance (Subject)	.010324	.002431

Note: Interm. = intermediate; Adv = Advanced; NS = Native speakers; bil = bilabial

Type III tests of fixed effects showed that there was a significant interaction between *height* and *group* ($F(2, 264) = 3.8$, $p < 0.05$). Mid vowels yielded higher accuracy (on average 94% correct) than high vowels (87% correct), a significant difference for both learner groups ($p < 0.01$), but only marginal for the native speakers ($p = 0.067$).

There was also a significant interaction between *height* and *context* ($F(1, 264) = 10.8$, $p < 0.01$). Accuracy was higher in the bilabial context than in the coronal context, but for high vowels only (91% vs. 83%, $p < 0.01$), not for mid vowels (95% vs. 94%, $p > 0.1$). There was no interaction between *group* and *context* ($F(2, 264) = 1.8$, $p > 0.1$), nor a significant triple interaction ($F < 1$).

To sum up, this pattern of results indicates that learners displayed lower accuracies in very specific conditions only, such as for high vowels in the coronal context (see Figure 3). In addition, the lack of triple interaction indicated that overall, learners' behavior was not fundamentally different from native speakers' across conditions. Binomial tests confirmed that the scores of the learners in every subcondition (high vs. mid vowels in bilabial vs. coronal contexts) were significantly different from chance (all $p < 0.01$).

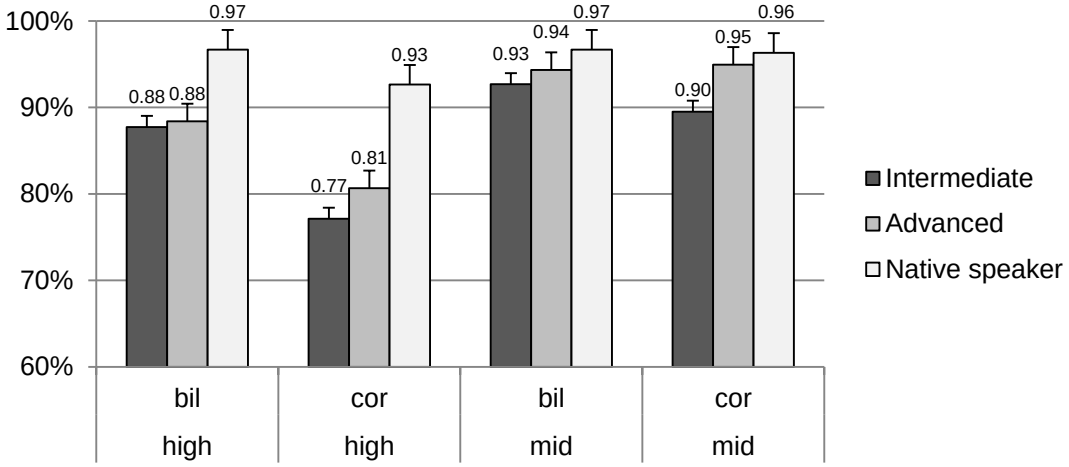

Figure 3: Mean accuracy on the test condition as a function of vowel height and context, for each group. Error bars represent +/- 1 SE.

Taken together, and despite lower accuracy levels in one specific comparison, the overall high accuracy levels in these data allows us to conclude that learners in our experiment did not have truly serious difficulties in differentiating back from front rounded vowels. We now turn to the examination of lexical encoding of these contrasts among the same learners.

4.2. Experiment 4: Lexical decision

4.2.1 Method and Procedure

Eighty common German words were selected and matched for frequency and familiarity. They were either mono- or disyllabic words containing one of the vowels under scrutiny. Eighty corresponding non-words were created by exchanging e.g. a back vowel for its front-rounded counterpart (*Honig* /hoːniç/ 'honey' vs. **Hönig* /høːniç/) or vice-versa (*König* /køːniç/ 'king' vs. **Konig* /koːniç/). There were 160 test items, and 128 filler items (64 words and 64 similar non-words, e.g. *Kanne* /kanə/ 'teapot', **Blanne* /blanə/, *Pflaume* /pflaʊmə/ 'plum', **Pfeude* /pfɔʏdə/). This resulted in a total number of 288 trials, which were randomized.

Stimuli were divided into two blocks, so that participants heard both the word and the non-word of a word/non-word pair, but never in the same block. Participants listened to stimuli and had to decide whether each token was a real word or a fake word of German by pressing buttons labeled "yes" or "no" as fast as possible. The "yes" button was always the dominant hand of the participant. Participants had 2500 ms to make their response, before the next trial was initiated.

In selecting the target words, we tried as much as possible to match the frequency among words containing the same vowel, and across vowel types

(old and new). Therefore, we avoided potential items with extreme frequency values (either very high or very low) that would stand out from the rest of the items. For all the selected word forms (that is, all words including all their homophones), the spoken and written frequency was looked up in the Celex Database (Baayen, Piepenbrock & Gulikers, 1995), and their average (over spoken and written values) and additive frequency was calculated. In addition to frequency measurements, we also obtained familiarity ratings for these German words from our participants at the end of the experiment in order to verify that the items were known words to them. These averaged measures are included in Table 15 below. A series of two-tailed t-tests revealed that the various measures were very similar for words containing old vs. new vowels.

Table 15: Average frequency and familiarity by contrast for the German words used in lexical decision

	Frequencies				**Familiarity**
	spoken	**written**	**mean**	**sum (S + W)**	**average**
Old	36.1	45.1	40.6	81.2	4.52
New	33.9	35.8	34.9	69.8	3.68
$t(78) =$	0.15, $p > 0.1$	0.72, $p > 0.1$	0.45, $p > 0.1$	0.45, $p > 0.1$	1.99, $p = 0.05$

Stimuli were recorded by a female native speaker of German (who also recorded tokens for the ABX task) in a sound-isolated recording booth. In order to further ensure that L2 learners even at intermediate levels would have established some lexical representations for the words used in the task, we developed a 917-word long “Märchenkrimi” (a detective fairy tale) called “Der schlaue Heinrich” (‘the smart Henry’). All the words selected for the experiment were embedded in the story. Students in German classes from which recruitment would later be made for this study were presented with the text during their regular classes and worked with it for a few weeks before recruitment began. Students did not know about the upcoming study while they worked with the text. Vocabulary items were explained by teachers, and also used in other exercises. No attention was specifically drawn to the items used as stimuli.

4.2.2 Participants

Three groups participated in this experiment: German native speakers, advanced learners and intermediate learners of German. Participants in each of these groups were the same as in Experiment 3.

4.2.3 Results

Accuracy on all items was screened to determine if some non-words were accepted as words by a majority of native speakers or vice-versa. Those items for which accuracy was below 2 *SD* of the mean for this group were excluded, separately for words and non-words. Four words and three non-words were excluded. These items did not yield high enough convergence in lexical decision responses among the native speakers for various reasons (e.g. indistinct pronunciation or dialectal differences in lexical status). The resulting mean accuracy of native speakers in this task was 95%. For participants, no further outlier was excluded.

The proportion of accurate answers (%) and mean RT for each subject were computed across items for each condition within the combination of lexical status (word vs. non-word), and vowel (old vs. new). They are presented in Table 16.

Table 16: Mean accuracy (%) and RT (ms) in the control vs. test conditions in lexical decision for German words and non-words

	Lexical status	condition	NS		Advanced		Interm.	
			mean	*SE*	mean	*SE*	mean	*SE*
Accuracy	word	control	96	*1.4*	90	*1.7*	67	*1.4*
	non-word	control	95	*1.4*	87	*1.7*	72	*1.4*
	word	test	96	*1.4*	89	*1.7*	70	*1.4*
	non-word	test	91	*1.4*	72	*1.7*	45	*1.4*
RT	word	control	943	*36.2*	1092	*33.6*	1096	*20.7*
	non-word	control	1051	*36.2*	1267	*33.6*	1247	*20.7*
	word	test	918	*36.2*	1131	*33.6*	1105	*20.7*
	non-word	test	1062	*36.2*	1353	*33.6*	1283	*20.7*

Note: mean RT is computed over correct responses only

A linear mixed model was conducted in SPSS 20 on the accuracy means. It declared the factor *group* (native speakers, advanced, intermediates), *condition* (control vs. test) and *lexical status* (word vs. non-word) as fixed effects. The factors *condition*, *lexical status*, as well as *vowel* (old vs. new) were declared as repeated effects within subjects. Parameter estimates are presented in Table 17.

Table 17: Parameter estimate, standard error, t-value, p-value, and 95% Confidence Interval of the predictors for the lexical decision accuracy

Fixed Effects	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						lower	upper
Intercept	.956771	.017667	384.623	54.154	.000	.922034	.991508
[condition = control]	.005154	.028797	461.000	.179	.858	-.051435	.061743

[Group = Adv.]	-.066096	.024077	384.623	-2.745	.006	-.113435	-.018758
[Group = Interm.]	-.252225	.020354	384.623	-12.392	.000	-.292245	-.212206
[lexical status = non-word]	-.042708	.023512	461.000	-1.816	.070	-.088913	.003496
[condition = control] * [Group = Adv.]	.004910	.039243	461.000	.125	.900	-.072208	.082028
[condition = control] * [Group = Interm.]	-.038275	.033176	461.000	-1.154	.249	-.103470	.026919
[lexical status = non-word] * [Group = Adv.]	-.123859	.032042	461.000	-3.866	.000	-.186826	-.060893
[lexical status = non-word] * [Group = Interm.]	-.208542	.027088	461.000	-7.699	.000	-.261773	-.155310
[condition = control] * [lexical status = non-word]	.034046	.040725	461.000	.836	.404	-.045983	.114075
[condition = control] * [lexical status = non-word] * [Group = Adv.]	.106237	.055498	461.000	1.914	.056	-.002825	.215298
[condition = control] * [lexical status = non-word] * [Group = Interm.]	.265316	.046918	461.000	5.655	.000	.173117	.357515

Covariance Parameters	Estimate	Std. Error
CS diagonal offset (Residual)	.009951	.000655
CS covariance (Subject)	.000643	.000358

Note: Interm. = intermediate; Adv. = advanced

Type III tests of fixed effects indicate that there was a main effect of group on accuracy (native speakers, 95%, advanced, 85%, intermediates, 64%, $F(2, 107.9) = 317$, $p < 0.01$). Performance for words (85%) was more accurate than for non-words (77%, *lexical status*: $F(1, 461) = 55.2$, $p < 0.01$). Accuracy was also higher in the control condition (85%) compared to the test (77%) condition (*condition:* $F(1, 461) = 53.0$, $p < 0.01$). All interactions were significant (all $p < 0.01$), including the triple interaction between group, lexical status and condition ($F(2, 461) = 18.4$, $p < 0.01$). Condition had no effect on the native speaker performance only ($p > 0.1$), whereas both learner groups were significantly more accurate in the control over the test condition (both $p < 0.01$). Similarly, lexical status did not influence native speakers' performance ($p > 0.1$), but learners were more accurate for words than non-words (both $p < 0.01$).

Of particular interest is the interaction between lexical status and condition ($F(1, 461) = 62.0$, $p < 0.01$), for which pairwise comparisons indicated that accuracy for words was similar in the test (85%) and the control (84%) condition ($p > 0.1$), but for non-words, accuracy was much higher in the

control condition (85%) compared to the test (70%) condition ($p < 0.01$), suggesting that the difficulty in rejecting non-words was not generalized to all non-words in the experiment but rather specific to the test non-words.

That this effect is mostly due to the learners (see Table 16), is confirmed in the triple interaction: for both learner groups, accuracy for non-words was higher on the control than on the test condition ($p < 0.01$ for both groups). For words, accuracy was similar in the two conditions ($p > 0.1$ for the advanced, $p = 0.045$ for the intermediates). For native speakers, neither condition nor lexical status influenced their performance (all $p > 0.1$).

The analysis of RTs was conducted similarly (Table 18). A linear mixed model was conducted on mean RT scores as the dependent variable. It declared the factors *group*, *condition* and *lexical status* as fixed effects, and *lexical status*, *condition* and *vowel* were declared as repeated factors within subjects.

Table 18: Parameter estimate, standard error, t-value, p-value, and 95% Confidence Interval of the predictors for the lexical decision RT

Fixed Effects	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						lower	upper
Intercept	918.128	35.536	107.475	25.837	.000	847.686	988.570589
[condition = control]	25.049	21.392	461.000	1.171	.242	-16.988	67.086171
[Group = Adv.]	212.538	48.427	107.475	4.389	.000	116.541	308.534122
[Group = Interm.]	187.246	40.940	107.475	4.574	.000	106.092	268.400497
[lexical status = non-word]	143.867	17.466	461.000	8.237	.000	109.544	178.190643
[condition = control] * [Group = Adv.]	-64.075	29.152	461.000	-2.198	.028	-121.362	-6.787526
[condition = control] * [Group = Interm.]	-34.702822	24.645	461.000	-1.408	.160	-83.133	13.727316
[lexical status = non-word] * [Group = Adv.]	77.817939	23.802	461.000	3.269	.001	31.043	124.592815
[lexical status = non-word] * [Group = Interm.]	34.913220	20.122	461.000	1.735	.083	-4.630	74.456262
[condition = control] * [lexical status = non-word]	-36.147842	30.252	461.000	-1.195	.233	-95.598	23.301988
[condition = control] * [lexical status = non-word] * [Group = Adv.]	-10.630492	41.227	461.000	-.258	.797	-91.647	70.385970
[condition = control] * [lexical status = non-word] * [Group = Interm.]	8.283788	34.853	461.000	.238	.812	-60.207	76.774346

Covariance Parameters	Estimate	Std. Error

CS diagonal offset (Residual)	5491.271	361.691
CS covariance (Subject)	19984.648	3098.988

Note: Interm. = intermediate; Adv. = advanced

According to the type III tests of fixed effects, there was a main effect of *group* (mean RT, native speakers = 993 ms; advanced = 1210 ms; intermediate = 1183 ms, $F(2, 92.0) = 13.7$, $p < 0.01$), of *lexical status* (mean RT, words = 1047 ms; non-words = 1210 ms, $F(1, 461) = 478.3$, $p < 0.01$), and of *condition* (mean RT, control = 1116 ms; test = 1142 ms, $F(1, 461) = 12.5$, $p < 0.01$). All interactions were significant (all $p < 0.05$), except the triple interaction between group, lexical status and condition ($F < 1$). Condition had no effect on the native speaker RTs only ($p > 0.1$), whereas both learner groups were significantly faster in the control over the test condition (both $p < 0.01$). Lexical status influenced performance as well: all groups responded faster to words over non-words (all $p < 0.01$). Of particular interest is the interaction between lexical status and condition ($F(1, 461) = 6.1$, $p < 0.05$), for which pairwise comparisons indicated that RTs for words were similar in the test (1051 ms) and the control (1043 ms) condition ($p > 0.1$), but for non-words, RTs were faster in the control condition (1188 ms) compared to the test (1233 ms) condition ($p < 0.01$), suggesting that in parallel to accuracy data, the difficulty in rejecting non-words was particularly pronounced for the test non-words. The same trends as for the accuracy data are observed again here, even though the triple interaction is not significant: native speakers' RTs are not influenced by condition or lexical status, whereas both factors clearly impact learners' performance. As can be seen in Table 16, the learners have slower RTs mainly on the test non-words.

We turn now to the asymmetries in mean accuracy rates as a function of vowel and lexical status. For the test condition only, the mean accuracy for words vs. non-words as a function of vowel type (new vs. old) is displayed in Figure 4. Average accuracy was low for both learner groups, a finding that contrasts with their performance in the ABX task.

For each group separately, a linear mixed model was conducted on the accuracy scores. The factors *vowel* (old vs. new) and *lexical status* (word vs. non-word) were declared as fixed effects. The factors *vowel* and *lexical status* were also entered as repeated effects within subjects. Parameter estimates are displayed in Tables 19, 20 and 21.

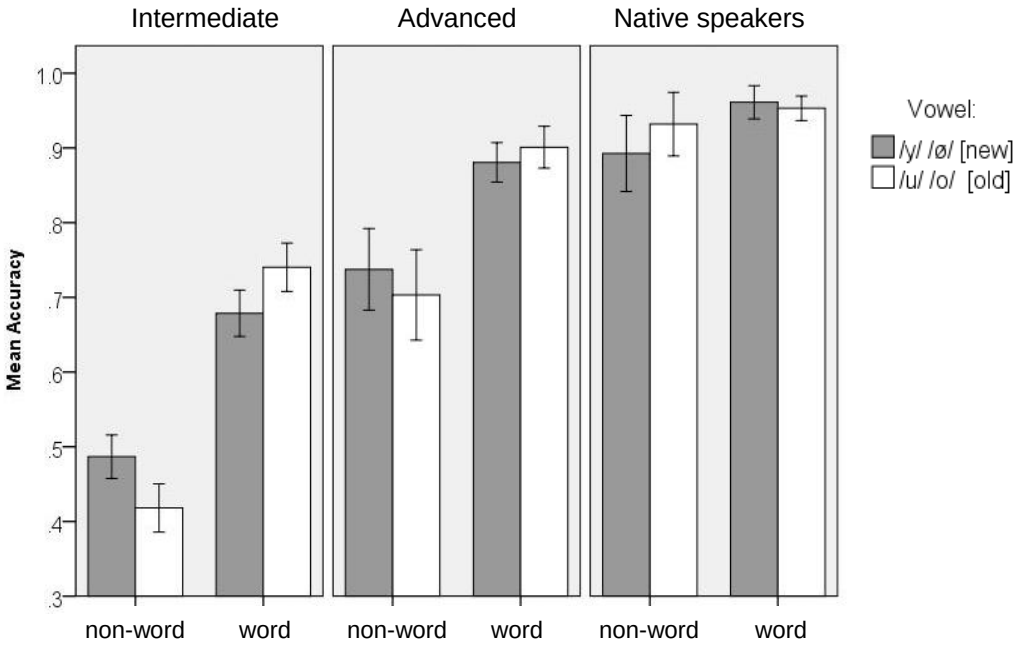

Figure 4: Mean accuracy in lexical decision as a function of vowel type and lexical status for each group. Error bars represent +/- 1 SE.

Table 19: Parameter estimate, standard error, t-value, p-value, and 95% Confidence Interval of the predictors for the intermediate group

Fixed Effects	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						lower	upper
Intercept	.740191	.015621	206.29	47.38	.000	.709393	.770989
[lexical status = non-word]	-.322091	.023435	162.0	-13.74	.000	-.368369	-.275814
[vowel = new]	-.061555	.023435	162.0	-2.627	.009	-.107833	-.015278
[lexical status = non-word] * [vowel = new]	.130119	.033142	162.0	3.926	.000	.064672	.195565

Covariance Parameters	Estimate	Std. Error
CS diagonal offset (Residual)	.015103	.001678
CS covariance (Subject)	-.001681	.000582

Table 20: Parameter estimate, standard error, t-value, p-value, and 95% Confidence Interval of the predictors for the advanced group

Fixed Effects	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower	Upper
Intercept	.900969	.022555	68.644	39.946	.000	.855969	.945968
[lexical status = non-word]	-.197825	.027902	60.000	-7.090	.000	-.253637	-.142012
[vowel = new]	-.020322	.027902	60.000	-.728	.469	-.076134	.035491
[lexical status = non-word] * [vowel = new]	.054629	.039459	60.000	1.384	.171	-.024301	.133559

Covariance Parameters	Estimate	Std. Error
CS diagonal offset (Residual)	.008174	.001492
CS covariance (Subject)	.002509	.001487

Table 21: Parameter estimate, standard error, t-value, p-value, and 95% Confidence Interval of the predictors for the native speaker group

Fixed Effects	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower	Upper
Intercept	.953058	.017960	39.061	53.066	.000	.916733	.989383
[lexical status = non-word]	-.021185	.018015	51.000	-1.176	.245	-.057351	.014981
[vowel = new]	.008053	.018015	51.000	.447	.657	-.028113	.044219
[lexical status = non-word] * [vowel = new]	-.047283	.025476	51.000	-1.856	.069	-.098430	.003863

Covariance Parameters	Estimate	Std. Error
CS diagonal offset (Residual)	.002921	.000578
CS covariance (Subject)	.002885	.001248

For intermediate learners, accuracy was higher for words than for non-words (71% vs. 45%, *lexical status*: ($F(1, 162) = 240.6$, $p < 0.01$), but there was no effect of *vowel* ($F < 1$). The interaction was significant ($F(1, 162) = 15.4$, $p < 0.01$): Intermediate learners were more accurate for words containing the old category (74%) than the new (68%) category ($p = 0.004$). Conversely, non-words containing a new category were more accurately rejected (49%) than those with an old (42%) category ($p = 0.009$). This pattern conforms to the predicted ordinal accuracy and suggests that there is an asymmetrical pattern in intermediate learners' lexical representations for these old vs. new vowels.

The pattern shown in Figure 4 for the advanced learners is comparable to that of the intermediate learners, but accuracy is overall higher. There was a main effect of *lexical status* ($F(1, 60) = 74.7$, $p < 0.01$): accuracy was higher for words (89%) over non-words (72%). There was again no effect of *vowel* ($F < 1$) but the interaction, this time, was not significant ($F(1, 60) = 1.9$, $p > 0.1$). Words containing the old category (90%) were more accurately recognized than those with a new (88%) category, a non-significant trend ($p > 0.1$). Conversely, non-words containing a new category were more accurately rejected (74%) than those with an old category (70%), again a non-significant trend ($p > 0.1$). This pattern indicates that advanced learners have perhaps resolved a former asymmetry, which is suggested by the accuracy pattern which trends towards the order of accuracy observed in the intermediate learners group.

Native speakers show a different pattern. Again, accuracy was higher for words (96%) than for non-words (91%, $F(1, 51) = 12.4$, $p < 0.01$), and there was no effect of *vowel* ($F(1, 51) = 1.5$, $p > 0.1$). The interaction was marginal ($F(1, 51) = 3.4$, $p = 0.07$), but not due to the same pattern of accuracy found in the learner groups. Rather, as shown in Figure 4, non-words containing a new category were *less* accurately rejected (89%) than those with an old category (93%, $p = 0.04$), the opposite pattern than the one observed for learners. For words, there was no difference. While the difference found for non-words clearly drives the marginal interaction, the fact that it is against the predicted accuracy pattern does not make the presence of this interaction problematic for our hypothesis. This pattern suggests that there is no such asymmetry in native speakers' lexical representations for each type of vowels as the one found for intermediate learners.

4.3. Discussion: Experiments 3 and 4

The results of Experiment 3 demonstrated the overall highly accurate performance of both learner groups in the ABX categorization task, and indicate that the learners in this experiment were able to clearly perceive a difference between back and front rounded German vowels.

In Experiment 4, where the same learners were asked to recognize real German words, we observed a very high error rate on non-words that contain difficult phonemes: just as was the case for the long obstruent consonants [p:, t:, k:] in Japanese, this was observed again for the front rounded vowels (e.g., [y, ø]) in German. This high error rate could be due to a generalized bias to say "yes" in this lexical decision experiment, but this is ruled out by the performance on control items, for which accuracy was much more similar for the word and non-word conditions than for test items. Another possible explanation is based on the fact that the test items contain difficult contrasts whereas the control items do not, perhaps implying that listeners did not actually perceive the non-words containing the difficult contrasts accurately. However, the fact that phonetic perception was rather accurate in most conditions in these same listeners (Experiments 1 and 3) makes the possibility that these non-words were not perceived accurately very remote. Therefore, a more likely explanation for the high error rate in rejecting these non-words is that the lexical representations which are being contacted by the stimuli do not allow their rejection as easily in the case of non-words. The results of Experiment 4 again reproduced the asymmetry of Experiment 2, suggesting that learners are able to maintain a contrast, at the lexical level, between lexical representations containing the German categories under scrutiny.

Combined results from Experiments 3 and 4 reveal a striking picture: given accurate phonetic discrimination between both back and front rounded vowels in German, the asymmetric pattern we found in lexical decision suggests that the lexical encoding of new vowels categories is not target-like. The asymmetry itself shows that, early on, learners are able to establish a contrast at the lexical level, but their specific encoding of the phonological contrast in these lexical representations is not fully target-like and most likely references L1 categories, as suggested by Cutler et al. (2006).

Interestingly, the interaction was not significant for advanced learners, even though there was a trend for accuracy rates to conform to the asymmetrical pattern. This suggests that perhaps advanced learners have recovered from asymmetric lexical access because their representations for each category are now less dependent on L1/dominant categories.

5. General Discussion

The goal of this study was to expand on recent reports of asymmetries in lexical access for L2 learners, and to explicitly investigate whether lexical representations are target-like or not, even if they are separate. We examined the degree to which a novel contrast is target-like in learners' lexical representations by looking at asymmetries in lexical decision patterns, combined with phonetic categorization tasks, in two different languages and different groups of L2 learners. This study is the first to examine both perceptual categorization ability and asymmetries in lexical access within the same listeners. By establishing categorization and lexical decision patterns in the same participants, our experiments allowed us to tease apart the level at which the contrast is not adequately represented: phonetic or lexical. The findings are very clear and consistent in two very different learner populations. Asymmetric lexical decision patterns, coupled with highly accurate categorization performance, suggest that L2 learners' lexical encoding of difficult phonological contrasts remains imprecise for some time, and offers rather clear support for the *lexical coding deficiency* hypothesis.

We were able to replicate and strengthen previously reported findings of asymmetric mapping from phonetic to lexical representations (Weber & Cutler, 2004, and Cutler, Weber & Otake, 2006) during lexical processing, and we also expand these findings to two different languages, German and Japanese, and with two different sets of contrasts: vowels and consonants. We also used a different method, lexical decision, which strengthens the robustness of the findings.

Our findings indicate that L2 learners' lexical representations are, in fact, quite detailed, even if their lexical encoding of difficult contrasts still makes

reference to dominant categories due to L1 influence. This conclusion contrasts with findings of homophonous lexical representations, for instance with those of Pallier et al. (2001), or with some findings reported in Darcy et al. (2012). While it is possible that these studies were unable to uncover asymmetries because of their specific experimental design, it is also possible that such cases where lexical representations fail to clearly separate contrasts exist (see also Ota et al., 2009, for further evidence). The question then arises whether it is more difficult to establish separate lexical representations for certain contrasts than for others, perhaps because of how well they map onto L1 categories. In other words, if a contrast cannot be mapped as a dominant vs. non-dominant category in a given L1 (perhaps because both phones are allophones or equally good exemplars of the L1 category, as is likely the case for the Catalan /e/-/ɛ/ contrast for Spanish listeners; Bosch, Costa & Sebastián-Gallés, 2000), it is possible that establishing a lexical contrast in any form will be more difficult. Purely phonetic mapping explanations aside, it is also possible that the degree to which explicit instruction, metalinguistic representations, and/or orthographic support are available for a given contrast or linguistic dimension will influence the difficulty with which a lexical contrast can initially be established.

In addition, this study is also the first to provide new evidence that such asymmetries can be resolved with more experience in an L2. In the case of learners of Japanese (experiments 1 and 2), both intermediate and advanced learners displayed asymmetric mapping, suggesting that the phonological geminate/singleton contrast is not fully accurately encoded lexically, even though contrast is maintained between the two. In the case of learners of German however (experiments 3 and 4), advanced learners did not show any significant interaction, which we interpreted as the absence of asymmetry in lexical processing, suggesting that they were in the process of establishing a more efficient, native-like lexical access, and that their lexical representations were gradually encoding all phonological contrasts accurately.

This difference between Japanese and German learners can be due to a number of factors, such as the contrast examined: a consonantal contrast involving duration, and a vocalic contrast involving mainly spectral differences. As mentioned above, it is possible that different contrasts lead to different acquisition speeds. Another possibility is that the learner groups are different. They cannot be directly compared, even though all the learners recruited for these studies were from very similar populations, mainly young college-educated adults. The advanced learners of Japanese might be slightly less advanced than the advanced learners of German. Their overall intensity of exposure to the language might differ, and class-level is obviously not a sufficiently objective basis for comparison. Interestingly, the advanced

learners of Japanese and of German in our study had spent overall a similar amount of time abroad (17.8 vs. 15.1 months) on average. It is striking that the advanced learners of German appeared to resolve the asymmetry, while the advanced learners of Japanese in our study (or the advanced learners of English in Weber and Cutler, 2004) did not. Nevertheless, it is possible that the length of time spent abroad is an important factor in acquisition.

Finally, while phonetic accuracy during perception is ultimately important for efficient word recognition in an L2, it is clear from our data and others that even a high accuracy in phonetic categorization will not guarantee accurate lexical encoding of a difficult contrast. The exact mechanisms L2 learners used to resolve initial asymmetries, as well as the factors that facilitate this acquisition, remain mysterious. Our experiments do not allow us to make strong claims about this point yet. A logical possibility is that length or intensity of exposure and perhaps vocabulary size in the L2 eventually promote accuracy, including the ability to lexically encode the new category in a target-like fashion. Similarly, orthographic and explicit instruction might provide first indications to learners to bootstrap the contrast separation lexically – but we still need to understand how phonological representations at the lexical level are updated. Indeed, a consequence of such lexical "fuzziness" is to increase lexical competition during spoken word recognition in L2 learners (Broersma, 2012; Broersma & Cutler, 2011). Thus, further studies should investigate what triggers phonological update at the lexical level in the course of second language learning, in order to understand when and how unwelcome lexical competition can be reduced, that is, when and how the input /konig/ stops being an acceptable rendition of the word *König*.

References

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX Lexical Database (CD-ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Banno, E., Ohno, Y., Sakane, Y., and Shinagawa, C. (1999). *Genki: An Integrated Course in Elementary Japanese [volume 1,2]*. Tokyo: Japan Times.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second language speech perception. Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13-34). Philadelphia: John Benjamins.

Bosch, L., Costa, A., & Sebastián-Gallés, N. (2000). First and second language vowel perception in early bilinguals. *European Journal of Cognitive Psychology*, *12*, 189–221.

Broersma, M. & Cutler, A. (2008). Phantom word activation in L2. *System, 36*, 22-34.

Broersma, M. (2002). Comprehension of non-native speech: Inaccurate phoneme processing and activation of lexical competitors. In *Proceedings of the 7th international conference on spoken language processing* (pp. 261–264).

Broersma, M. (2012). Increased lexical activation and reduced competition in second-language listening. *Language and Cognitive Processes*, *27*, 1205-1224

Broersma, M., & Cutler, A. (2011). Competition dynamics of second-language listening. *The Quarterly Journal of Experimental Psychology*, *64*, 74-95.

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. A., Glover, J., Kaden, C., McGuire, M., & Scott, J. H. G. (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.

Dijkstra, A. & Van Heuven, W. J. B. (1998). The BIA model and bilingual word recognition. In J. Grainger & A.M. Jacobs (eds.), *Localist connectionist approaches to human cognition* (pp. 189-225). Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12*, 627-635.

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows Display Program with Millisecond Accuracy. *Behavior Research Methods Instruments and Computers, 35*, 116-124.

Frauenfelder, U. H., Scholten, M., & Content, A. (2001). Bottom-up inhibition in lexical selection: Phonological mismatch effects in spoken word recognition. *Language and cognitive processes, 16*, 583-607.

Gaskell, M. G., & Marslen-Wilson, W. D. (2001). Lexical Ambiguity Resolution and Spoken Word Recognition: Bridging the Gap. *Journal of Memory and Language, 44*, 325-349.

Han, M. (1992). The timing control of geminate and single stop consonants in Japanese: a challenge for non-native speakers, *Phonetica*, *49*, 102 - 127.

Hardison, D. M., & Motohashi Saigo, M. (2010). Development of perception of second language Japanese geminates: Role of duration, sonority, and segmentation strategy. *Applied Psycholinguistics*, *31*, 81-99.

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

Jared, D., & Szucs, C. (2002). Phonological activation in bilinguals: Evidence from interlingual homograph naming. *Bilingualism: Language and Cognition, 5*, 225-239.

Ju, M., & Luce, P. A. (2004). Falling on sensitive ears: constraints on bilingual lexical activation. *Psychological Science, 15*, 314-318.

Ju, M., & Luce, P. A. (2006). Representational specificity of within-category phonetic variation in the long-term mental lexicon. *Journal of Experimental Psychology: Human Perception and Performance, 32*, 120-138.

Marian, V., & Spivey, M. J. (2003). Competing activation in bilingual language processing: Within- and between-language competition. *Bilingualism: Language and Cognition, 6*, 97-115.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*, 71-102.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86*, B33-B42.

Ota, M., Hartsuiker, R. J., & Haywood, S. L. (2009). The KEY to the ROCK: Near-homophony in nonnative visual word recognition. *Cognition, 111*, 263-269.

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

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

Polka, L. (1995). Linguistic influences in adult perception of non-native vowel contrasts. *Journal of the Acoustical Society of America, 97*, 1286-1296.

Sebastian-Gallés, N. (2005). Cross-Language Speech Perception. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 546-566). Malden, MA: Blackwell Publishing.

Sebastián-Gallés, N., Echeverría, S., & Bosch, L. (2005). The influence of initial exposure on lexical representation: Comparing early and simultaneous bilinguals. *Journal of Memory and Language, 52*, 240-255.

Sebastián-Gallés, N., Rodríguez-Fornells, A., de Diego-Balaguer, R., & Díaz, B. (2006). First- and Second-language Phonological Representations in the Mental Lexicon. *Journal of Cognitive Neuroscience, 18*, 1277-1291.

Spivey, M. J., & Marian, V. (1999). Cross Talk Between Native and Second Languages: Partial Activation of an Irrelevant Lexicon. *Psychological Science, 10*, 281-284.

Strange, W., & Shafer, V. (2008). Speech perception in second language learners. The re-education of selective perception. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and Second Language Acquisition* (pp. 153-191). Philadelphia: John Benjamin.

Strange, W., Weber, A., Levy, E., Shafiro, V., Hisagi, M., & Nishi, K. (2007). Acoustic variability within and across German, French, and American English vowels: Phonetic context effects. *The Journal of the Acoustical Society of America, 122*, 1111-1129.

Sumner, M., & Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language, 60*, 487-501.

Tajima, K., Kato, H., Rothwell, A., Akahane-Yamada, R., and Munhall, K. G. (2008). Training English listeners to perceive phonemic length contrasts in Japanese, *Journal of Acoustical Society of America, 123*, 397-413.

Trofimovich, P., & John, P. (2011). When three equals tree. In P. Trofimovich & K. McDonough (Eds.), *Applying priming methods to L2 learning, teaching and research: Insights from Psycholinguistics* (pp. 105–129). Philadelphia, PA: John Benjamins.

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

Acknowledgements:
We thank John H.G. Scott for creating the *Märchenkrimi* and for substantial help with data collection and coding, as well as Christiane Kaden, Franziska Krüger, Justin Glover for help with running participants, Laurent Dekydtspotter, Rex Sprouse for discussion and feedback, Stephanie Dickinson and Pan Yi from the IU Statistical Consulting Center for their invaluable help with statistical analysis, the Department of Germanic Studies and the Department of Second Language Studies at Indiana University, and the Second Language Psycholinguistics Lab members for comments, help and support. We also thank audiences at the 18th Germanic Linguistics Annual Conference (2012), the 8th International Conference on the Mental Lexicon (2012), the Second Language Research Forum (2012), as well as the Princeton Japanese Pedagogy Forum (2012).