Difficulty in learning similar-sounding words: a developmental stage or a general property of learning?

Bozena Pajak[1]

Duolingo, Inc.

Sarah C. Creel and Roger Levy

University of California, San Diego

13799 words (text + references + appendices)

Author Note

For their helpful feedback the authors thank Eric Bakovic, Klinton Bicknell, the Computational Psycholinguistics Lab at UC San Diego, the audience of LSA 86, and four anonymous CogSci 2012 reviewers. Rafi Feliciano, K. Michael Brooks, and Hannah Byers-Straus helped with data collection. We are grateful to Eugene Carsey for permission to use his photographs of mushrooms (www.eugenecarsey.com). This research was supported by NIH Training Grant T32-DC000041 from the Center for Research in Language at UC San Diego to the first author, NIH Training Grant T32-DC000035 from the Center for Language Sciences at University of Rochester to the first author, and research grants from the UC San Diego Academic Senate and the Alfred P. Sloan Foundation to the third author.

[1] Address correspondence to Bozena Pajak, Duolingo, Inc., 5533 Walnut Street, 3rd floor, Pittsburgh, PA 15232, USA. Phone: (+1) 412-567-6602. Fax: (+1) 412-436-0631. E-mail: bozena@duolingo.com

Abstract

How are languages learned, and to what extent are learning mechanisms similar in infant native-language (L1) and adult second-language (L2) acquisition? In terms of vocabulary acquisition, we know from the infant literature that the ability to discriminate similar-sounding words at a particular age does not guarantee successful word-meaning mapping at that age (Stager & Werker, 1997). However, it is unclear whether this difficulty arises from developmental limitations of young infants (e.g., poorer working memory) or whether it is an intrinsic part of the initial word learning, L1 and L2 alike. Here we show that adults of particular L1 backgrounds—just like young infants—have difficulty learning similar-sounding L2 words that they can nevertheless discriminate perceptually. This suggests that the early stages of word learning, whether L1 or L2, intrinsically involve difficulty in mapping similar-sounding words onto referents. We argue that this is due to an interaction between two main factors: (1) memory limitations that pose particular challenges for highly similar-sounding words, and (2) uncertainty regarding the language's phonetic categories, as these are being learned concurrently with words. Overall, our results show that vocabulary acquisition in infancy and in adulthood share more similarities than previously thought, thus supporting the existence of common learning mechanisms that operate throughout the lifespan.

Keywords: word learning, spoken word recognition, non-native speech perception, second language acquisition

Difficulty in learning similar-sounding words: a developmental stage or a general property of learning?

Humans are able to learn languages throughout their lifespan. But how similar are the learning mechanisms for infants acquiring their native language (L1) and adults learning a second language (L2)? There has been little work trying to connect these two literatures, reflecting the underlying assumption of a lack of developmental continuity in terms of language learning (see, for example, discussion in White, Yee, Blumstein, & Morgan, 2013). Instead, infants and adults have been assumed to use qualitatively different mechanisms to process and learn languages, largely following the critical period hypothesis (Lenneberg, 1967; Johnson & Newport, 1989). However, recent work has shown that while age of L2 acquisition negatively correlates with achieved proficiency, there are signs of developmental continuity in language learning and similarities between infant and adult acquisition (Birdsong, 2009; Birdsong & Molis, 2001; Hakuta, Bialystok, & Wiley, 2003; Werker & Tees, 2005; White et al., 2013). For example, it has been shown that infants and adults rely on similar statistical learning mechanisms to segment words out of a continuous speech stream (e.g., Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996) or to learn phonetic categories (e.g., Maye & Gerken, 2000; Maye, Weiss, & Aslin, 2008; Maye, Werker, & Gerken, 2002; Pajak & Levy, 2011; for a review and further discussion see Pajak, Fine, Kleinschmidt, & Jaeger, 2015), and are similarly affected by word familiarity during lexical processing of newly learned words (White et al., 2013). White et al. argued that these parallel results for infants and adults might reflect common mechanisms that operate throughout development, thus highlighting the need for greater interaction between the infant and the adult language learning literatures.

Here we pursue a comparison between infant and adult language learning by considering one aspect that is crucial at the initial stages of acquisition: the encoding of phonetic detail during word learning. Learning words requires not only remembering a label for a given referent, but also forming a phonetically-rich representation of that label by segmenting the word into individual sounds. The detailed phonetic representation is especially important for similar-sounding words (e.g., *bin* vs. *pin*), because successful learning crucially relies on the ability to distinguish between the words based on subtle acoustic-phonetic cues. Thus, the learner must be able to perceptually discriminate the sounds that distinguish between the words (e.g., [b] vs. [p]), but also to ignore any irrelevant variability between instances of the same phonetic category (e.g., multiple exemplars of the word *bin*). The ability to discriminate among similar sounds is thus a necessary condition for successfully learning words distinguished by those sounds. But is it a sufficient condition? In this paper we investigate this question for adult learners, by taking advantage of the influence of L1 background (here, Mandarin and Korean) on adult perceptual discrimination abilities. We examine to what extent the L1-driven differences in adult speech sound discrimination are associated with differences in word learning ability, building on a small existing body of work in this area (Creel & Dahan, 2010; Creel, Aslin, & Tanenhaus, 2006; Silbert, Smith, Jackson, Campbell, Hughes, & Tare, 2015).

In the case of infant language learning, we know from the literature that discrimination does not guarantee successful learning of similar-sounding words: despite the ability at age 14 months to perceptually discriminate between similar sounds (e.g., *b* and *d*), 14-month-olds have been shown to confuse newly-learned words differentiated by those sounds (e.g., *bih* and *dih*; Pater, Stager, & Werker, 2004; Stager & Werker, 1997), unless there is additional contextual information, or less demanding learning conditions (Ballem & Plunkett, 2005; Fennell, Waxman,

& Weisleder, 2007; Fennell & Werker, 2003; Rost & McMurray, 2009; Swingley & Aslin, 2002; Thiessen, 2007; Yoshida, Fennell, Swingley, & Werker, 2009). The initial explanation proposed for this result was a limited resource hypothesis (Stager & Werker, 1997; Werker, Fennell, Corcoran, & Stager, 2002): since attending to fine phonetic detail while learning new words is computationally very demanding, young infants—who have limited attentional and cognitive resources—might have difficulty accessing full phonetic detail when focusing their attention on learning meaning. Other explanations have emphasized the role of increased lexical competition in learning similar-sounding words (Swingley & Aslin, 2002, Swingley & Aslin, 2007), or suggested that the difficulty might arise from poorly defined phonetic category boundaries at that stage of infant development (Rost & McMurray, 2009) and limited experience with phonological categorization (Yoshida et al., 2009).

Regardless of the exact explanation, the consensus is that children outgrow this initial difficulty, and by 17-20 months of age succeed at learning new similar-sounding words (Werker et al., 2002). However, despite this acquired sensitivity to minimal differences between words in the learners' L1, phonological similarity continues to play a role in lexical processing in both older children and adults. This is indicated, for example, by robust and automatic activation of words that sound similar to the target word (e.g., Andruski, Blumstein, & Burton, 1994; Allopenna, Magnuson, & Tanenhaus, 1998; Magnuson, Dixon, Tanenhaus, & Aslin, 2007; Mani & Plunkett, 2011; Swingley & Aslin, 2000; White & Morgan, 2008). Adults are also slower at processing words that have a high neighborhood density (i.e., have a large number of similar-sounding words, generally defined as a one-phoneme distance; e.g., Luce & Pisoni, 1998) compared to words in sparse lexical neighborhoods (e.g., Vitevitch & Luce, 1998), and have increasing difficulty distinguishing phonologically native-like nonsense words as the word similarity increases (Creel & Dahan, 2010; Creel et al., 2006). All these results suggest gradient effects of phonological similarity, where the encoding and the retrieval of similar-sounding minimal-pair words is impaired relative to dissimilar words. (But see, e.g., Storkel, 2004; Storkel, Armbruster, & Hogan, 2006 for evidence that children and adults learn new dense-neighborhood words in their native language more readily than new sparse-neighborhood words, suggesting that partial phonological overlap with known words may help strengthen newly formed lexical representations.)

Thus, both children and adults are known to have difficulty learning novel similar-sounding words whose phonological form resembles their native language. These results do not, however, answer the question of whether adults are affected by phonological similarity during learning of an *unfamiliar* language, a situation more parallel to the case of 14-month-old infants learning their native language.

As we mentioned earlier, it is known that adult L2 learners have extreme difficulty distinguishing—and therefore also learning—similar-sounding words that involve novel sound contrasts not found in their native language, such as *rake* vs. *lake* for native speakers of Japanese (e.g., Escudero, Broersma, & Simon, 2013; Escudero, Hayes-Harb, & Mitterer, 2008; Hayes-Harb & Masuda, 2008; Weber & Cutler, 2004). In those cases, L2 learners have to override the L1 phonetic-category information that is incompatible with the L2 information (e.g., the acoustic-phonetic range occupied by the English 'r' and 'l' roughly corresponds to a single category in Japanese; e.g., Miyawaki, Strange, Verbrugge, Liberman, Jenkins, & Fujimura, 1975; see also, e.g., Escudero, Simon, & Mulak, 2014 for how orthography may help or hinder learning in these cases). Indeed, it is already known that, within a given L1, listeners' ability to discriminate a non-native contrast predicts how well they learn words that differ by that contrast (Silbert et al., 2015). It is also known that individual differences in learning may in these cases arise from variability in purely auditory abilities (Kidd, Watson, & Gydi, 2007), as well as variability in phonological short term memory (Silbert et al., 2015). We know less, however, about listeners' overall abilities—as a group—to learn similar-sounding words when (i) they can reliably perceive the perceptual contrast, and (ii) the L1 phonetic-category information does not strongly interfere with L2 perception (cf. Pajak & Levy, 2014). This is the topic of the present paper. That is, instead of trying to predict *an individual's* ability to learn words from that individual's ability to discriminate those words, as in prior work (e.g., Silbert et al., 2015), we investigate the relationship between perceptual discrimination and word learning in listeners as a function of one of two different native language backgrounds. We use two native-language populations which we know have complementary expertise in perceptual discrimination: Mandarin speakers, who are sensitive to non-native sibilant place-of-articulation distinctions, and Korean speakers, who are sensitive to non-native consonant length distinctions (Pajak & Levy, 2014).

In particular we examine two specific questions. First, is the mismatch between discrimination and word learning a developmental phenomenon, or is it driven by the information being learned? That is, will adult L2 learners—who, like infants, are concurrently learning the language's phonological categories, but unlike infants, have vastly greater working memory capacity—show greater difficulty in word-learning tasks relative to discrimination tasks when acquiring new L2 vocabulary? Second, how does phonological similarity moderate discrimination vs. learning of similar-sounding L2 words?

Answers to these questions for the two different L1 populations we study can potentially provide a key missing link connecting theories of adult and infant language learning. If adults of a particular L1 background—just like 14-month-old infants—are found to have difficulty learning similar-sounding words that they can nevertheless discriminate perceptually, then this would provide evidence in favor of the existence of common language learning mechanisms that operate throughout development and into adulthood. Furthermore, a more detailed examination of the role of phonological similarity in word discrimination and learning, and how similarity interacts with different task demands, can help us shed more light onto the nature of those common mechanisms that underlie language learning.

Before we continue, however, we first examine the factors that might potentially contribute to the difficulty of learning the correct label/referent pairing. One such factor is that beginner learners, both infant and adult, might have noisy phonetic representations, reflecting low confidence in the fidelity of phonetic encoding of the newly-learned words or in the exact location of phonetic category boundaries (Rost & McMurray, 2009; Yoshida et al., 2009). Another factor might involve task-specific memory limitations, which pose particular challenges for highly similar-sounding words. Below we describe in more detail how these factors might lead to potential difficulty in word learning relative to discrimination.

A conceptual model of word-referent mapping difficulties

Learners acquiring words need to rely on their memory representations of label/referent pairs, where each label can be described as a sequence of sounds sampled from the language's phonetic categories. Precise encoding of the label’s phonetic form thus requires establishing what categories the sounds were sampled from, a difficult task at the early stages of language learning. Word learning is then likely affected by two sources of noise: (1) noise and uncertainty associated with categorization of each individual sound, and (2) noise associated with retrieving a memory trace of the phonetic input and of the label/referent pairing. Thus, one way of thinking about the difference between the discrimination and the word-learning tasks is that the quality of phonetic representations for individual input exemplars is lower for word learning than for discrimination due to heavier long-term memory demands in the word-learning task: in word learning, listeners have to simultaneously keep track of the referent and try to form phonetic representations, while a discrimination task only requires comparing short-term memory traces of phonetic input, without the need to link them to referents. As more data are obtained over time, successful word learning requires integrating multiple memory traces to arrive at the correct label/referent pairing. This would correspond to narrowing down the effective variance around the memory representation of the label's phonetic form.

In cases when the words being learned are composed of highly dissimilar sequences of sounds—that is, the phonetic distance between sound categories is large relative to the variance of the label's phonetic form—the learner's performance should not be impeded because a small number of samples would be sufficient to learn the distinctions among the categories. However, when the words are highly similar—that is, the phonetic form variance is large relative to the distance between sound categories—it should be much harder for learners to separate the sounds into categories and pinpoint the right label/referent pairings. It is expected that much more data (i.e., more learning instances) is needed in this case before learners can accumulate a sufficient number of exemplars to learn phonetic category distinctions and form correct label/referent mappings. Furthermore, if the learner is unable to integrate the information accumulated from a number of exemplars of the label/referent pairing, he or she will have great difficulty in learning the pairing reliably at all.

This conceptual model is consistent with prior suggestions that additional cognitive load, such as simultaneous presentation of visual stimuli, lowers the resolution of auditory processing of phones (Mattys & Palmer, 2015). Such lower-resolution processing may be due to missing some temporal pulses in the auditory signal (Casini, Burle, & Nguyen, 2009) or to reduced cochlear sensitivity (Lukas, 1980; Puel, Bonfils, & Pujol, 1988) during concurrent attention to visual stimuli. Interestingly, perceptual sensitivity seems to linearly decrease as the effort involved in the simultaneous visual task increases (Mattys, Barden, & Samuel, 2014). This type of disruption in auditory processing may be understood as an increased tolerance to imprecise acoustic encoding, and, as a result, to an increased perceptual overlap between similar-sounding phones (Mattys et al., 2014; Mattys & Palmer, 2015).

The current study

In the current study, we examine how adults learn vocabulary in a new language that is phonologically unfamiliar (i.e., an L2), but composed of discriminable speech-sound categories. In particular, we compare two populations of participants with differential perceptual sensitivities to certain speech sound contrasts that are due to their different L1 backgrounds. We test one participant group on a discrimination task and another group on a word-learning task, and examine whether the known L1-background-driven differences in sound discrimination are also observed in the word-learning task when participants learn words that differ by those sounds (as described in more detail below). The situation of learning phonologically novel words that are similar-sounding, but that adults can nevertheless discriminate perceptually, is analogous to the situation of 14-month-old infants observed in Stager and Werker (1997). This allows us to assess whether the good-discrimination-without-learning pattern observed in Stager and Werker (1997) reflects a purely developmental phenomenon, or instead reflects general mechanisms of (language) learning. In addition, we include multiple sets of word pairs that differ in their degree of similarity, which lets us investigate how phonological similarity modulates discrimination and learning of L2 words.

We constructed a miniature language with pairs of words at three levels of similarity: (1) *dissimilar* (e.g., [tala]-[kenna]), (2) *similar* (e.g., [tala]-[taja]), and (3) *highly-similar*, where the words differed either in consonant *length* (e.g., [taja]-[tajja]) or in *place* of articulation between alveolo-palatal and retroflex sounds (e.g., [gotɕa]-[gotʂa]). We chose the *length* and the *place* dimensions because they have been shown to be differentially discriminable by two different L1-speaker populations: L1-Korean and L1-Mandarin (Pajak, 2012; Pajak & Levy, 2014). In particular, Korean speakers have an advantage over Mandarin speakers in discriminating consonant length contrasts, while Mandarin speakers have an advantage over Korean speakers in discriminating alveolo-palatal and retroflex consonant contrasts.[2] Therefore, we can investigate whether these differential L1-based perceptual advantages on *highly-similar* word pairs occur not only in discrimination but also in word learning. Note that we are not asking whether performance on the discrimination task *predicts* performance in the word-learning task at the individual level. Rather, we are investigating the question of whether between-group differences in discrimination ability arising from differences in native language would also be reflected in between-group differences in word-learning ability. There might certainly be some individual variation in how well learners take advantage of their L1-based perceptual abilities when learning words, but the group-level comparisons reveal the overall trends in the population as a whole, and this is the question that we are addressing here. We did not expect differences in word-learning performance between L1-Korean and L1-Mandarin participants for *dissimilar* and *similar* items because those contrasts were acoustically more salient (relative to the contrasts in *highly-similar* pairs) and there is no reason to believe that participants' language background would affect their discrimination in a differential way. Crucially, the language was phonologically novel to all participants in that all phonetic properties of the stimuli (e.g., voice onset time, vowel quality, stress, etc.) were taken from an unrelated language, Polish. Therefore,

[2] Pajak & Levy (2014) argued that these differential perceptual sensitivities follow from the fact that Korean has some length distinctions, while Mandarin has none, whereas Mandarin has some alveolo-palatal and retroflex sounds, while Korean only has alveolo-palatal but no retroflex sounds (Lin, 2001; Sohn, 1999).

this scenario was more comparable to the situation of the 14-month-old infants than many previous studies in which adults learn phonologically native-like words.

We had two main sets of predictions. The first concerned the overall effect of phonological similarity on discrimination vs. learning of L2 words that are all discriminable by learners, either because the differences are salient (*dissimilar*, *similar*) or because a related distinction is used in the learners' L1 (*highly-similar*: *length* for Korean, *place* for Mandarin). Given the gradient acoustic similarity between the different sets of word pairs, we expect discrimination to also be gradient: best for *dissimilar* words, intermediate for *similar* words, and poorest for *highly-similar* words. As for word learning, prior work using native-like words has shown gradience in performance as a function of words' phonological similarity (e.g., Creel & Dahan, 2010; Creel et al., 2006). But how does phonological similarity interact with word learning in the case of learning a new language with overall non-native phonology? If it works similarly to learning vocabulary in an L1 phonological system, then we would expect learners' performance to change as a function of similarity between the word pairs, matching the discrimination performance: best when identifying the referent in the context of two *dissimilarly*-named possible referents, intermediate for *similarly*-named referents, and poorest for *highly-similar* pairs. On the other hand, it is possible that learning words with an unfamiliar phonology is not affected by similarity in the same way that discrimination is. That is, we might expect a mismatch between the discrimination vs. the word-learning task: for example, gradient performance for the participants in the discrimination task, but no differences in performance for the participants identifying word referents; or more exaggerated gradient effects in one task than the other.

Our second set of predictions concerned *highly-similar* words. These predictions are in some sense a more focused version of the first set of predictions, as both are examining the influence of perceptual similarity. However, here we focus on the *highly-similar* words whose underlying speech sound categories are acoustically overlapping, thus being most comparable to infants' nascent speech sound categories. More specifically, we compare two populations of speakers with differential perceptual sensitivities to these word differences: Korean speakers better on *length*, and Mandarin speakers better on *place*. We make the Korean- vs. Mandarin-speaker comparison for two different tasks, word learning and discrimination, where each task was completed by independently recruited subjects. Therefore, we test word learning vs. discrimination of *highly-similar* words that either (a) are relatively easily discriminable because the differences are based on a phonetic dimension informative in L1 (even though the learners do not actually know any words distinguished by some of the specific sounds used, thus resembling the situation of young infants), or (b) are not easily discriminable because the differences are based on a phonetic dimension not informative in L1. The contrasts in (a) are *length* for Korean speakers and *place* for Mandarin speakers, while the contrasts in (b) are *place* for Korean speakers and *length* for Mandarin speakers.

First of all, we expect the word learning task to be overall harder than discrimination, which means that we should find overall worse performance for the group performing the word learning task. However, overall task differences do not tell us anything about participants' *use* of their native-language-based perceptual abilities in word learning. Our discrimination task is aimed to replicate prior results (Pajak, 2012; Pajak & Levy, 2014)—with materials that are more comparable to our word-learning task materials—and it should show us how much perceptual advantage Korean speakers have over Mandarin speakers in *length* contrasts, and how much perceptual advantage Mandarin speakers have over Korean speakers in *place* contrasts. Comparing the *degree* of this group-level asymmetry in the discrimination vs. the word-learning tasks will let us assess how much participants are using their L1-based perceptual abilities in word learning. Therefore, we expect both Korean and Mandarin speakers to perform worse on highly-similar trials in word learning than in discrimination (i.e., overall task differences). However, the critical question is the following: Is the relative difference between the two L1 populations (Korean better at *length*; Mandarin better at *place*) also observed in the word-learning task? If it is, then it would suggest that L1-based perceptual abilities are used in word learning (regardless of whether the overall performance is lower than in discrimination). If it is not, then it would suggest that the word-learning task makes it difficult for participants to use their L1-based perceptual abilities. The extreme version of the latter would be no group-level difference between Korean and Mandarin speakers on length or place in the word-learning task, showing that *the whole* relative perceptual advantage observed at the group level has been eliminated during word learning. It is also the latter case (no use of perceptual advantages evident during word learning) that would be most analogous to the results reported for L1-learning 14-month-olds.

Method

Participants. Ninety undergraduate students at UC San Diego participated in the experiment for course credit or payment. Half were speakers of Korean, and the other half were speakers of Mandarin. We recorded participants' language background information, including self-reported proficiency in both Korean/Mandarin and English, current language exposure, as well as a measure of English proficiency through the Shipley Vocabulary Test (Shipley, 1967).

Participants varied in terms of their length of residence in the US: some were born in the US, while others immigrated at some point after birth or were international students who arrived very recently. Consequently, they varied in English proficiency. Importantly, however, they all learned Korean or Mandarin from birth, reported high proficiency in those languages, and still used them regularly, predominantly with family. In most cases they had some high school and/or college exposure to Spanish or French. Some Mandarin speakers were also familiar with Taiwanese Hokkien, mostly through family exposure. No exposure to any other languages was reported, including Polish, the language that provided source material for novel words. All participants reported no history of speech or hearing problems. We collected individual measures of participants' nonverbal IQ using the Matrices subtest of the Kaufman Brief Intelligence Test (KBIT-2, Kaufman & Kaufman, 2004). All participant characteristics are shown in Table 1. More detailed comparisons between participants depending on language background are provided in Appendix A.

[insert Table 1 about here]

In order to avoid potential carryover effects from one task to the other task, we chose to test discrimination and word learning in a between-participant design, investigating the question whether between-L1-population differences in discrimination ability arising from differences in native language would also be reflected in between-L1-population differences in word-learning ability. Task cross-contamination in a within-subjects design would be a serious barrier in interpreting the results: whichever order of tasks we would choose, participants would be biased in the second task because (1) their attention would be directed to the tested contrasts, and (2) they would have received a great deal of perceptual exposure to those distinctions. In fact, there is ample evidence from the perceptual learning literature that even relatively brief exposure can affect adults' perception (e.g., Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Kraljic & Samuel, 2005; Norris, McQueen, & Cutler, 2003). This design choice is analogous to infant studies, where the results regarding the dissociation of discrimination and word-learning abilities are based on differences between groups.

Fifty-four participants were assigned to the word-learning task, and 36 to the discrimination task.[3] In each group, half were speakers of Korean, and the other half were speakers of Mandarin. Comparing participants assigned to the discrimination vs. the word-learning task revealed no significant differences in any of the measures we collected, as shown in Table 1.

Materials. The materials consisted of 16 bisyllabic consonant-vowel-consonant (CVCV) nonce words (see Table 2; a subset of contrasts tested by Pajak, 2012; Pajak & Levy, 2014). The words were constructed in such a way that there were 8 minimal pair words differing only in the middle consonant; these were the *highly-similar* word pairs. The complete list of trial types is described below in the *Procedure* section. More specifically, the minimal pairs differed either in *length* (a short vs. a long middle consonant) or *place* of articulation (an alveolo-palatal vs. a retroflex sibilant consonant in the middle position).[4] The materials were constructed using the

[3] The difference in the number of participants in the two tasks was due to the fact that we have had more experience with discrimination experiments, and so we had a better sense of how many participants we would need in order to obtain good statistical power. Studying word learning in this type of task was relatively novel to us, and we expected more between-subject variability, which is why we decided to collect data from more participants.

[4] We chose the middle consonants in our stimuli in such a way that half of the corresponding sound distinctions had their analog in the listener's L1, and the other half did not. This was done in order to allow for a comparison in performance between distinctions that were relatively familiar to our participants from their L1s vs. completely unfamiliar distinctions that yet varied along familiar dimensions. In the previous work on *length* and *place* discrimination by Korean and Mandarin speakers (Pajak, 2012; Pajak & Levy, 2014), Korean speakers outperformed Mandarin speakers on discriminating all *length* contrasts, whether familiar or not, while Mandarin speakers outperformed Korean speakers on discriminating all *place* contrasts (note that the stimuli in that study

sound inventory and other phonological properties of Polish, and were recorded by a phonetically-trained Polish native speaker.

[insert Table 2 about here]

The inventories of Korean and Mandarin include some sound distinctions along the dimensions of *length* and *place*, respectively, that are similar but not identical to the distinctions used in the experiment. Korean employs the dimension of *length*, distinguishing between short and long sounds, but not the dimension of *place*. We found previously that Korean speakers are better than Mandarin speakers at discriminating consonant *length* contrasts, while Mandarin speakers are better than Korean speakers at discriminating alveolo-palatal vs. retroflex *place* contrasts (Pajak, 2012; Pajak & Levy, 2014). This follows from the fact that Korean has some length distinctions, while Mandarin has none; whereas Mandarin has some alveolo-palatal and retroflex sounds, while Korean does not (Sohn, 1999; Lin, 2001).

More specifically, Korean uses length distinctions mostly on vowels (e.g., [pul] ‘fire’ vs. [pu:l] ‘blow’), but some long consonants ([ll], [nn], [mm]) arise from phonological assimilation processes (Sohn, 1999), and Korean tense obstruents ([p’], [t’], [k’], [s’], [tɕ’]) have sometimes

were also based on the Polish *length* and *place* contrasts). However, there was a trend in that earlier study for both groups to perform slightly better at the distinctions familiar from their L1s compared to the unfamiliar distinctions that varied along familiar dimensions (e.g., familiar [m]-[mm] > unfamiliar [j]-[jj] for Korean speakers). Therefore, it is possible that a similar difference will hold in a word learning task: that is, both Korean and Mandarin speakers will be better at learning similar-sounding words that include familiar categories than those that include unfamiliar categories that nevertheless vary along a familiar dimension (see Table 2 for the list of word contrasts based on (a) familiar categories, where the specific distinction exists in Korean/Mandarin, and (b) unfamiliar category contrasts, but familiar phonetic dimensions). In order to obtain enough power for such a comparison, but at the same time keeping the total number of words relatively small in order to assure their learnability in a single experimental session, we decided to focus our analysis on *length* words (12 words in total), and included a much smaller number of *place* words (4 words in total). Length was chosen as the dimension of main interest because length contrasts are possible for many more types of segments than the alveolo-palatal vs. retroflex contrasts (of which Polish only has four). No difference between familiar categories vs. unfamiliar category contrasts but familiar dimension was borne out in the current results.

been analyzed as long (Choi, 1995). In terms of place of articulation, Korean has some alveolo-palatal sounds ([ɕ], [tɕ]), but no retroflex sounds, thus lacking the *place* contrast as defined in this paper.

Mandarin, on the other hand, has both alveolo-palatal and retroflex sounds that are distinguished by spectral shape in the frication noise (the *place* dimension), but does not use the *length* dimension. In particular, Mandarin has voiceless alveolo-palatals ([ɕ], [tɕ]) and retroflexes ([ʂ], [tʂ]) as allophones of the same phonemic category. In addition, the voiced retroflex fricative ([ʐ]) is a between-speaker variant of the retroflex approximant ([ɻ]). Other voiced sibilants are assumed to be absent because Mandarin has obstruent distinctions in aspiration, not in voicing (Lin, 2001). Note, however, that the analogous place distinction in Polish, which we used in the stimuli, is not exactly the same as the one in Mandarin, differing somewhat in the placement of the tongue tip.

Note that all participants spoke American English, where *length* and alveolo-palatal vs. retroflex *place* are not used contrastively. While vowel length varies in English, it correlates with other cues (e.g., the tense-lax distinction), and native speakers of English identify vowels relying predominantly on spectral properties (e.g., Hillenbrand, Clark, & Houde, 2000). Long consonants are sometimes attested, but only at morpheme boundaries (e.g., *dissatisfied*; Benus, Smorodinsky, & Gafos, 2003), and only produced as long by some speakers (Kaye, 2005). English has neither alveolo-palatal nor retroflex obstruents, although some speakers produce the alveolar approximant [ɹ] as retroflex (Ladefoged & Maddieson, 1996; Westbury, Hashi, & Lindstrom, 1998). While it is possible that the knowledge of English might affect discrimination of the Polish *length* and *place* contrasts (and, in particular, help with the *length* contrasts), Pajak and

Levy (2014) found no evidence in support of that hypothesis when testing discrimination of these contrasts by a variety of bilingual listeners (English-Korean, English-Vietnamese, English-Cantonese, and English-Mandarin), whether their dominant language was English or their native language.

The materials were recorded in a soundproof booth by a phonetically-trained native speaker of Polish. There were 10 tokens recorded for each word. For *length* words, two tokens of each word with long consonants were chosen for the experiment. Subsequently, words with short consonants were created by shortening the tokens with long consonants in a way that, for each word and each recording, the naturally-recorded long consonant was reduced to half its duration so as to maintain a constant 2:1 duration ratio (cross-linguistically, the long-to-short consonant ratio varies between 1.5 to 3; Ladefoged and Maddieson, 1996). For *place* words, given that the alveolo-palatal vs. retroflex distinction is intrinsically already very subtle, even in natural speech (Nowak, 2006), we used natural recordings of both alveolo-palatals and retroflexes with no additional manipulations. Two tokens each were chosen for the experiment with the goal of maximizing the similarity between the words in minimal pairs with regards to how vowels were pronounced, but at the same time choosing tokens with clearly enunciated sibilants. This was a departure from how the stimuli were constructed by Pajak (2012) and Pajak and Levy (2014), where both alveolo-palatals and retroflexes were spliced into an identical word frame. Pajak and Levy's procedure removed one of the cues to the contrast (the formant transition into the following vowel), thus making it extremely subtle. In the current study, we left this cue intact so that stimuli in the word-learning task were not overly difficult.

These auditory stimuli were used for both the discrimination and the word-learning task. For the word-learning task, each word was paired with a picture of a different kind of mushroom (see two examples in Figure 1), which were chosen in order to include objects that were unfamiliar to our participants, but not so unfamiliar that participants would find them bizarre and hard to remember. We selected pictures that varied in shape and color so as to maximize visual differences between them. We created four different one-to-one word-to-picture mappings that were counterbalanced between participants in order to make sure that the results were not driven by any peculiarities in the mappings we chose.

[insert Figure 1 about here]

Procedure. Participants sat in front of a computer, and responded by using a mouse. They were instructed that in this experiment they would be listening to a novel language, and, specifically, either (i) learn to distinguish this language's sounds (in the discrimination task), or (ii) learn the language's words for different types of mushrooms (in the word-learning task). The experiment was completed in a single session, and each participant took part in only one of the tasks. The discrimination and the word-learning tasks were made equal in the total auditory exposure to each stimulus in order to keep them as parallel as possible.

Discrimination task. Discrimination was tested in an ABX task. In each trial, three words were presented auditorily through headphones: A *[500ms]* B *[750ms]* X (e.g., [taja_1] [tajja_1] [taja_2]). The task was to assess whether X sounded more like A or more like B. As indicated by subscripts, the X word was always acoustically different (i.e., a physically different recording) from both A and B words to make sure that the matching of X onto A or B was not based on pure acoustical identity of two tokens. This was a different procedure than in Pajak (2012) and Pajak and Levy (2014), where an AX task was used instead. In this study, however, we wanted to maintain a close parallel between the discrimination and the word-learning tasks, which was achieved with the ABX procedure. The AB word order was counterbalanced, and the trial order was randomized for each participant. There were 4 blocks, each with 64 trials and lasting about 5 minutes. Note that this means that each block included exposure to 192 words (64 trials x 3 words per trial). Blocks were separated by self-terminated breaks. There were four types of trials depending on the AB contrast, as illustrated in Table 3: (i) *dissimilar* word pairs (e.g., [tala]-[kenna]), which differed in all sounds but the last vowel (16 trials; 8 AB word pairs x 2 trials: 1 trial with X = A and 1 trial with X = B), (ii) *similar* word pairs (e.g., [tala]-[taja]), which shared the initial CV sequence, but the middle consonants differed along multiple phonetic dimensions (16 trials; 8 AB word pairs x 2 trials: 1 trial with X = A and 1 trial with X = B), (iii-iv) *highly-similar* (e.g., *length*: [tala]-[talla] or *place*: [gotça]-[gotşa]), where the initial CV sequence was identical and the middle consonants differed minimally, either in *length* or in *place* (32 trials per block: 8 AB word pairs x 4 trials: 2 trials with X = A and 2 trials with X = B).

[insert Table 3 about here]

Word learning task. In this task, participants learned to associate auditorily presented words (1 word per trial) with pictures of mushrooms. There were 4 training blocks (each with 128 trials, about 10-15 minutes long) and 4 testing blocks (each with 64 trials, about 5 minutes long), interleaved. Thus, each train+test combination contained 128 training trials + 64 test trials = 192 auditory exposures to the words, the same number of exposures as a block of the discrimination task (64 trials x 3 words per trial = 192). Blocks were separated by self-terminated breaks. In each trial, two pictures were presented on a computer screen (see Figure 1), and after a delay of 500 ms, a word was played through headphones. Participants were asked to click on the picture that they thought went with the word. In training, feedback was provided following the response in the form of the correct picture staying on the screen. A mouse click triggered the start of the next trial. Because participants were learning via feedback presented after each response, the early responses were necessarily random. Participants were told to guess at first, and that through feedback they would eventually learn the correct word-to-picture mappings. In testing, no feedback was provided.

The training trial types consisted of picture pairs that were always associated with *dissimilar* word pairs (e.g., [taja]-[diwa], [gotça]-[kemma]; see Table 3) so that participants were not directly alerted to the distinctions of interest. Each word was played 8 times per training block (8 x 16 words = 128 total), and each time it was accompanied by a different set of two pictures. None of the training picture pairings appeared in later testing.

The testing trial types were always different from the training trials, and were completely analogous (in form and number) to trials in the discrimination task, as illustrated in Table 3. Specifically, each picture pair in the word-learning task test was an analog of an AB word pair in the discrimination task, and the auditorily presented word in the word-learning task corresponded to the X word in the discrimination task. This means that there were the following trial types: (i) *dissimilar* picture pairs (e.g., picture of [tala] and picture of [kenna]) (16 trials; 8 picture pairs x 2 trials: 1 trial where the auditorily presented word corresponded to the picture on the left, and 1 trial where the auditorily presented word corresponded to the picture on the right), (ii) *similar* word pairs (e.g., picture of [tala] and picture of [taja]) (16 trials; 8 picture pairs x 2 trials: 1 trial where the auditorily presented word corresponded to the picture on the left, and 1 trial where the auditorily presented word corresponded to the picture on the right), (iii-iv) *highly-similar* (e.g., *length*: picture of [tala] and picture of [talla], or *place*: picture of [gotçа] and picture of [gotʂa] (32 trials per block; 8 picture pairs x 4 trials: 2 trials where the auditorily presented word corresponded to the picture on the left, and 2 trials where the auditorily presented word corresponded to the picture on the right).

The picture position was counterbalanced for both training and testing trials. The trial order was pseudo-randomized: we created four randomized lists, and then altered them manually so that the same word was never repeated in two consecutive trials. Furthermore, the minimal-pair trials were always separated by at least two other trials. Each participant heard each list once, with a different list for each block. Block order was counterbalanced across participants.

Results

We investigated two main questions in our data analysis, based on the two sets of predictions that guided our study's design. First, how does phonological similarity moderate discrimination vs. learning of similar-sounding words in an L2 with clearly non-native phonology? Second, when testing phonologically highly similar word pairs, is an L1 background that confers good discrimination sufficient for good learning of similar-sounding words in adult L2 learners, or is there a disconnect between discrimination and word learning like that observed in infant L1 learners (Stager & Werker, 1997)?

To answer these questions, we analyzed accuracy scores from both discrimination and testing in word learning with mixed-effects logit models (Jaeger, 2008). Following Barr, Levy, Scheepers, & Tily (2013), who recommend maximal random-effects structures for mixed-effects models as best practice, we included random intercepts for participants and items, and random slopes for participants and items for all effects of interest (including interaction effects) that were respectively manipulated within participants or within items.[5] There was no step-wise model selection. We controlled for participants' nonverbal IQ and L1 proficiency and use (through a combined score of (a) proficiency of L1 in speaking, (b) proficiency in L1 understanding, and (c) the percent of time of current L1 exposure) by including them as fixed effects in the models.[6] All the binary and continuous predictor variables were centered; three-level variables were coded using successive differences contrast coding. The reported *p* values came from the *z* statistic. For *highly-similar* trials, the difference between familiar categories vs. familiar dimensions only (as shown in Table 2, p. 16, and discussed in footnote 6, p. 16) was not significant in either task, and so we do not report it in the analysis.

Gradient phonological competition in L2 learning? We begin by addressing the second question concerning the effects of phonological similarity on discrimination and learning of L2 words. We analyzed trials from both discrimination and word-learning tasks for word pairs that were all expected to be *discriminable* by learners due to their familiarity with the tested sound contrasts from their L1s. That is, the analysis included the following trials: *dissimilar* (e.g., [tala]-[kenna]), *similar* (e.g., [tala]-[taja]), and a subset of *highly-similar* pairs, depending on participants' L1: *length* (e.g., [tala-talla]) for Korean speakers and *place* (e.g., [gotça]-[gotşa]) for Mandarin speakers. Crucially, we withhold from this analysis the following *highly-similar* pairs: *place* for Korean speakers and *length* for Mandarin speakers, which, based on prior work, we know are not easily discriminable by the learners due to their native language backgrounds.

---

[5] Note that maximal random-effects structures are the most analogous to ANOVA procedures.

[6] We checked the fit of our two main models (TASK*TRIAL-TYPE*LANGUAGE and TASK*FEATURE-TYPE*LANGUAGE, as described in the analyses below) with and without the IQ and the L1 proficiency/use factors. In both cases, adding these factors significantly improved the model fit (*p*s<.05); the other effects remained unaffected across the models.

These trials are analyzed separately below where we address the main question regarding the discrimination vs. word-learning performance.

Based on prior work, we expected discrimination to be gradient: best for *dissimilar* words, intermediate for *similar* words, and poorest for *highly-similar* words. Of most interest here was the comparison between discrimination and the word learning in order to evaluate how phonological similarity moderates performance across different tasks.

The results are illustrated in Figure 2 (see Appendix Figure B1 for the results by block). We began the analysis by evaluating a model with fixed effects of TASK (*discrimination*, *word-learning*), TRIAL-TYPE (*dissimilar*, *similar*, *highly-similar*), and LANGUAGE (*Korean*, *Mandarin*). TRIAL-TYPE was coded with *similar* trials as the reference level so that we would be able to directly compare *similar* and *dissimilar* trials, as well as *similar* and *highly-similar* trials. As expected, there was a significant effect of TRIAL-TYPE in that accuracy, pooled across the discrimination and the word learning tasks, varied in accordance to the similarity between words: the responses on *dissimilar* trials were significantly higher than on the *similar* trials (p<.001), which in turn were higher than *highly-similar* trials (p<.001).

[insert Figure 2 about here]

Furthermore, there were significant interactions between TASK and TRIAL-TYPE (ps<.001; note that there were two interaction terms due to the contrast coding of TRIAL-TYPE), suggesting that accuracy on each type of trial was moderated by the task: discrimination vs. word learning. To examine this further, we directly compared performance in discrimination and word learning separately for each TRIAL-TYPE using models with the fixed effect of TASK (*discrimination*, *word-*

learning). For *dissimilar* trials, there were no main effects of TASK (p=.59), suggesting that performance did not differ across tasks (although it is possible that an underlying difference between the tasks was masked by ceiling effects, given that the overall performance was above 95%). However, for both *similar* and *highly-similar* trials, we found significant main effects of TASK (ps<.001): higher overall performance in discrimination than in the word-learning task.

No other effects in the full model reached significance, including effects or interactions involving LANGUAGE (ps>.2). This suggests that the two L1 populations had similar overall response patterns on the word pairs that were predicted to be relatively well discriminated by all participants (i.e., all word pairs excluding *place* trials for Korean speakers and *length* for Mandarin speakers).

In sum, these results suggest that there is a gradient effect of phonological similarity in both word discrimination and word learning: performance decreases as similarity grows. However, this effect is moderated by the specific task: relative to discrimination, performance in word learning suffers substantially more as similarity increases.

Discrimination vs. word learning: are L2 learners like infants? Addressing whether L2 learners show a discrimination-word learning asymmetry like that observed in L1-learning infants (Stager & Werker, 1997) entails a specific comparison between the two different *highly-similar* trial types—*length* and *place*—for both L1 populations and across the two different tasks. Given previous studies with similar stimuli (Pajak, 2012; Pajak & Levy, 2014), we expected differential discrimination of *length* and *place* contrasts by the two L1 populations: Korean speakers more accurate than Mandarin speakers on discriminating *length* trials, and Mandarin speakers more accurate than Korean speakers on discriminating *place* trials (a FEATURE-TYPE x

LANGUAGE interaction). The question was whether this interaction would also extend to the word-learning data.

The results are illustrated in Figure 3. We analyzed these data in a model with fixed effects of TASK (*discrimination*, *word-learning*), FEATURE-TYPE (*length*, *place*), and LANGUAGE (*Korean*, *Mandarin*). (See Appendix B for additional analyses that include BLOCK as a fixed effect, demonstrating that the main result of a FEATURE-TYPE x LANGUAGE interaction was consistent throughout the experiment.) First, there was a significant main effect of TASK (p<.001): performance was overall higher in discrimination than in word learning, indicating that the latter task was more difficult. Furthermore, there was a significant FEATURE-TYPE x LANGUAGE interaction (p<.001): as expected, Korean speakers performed better on *length* trials, and Mandarin speakers performed better on *place* trials. Critically, however, there was also a significant three-way TASK x FEATURE-TYPE x LANGUAGE interaction (p<.001), indicating that performance as a function of FEATURE-TYPE was different across the two tasks. In order to interpret this interaction, we continued by analyzing each task separately.

[insert Figure 3 about here]

For the discrimination task (Figure 3, left), there was a significant interaction between FEATURE-TYPE and LANGUAGE (p=.001): as predicted, Korean listeners performed better on *length* trials, while Mandarin listeners performed better on *place* trials. For the word-learning task (Figure 3, right), however, there was no interaction between the two variables (p=.21; nor any main effects, ps>.5), suggesting that the respective group-level perceptual advantages of Korean and Mandarin speakers did not translate into an advantage during word learning.

Therefore, the three-way interaction in the main model indicates that good discrimination does not necessarily yield better learning. Overall, these results suggest that learners did not take full advantage of their L1-based perceptual abilities in a word-learning task. This pattern is highly similar to 14-month old infants learning their native language (Stager & Werker, 1997), suggesting that discordant performance between discrimination and mapping is not a developmental phenomenon, but a more general feature of learning words in a new phonology.

Do individual differences affect use of perceptual advantages in word learning? Word learning is a complex task that involves a combination of cognitive abilities and attention. Therefore, it is possible that there is a high degree of individual differences in how efficiently learners make use of their L1-based perceptual abilities in the word-learning task. To answer that question we examined the word-learning data separately for two groups of participants: higher and lower performers.

We split participants into higher- and lower-performer groups based on their accuracy scores on *dissimilar* and *similar* trials. These trials consisted of more salient distinctions (see Table 3), and were independent from the variables of interest. The median accuracy on all these trials combined was 94.5%.[7] There were 7 participants who scored exactly at 94.5%. We performed two separate analyses of the data, where the 7 participants were either all included in the higher-performer group or in the lower-performer group, later referred to as split-1 and split-2, respectively. The results were equivalent in both cases, but for simplicity reasons, we only illustrate the split-1 results. The distribution of participants was fairly equal across language background (Table 4). The scores on *dissimilar* and *similar* trials for both higher and lower

[7] A reviewer points out that a median level of performance of 94.5% suggests ceiling effects such that distinguishing good learners from poor learners ceases to be very meaningful. We agree that this is in principle a valid concern, but since this median split reveals a significant and interpretable interaction, we do not think it has turned out to be a concern in practice.

performers are provided in Table 4. Both groups were highly accurate on these trials, but there was much more variability among lower performers, as indicated by the higher standard deviations.

[insert Table 4 about here]

Figure 4 illustrates the word-learning results for length and place trials split into higher and lower performers. By visual inspection alone, it can be seen that participants in the higher-performer group were clearly learning the minimal-pair words, as indicated by their much higher levels of accuracy. In the lower-performer group, on the other hand, participants' responses were close to chance.

[insert Figure 4 about here]

We analyzed these results with models with fixed effects of FEATURE-TYPE (*length*, *place*), LANGUAGE (*Korean*, *Mandarin*), and PERFORMER-TYPE (*high*, *low*), separately for split-1 and split-2. In both cases we found a significant effect of PERFORMER-TYPE (*p*s<.001), reflecting higher accuracy in the higher-performer group than in the lower-performer group. Critically, there were also significant three-way interactions (in both split-1 and split-2) between FEATURE-TYPE, LANGUAGE, and PERFORMER-TYPE (*p*s<.05), indicating distinct response patterns for Korean vs. Mandarin speakers on *length* and *place* trials depending on their overall success rate in learning, as measured by their accuracy on *dissimilar* and *similar* trials.

To assess the nature of the 3-way interaction, we analyzed the effects of FEATURE-TYPE and LANGUAGE separately for each PERFORMER-TYPE. Higher performers showed a pattern more consistent with taking advantage of their perceptual biases: Korean speakers were more accurate on *length* trials than Mandarin speakers, but the reverse was true on *place* trials, as indicated by significant interactions (in both split-1 and split-2) between FEATURE-TYPE and LANGUAGE (ps<.05).[8] The lower performers, on the other hand, showed no significant interactions between FEATURE-TYPE and LANGUAGE (and no other significant effects).

Therefore, analyzing the word-learning data split by accuracy on more salient distinctions revealed a difference between higher and lower performers, with the former group taking more advantage of their L1-based perceptual abilities than the latter group. What might underlie such individual differences? One possibility is that the lower performers in the word-learning task simply had lower discrimination abilities. Due to the between-subjects design we do not have any direct evidence about how discrimination related to word learning at the level of individual participants. However, we can examine a similar split of participants in the discrimination task to see whether the difference between Korean and Mandarin speakers disappears for lower performers. We found that with a similar split of participants (Figure 5), the FEATURE-TYPE x LANGUAGE interactions were significant for both higher-performer and lower-performer groups (p*s*<.01). Therefore, while we cannot rule it out, we did not find any evidence in support of the hypothesis that overall lower performers might have lower discrimination abilities. Instead, these results suggest that the higher-performer advantage in the word-learning task is driven by

[8] One might wonder if these results only showed up after better learners were alerted to the nature of the task. However, the interaction was already numerically present (marginal on split-1, p=.08; split-2, p=.11) in the first block of testing. This suggests that the effect of FEATURE-TYPE was present before test trials “tipped off” participants to the presence of the *length* and *place* minimal pairs, since the pictures corresponding to these minimal pairs were never shown together in training.

something other than (or perhaps in addition to) better discrimination abilities. We outline possible explanations in the discussion.

[insert Figure 5 about here]

To summarize, in the overall results there was no effect (interaction) of language background in the word-learning task, suggesting that discrimination is necessary but not sufficient for successful learning of similar-sounding words. While we found some evidence of language-background-based learning of highly-similar words in participants who scored better in the learning task overall, the effect was confined to better word-learners and far less robust than the effect we saw in the discrimination task.

Discussion

In this paper we asked two specific questions: (1) do adult L2 learners have difficulty learning similar-sounding words that they can nevertheless discriminate, just like 14-month-old infants?, and (2) how does phonological similarity moderate discrimination vs. learning of similar-sounding L2 words? To answer these questions we asked participants of two different language backgrounds (Korean and Mandarin) either to discriminate similar-sounding words in a new language, or to map similar-sounding words onto novel referents. That is, we investigated the relationship between perceptual discrimination and word learning across speaker populations varying in native-language background. We used three levels of word similarity: *dissimilar* word pairs (differing in all sound segments but the last vowel), *similar* word pairs (same first syllable, but salient differences in the middle consonant), and *highly-similar* word pairs (same first syllable and only subtle differences in the middle consonant). Furthermore, the *highly-similar* word pairs consisted of two distinction types that the two participant populations had differential perceptual sensitivities to (as shown by previous work, and confirmed by our discrimination task). In particular, some distinctions varied along dimensions familiar to participants from their L1, and others did not. Comparing the three levels of word pairs (*dissimilar*, *similar*, *highly-similar*) allowed us to assess the role of phonological similarity in discrimination vs. word learning (question 2), while investigating the differences between L1 populations on the two types of *highly-similar* word pairs allowed us to answer the question regarding the potential group-level mismatch between discrimination and learning of similar-sounding words in adult L2 learning (question 1).

Regarding the effects of phonological similarity on discrimination vs. word learning, we found—as in prior work on learning phonologically native-like words—gradient performance. In both discrimination and word-learning tasks, performance was best on *dissimilar* word pairs, intermediate for *similar* word pairs, and worst on *highly-similar* word pairs. However, we also found a mismatch between discrimination and learning: the gradient effects were more exaggerated for participants in the word-learning task than for participants in the discrimination task, with performance dropping disproportionally as word similarity increased. Therefore, our results suggest that word learning is particularly difficult when words are similar-sounding, even when the differences are fairly salient (as in our *similar* word pairs).

The results from *highly-similar* word pairs revealed that participants in the word-learning task were unable to use their L1-based perceptual abilities effectively during word learning: despite Korean- vs. Mandarin-speaker differences in discrimination of *highly-similar* word pairs, there were no analogous Korean- vs. Mandarin-speaker differences for participants in the word-learning task. However, there was some evidence that higher performers (as independently assessed by performance on unrelated trials) were somewhat more successful. This was in sharp contrast to the discrimination results, where L1-based perceptual advantages were observed for all participants. This result thus reveals an intermediate effect between failure to learn similar-sounding words (as observed for 14-month-old infants) and a full ability to use existing L1-based perceptual abilities in learning (which should mimic the discrimination data).

What factors led to better performance in learning highly-similar-sounding words? It is possible that the observed differences in the word-learning task performance for higher vs. lower performers simply arise to due better discrimination abilities of the higher performers: that is, higher performers might be better at detecting subtle differences, especially when given the chance to directly compare the highly-similar words during test trials. However, our data do not provide any evidence that the lower performers were actually perceptually worse. In the conceptual model we discussed in the introduction, the difference between higher and lower performers can be explained by higher performers being able to keep track of and/or integrate evidence from more exemplars relative to the lower performers. In addition, higher performers might be able to store exemplars better (i.e., with higher phonetic fidelity) than the lower performers from the outset of language exposure, and thus be faster at forming new phonetic categories. This would not hurt lower performers when asked to identify dissimilar words, but would impact them more substantially when identifying label-referent pairings of similar-sounding words. This account means that the observed differences between higher and lower performers can, at least in part, be attributed to individual differences between learners. In order to investigate this further, we examined more closely the individual measures collected from our participants, as shown in Table 5 (split-1 only; minor differences between split-1 and split-2 are reported in table footnotes).

[insert Table 5 about here]

A series of t-tests revealed that the higher-performer and the lower-performer groups did not significantly differ in nonverbal IQ or L1 proficiency. However, they did differ in English proficiency: higher performers reported higher overall proficiency than lower performers. One possibility, thus, is that better knowledge of English was what was beneficial to the higher performers in this task. However, as outlined earlier in describing the stimuli, it is not entirely clear what properties of English would produce this type of benefit. It is true that English uses segmental length as a secondary cue to some vowel contrasts (as well as coda voicing), but our stimuli only included length differences for consonants, and prior work (Pajak & Levy, 2014) found no benefit of higher English proficiency on discriminating consonant length contrasts. Another possibility is that the better performance of some participants in the word-learning task is an effect of early bilingualism: more balanced bilinguals might be better equipped to use their L1-based perceptual abilities when learning novel words. Indeed, higher English proficiency of the higher performers in our experiment was likely the consequence of their earlier age of arrival to the USA and a longer length of residence relative to the lower-performer group. Therefore, the higher-performer group might have been composed of early bilinguals who were perhaps more balanced in both languages (i.e., fluent in both Korean/Mandarin and English), while the lower-performer group included more speakers strongly dominant in their L1. This is consistent with other findings suggesting that early simultaneous bilingual adults have a general cognitive advantage over monolinguals for novel word learning, whether phonologically familiar or unfamiliar, and independently of phonological memory capacity (Kaushanskaya, 2012; Kaushanskaya & Marian, 2009). Yet another alternative explanation is that higher English proficiency of some participants might be an indicator (more precise than IQ) of better language-learning skills that led to superior performance in our word-learning task. However, this explanation seems unlikely given that the differences in self-reported English proficiency between the higher- and lower-performer groups go away after controlling for age of arrival or length of residence in the USA.

Why is word learning difficult?

Irrespective of individual differences, the results reported in this paper suggest that there is something inherently hard about the early stage of word learning that precludes attention to fine phonetic detail that is otherwise available during phonetic processing – this is true for adult L2 learners in the same way it seems to be true for 14-month-old L1 learners. Together with the results of White et al. (2013), who showed that infants and adults are similarly affected by lexical familiarity during word learning, these findings provide evidence that there might be common learning mechanisms operating throughout development. But what are these mechanisms, and what is the source of difficulty in learning similar-sounding words?

One answer is that learning novel words is simply a highly complex task that leads to information processing overload. This explanation was originally proposed by Stager & Werker (1997) and Werker et al. (2002), who argued that young infants struggle with accessing sufficient phonetic detail in their lexical representations to successfully differentiate between words that are phonetically highly similar. This means that—relative to task difficulty—only individuals with better attentional or general cognitive abilities might effectively manage concurrent information at multiple levels of processing. This would explain why older, 17 to 20-month-old infants outperform 14-month-olds, as well as why adults with some general cognitive advantages (such as early simultaneous bilinguals; e.g., Bialystok, 1999) might outperform other adults, as with the results reported in this paper. This would also suggest that encoding and using phonetic detail when learning similar-sounding words would be improved by decreasing memory demands in the word-learning task (consistent with prior work: e.g., Fennell & Werker, 2003), as well as providing additional support for storing the phonetic detail of word exemplars and forming new phonetic categories. The latter idea is consistent with recent findings that auditory training on difficult phonetic distinctions improves subsequent word learning, and is particularly beneficial for learners with low pre-training auditory sensitivity (Cooper & Wang, 2013; Ingvalson, Barr, & Wong, 2013).

Note, however, that it does not seem to be the case that word learning is simply overall harder than discrimination – our results indicate that for highly dissimilar word pairs, performance is equally high in both tasks (although an underlying difference might be hidden due to ceiling effects). Instead, performance in word learning drops disproportionately as the word pairs get phonologically more similar, suggesting that the task is especially taxing when detailed phonetic representations are needed to distinguish between words, even when the words are perceptually easily discriminable (as our *similar* word pairs). This is consistent with other work showing that confusability between newly-learned—but phonologically native-like—words is modulated by the phonetic distance between the sounds that differentiate between them (e.g., Creel & Dahan, 2010; Creel et al., 2006; White et al., 2013). Thus, the overall difficulty of word learning in general lies most likely not just in the task itself, but also in pinning down the correct label/referent pairings, as outlined in the conceptual model we proposed in the introduction.

Implications for L2 acquisition

The results presented in this paper contribute to our understanding of L2 learning, as well as provide more practical implications for second-language teaching. The current view is that L2 phonetic category learning is largely hindered due to perceptual difficulties arising from prior acquisition of L1 phonology (e.g., Best, 1995, Flege, 1995, Kuhl & Iverson, 1995). While we do not dispute the importance of L1 influence on L2 phonological acquisition, our results suggest that at least some of the learners' difficulties in distinguishing between novel L2 sounds might be due to their introduction in the context of highly similar-sounding lexical items. In such a context, when attention is directed at forming label-referent mappings, learners might be unable to properly separate similar-sounding categories. In fact, it has been shown that, when discrimination between some L2 sounds is initially present but fragile, the mere act of learning the word-referent mappings for similar-sounding words that differ by those sounds makes their perceptual discrimination even worse (Dobel, Lagemann, & Zwitserlood, 2009). It thus seems that the introduction of highly similar L2 phonetic categories might be more effective either outside of a word-learning task (e.g., as a focused non-native sound discrimination practice) or when learning words that are overall fairly dissimilar. The latter conclusion is independently supported by the results from infant studies, showing that prior experience with sounds in non-minimal-pair lexical contexts improves infants' ability to later learn minimal-pair words distinguished by those sounds (Thiessen, 2007). A similar conclusion can be drawn from both behavioral and computational modeling work, showing that non-minimal-pair lexical contexts improve distributional learning of overlapping sound categories for both infants and adults (Feldman, Griffiths, Goldwater, & Morgan, 2013; Feldman, Myers, White, Griffiths, & Morgan, 2013).

Conclusion

The results presented in this paper show that adults, just like young infants, have difficulty learning similar-sounding words that they can nevertheless distinguish perceptually, showing that discrimination is necessary but not sufficient for successful learning of similar-sounding words. This parallel between infants and adults point to a common mechanism underlying the initial stage of lexical acquisition throughout development, whether in the native language or any additional language acquired in adulthood. Together with recent results showing other parallels between infant and adult lexical acquisition (White et al., 2013), these findings highlight the necessity for greater interaction between infant and adult language learning literatures that would investigate the commonalities and the differences between native-language development in infancy and second-language learning in adulthood, thus shedding more light onto the degree of developmental continuity in language learning.

Appendix A

Individual measures. This appendix includes a more detailed version of the individual measures collected from participants that were provided in Table 1.

Tables A1 and A2 compare participants of the discrimination vs. the word-learning tasks, separately for each language background (Table A1: Korean speakers; Table A2: Mandarin speakers). Just as what was shown in Table 1, there were no significant differences between the participants within each L1 population, with the minor exception of the word-learning task Korean speakers reporting slightly higher proficiency in understanding English than the discrimination task Korean speakers.

Tables A3 and A4 compare Korean vs. Mandarin speakers separately for each task (Table A3: discrimination; Table A4: word-learning). There were no significant differences between Korean and Mandarin speakers assigned to the discrimination task. There were, however, some differences between the participants assigned to the word-learning task. Namely, relative to speakers of Mandarin, Korean speakers on average immigrated to the USA earlier, had a longer length of residence in the USA, and reported higher proficiency in English and slightly lower proficiency in their L1. At the same time, they did not differ on an objective measure of English proficiency, the Shipley Vocabulary Test, and they were all fluent speakers of both English and Korean/Mandarin, with similar amounts of current exposure to each language. While it is possible that these differences might have affected our results, prior work with subjects from the same populations has shown that, when subjects use both languages regularly, their relative dominance in English vs. Korean/Mandarin does not affect the discrimination of the contrasts tested in this study (Pajak & Levy, 2014). Furthermore, and most critically, within each L1 population, participants did not differ across the two tasks, as shown in Tables A1 and A2.

[insert Tables A1, A2, A3, and A4 about here]

Appendix B

Learning during the task. In both discrimination and word learning it is reasonable to expect that performance could improve with exposure to the materials and greater experience in the test task; hence we analyzed the results including test block as an additional variable (see Figures B1-B4). Throughout the analysis BLOCK was treated as a mean-centered continuous covariate, which is a simple way of imposing the constraint that temporal changes should be monotonic.

First, we examined the word learning results for *dissimilar*, *similar*, and *highly-similar* (Korean: *length*, Mandarin: *place*) trials in order to test whether there is differential improvement over the course of the experiment depending on the word similarity. In particular, we expected relatively stable, good performance on *dissimilar* trials throughout the experiment, with most dramatic improvements observed for *highly-similar* trials. We tested this by evaluating a model with fixed effects of TRIAL-TYPE (*dissimilar*, *similar*, *highly-similar*), LANGUAGE (*Korean*, *Mandarin*), and BLOCK (as a continuous predictor). These results are illustrated in Figure A1. Just as in the main analysis, TRIAL-TYPE was coded with *similar* trials as the reference level, and there was a significant effect of TRIAL-TYPE in that accuracy varied in accordance to the similarity between words: the responses on *dissimilar* trials were significantly higher than on the *similar* trials ($p<.001$), which in turn were higher than *highly-similar* trials ($p<.001$). We found a significant main effect of BLOCK ($p<.001$) indicating that participants improved throughout the experiment. There was also a significant interaction between TRIAL-TYPE and BLOCK ($p<.05$) in that the relative difference between *dissimilar* and *highly-similar* trials varied as a function of block: the difference was largest at the beginning of the experiment, and it gradually decreased with time (no such interaction was found for the difference between *similar* and *highly-similar* trials). This is in line with our prediction that learners' performance should not be impeded in cases of learning very dissimilar words; however, learning is expected to be slower for highly similar words because more data is needed before learners can accumulate a sufficient number of exemplars to learn phonetic category distinctions and form correct label/referent mappings.

[insert Figure B1 about here]

Next, we analyzed both discrimination and word-learning sets of *highly-similar* trials (see Figure B2). For both tasks we found significant main effects of BLOCK (discrimination: $p<.05$; word-learning: $p<.001$), indicating that participants improved throughout the experiment. There were also significant interactions between FEATURE-TYPE and BLOCK (discrimination: $p<.05$; word-learning: $p<.01$): in word learning, the improvement was more prominent for the *length* trials than for the *place* trials; the opposite seemed to be the case for discrimination – more improvement on *place* than on *length* trials. Crucially, adding the block information revealed that the main result – a difference between Korean and Mandarin speakers in discrimination, but not in word learning – was consistent throughout the experiment. Figures B3 and B4 illustrate the by-block results split by higher and lower performers in the word-learning and the discrimination tasks, respectively.

[insert Figures B2, B3, and B4 about here]

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439.

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition*, *52*, 163–187.

Ballem, K. D., & Plunkett, K. (2005). Phonological specificity in children at 1;2. *Journal of Child Language*, *32*, 159–173.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Benus, S., Smorodinsky, I., & Gafos, A. (2003). Gestural coordination and the distribution of English 'geminates'. In Proceedings of the 27th Annual Penn Linguistics Colloquium (pp. 33–46). Philadelphia, PA: University of Pennsylvania.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: issues in cross-language research* (pp. 171–204). Timonium, MD: York Press.

Bialystok, E. (1999). Cognitive complexity and attentional control in the bilingual mind. *Child Development*, *70*, 636–644.

Birdsong, D. (2009). Age and the end state of second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *The new handbook of Second Language Acquisition* (pp. 401–424). Amsterdam: Elsevier.

Birdsong, D., & Molis, M. (2001). On the Evidence for Maturational Constraints in Second-Language Acquisition. *Journal of Memory and Language*, *44*, 235–249.

Casini, L., Burle, B., and Nguyen, N. (2009). Speech perception engages a general timer: Evidence from a divided attention word identification task. *Cognition*, *112*, 318–322.

Choi, D.-I. (1995). Korean “tense" consonants as geminates. *Kansas Working Papers in Linguistics*, *20*, 25–38.

Clayards, M. A., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–9. doi:10.1016/j.cognition.2008.04.004

Cooper, A., & Wang, Y. (2013). Effects of tone training on Cantonese tone-word learning. *Journal of the Acoustical Society of America*, *134*(2), EL133–EL139.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2006). Acquiring an artificial lexicon: segment type and order information in early lexical entries. *Journal of Memory and Language, 54,* 1-19.

Creel, S. C., & Dahan, D. (2010). The influence of the temporal structure of spoken words on paired-associate learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 110-122.

Dobel, C., Lagemann, L., & Zwitserlood, P. (2009). Non-native phonemes in adult word learning: evidence from the N400m. *Philosophical Transactions of the Royal Society B*, *364*, 3697–3709. doi:10.1098/rstb.2009.0158

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

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

Escudero, P., Simon, E., & Mulak, K. E. (2014). Learning words in a new language: Orthography doesn't always help. *Bilingualism: Language and Cognition*, *17*(2), 384–395. doi:10.1017/S1366728913000436

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*(4), 751–778.

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., & Morgan, J. L. (2013). Word-level information influences phonetic learning in adults and infants. *Cognition*, 127, 427–438.

Fennell, C. T., Waxman, S. R., & Weisleder, A. (2007). With referential cues, infants successfully use phonetic detail in word learning. In *Proceedings of the 31st Boston University Conference on Language Development*. Cascadilla Press.

Fennell, C. T., & Werker, J. F. (2003). Early word learners' ability to access phonetic detail in well-known words. *Language and Speech*, *46*(2-3), 245–264.

Flege, J. E. (1995). Second-language speech learning: theory, findings and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: issues in cross-language research* (pp. 229–273). Timonium, MD: York Press.

Hakuta, K., Bialystok, E., & Wiley, E. (2003). Critical Evidence: A Test of the Critical-Period Hypothesis for Second-Language Acquisition. *Psychological Science*, *14*(1), 31–38.

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

Hillenbrand, J. M., Clark, M. J., & Houde, R. A. (2000). Some effects of duration on vowel recognition. *Journal of the Acoustical Society of America*, *108*(6), 3013–3022.

Ingvalson, E. M., Barr, A. M., & Wong, P. C. M. (2013). Poor phonetic perceivers show greater benefit in phonetic-phonological speech learning. *Journal of Speech, Language, and Hearing Research*, *56*, 1045–1050.

Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed effects models. *Journal of Memory and Language*, *59*, 434–446.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*, 60–99.

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Brief Intelligence Test – Second Edition (KBIT-2)*. Circle Pines, MN: American Guidance Service.

Kaushanskaya, M. (2012). Cognitive mechanisms of word learning in bilingual and monolingual adults: The role of phonological memory. *Bilingualism: Language and Cognition*, *15*(3), 470–489.

Kaushanskaya, M., & Marian, V. (2009). The bilingual advantage in novel word learning. *Psychonomic Bulletin & Review*, *16*(4), 705–710.

Kaye, A. (2005). Gemination in English. *English Today*, *21*, 43–55.

Kidd, G. R., Watson, C. S., Gygi, B. (2007). Individual differences in auditory abilities. *The Journal of the Acoustical Society of America*, *122*(1), 418–435. doi:dx.doi.org/10.1121/1.2743154

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–78. doi:10.1016/j.cogpsych.2005.05.001

Kuhl, P. K., & Iverson, P. (1995). Linguistic experience and the "perceptual magnet effect". In W. Strange (Ed.), *Speech perception and linguistic experience: issues in cross-language research* (pp. 121–154). Timonium, MD: York Press.

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Oxford, UK; Cambridge, MA: Blackwell.

Lenneberg, E. (1967). *Biological foundations of language*. New York: Wiley.

Lin, H. (2001). *A grammar of Mandarin Chinese*. München: Lincom Europa.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*, 1–36.

Lukas, J. H. (1980). Human auditory attention: The olivocochlear bundle may function as a peripheral filter. *Psychophysiology*, *17*, 444–452.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science*, *31*, 1–24.

Mani, N., & Plunkett, K. (2011). Does size matter? Subsegmental cues to vowel mispronunciation detection. *Journal of Child Language*, *38*, 606–627.

Mattys, S. L., Barden, K., and Samuel, A. G. (2014). Extrinsic cognitive load impairs low-level speech perception. *Psychonomic Bulletin and Review*, *21*, 748–754.

Mattys, S. L., & Palmer, S. D. (2015). Divided attention disrupts perceptual encoding during speech recognition. *The Journal of the Acoustical Society of America*, *137*(3), 1464–1472.

Maye, J., & Gerken, L. (2000). Learning phonemes without minimal pairs. In S. C. Howell, S. A. Fish, & T. Keith-Lucas (Eds.), *Proceedings of the 24th Annual Boston University Conference on Language Development* (pp. 522–533). Somerville, MA: Cascadilla Press.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental Science*, *11*(1), 122–134.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101–B111.

Miyawaki, K., Strange, W., Verbrugge, R. R., Liberman, A. M., Jenkins, J. J., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [I] by native speakers of Japanese and English. *Perception and Psychophysics*, *18*, 331–340.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. doi:10.1016/S0010-0285(03)00006-9

Nowak, P. M. (2006). The role of vowel transitions and frication noise in the perception of Polish sibilants. *Journal of Phonetics*, *34*, 139–152.

Pajak, B. (2012). *Inductive inference in non-native speech processing and learning*. PhD dissertation, University of California, San Diego.

Pajak, B., & Levy, R. (2011). Phonological generalization from distributional evidence. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2673–2678). Austin, TX: Cognitive Science Society.

Pajak, B., & Levy, R. (2014). The role of abstraction in non-native speech perception. *Journal of Phonetics*, *46*, 147–160.

Pajak, B., Fine, A. B., Kleinschmidt, D. F., & Jaeger, T. F. (2015). Learning additional languages as hierarchical probabilistic inference: insights from L1 processing. *Language Learning*, accepted.

Pater, J., Stager, C., & Werker, J. (2004). The perceptual acquisition of phonological contrasts. *Language* , *80*(3), 384–402.

Puel, J. L., Bonfils, P., and Pujol, R. (1988). Selective attention modifies the active micromechanical properties of the cochlea. *Brain Research*, *447*, 380–383.

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349.

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

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.

Shipley, W. C. (1967). *Shipley Institute of Living Scale: Vocabulary Test and Abstraction Test*. Los Angeles, CA: Western Psychological Services.

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

Sohn, H.-M. (1999). *The Korean language*. Cambridge, UK: Cambridge University Press.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*, 381–382.

Storkel, H. L. (2004). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, *25*, 201–221.

Storkel, H. L., Armbruster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research*, *49*, 1175–1192.

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, *76*, 147–166.

Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science*, *13*(5), 480–484.

Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive Psychology*, *54*, 99–132.

Thiessen, E. D. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language*, *56*, 16–34.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: levels of processing in perception of spoken words. *Psychological Science*, *9*(4), 325–329.

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

Werker, J. F., Fennell, C. T., Corcoran, K. M., & Stager, C. L. (2002). Infants' ability to learn phonetically similar words: effects of age and vocabulary size. *Infancy*, *3*(1), 1–30.

Werker, J. F. & Tees, R. C. (2005). Speech perception as a window for understanding plasticity and commitment in language systems of the brain. *Developmental Psychobiology*, *46*(3), 233–251.

Westbury, J. R., Hashi, M., & Lindstrom, M., J. (1998). Differences among speakers in lingual articulation for American English /ɹ/. *Speech Communication*, *26*(3), 203–226.

White, K. S., & Morgan, J. L. (2008). Sub-segmental detail in early lexical representations. *Journal of Memory and Language*, *59*, 114–132.

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

Yoshida, K. A., Fennell, C. T., Swingley, D., & Werker, J. F. (2009). Fourteen-month-old infants learn similar-sounding words. *Developmental Science*, *12*(3), 412–418.

Table 1. Individual measures: participants in the discrimination vs. the word-learning task.

MEASURE	DISCRIMINATION TASK PARTICIPANTS		WORD-LEARNING TASK PARTICIPANTS		T-TEST
	mean	sd	mean	sd	
Age	20	1.6	21	2.0	t(84.1)=-1.23, p=.22
L1 proficiency: speaking[a]	8.1	1.7	7.8	1.9	|t|<1
L1 proficiency: understanding[a]	8.3	1.6	8.3	1.5	|t|<1
% time current L1 exposure	33	19.0	30	19.3	|t|<1
Age when began regular English exposure	6.3	4.0	5.5	4.1	|t|<1
Age of arrival in the USA	8.6	7.0	7.5	7.5	|t|<1
Length of residence in the USA[b]	11.6	7.6	13.1	6.7	|t|<1
English proficiency: speaking[a]	8.1	1.7	8.3	1.9	|t|<1
English proficiency: understanding[a]	8.4	1.4	8.8	1.5	t(78.8)=-1.24, p=.22
English Vocabulary Test (% correct)	72	10.8	72	13.7	|t|<1
% time current English exposure	66	19.1	69	19.5	|t|<1
Nonverbal IQ test (% correct)	88	7.0	88	7.8	|t|<1

[a] On a 0-10 scale (0-none, 10-perfect).
[b] If born in the USA, coded as 0.

Table 2. Stimuli (in IPA).

LENGTH WORDS			PLACE WORDS		
short	*long*	*specific distinction exists in Korean?*	*alveolo-palatal*	*retroflex*	*specific distinction exists in Mandarin?*
tala	talla	yes(?)[a,b]			
kema	kemma	yes[b]	gotɕa	gotʂa	yes[c]
kena	kenna	yes[b]			
diwa	diwwa	no, but familiar dimension			
difa	diffa	no, but familiar dimension	goʑa	goʐa	no, but familiar dimension
taja	tajja	no, but familiar dimension			

[a] In Korean, intervocalic [l] is realized as a flap.
[b] In Korean, the long sound results almost exclusively from phonological assimilation.
[c] While the general distinction exists, the exact place of articulation in Mandarin is different for both the alveolo-palatal and the retroflex, and the distinction is only allophonic.

Table 3. Trial types in discrimination (AB=words presented auditorily) and testing in word learning (AB=labels for visually presented pictures). For each pair, both orders of presentation were tested.

DISSIMILAR	SIMILAR	HIGHLY-SIMILAR	
A B	A B	A B	
tala-kenna	diwa-difa	diwa-diwwa	
talla-goʐa	diwwa-diffa	difa-diffa	
taja-gotɕa	taja-tala	taja-tajja	*length*
tajja-kema	tajja-talla	tala-talla	
diwa-kemma	kema-kena	kema-kemma	
diwwa-goʑa	kemma-kenna	kena-kenna	
diffa-kena	gotʂa-goʐa	goʑa-goʐa	*place*
difa-gotʂa	gotɕa-goʑa	gotɕa-gotʂa	

Table 4. *Dissimilar* and *similar* trial scores from the word-learning task split by higher and lower performers (standard deviations are in parentheses)

TRIAL TYPE	HIGHER PERFORMERS		LOWER PERFORMERS	
	Korean	*Mandarin*	*Korean*	*Mandarin*
	n = 16 / 11[a]	*n* = 15 / 13	*n* = 11 / 16	*n* = 12 / 14
dissimilar	.99 (.10) / .99 (.07)	.98 (.11) / .99 (.09)	.93 (.26) / .93 (.25)	.90 (.30) / .92 (.27)
similar	.95 (.21) / .96 (.20)	.94 (.24) / .95 (.22)	.79 (.41) / .81 (.39)	.78 (.41) / .82 (.38)

[a] Split-1 / Split-2.

Table 5. Higher vs. lower performers in the word-learning task (significant differences in **bold**).

MEASURE	HIGHER PERFORMERS		LOWER PERFORMERS		T-TEST
	mean	sd	mean	sd	
Age	20	1.6	21	2.3	|t|<1
L1 proficiency: speaking[a]	7.5	1.9	8.1	1.9	t(47.3)=1.02, p=.31
L1 proficiency: understanding[a]	8.2	1.4	8.4	1.6	|t|<1
% time current L1 exposure	**23**	**16.6**	**39**	**19.2**	**t(43.3)=3.03, p<.01**[c]**
Age when began regular English exposure	4.8	3.9	6.3	4.1	t(46.4)=1.34, p=.18
Age of arrival in the USA	**5.6**	**6.5**	**10.2**	**8.0**	**t(41.1)=2.19, p<.05***
Length of residence in the USA[b]	**14.9**	**6.2**	**10.8**	**6.7**	**t(45.4)=-2.23, p<.05***
English proficiency: speaking[a]	**8.9**	**1.4**	**7.4**	**2.1**	**t(35.4)=-2.83, p<.01****
English proficiency: understanding[a]	**9.2**	**1.2**	**8.1**	**1.7**	**t(38.2)=-2.60, p<.05***
English Vocabulary Test (% correct)	75	12.7	68	14.1	t(44.4)=-1.69, p=.10
% time current English exposure	**75**	**16.2**	**62**	**21.0**	**t(39.8)=-2.38, p<.05*[d]**
Nonverbal IQ test (% correct)	89	7.6	86	7.9	t(46.5)=-1.26, p=.21

[a] On a 0-10 scale (0-none, 10-perfect).
[b] If born in the USA, coded as 0.
[c] Effect only marginal in split-2: t(51.4)=1.74, p=.087.
[d] Effect not significant in split-2: t(51.8)=-1.36, p=.181.

Table A1. Individual measures: Korean speakers by task (significant differences in **bold**).

MEASURE	DISCRIMINATION TASK PARTICIPANTS		WORD-LEARNING TASK PARTICIPANTS		T-TEST
	mean	sd	mean	sd	
Age	20	1.7	21	1.4	|t|<1
L1 proficiency: speaking[a]	7.6	1.6	7.2	2.0	|t|<1
L1 proficiency: understanding[a]	7.8	1.5	8.1	1.6	|t|<1
% time current L1 exposure	32	15.4	29	16.8	|t|<1
Age when began regular English exposure	5.7	3.1	4.8	3.4	|t|<1
Age of arrival in the USA	6.7	5.4	5.5	6.0	|t|<1
Length of residence in the USA[b]	13.7	5.7	15.2	6.0	|t|<1
English proficiency: speaking[a]	8.5	1.3	8.9	1.4	t(36.6)=-1.01, p=.31
English proficiency: understanding[a]	**8.6**	**1.1**	**9.4**	**0.8**	**t(28.8)=-2.58, p<.05***
English Vocabulary Test (% correct)	74	8.5	74	11.5	|t|<1
% time current English exposure	67	16.0	70	16.8	|t|<1
Nonverbal IQ test (% correct)	87	5.0	86	6.1	|t|<1

[a] On a 0-10 scale (0-none, 10-perfect).
[b] If born in the USA, coded as 0.

Table A2. Individual measures: Mandarin speakers by task.

MEASURE	DISCRIMINATION TASK PARTICIPANTS		WORD-LEARNING TASK PARTICIPANTS		T-TEST
	mean	sd	mean	sd	
Age	20	1.5	21	2.4	t(42.8)=-1.16, p=.25
L1 proficiency: speaking[a]	8.5	1.6	8.3	1.6	|t|<1
L1 proficiency: understanding[a]	8.8	1.5	8.5	1.4	|t|<1
% time current L1 exposure	34	22.0	30	22.0	|t|<1
Age when began regular English exposure	6.9	4.7	6.1	4.6	|t|<1
Age of arrival in the USA	10.4	7.9	9.6	8.3	|t|<1
Length of residence in the USA[b]	9.5	8.7	11.0	6.7	|t|<1
English proficiency: speaking[a]	7.6	2.0	7.6	2.1	|t|<1
English proficiency: understanding[a]	8.2	1.6	8.1	1.8	|t|<1
English Vocabulary Test (% correct)	69	12.2	70	15.3	|t|<1
% time current English exposure	65	21.7	69	21.8	|t|<1
Nonverbal IQ test (% correct)	90	8.4	90	9.0	|t|<1

[a] On a 0-10 scale (0-none, 10-perfect).
[b] If born in the USA, coded as 0.

Table A3. Individual measures: participants in the discrimination task by language background.

MEASURE	L1-KOREAN PARTICIPANTS		L1-MANDARIN PARTICIPANTS		T-TEST
	mean	sd	mean	sd	
Age	20	1.7	20	1.5	|t|<1
L1 proficiency: speaking[a]	7.6	1.6	8.5	1.6	t(34.0)=-1.61, p=.12
L1 proficiency: understanding[a]	7.8	1.5	8.8	1.5	t(33.9)=-1.83, p=.08
% time current L1 exposure	32	15.4	34	22.0	|t|<1
Age when began regular English exposure	5.7	3.1	6.9	4.7	|t|<1
Age of arrival in the USA	6.7	5.4	10.4	7.9	t(30.1)=-1.60, p=.12
Length of residence in the USA[b]	13.7	5.7	9.5	8.7	t(29.2)=1.68, p=.10
English proficiency: speaking[a]	8.5	1.3	7.6	2.0	t(30.0)=1.53, p=.14
English proficiency: understanding[a]	8.6	1.1	8.2	1.6	|t|<1
English Vocabulary Test (% correct)	74	8.5	69	12.2	t(30.5)=1.46, p=.15
% time current English exposure	67	16.0	65	21.7	|t|<1
Nonverbal IQ test (% correct)	87	5.0	90	8.4	t(27.5)=-1.03, p=.31

[a] On a 0-10 scale (0-none, 10-perfect).
[b] If born in the USA, coded as 0.

Table A4. Individual measures: participants in the word-learning task by language background (significant differences in **bold**).

MEASURE	L1-KOREAN PARTICIPANTS		L1-MANDARIN PARTICIPANTS		T-TEST
	mean	sd	mean	sd	
Age	21	1.4	21	2.4	|t|<1
L1 proficiency: speaking[a]	**7.2**	**2.0**	**8.3**	**1.6**	**t(49.2)=-2.20, p<.05***
L1 proficiency: understanding[a]	8.1	1.6	8.5	1.4	|t|<1
% time current L1 exposure	29	16.8	30	21.5	|t|<1
Age when began regular English exposure	4.8	3.4	6.1	4.6	t(47.9)=-1.14, p=.26
Age of arrival in the USA	**5.5**	**6.0**	**9.6**	**8.3**	**t(47.5)=-2.07, p<.05***
Length of residence in the USA[b]	**15.2**	**6.0**	**11.0**	**6.7**	**t(51.3)=2.4, p<.05***
English proficiency: speaking[a]	**8.9**	**1.4**	**7.6**	**2.1**	**t(44.0)=2.6, p<.05***
English proficiency: understanding[a]	**9.4**	**0.8**	**8.1**	**1.8**	**t(36.8)=3.25, p<.01****
English Vocabulary Test (% correct)	74	11.5	70	15.3	t(48.2)=1.11, p=.27
% time current English exposure	70	16.8	69	21.8	|t|<1
Nonverbal IQ test (% correct)	86	6.1	90	9.0	t(45.8)=-1.59, p=.12

[a] On a 0-10 scale (0-none, 10-perfect).
[b] If born in the USA, coded as 0.

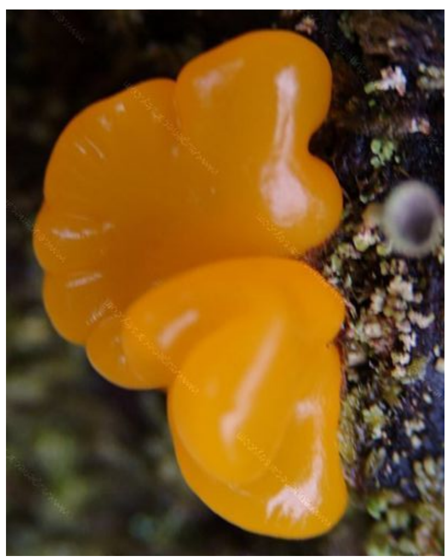
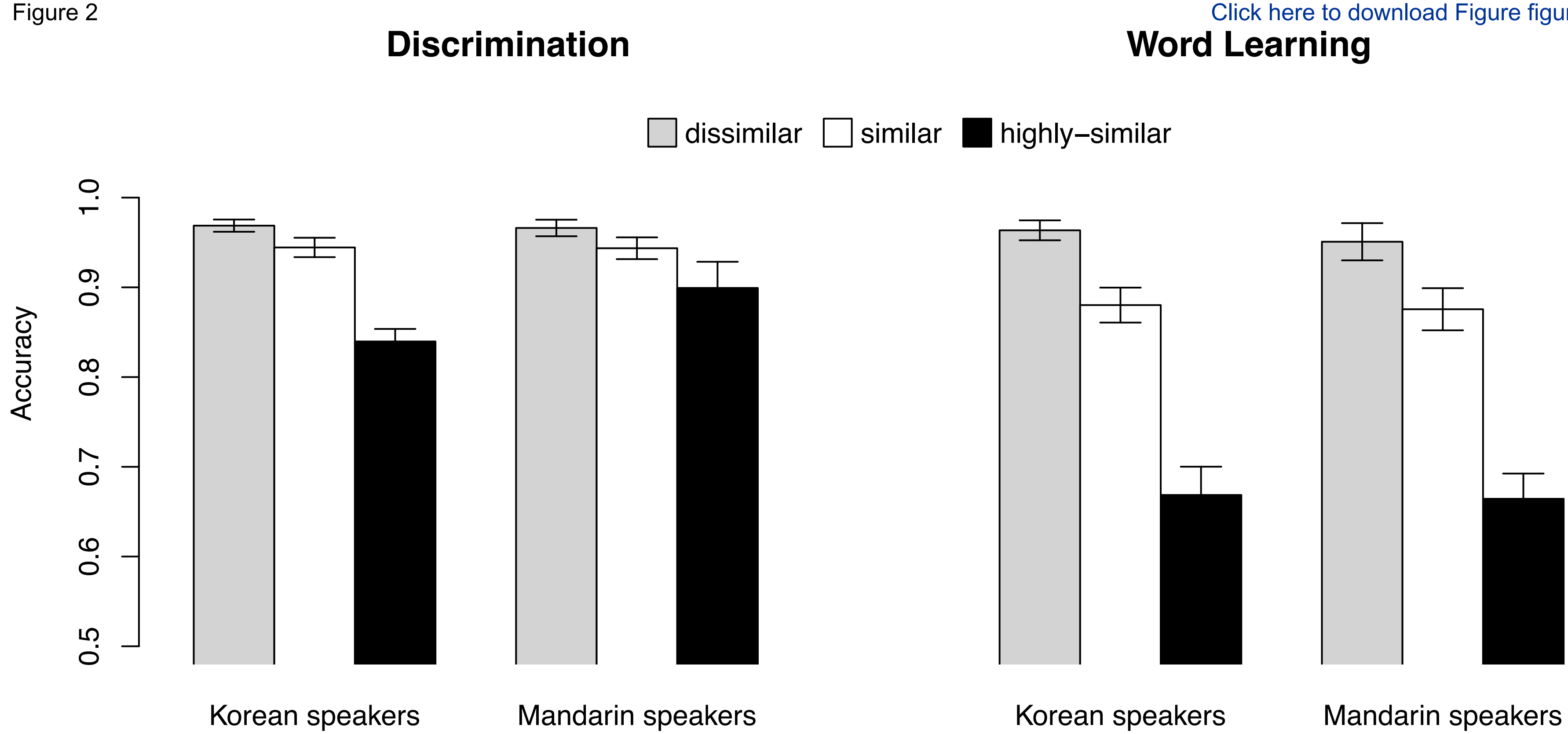
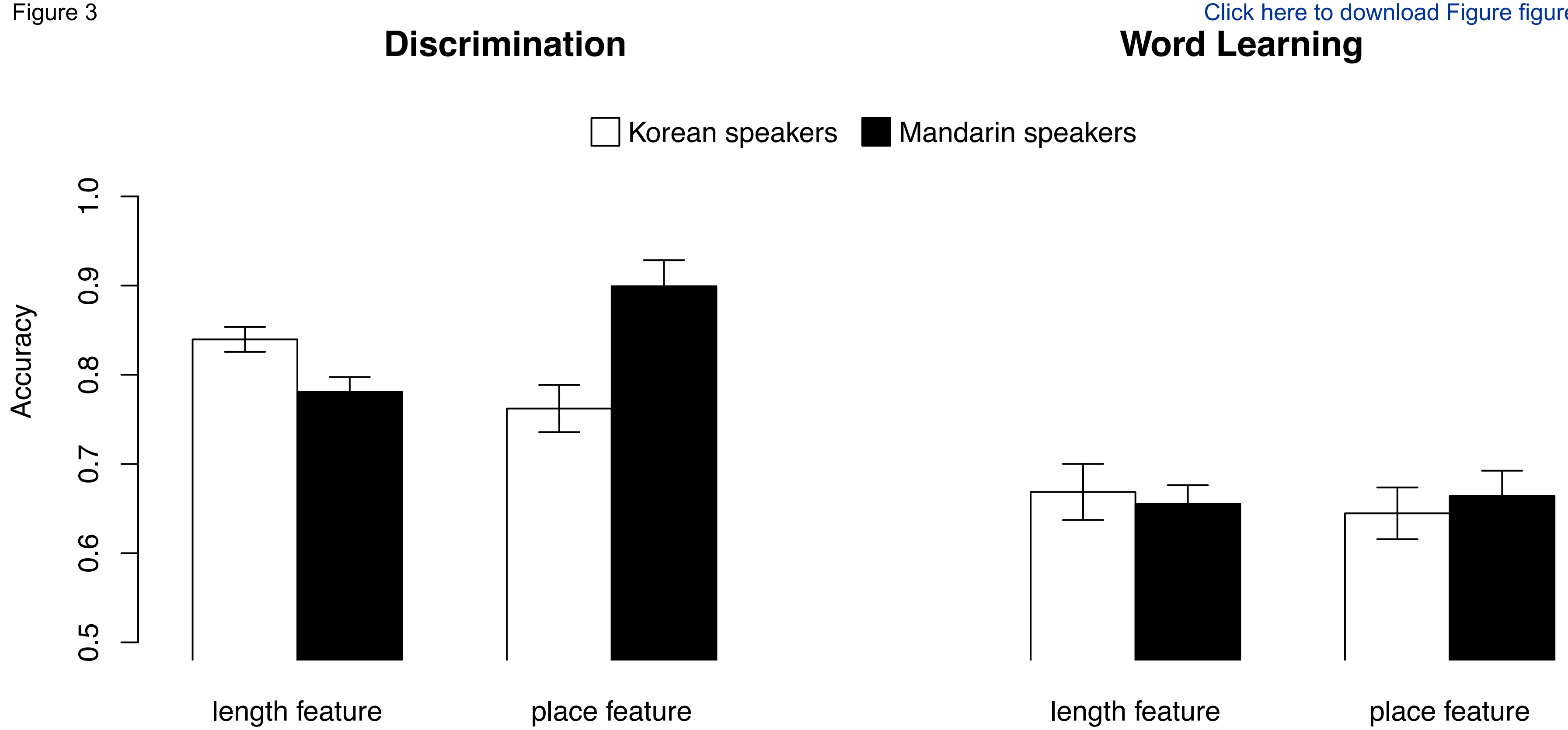


Figure 2



Discrimination
Word Learning
dissimilar
similar
highly–similar
Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
Korean speakers
Mandarin speakers
Korean speakers
Mandarin speakers

Figure 3



Discrimination
Word Learning
Korean speakers
Mandarin speakers
Accuracy
0.5
0.6
0.7
0.8
0.9
1.0
length feature
place feature
length feature
place feature

Figure 4

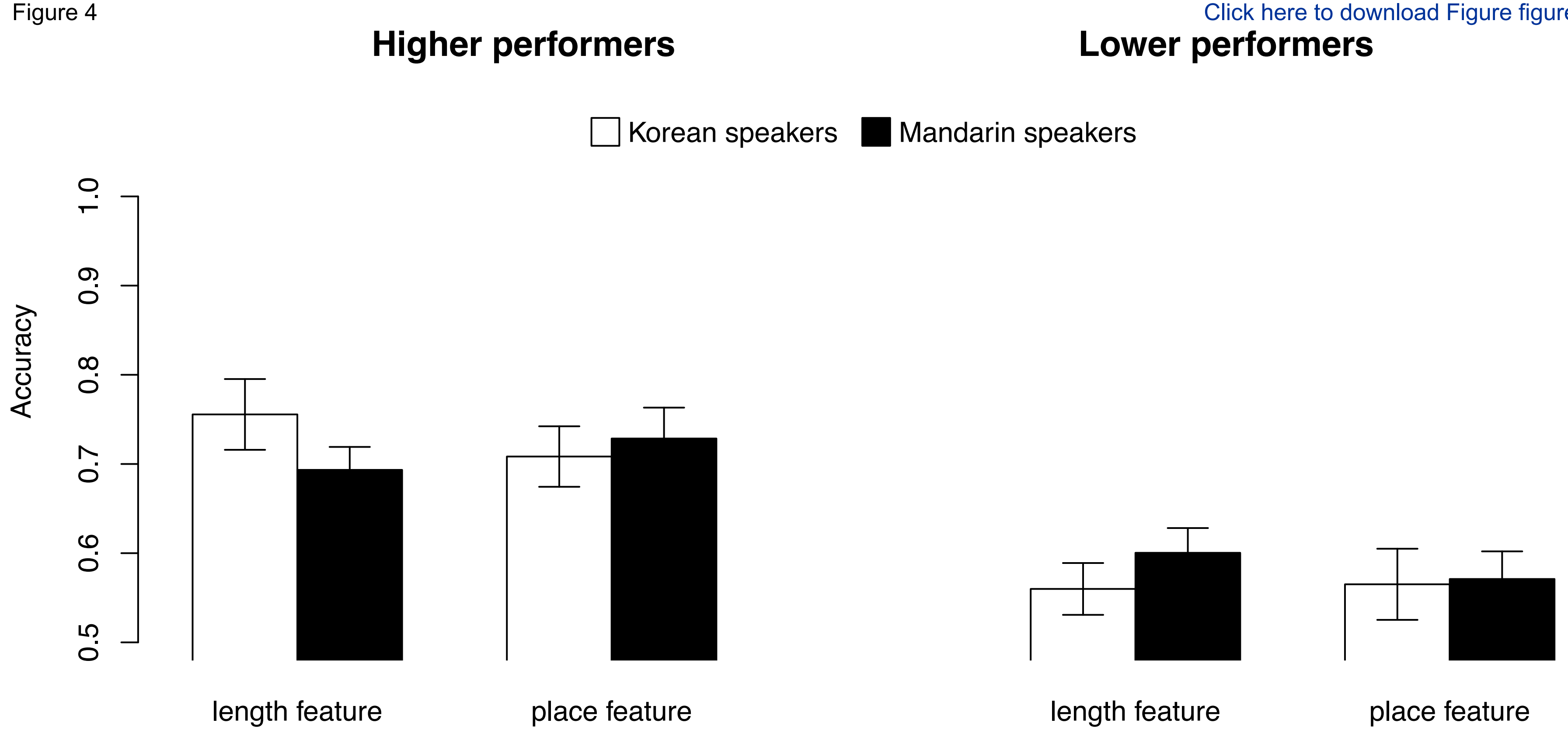

Higher performers
Lower performers
Korean speakers
Mandarin speakers
Accuracy
0.5
0.6
0.7
0.8
0.9
1.0
length feature
place feature
length feature
place feature

Figure 5

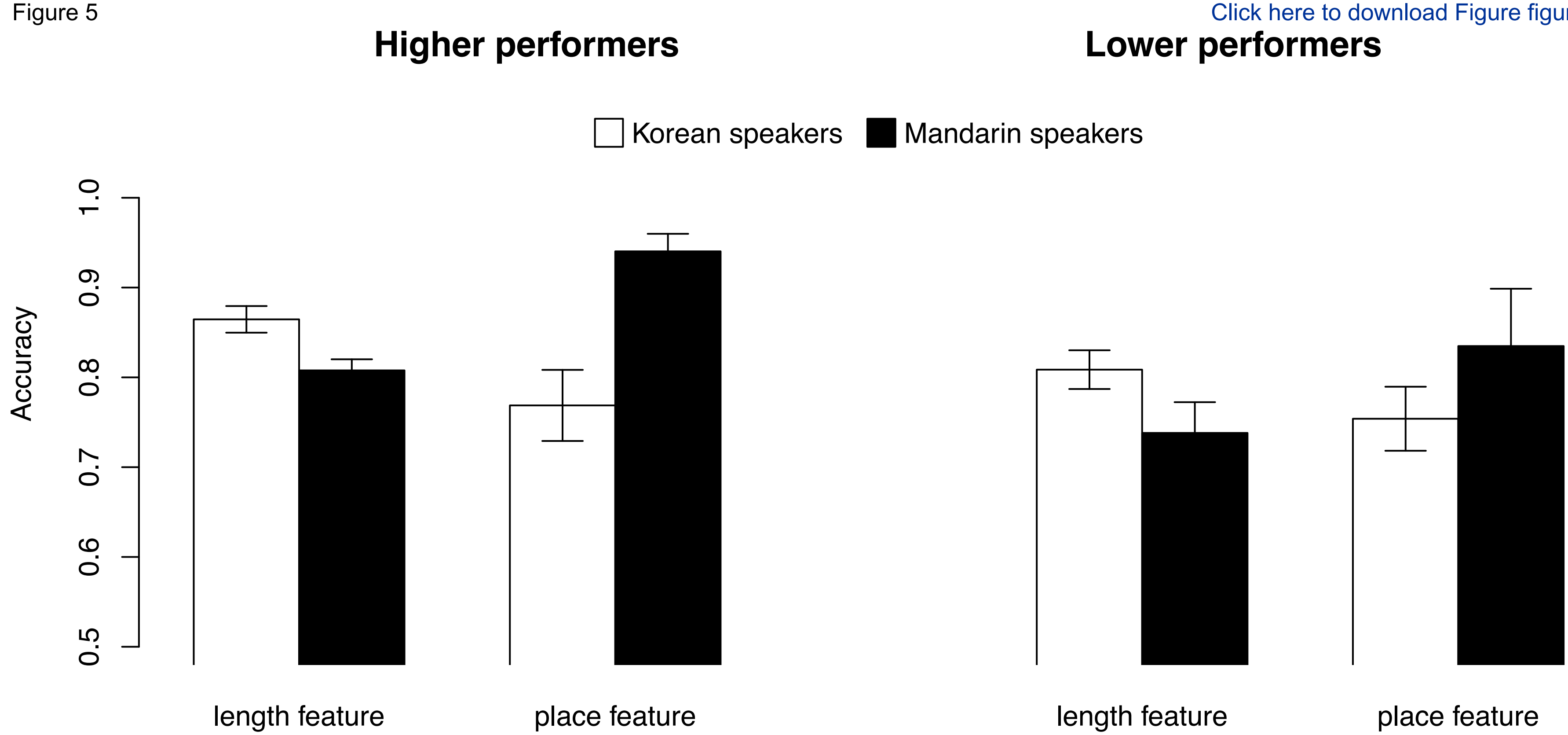

Higher performers
Lower performers
Korean speakers
Mandarin speakers
Accuracy
0.5
0.6
0.7
0.8
0.9
1.0
length feature
place feature
length feature
place feature

Figure B1

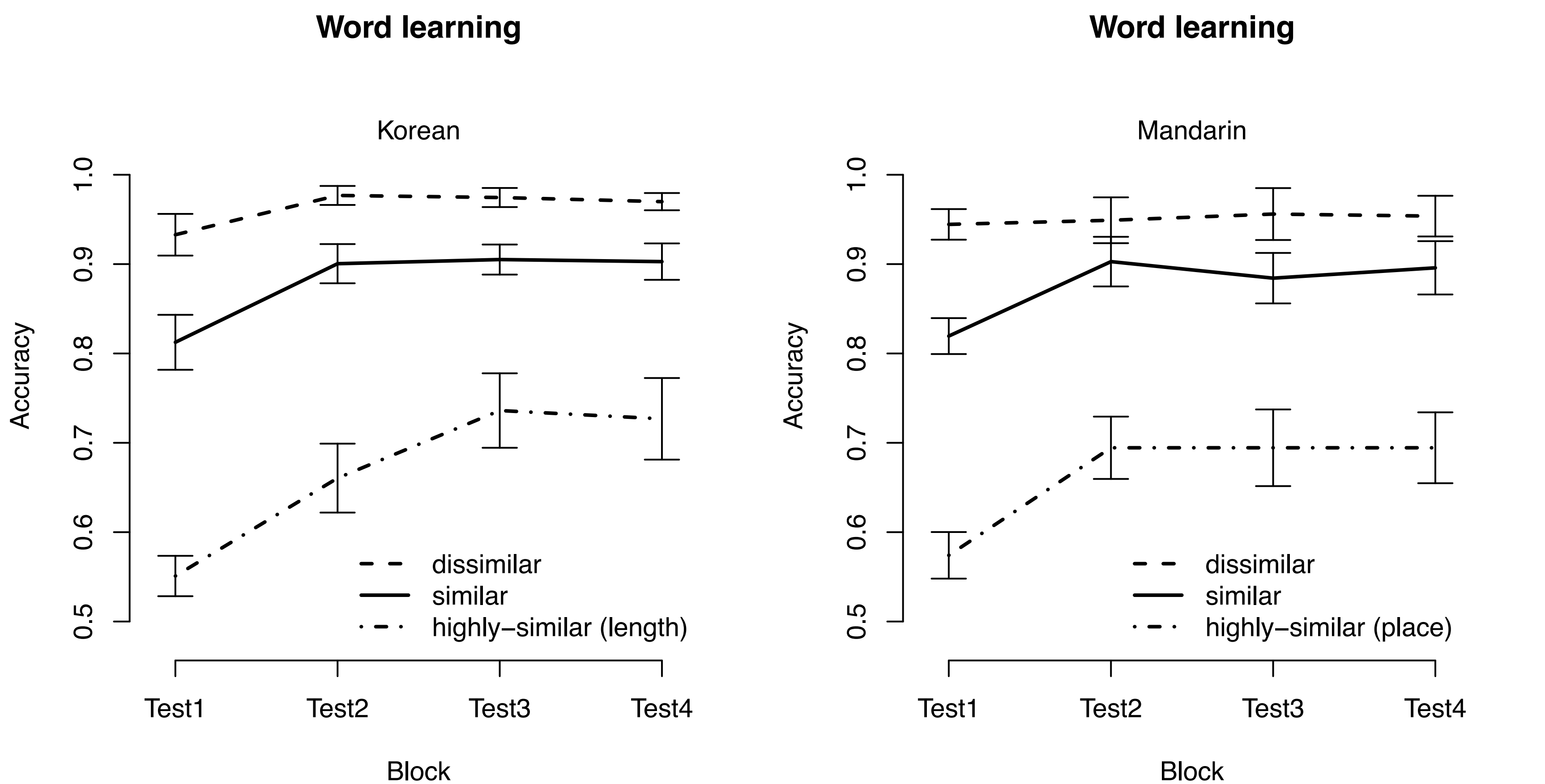

Word learning
Korean
Accuracy
0.5
0.6
0.7
0.8
0.9
1.0
Test1
Test2
Test3
Test4
Block
dissimilar
similar
highly–similar (length)
Word learning
Mandarin
Accuracy
0.5
0.6
0.7
0.8
0.9
1.0
Test1
Test2
Test3
Test4
Block
dissimilar
similar
highly–similar (place)

Figure B2

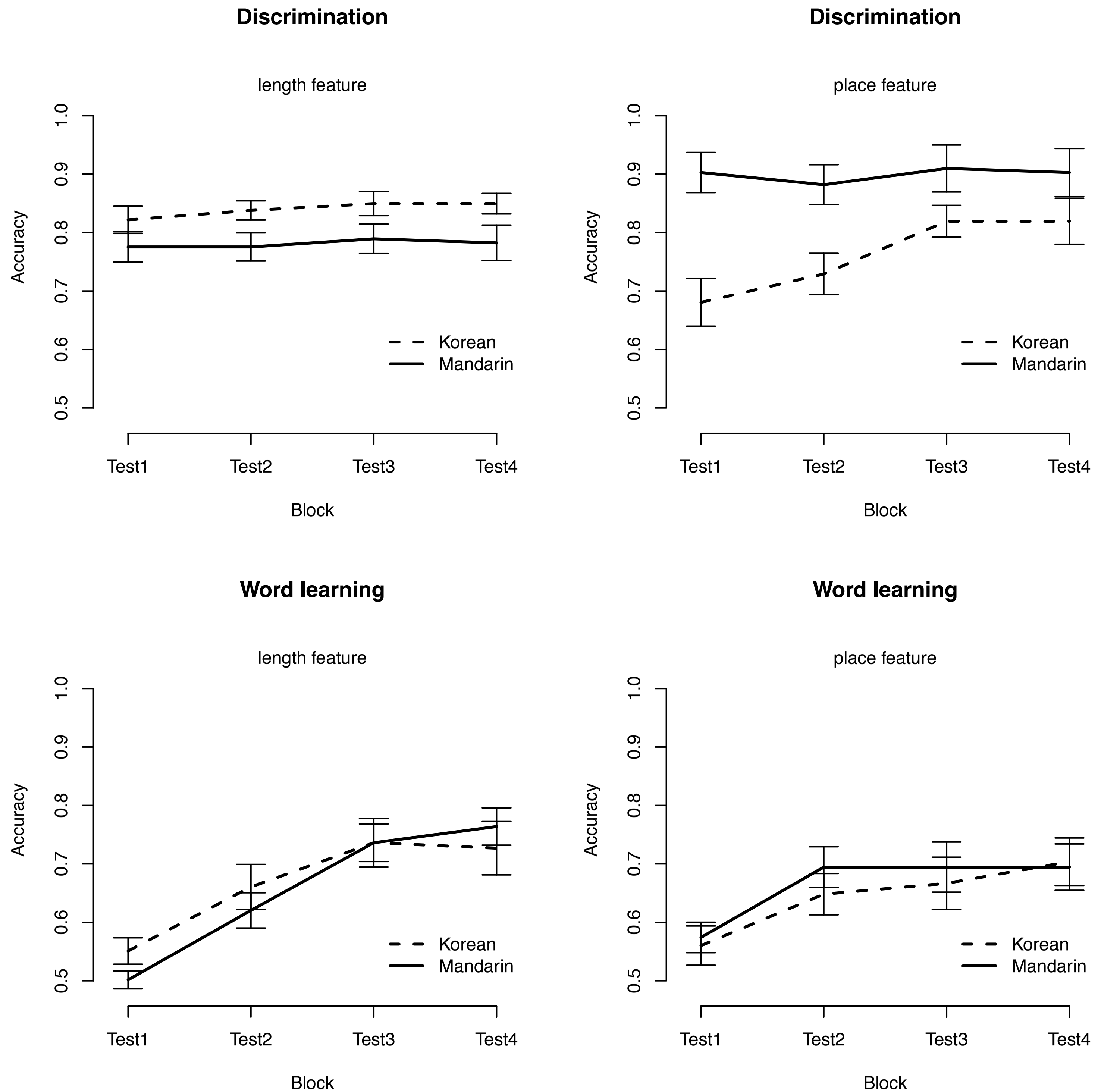

Discrimination
length feature
Accuracy
0.5
0.6
0.7
0.8
0.9
1.0
Test1
Test2
Test3
Test4
Block
Korean
Mandarin
Discrimination
place feature
Accuracy
Block
Korean
Mandarin
Word learning
length feature
Accuracy
Block
Korean
Mandarin
Word learning
place feature
Accuracy
Block
Korean
Mandarin

Figure B3

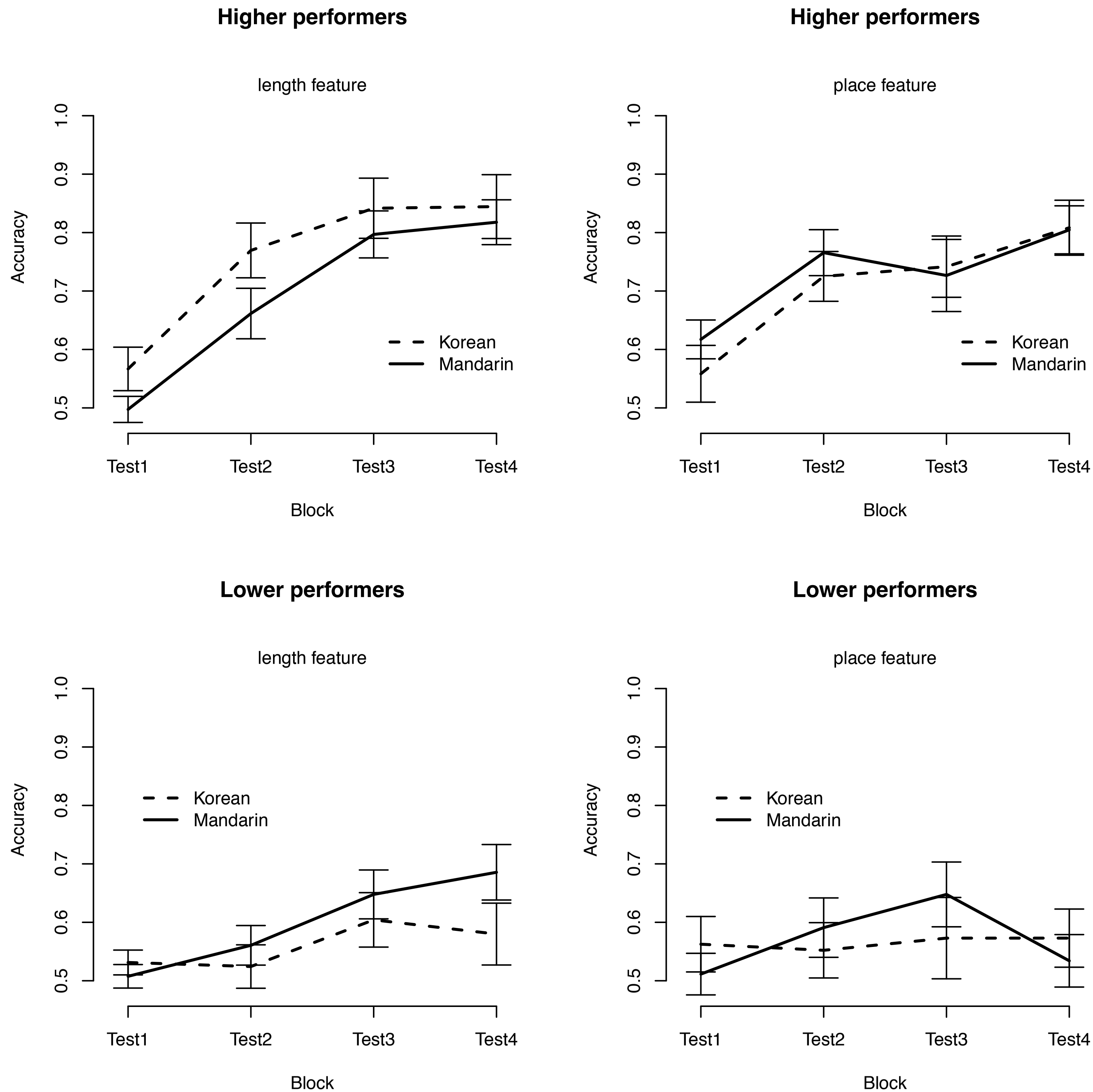

Higher performers
length feature
Accuracy
0.5
0.6
0.7
0.8
0.9
1.0
Korean
Mandarin
Test1
Test2
Test3
Test4
Block
Higher performers
place feature
Accuracy
0.5
0.6
0.7
0.8
0.9
1.0
Korean
Mandarin
Test1
Test2
Test3
Test4
Block
Lower performers
length feature
Accuracy
0.5
0.6
0.7
0.8
0.9
1.0
Korean
Mandarin
Test1
Test2
Test3
Test4
Block
Lower performers
place feature
Accuracy
0.5
0.6
0.7
0.8
0.9
1.0
Korean
Mandarin
Test1
Test2
Test3
Test4
Block

Figure B4

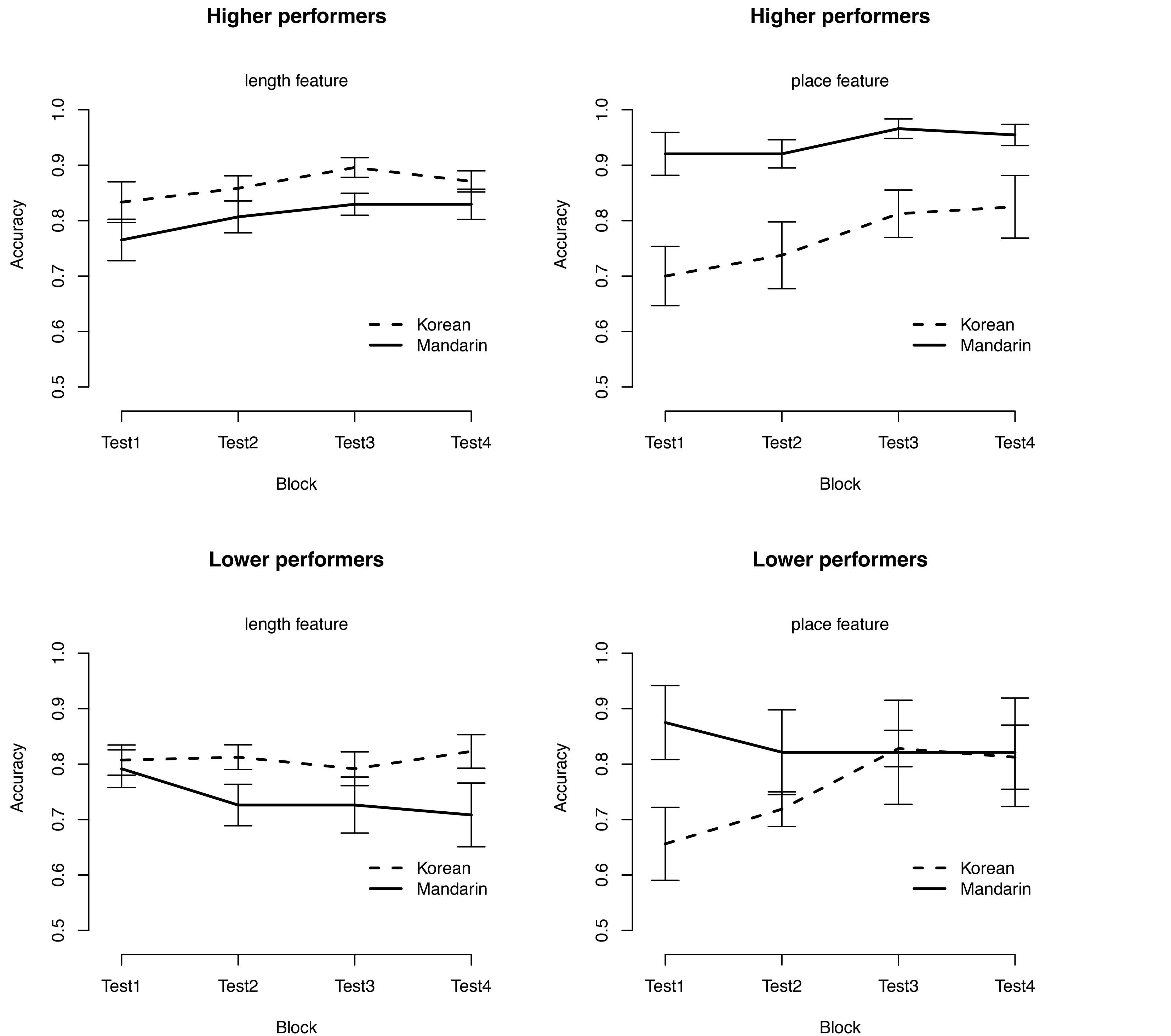

Higher performers
length feature
Accuracy
0.5
0.6
0.7
0.8
0.9
1.0
Korean
Mandarin
Test1
Test2
Test3
Test4
Block
Higher performers
place feature
Accuracy
Korean
Mandarin
Block
Lower performers
length feature
Accuracy
Korean
Mandarin
Block
Lower performers
place feature
Accuracy
Korean
Mandarin
Block

Figure 1. Example of a screen shot from the word-learning task.

Figure 2. Results for *dissimilar*, *similar* and *highly-similar* (*length* for Korean, *place* for Mandarin) trials. Accuracy scores indicate proportion of correct responses and error bars are standard errors.

Figure 3. Results for all *highly-similar* trials: *length* and *place*. Accuracy scores indicate proportion of correct responses and error bars are standard errors.

Figure 4. Word-learning task results (*length* and *place* trials) split by higher and lower performers (split-1; results for split-2 were equivalent). Accuracy scores indicate proportion of correct responses and error bars are standard errors.

Figure 5. Discrimination task results (*length* and *place* trials) split by higher and lower performers. Accuracy scores indicate proportion of correct responses and error bars are standard errors.

Figure B1. Results for *dissimilar*, *similar* and *highly-similar* (*length* for Korean, *place* for Mandarin) trials by block. Accuracy scores indicate proportion of correct responses and error bars are standard errors.

Figure B2. Results for all *highly-similar* trials by block: *length* and *place*. Accuracy scores indicate proportion of correct responses and error bars are standard errors.

Figure B3. Word-learning task results by block (*length* and *place* trials) split by higher and lower performers (split-1; results for split-2 were equivalent). Accuracy scores indicate proportion of correct responses and error bars are standard errors.

Figure B4. Discrimination task results by block (*length* and *place* trials) split by higher and lower performers (split-1; results for split-2 were equivalent). Accuracy scores indicate proportion of correct responses and error bars are standard errors.