Blocking effects in the learning of Chinese classifiers

Jing Z. Paul[1] and Theres Grüter[2]

[1]Agnes Scott College

[2]University of Hawai‘i at Mānoa

Author Note

We gratefully acknowledge support from the Chong-fong and Grace Ning Studies Fund to J. Paul for paying participants in Experiment 1. Our thanks to our annotators Xiaoyu Chen and Fengqun Zhu, to Xue Xia, Jing Wu, I-Chun Peir and Judy Shoaf for helping with participant recruitment and data collection, and to Nicole Ziegler, the audiences of SLRF 2014 and BUCLD 39 for helpful feedback. We thank the three anonymous reviewers and the associate editor, Kara Morgan-Short, for their insightful comments. This research was conducted while the first author was a graduate student at the University of Hawai‘i (Experiment 1) and then faculty at the University of Florida (Experiment 2).

Corresponding author: Theres Grüter, Department of Second Language Studies, University of Hawai‘i at Mānoa, 1890 East-West Road, Moore Hall, room 570, Honolulu, HI 96822, U.S.A.; email: theres@hawaii.edu

ACCEPTED for publication in *Language Learning*, 1/25/2016

Abstract

This paper investigates order-of-learning effects on the acquisition of classifier-noun associations in Chinese in two experiments modeled after an artificial language learning study by Arnon and Ramscar (2012). In Experiment 1, learners with no prior exposure to Chinese showed better learning of classifier-noun associations when exposed to larger units (sentences) before smaller ones (words) than vice versa, replicating Arnon and Ramscar's findings in a natural language. In Experiment 2, learners with 5-7 weeks of classroom exposure to Chinese completed the same experimental procedure. No order-of-learning effects emerged, suggesting even very basic prior knowledge of Chinese eliminated the advantage of initial exposure to larger units. These findings shed light on the extent to which blocking effects arising from carefully controlled training in the laboratory scale up, both from artificial to natural language learning, and to learning contexts involving relevant experience beyond a controlled training phase. Implications for L2 instruction and curriculum design are discussed.

Keywords: blocking, Chinese, classifiers, prior knowledge, adult L2 learning

Introduction

The undeniable differences in ultimate attainment between child and adult language learners have given rise to theoretical debate for decades. Various proposals have been advocated, including perhaps most prominently the Critical Period hypothesis (Penfield & Roberts, 1959; Lenneberg, 1967; see Long, 2013, and DeKeyser, 2013, for recent discussions), but also accounts emphasizing differences in properties of the input and language environment (e.g., Jia & Aaronson, 2003; Llanes & Muñoz, 2013). An alternative line of argument, associated most prominently with the work of N. C. Ellis and Brian MacWhinney, has drawn on the principles of associative learning theory, which aims to explain animal and human learning more generally (Rescorla & Wagner, 1972; Shanks, 1995), in order to account for limited attainment in second language (L2) acquisition (Ellis, 2006a,b; Ellis & Sagarra, 2010, 2011; Ellis, Hapeez, Martin, Chen, Boland, & Sagarra, 2014; MacWhinney, 1987, 2001, 2008; MacWhinney, Bates, & Kliegl, 1984). These accounts have focused specifically on the role of learners' prior linguistic and learning experience, and the consequences of this previous experience for the allocation of attention when learning a new language (see also Luk & Shirai, 2009). In this paper, we hope to contribute to this line of inquiry by further probing the phenomenon of blocking, or "learned inattention" (Ellis, 2006b; Kruschke & Blair, 2000). We present findings from two experiments aimed to replicate an artificial language learning study (Arnon & Ramscar, 2012) in a natural language (Mandarin Chinese) with two different learner groups: learners with no prior experience with Chinese (Experiment 1), and learners with five to seven weeks of classroom exposure to Chinese (Experiment 2). Blocking as a result of order-of-learning is observed in Experiment 1, replicating the results of Arnon and Ramscar (2012), but not in Experiment 2. These findings shed light on the extent to which effects of carefully controlled prior experience

in a laboratory setting can be expected to scale up, both from artificial to natural language learning, and to learning contexts where relevant prior experience cannot be as tightly controlled as in the laboratory, as is typically the case in classroom instructed foreign language learning.

Blocking and order-of-learning effects

The phenomenon of blocking was first described and formalized within behaviorist approaches to animal learning (Kamin, 1969; Rescorla, 1968; Rescorla & Wagner, 1972), and further explored in current models of associative learning more generally (e.g., Kruschke & Blair, 2000; Haselgrove, Esber, Pearce, & Jones, 2010; Le Pelley, Haselgrove, & Esber, 2012; Ramscar, Dye, Popick, & O'Donnell-McCarthy, 2011). Blocking arises when previous knowledge leads a learner to disregard, or not attend to, a certain cue due to the fact that this cue does not provide additional information with regard to a specific association that is to be learned; consequently the learner fails to learn information associated with that cue. For example, if a rat has learned that a conditioned stimulus (CS, e.g., a tone) is reliably associated with an unconditioned stimulus (US, e.g., a mild electric shock), it will fail to pay attention to, or learn, an additional CS (e.g., a light) that is associated, equally reliably, with the US. Since knowledge of the association between the second CS (light) and the US (shock) does not increase the probability of predicting the US, it constitutes a redundant cue. In other words, the rat does not gain anything by paying attention to the light, hence attention to the light and its association with the shock, is *blocked*.

The role of blocking in second language acquisition (SLA) was discussed from a learning theoretical perspective in Ellis (2006b), and has been examined empirically in a series of studies by Ellis and colleagues on the acquisition of temporal reference in Latin (Ellis & Sagarra, 2010,

2011; Ellis et al., 2014). In these studies, participants were divided into three groups. In a pretraining phase, one group learned adverbs and their temporal reference (e.g., *heri* – 'yesterday'), another learned inflected verbs and their temporal reference (*cogitavi* – 'I thought'), while a third (control) group received no pretraining. In a second phase, everyone was exposed to congruent adverb-verb combinations (*heri cogitavi* – 'yesterday I thought') and their temporal reference. In a third (testing) phase, participants were asked to indicate the temporal reference of incongruent adverb-verb combinations (*heri cogitabo* – 'yesterday I will think'). Results showed that participants in the adverb pretraining group relied primarily on the adverb in their interpretation of these incongruent combinations, whereas the opposite (to a lesser extent) was the case for the verb pretraining group; participants in the control group used both cues to a similar extent. These findings indicate that prior exposure to one cue blocked learners' attention to the other cue. Ellis et al. (2014) further explored the effect of learned attention through the use of an eye-tracking measure, which showed that prior knowledge of a cue led to overt attention (i.e., longer dwell time) to this cue and resulted in covert attentional biases in subsequent comprehension and production tasks. Furthermore, and of particular relevance to the present study, Ellis and Sagarra (2011, Experiment 2) presented results from native speakers of Chinese who completed the same procedure as the English-speaking participants in the control condition of the experiment described above (i.e., no pretraining). A comparison of the Chinese- and the English-speaking participants in this condition showed that the Chinese speakers relied on adverb cues more than the English speakers did, suggesting that their experience with a first language (L1) that lacks verbal morphology influenced their attention to the two available cues to temporal reference in the Latin stimuli in the experiment. This is noteworthy because it

demonstrates that real-life experience, not just controlled pretraining exposure, can affect attentional biases in language learning.

At the same time, it is relevant to note that the three groups in the experiments by Ellis and colleagues differed from each other not only in which cue they were exposed to first, but also in their cumulative experience with each cue. In other words, by the beginning of the test phase, participants in the adverb pretraining group, for example, had seen more adverbs than those in the other groups. This leaves open the possibility that the greater attention they allocated to this cue stemmed not (only) from being exposed to it *first* – the critical assumption of a blocking account – but from being exposed to it *more.* While the findings from these studies are fully consistent with a blocking account, cumulative frequency of exposure to each cue was not controlled between groups in the experimental design that was employed. Hence it remains open to what extent the observed group differences are due to blocking as a result of order of learning, or to cumulative frequency of exposure.

In a study inspired by formal learning theory, Arnon and Ramscar (2012, henceforth A&R) presented an experiment that elegantly eliminates this potential confound. We describe their study in some detail here as the two experiments we present below replicate A&R's method and procedure. In A&R, native English speakers were exposed to an artificial language in one of two learning conditions: sentence-first or word-first. In the sentence-first condition, participants were exposed to a block of sentences *before* a block of nouns. In the word-first condition, they were exposed to a block of sentences *after* a block of nouns. Importantly, by the end of the training phase, cumulative frequency of exposure to all stimuli was identical across groups. In the noun block, stimuli consisted of an (aurally presented) individual word in the artificial language (e.g., *viltord*) accompanied by a visual stimulus depicting an object (e.g., airplane). In

the sentence block, linguistic stimuli consisted of an entire "sentence" in the artificial language, i.e., a string of sounds consisting of an invariant carrier phrase, an article, and a noun (e.g., *os ferpel en* + *bol* + *viltord*). Critically, the artificial language contained two articles, each associated with a separate set of nouns, thus mimicking the distribution of articles in a gender-marking language like Spanish. The visual stimuli in the sentence block consisted of an (invariant) image of a man gesturing towards the object associated with the noun (i.e., the last word in the sentence).

A&R predicted that learning of article-noun associations – framed in terms of the learning of grammatical gender, a notorious example of non-native attainment in L2 acquisition (Guillelmon & Grosjean, 2001; Grüter, Lew-Williams, & Fernald, 2012, *inter alia*) – would be superior in the sentence-first compared to the word-first group, as per the following rationale: For participants in the sentence-first condition, the relationship between the gesturing figure and the object (e.g., the airplane) in the visual stimulus on the one hand, and the complex acoustic stimulus (carrier phrase + article + noun) on the other, is initially entirely open in terms of which part of the acoustic stimulus should be associated with which part of the visual stimulus. Over time, participants will notice that the carrier phrase and the gesturing figure stay the same across trials, while the latter part of the sentence (article + noun) and the object in the visual display change. Thus evidence accumulates that the carrier phrase is unlikely to refer to the object, while the relationship between the article-noun sequence and the object is strengthened. Yet it is only as trials proceed that the occurrence of the same article with multiple nouns could become apparent. Until that point, and possibly not until they encounter nouns on their own in the noun block, participants in the sentence-first condition may continue to entertain the hypothesis that the article-noun sequence as a whole constitutes the label for the object, thus strengthening the

representation of the associative relationship between article and noun. In the word-first condition, on the other hand, learners can easily and quickly make the association between the noun they hear and the object they see in the noun block. When subsequently exposed to the sentence block, their prior knowledge of noun labels will then allow them to orient immediately to the noun within the sentence they hear, thus blocking attention to the rest of the sentence, including the article preceding the noun, which at no point would present a potentially competing cue for association with the visual stimulus. As a consequence, a much weaker associative relationship between articles and nouns is built by learners in the word-first, compared to those in the sentence-first condition. Results from a forced-choice comprehension task and an oral production task, conducted after the two learning blocks (and a brief distractor block), both supported A&R's prediction that participants in the sentence-first group would be better at learning article-noun associations than those in the word-first group. Participants in the sentence-first group were faster and more accurate in the forced choice task, and more likely to produce a correct article-noun sequence in the production task.

In a follow-up study, Siegelman and Arnon (2015) presented additional and more direct evidence for the benefits of initial exposure to multiword units for the learning of associative relations. In a similar paradigm to that in A&R, participants were initially exposed to either acoustically unsegmented or segmented (by pauses between words) sentences. Participants in the unsegmented-first condition, similar to the sentence-first group in A&R, were more successful at learning article-noun sequences. Moreover, those participants who were more likely to type article-noun sequences as one word (without a space) as opposed to two during the training phase were also more successful at learning article-noun sequences. These findings provide further evidence for a causal link between the use of multiword units and successful learning of

associative relations. In a second experiment, Siegelman and Arnon (2015) also illustrated the limitations of this link: when the relation between the article and the noun/object was no longer purely associative but semantically informative (the two articles were paired with animate and inanimate objects respectively), initial exposure to unsegmented input no longer conferred learning benefits. This is precisely what a blocking account predicts: the relation between the (animacy-marking) article and the object is no longer entirely subsumed by the association between the noun and the object; hence the article does not lack informativity, and consequently should continue to be attended to.

These findings present evidence for blocking in (artificial) language learning in the absence of any confounds of frequency, thus adding to the growing body of evidence that prior experience and order of learning affect the allocation of attention and subsequent learning in (second) language acquisition. The goal of the present study is to further reconcile these converging approaches to language learning from applied linguistics and cognitive psychology by employing A&R's experimental design with stimuli from a natural language, and with learners who have prior experience with the language – albeit minimal and as tightly controlled as possible outside a laboratory. Exploring whether findings from artificial language learning studies scale up to natural language and learning contexts is important for understanding the generalizability of such findings for language learning and instruction more generally. Moreover, the use of natural language stimuli allows for the possibility to extend the paradigm to learners who have had exposure to that language prior to the experiment in the lab. At a theoretical level, this allows for a more extensive investigation of the effects of prior experience and the variation therein. At an applied level, it provides for an ecologically more valid setting, given that in the real world, L2 learners almost never approach a learning task without any potentially relevant

prior experience. If we are to draw inferences about L2 learning from findings of artificial language learning experiments, it is thus essential that we examine carefully whether effects such as those observed by A&R continue to arise under circumstances that are more directly comparable to language learning in the real world.

This Study

Overview

The present study is essentially a replication of A&R, with one important difference: We used stimulus materials from a real language, Chinese, instead of an artificial one. This allowed us to further explore the role of prior knowledge in order-of-learning effects by including participants who had no prior knowledge of Chinese (Experiment 1) as well as participants who had had 5-7 weeks of classroom exposure to Chinese immediately prior to the experiment (Experiment 2). The primary goal of Experiment 1 was to test whether the effects observed by A&R with carefully constructed artificial language stimuli could be replicated with natural language materials, which, however carefully selected, will contain additional complexities inherent to a natural language. In other words, Experiment 1 probes to what extent A&R's findings scale up to a natural language context. Experiment 2 takes the next step by moving towards more natural *learning* contexts. Participants in Experiment 2 came to the experimental learning task with some basic prior knowledge of Chinese through classroom instruction. The specific nature of this prior knowledge, particularly as it pertains to classifiers, was controlled and assessed to the extent possible given the circumstances. We report on these measures in detail in the context of Experiment 2. This prior knowledge may affect the allocation of attention and the blocking effects that arise as a consequence, as we will discuss in more detail below.

Experiment 2 thus presents a first step towards assessing the robustness of the order-of-learning effects observed in A&R's laboratory experiment once minimal prior knowledge of the target language learned beyond the laboratory comes into the picture.

The materials and procedures used in Experiments 1 and 2 were identical. We thus begin by describing the methods common to both experiments, followed by the description of participants and presentation of results for each experiment separately.

Stimuli

Our goal was to create a linguistic stimulus set from existing Chinese vocabulary that would be as comparable as possible to the mini artificial language used in A&R. Chinese provides an ideal opportunity for implementing the noun-class distinction in A&R's artificial language through prenominal classifiers. Individual classifiers in Mandarin Chinese categorize nouns based on inherent properties of the objects they denote, and they obligatorily precede a noun when the noun phrase includes a numeral, as in (1). For the purpose of this experiment, we selected two individual classifiers (*ba, gen*) and 14 nouns denoting familiar concrete objects, 7 associated with *ba* and 7 with *gen*. The classifier *gen* is associated with rigid long-shape objects, whereas *ba* is used to refer to objects with a handle (Liang, 2009). The seven nouns in the experiment associated with *gen* were *xiangjiao* ('banana'), *yan* ('cigarette'), *yumao* ('feather'), *maisui* ('wheat head'), *shuzhi* ('twig'), *guaizhang* ('cane'), and *gutou* ('bone'); the nouns associated with *ba* were *yizi* ('chair'), *yaoshi* ('key'), *san* ('umbrella'), *jian* ('sword'), *jiandao* ('scissors'), *qiang* ('gun'), and *shuzi* ('comb'). These items were selected because they are easily depictable and, critically, not included in the vocabulary list of a leading introductory Chinese textbook for L2 learners at the college level, *Integrated Chinese* Level 1 Part 1 (Liu, Yao, Bi, Ge,

& Shi, 2009), thus making it highly unlikely that they would be familiar to participants in Experiments 1 or 2. Classifier-noun pairs were combined with an invariant carrier phrase (*zhe shi yi*, 'this is one') to create simple sentences consisting of the same three components as those in A&R, i.e., a carrier phrase, a noun-class-marker (here the classifier) and a noun, as illustrated in (1). The carrier phrase was selected from the vocabulary lists of the Introduction, Lesson 1 and Lesson 2 in *Integrated Chinese* Level 1 Part 1 (used in the class Experiment 2 participants were enrolled in), so that it would be familiar to participants in Experiment 2 (who were studying Lesson 4 & 5 at the time of data collection), but not to those in Experiment 1.

(1)	*zhe*	*shi*	*yi*	*gen*	*guaizhang*.
	this	is	one	CL	cane
	CARRIER PHRASE			CLASSIFIER	NOUN

A female native speaker of Chinese recorded the stimuli. The carrier phrase and the classifiers were recorded in sentential contexts where they were not subject to tone sandhi, a phonological process common in tone languages (Yip, 2002). In Chinese, the tone of a small set of characters (always one syllable for a character) changes depending on the tones of adjacent syllables (Li & Thompson, 1981). For example, the syllable *yi* ('one'), which carries the first tone when it stands alone, changes to the fourth tone when preceding a first-tone or a third-tone syllable (e.g., *yī* to *yì gēn* or *yì bǎ*). In naturalistic spoken Chinese, tone sandhi would affect the tone of the numeral *yi* ('one') in the carrier phrase, the classifier *ba*, and some of the nouns, such that these morphemes would be realized with different tones in different experimental items. These differences in tone would constitute a separate and potentially confounding cue here. For

this reason, we decided to create the acoustic stimuli such that each syllable retained its original tone, i.e., without undergoing tone sandhi. Thus recordings of the carrier phrase and the classifiers in neutral contexts were segmented to extract tokens of the carrier phrase and of each classifier. We selected one token of each, and concatenated these individual segments with recordings of the nouns to create 'sandhi-free' linguistic stimuli for the sentence condition. All recording and editing were done in Praat (Boersma & Weenink, 2014).

Although the participants in the experiments reported here had no (Experiment 1) or only minimal (Experiment 2) prior experience with Chinese, and were thus unlikely to be sensitive to the absence of tone sandhi, this modification nevertheless raised a potential concern that the stimuli might sound unnatural. In order to address this concern, the concatenated sound files were sent to seven native speakers of Chinese in China for evaluation before the experiments were conducted. All raters were blind to the purpose of the study. They were asked to rate each sound file on a scale of 1-5, with 1 as "not natural" and 5 as "very natural". The files were rated at an average of 3.6 (SD = 1.41). Two of the raters indicated that the main reason they rated the sound files as "3 out of 5" was that the speed of the sentences was slower than that of naturalistic conversation among native Chinese speakers. None of the raters commented on the absence of tone sandhi. We thus assumed that the absence of tone sandhi in the acoustic stimuli was unlikely to negatively impact the performance of the participants in our experiments.

Visual stimuli were created to be analogous to those in A&R. As in A&R, 14 black-and-white images depicting the 14 nouns were selected and modified so that all images were approximately similar in size. For the sentence block, these images were combined with an additional (invariant) image of a girl gesturing towards the object. Figure 1 presents examples of visual stimuli in the noun and sentence learning blocks. In the noun block, visual stimuli such as

that in panel A were paired with auditory stimuli consisting of a bare noun only (e.g., *san*, 'umbrella'); in the sentence block, visual stimuli such as that in panel B were paired with auditory stimuli consisting of sentences like that in (1).

[INSERT FIGURE 1 ABOUT HERE]

In addition to the experimental items, a set of distractor items was created (as in A&R). Items in the distractor set were constructed from the same carrier phrase, two different noun class markers, and an additional set of nouns. Unlike in the experimental items (and as in A&R), noun class markers and nouns in the distractor block were not consistently associated with each other; in other words, a noun could appear with either of the two noun class markers. The Chinese classifier system offers a natural implementation of this scenario: group classifiers such as *shuang* ('pair') and *fu* ('pair/pack/deck'), can occur with an overlapping set of nouns. In the distractor set, these two classifiers were thus both paired with the same four nouns (*chibang* ('wings'), *kuaizi* ('chopsticks'), *erhuan* ('earrings'), and *shoutao* ('gloves')). In addition, reflecting the natural occurrence of these classifiers in Chinese, only *shuang* was also paired with *yanjing* ('eye'), *xie* ('shoe'), and *jiao* ('foot'), and only *fu* was also paired with *xiaolian* ('smiley face'), *yanjing* ('glasses'), and *wankuai* ('bowl and chopsticks'). A total of 35 distractor sentences (as in A&R) was created following these constraints, and paired with relevant images.

Procedure

As in A&R, the experiment consisted of two phases: a learning phase and a test phase. Table 1 provides an overview of the general procedure. In the learning phase, participants were

exposed to two types of stimuli: a block of nouns and a block of sentences. In the noun block, for each trial, an image of the object was presented on the screen, and participants heard the corresponding noun (Fig. 1A). In the sentence block, an image of the object together with a girl gesturing to it was displayed, and participants heard the corresponding sentence (Fig. 1B). Participants were instructed to repeat each word or sentence aloud to reinforce learning, and were told that they would later be tested on what they had learned. The task was presented in PowerPoint using the Full Screen View display. As in A&R, each learning block was comprised of 70 items, with each object named five times. All participants were exposed to exactly the same 2 x 70 items during training, thus there were no differences in frequency of exposure to different items or item types between participants. The only difference between participants was that those in the word-first condition completed the noun block first, followed by the sentence block, while those in the sentence-first condition completed the two learning blocks in the opposite order. Following the two experimental learning blocks, all participants were exposed to the same distractor block consisting of 35 items (described above), following A&R's procedure and rationale of interspersing a brief distractor phase before testing in order to control for potential effects of recency on participants' performance in the test phase.

[INSERT TABLE 1 ABOUT HERE]

As in A&R, the test phase immediately followed the learning phase and consisted of two separate tests: a forced-choice task and an oral production task. In the forced-choice task, participants saw an image of an object from the learning phase and heard two sentences; their task was to indicate which of the two sentences was a better description of the image. As in

A&R, the task comprised 28 trials, half of which tested participants' knowledge of nouns (=noun trials), the other half their knowledge of classifier-noun pairings (=classifier trials). On noun trials, the two sentences differed only in the noun label. For example, participants saw an image of a twig and heard: **Zhe shi yi gen **yaoshi*** ('this is one CLASSIFIER-FOR-RIGID-LONG-SHAPE-OBJECTS **key**') vs. *Zhe shi yi gen **shuzhi*** ('this is one CLASSIFIER-FOR-RIGID-LONG-SHAPE-OBJECTS **twig**'). On classifier trials, the two sentences differed only in the classifier. For examples, participants saw an image of a twig and heard: *Zhe shi yi **gen** shuzhi* ('this is one **CLASSIFIER-FOR-RIGID-LONG-SHAPE-OBJECTS** twig') versus **Zhe shi yi **ba** shuzhi* ('this is one **CLASSIFIER-FOR-OBJECTS-WITH-A-HANDLE** twig'). Each of the 14 images appeared once in a noun trial and once in a classifier trial in pseudorandomized order. The order in which the two sentences were presented was counterbalanced between participants. Participants listened to the sentences by clicking on the sound icon on the screen, which was linked to a single audio file consisting of both sentences, prefaced by "A" and "B". Participants were told that only one sentence was correct and they must select one. The task was presented in PowerPoint using the Normal View display. Participants were instructed to click on each sound icon only once, and to type "A" or "B" into the 'notes' field to make their response.

In the production task, participants saw each of the 14 images once, and were asked to describe it, as in A&R. They were instructed to try and produce complete sentences, to the best of their abilities. This task was audio recorded for later transcription and analysis.

Each task was preceded by instructions displayed on the screen along with two practice items (including linguistic and visual stimuli not included in the experimental materials). No time constraints were imposed in either the learning or the test phase, nor were any measures of reaction time collected.[1] It generally took participants approximately 15 minutes to complete the

learning phase and 10 minutes to complete the test phase (7 minutes for the forced-choice task and 3 minutes for the production task).

Data coding

Participants' responses in the production task were transcribed and coded by two independent research assistants, both native Chinese-speaking graduate students at a university in China. The coding scheme was created to reflect A&R's criteria as closely as possible, but complexities associated with natural language materials had to be taken into consideration. Classifiers and nouns were coded as correct or incorrect (1/0); as in A&R, they were coded as correct if they did not differ from the target in more than one phoneme. Coders were instructed to disregard tone, as this constituted an additional layer of representation not present in A&R, and one that it would be unrealistic for participants to learn to produce after only 15 minutes of exposure. Since the key interest in this study was whether participants would be able to produce correct classifier-noun sequences, each response was additionally coded for whether this sequence was produced correctly. Classifier-noun pairs were coded as correct only if both the classifier and the noun were coded as correct. In addition, the accuracy of the carrier phrase was coded on a scale slightly modified from that used in A&R (0-no carrier phrase, 1-not accurate, 2-partially accurate, 3-fully accurate). The 0 was added to the scale used by A&R to distinguish between responses where no carrier phrase was attempted, and ones where an attempt was made but the resulting production was far from the target.

Experiment 1

The goal of this experiment was to replicate A&R's artificial language experiment with natural language stimuli. Participants had no prior experience with Chinese, thus the Chinese materials in this experiment were as novel to them as the artificial language stimuli were to the participants in A&R's study. We thus expect to find the same pattern of results observed by A&R, namely an advantage for the participants in the sentence-first compared to those in the word-first condition with regard to the learning of classifier-noun associations. This should manifest as a significant effect of learning condition on responses in both the forced-choice and the production task.

Participants

Forty-eight native speakers of English from the University of Hawai'i community participated in Experiment 1; half of them were assigned to the word-first, the other half to the sentence-first condition. An additional 8 were tested but excluded from analysis because they indicated extensive exposure to a language other than English during childhood (5), did not follow instructions (2), or reported having prior experience with Chinese (1). All 48 participants in the final sample (29 female, 19 male; mean age 24 years) indicated that they had at least some knowledge of a language other than English through classroom instruction during high school or college. 33 participants (16 in the sentence-first, 17 in the word-first condition) indicated some knowledge of a gender-marking language (Romance, Slavic or Germanic).[2] None of them reported any prior exposure to Chinese. Participants were paid $5 or received course credit for their participation.

Results

Forced-choice task

Results from the forced-choice task are illustrated in Figure 2 in terms of accuracy by learning condition (sentence-first, word-first) and trial type (classifier, noun). Participants were generally more accurate on noun (sentence-first: M = 89% (SD = 17), word-first: M = 87% (SD = 10)) than on classifier trials (sentence-first: M = 57% (SD = 15), word-first: M = 48% (SD = 15)). In order to test the critical hypothesis regarding the effect of learning condition, we used mixed-effects logistic regression modeling (implemented in R using the lme4 package; Bates, Maechler, & Bolker, 2011). Response (correct/incorrect) was modeled as the binary outcome. Learning condition (sentence-first/word-first) and trial type (noun/classifier) were treated as fixed effect predictors, and centered using deviation coding (-.5, .5) in order to make interactions interpretable (Barr, 2013). Participants and items were treated as random effects, using the maximal random effects structure justified by the design (Barr, Levy, Scheepers, & Tily, 2013; here: (1+ trial type | participant) + (1 + trial type * learning condition | item)). We used a forward-fitting strategy in model building, starting with a simple model with only learning condition as a single fixed effect (Model 1). In Model 2, trial type was added as a second fixed effect; Model 3 additionally contained the interaction term. Model fit was evaluated using the *anova* function in R. Model 2 – with two fixed effects but no interaction – emerged as the best fitting model, and revealed significant effects of both learning condition (B = .36, Z = 2.3, p = .02) and trial type (B = 2.56, Z = 7.2, p < .001), indicating that, as in A&R, participants in the sentence-first condition performed better overall than those in the word-first condition, and that participants overall were better at identifying nouns than classifier-noun associations.

Based on preliminary data inspection and anecdotal evidence, we suspected that having previous knowledge of a language with grammatical gender could lead to enhanced performance

on this task, regardless of learning condition (see also note 2). In an attempt to control for this potential confound, we tried to balance the proportion of participants with such knowledge evenly between the two learning conditions. In addition, to test for an effect statistically, we added 'knowledge of a gender-marking language' as a binary predictor to Model 2. The effect of this predictor was not significant (B = -.02, Z = -.1, p = .90), nor did it improve model fit. Indeed, accuracy among the 33 participants with knowledge of a gender-marking language on both classifier (M = 52%, SD = 17) and noun (M = 88%, SD = 14) trials was almost identical to that among the 15 participants without such knowledge (classifiers: M = 53%, SD = 13; nouns: M = 86%, SD = 14). Thus contrary to our initial concerns, knowledge of a gender-marking language did not appear to affect participants' performance in this experiment.

Following A&R, we also conducted one-sample t-tests comparing accuracy in each condition to chance (50%). This indicated that on classifier trials the sentence-first (57%, $t(23)$ = 2.282, p =.03, d = .47), but not the word-first (48%, $t(23)$= -.689, p =.50, d = .13) group performed above chance. Both groups were significantly above chance on noun trials (sentence-first: 89%, $t(23)$ = 11.256, p < .001, d = 2.29; word-first: 87%, $t(23)$ = 18.05, p < .001, d = 3.70). This pattern again mirrors the findings from A&R's artificial language experiment.

Production task

Interrater agreement for the coding of correct/incorrect classifier-noun pairs was 96%, and 81% for accuracy of the carrier phrase coded on a 0-3 scale. When there was disagreement between the two coders, the first author acted as a third coder to resolve the difference.

The production task was challenging for participants, as was to be expected after only about 15 minutes of exposure to a new language. Figure 3 illustrates the mean percentage of

correct classifier-noun pairs produced by participants in the sentence-first (M = 28%, SD = 25) and the word-first (M = 16%, SD = 19) conditions. Mixed-effects logistic regression modeling was used to probe for effects of learning condition on the production of correct classifier-noun pairs. Learning condition was entered as a fixed-effect predictor using deviation coding (-.5, .5). Participants and items were treated as random effects; the best fitting model included random intercepts only. The effect of learning condition was significant (B = -1.25, Z = -2.15, p = .03), indicating that participants in the sentence-first condition were more likely to produce a correct classifier-noun pair than those in the word-first condition. This mirrors the findings from the forced-choice task.

Following A&R, an independent-sample t-test was used to compare participants' accuracy in producing the carrier phrase across learning conditions. This revealed no significant difference between the sentence-first (M = 2.37, SD = 0.78) and the word-first (M = 1.89, SD = 1.14) groups, $t(46) = 1.68$, p = .10, d = .49, again mirroring the pattern observed in A&R's study.

[INSERT FIGURES 2 AND 3 ABOUT HERE]

In sum, the findings obtained in Experiment 1 fully replicate those observed by A&R in showing an advantage for the sentence-first group on both the forced-choice and the elicited production task, indicating that exposure to larger units before smaller units was beneficial for the learning of classifier-noun associations in the absence of any prior experience with the language.

Experiment 2

The goal of Experiment 2 was to test whether the learning advantage of the sentence-first group observed by A&R and replicated in Experiment 1 would continue to emerge when learners came to the experiment with some prior knowledge of the target language learned outside the laboratory, as is typically the case in real-life language learning. From an applied perspective, it is important to understand to what extent the order-of-learning effects observed in laboratory-based experiments such as these may be relevant in real-life learning contexts, such as classroom based foreign language instruction. A possible implication of the advantage observed for the sentence-first groups in A&R and Experiment 1 could be that curricula should be structured so as to present beginning learners with noun-class markers – gender-marked determiners or classifiers – in less segmented input, that is, together with associated nouns and without separate, explicit explanation of what they may denote. However, the existing laboratory-based findings are not sufficient to allow us to conclude that this would indeed be beneficial. Thus the applied goal of Experiment 2 is to test whether actual beginning learners of Chinese, otherwise exposed to typical college-level instructional materials, would also experience a learning benefit from the sentence-first condition.

The theoretical goal of this experiment is to extend the investigation of blocking in language learning to a context where the nature of prior knowledge is more complex. To this end, it is critical to capture the specific knowledge that participants in Experiment 2 had prior to the experiment as precisely as possible. We took two major steps to achieve this. First, all participants completed an 'exit survey' (described in more detail below) designed to gauge the knowledge they had prior to the experiment of the critical 2 classifiers and 14 nouns. Second, all participants were recruited from sections of the same first-semester Chinese class, and tested after 5-7 weeks of instruction, when they were studying Lessons 4 and 5 in *Integrated Chinese*

Level 1 Part 1 (Liu et al., 2009). At that point, the vocabulary used in the carrier phrase (*zhe shi yi*, 'this is one'), but neither the two classifiers nor the 14 nouns, had been introduced. In Lesson 2, students were, however, introduced to the concept of measure words (a term typically used for classifiers in L2 Chinese textbooks), and to the general classifier *ge.* The relevant passage from the textbook is reproduced in (2).

(2) **Measure Words**

In Chinese a numeral is usually not followed immediately by a noun. Rather, a measure word is inserted between the number and the noun, as in (1), (2), and (3) below. Similarly, a measure word is often inserted between a demonstrative pronoun and a noun, as in (4) and (5) below. There are over one hundred measure words in Chinese, but you may hear only two or three dozen in everyday speech. Many nouns are associated with special measure words, which often bear a relationship to the meaning of the given noun.

个 (gè/ ge) is the single most common measure word in Chinese. It is also sometimes used as a substitute for other measure words.

1) 一个人

yí ge rén

(a person)

(Liu et al., 2009, p. 45)

This particular prior knowledge about Chinese could, in principle, allow participants in the sentence-first block to do the following: (i) immediately eliminate the carrier phrase as a potential label for the object, (ii) posit that the first syllable after the carrier phrase may be a

measure word, and thus (iii) identify the remainder of the sentence as the noun labeling the object. In other words, this prior knowledge would be sufficient for segmenting the input to allow extraction of the noun. If this is the case, the learning experience in the sentence block should become more similar to that in the noun block (and to Siegelman and Arnon's segmented condition). As a result, we should see differences between the sentence-first and word-first groups diminish in Experiment 2.

Participants

A total of 55 students from beginner level Chinese classes at the University of Florida participated in Experiment 2; 48 were retained in the final sample. The 7 exclusions were due to equipment failure (3), substantial exposure to a gender-marking language during childhood (1), or prior knowledge of one or both classifiers used in the experiment, as indicated by the exit survey described below (3). Of the 48 participants in the final sample (27 female, 21 male; mean age 20 years), 37 (16 in the sentence-first, 21 in the word-first condition) indicated some knowledge of a gender-marking language, and all indicated that English was their dominant language. Ten participants reported some exposure to a language other than English during childhood, none of which were a gender-marking language. For 4 of these 10 (1 in the sentence-first, 3 in the word-first condition), the early exposure was to a variety of Chinese. Since their responses on the exit survey indicated that they had little or no knowledge of the experimental items in this study, and given the difficulty of replacing participants due to the narrow time window for testing necessary to maintain amount of classroom instruction comparable across participants, these participants were retained in the final sample. For the same reason, we also retained the 4 participants (2 in the sentence-first, 2 in the word-first condition) who indicated

some prior instructional experience with Chinese, mainly through introductory Chinese classes during high school; their responses on the exit survey indicated little or no prior knowledge of the experimental items in the study. Nevertheless, all analyses reported below were repeated on a subset of the data excluding the 8 participants with prior exposure to Chinese; unless we note otherwise, the pattern of results was the same.

Materials and Procedure

Materials and procedure in Experiment 2 were identical to those in Experiment 1, with the exception of an additional 'exit survey' added at the very end. The exit survey consisted of a list of the 14 nouns and 2 classifiers used in the experimental trials. For each of these 16 words, participants were asked to indicate on a scale of 0 to 2 whether they had had any knowledge of them before the experiment (0: did not know the word; 1: heard the word before but wasn't sure what it meant; 2: knew the word). Those who marked 2 for either of the two classifiers were not included in the final sample. A small number of participants (3 in the sentence-first, 2 in the word-first condition) marked 1 for one or both classifiers. Three participants in each learning condition indicated knowledge ('2') of one or more nouns, with the overall percentage equally low in both groups (sentence-first: M = 2.7% (SD = 9.1), word-first: M = 2.1% (SD = 7.4)). When including both '1' and '2' ratings, the percentage of affected nouns was again comparable across groups (sentence-first: M = 9.2% (SD = 17.2), word-first: M = 9.2% (SD = 16.2)). For this reason, and in the interest of balanced samples, no further exclusions were made based on results from the exit survey. The results provide an illustration of the variability that is inevitable when we start looking at prior experience in more naturalistic contexts. Despite our best efforts to capture learners at a precise point in their instructional experience, and to construct experimental

stimuli based on material (not) introduced in their textbook, complete control over their prior experience was impossible. This reality will have to be borne in mind when using findings from laboratory induced prior experience to inform the design of classroom curricula.

Results

Forced-choice task

Results from the forced-choice task are illustrated in Figure 2 in terms of accuracy by learning condition (sentence-first, word-first) and trial type (classifier, noun). Participants were generally more accurate on noun (sentence-first: M = 96% (SD = 8), word-first: M = 93% (SD = 10)) than on classifier trials (sentence-first: M = 58% (SD = 18), word-first: M = 63% (SD = 20)).[3] In order to test for effects of learning condition, the same model building procedure and random effect structures as in Experiment 1 were applied. Model 1 contained only learning condition as a single fixed effect. In Model 2, trial type was added as a second fixed effect; Model 3 additionally contained the interaction term. Model fit was evaluated using the *anova* function in R. Model 3 emerged as the best fitting model, and revealed a significant effect of trial type (B = 3.41, Z = 7.8, p < .001), no main effect of learning condition (B = -.37, Z = -.9, p = .39), and a significant interaction (B = -1.17, Z = -2.0, p = .045).[4] Adding 'knowledge of a gender-marking language' as an additional predictor did not improve model fit. In order to further explore the significant interaction term, separate models were fit to the data from the classifier and the noun trials respectively, with learning condition as the single fixed effect predictor. Learning condition was not significant in either of these models (ps > .1). Thus while the sentence-first group was numerically somewhat more accurate than the word-first group on noun

trials, and the opposite was the case for classifier trials, no clear group differences emerged in these follow-up models.

Following A&R and Experiment 1, we again conducted one-sample *t*-tests comparing accuracy in each condition to chance (50%). This indicated that both groups performed above chance on classifier (sentence-first: $M = 58\%$, $t(23) = 2.07$, $p = .050$, $d = .44$; word-first: $M = 63\%$, $t(23) = 3.01$, $p = .006$, $d = .65$) as well as on noun trials (sentence-first: $M = 96\%$, $t(23) = 27.88$, $p < .001$, $d = 5.75$; word-first: $M = 93\%$, $t(23) = 22.18$, $p < .001$, $d = 4.30$).[5]

In sum, as in Experiment 1, participants in Experiment 2 were better at learning noun labels than classifier-noun associations. Unlike in Experiment 1, however, results from the forced-choice task showed no effects of learning condition on participants' performance on either trial type. In order to examine whether participants' prior knowledge of Chinese affected performance on the forced-choice task, we compared results from Experiments 1 and 2 by combining the two datasets and adding Experiment as an additional fixed effect predictor. In a first model, we included the three predictors (experiment, learning condition, trial type) and no interactions. In a second model, we added the interaction terms between the three predictors. Model comparison indicated that the second model was a better fit. This model revealed significant main effects of experiment ($B = .80$, $Z = 3.5$, $p < .001$) and trial type ($B = 2.99$, $Z = 10.1$, $p < .001$), but not of learning condition ($B = -.35$, $Z = -1.6$, $p = .12$). The only interaction term that reached significance was that between experiment and trial type ($B = .92$, $Z = 2.3$, $p = .02$).[6]

Thus, it appears that participants in both experiments were more successful at learning noun labels than classifier-noun associations, and participants in Experiment 2 performed better than those in Experiment 1. This was still the case after exclusion of the 8 participants in

Experiment 2 who had some experience with Chinese prior to the Chinese class they were enrolled in at the time of testing. The overall increase in performance in Experiment 2 indicates that the five to seven weeks of classroom exposure to Chinese that Experiment 2 participants had experienced facilitated their learning of both new nouns and classifier-noun associations. Notably, their prior knowledge did not appear to block their attention to classifiers entirely, as indicated by above-chance performance on classifier trials by participants in both learning conditions in Experiment 2. We hypothesize that their general knowledge of measure words and the syntactic frames in which they appear allowed these learners to identify the syllable immediately following the carrier phrase as a classifier, thus facilitating the learning of the classifier, and potentially its association with the subsequent noun. At the same time, and critically, the advantage for the sentence-first group disappeared in Experiment 2, indicating that the order-of-learning effects observed in A&R and in Experiment 1 may be neutralized when participants have even minimal additional knowledge about the target language.

Production task

The same coders as in Experiment 1 coded the production results from Experiment 2. Interrater agreement was 91% for the coding of correct/incorrect classifier-noun pairs, and 96% for accuracy of the carrier phrase coded on a 0-3 scale. As in Experiment 1, the first author acted as a third coder in case of discrepancies.

Figure 3 illustrates the mean proportion of correct classifier-noun pairs produced by participants in the sentence-first (M = 42%, SD = 20) and the word-first (M = 40%, SD = 22) conditions. Mixed-effects logistic regression modeling was used to probe for effects of learning condition on the production of classifier-noun pairs (coded as correct/incorrect). Learning

condition was entered as a fixed-effect predictor using deviation coding (-.5, .5). Participants and items were treated as random effects; as in Experiment 1, the best fitting model included random intercepts only. Unlike in Experiment 1, the effect of learning condition was not significant (B = -.13, Z = -.40, p = .69), thus mirroring the findings from the forced-choice task. The independent-sample t-test conducted to compare participants' accuracy in producing the carrier phrase across learning conditions also revealed no significant difference between the sentence-first (M = 2.67, SD = .79) and the word-first (M = 2.69, SD = .54) groups, $t(46)$ = -.09, p = .93, d = .03.

In order to examine whether prior knowledge of Chinese affected performance on the production task, we compared results from Experiments 1 and 2 by combining the two datasets and adding experiment as an additional fixed effect predictor. Using the same random effects structure as in the models for the individual experiments (intercepts only), a first model was constructed with experiment and learning condition as fixed effects. In a second model, the interaction term was added. Adding the interaction term did not improve model fit, hence we report the outcome of the first model. This revealed a highly significant main effect of experiment (B = 1.29, Z = 4.3, p < .001), and a marginally significant effect of learning condition (B = -.57, Z = -1.9, p = .06). Participants in Experiment 2 (mean accuracy = 41%, SD = 21) were overall more likely to correctly produce a newly learned classifier-noun pair than those in Experiment 1 (M = *22*%, SD = 22), who had not had any experience with Chinese prior to the experiment. Similarly, participants in Experiment 2 (M = 2.68, SD = .67) were more accurate at producing the carrier phrase than those in Experiment 1 (M = 2.13, SD = 1.00), $t(94)$ = 3.17, p = .002, d = .65. (None of these patterns changed when the 8 participants with prior exposure to Chinese were excluded from the Experiment 2 data.)

In sum, the results from the production task align with those from the forced-choice task in that (i) participants in Experiment 2 performed better overall than those in Experiment 1, indicating that their prior knowledge of Chinese was facilitative overall, and (ii) the effect of learning condition observed in Experiment 1 was no longer present in Experiment 2.

Discussion and Conclusion

Our goals in this study were (a) to examine whether the order-of-learning effects observed by A&R with carefully constructed artificial language stimuli could be replicated with natural language materials, and (b) to explore how prior language experience acquired outside the laboratory may affect participants' allocation of attention and subsequent learning in a controlled laboratory experiment. The results from Experiment 1 consistently replicated the effects observed by A&R: Native speakers of English with no prior experience with Chinese showed better learning of classifier-noun associations when they were exposed to larger units (sentences) before smaller ones (words) than vice versa. Importantly, participants' learning experience in this experiment differed only in the *order* in which different input types (sentences vs. words) were encountered, with no differences between groups in the cumulative frequency of different cues. These results align with the findings reported by Ellis, Sagarra and colleagues in demonstrating that prior knowledge guides learners' (in)attention and affects which properties of the input are (not) learned. In addition, our findings provide unique evidence for blocking in the absence of potential confounds of cumulative frequency in L2 learning. The results from this study also provide some validation for the use of artificial language paradigms to investigate language processing and learning more generally. The results we obtained in Experiment 1 using Chinese stimulus materials were remarkably similar to those obtained by A&R, who used

artificial language materials. We hope that this finding may serve to inspire confidence within the field of Applied Linguistics that insights gained from studies using artificial language paradigms – more commonly used in neighboring fields such as Cognitive Science and Psychology – can be of relevance to the investigation of key concerns in applied linguistics and second language acquisition, such as selective attention and learning.

Meanwhile, the results from Experiment 2 contribute to our understanding of the robustness of order-of-learning effects when additional prior knowledge is present, as it typically is in real-life language learning contexts. The only difference between Experiments 1 and 2 was in the prior knowledge of Chinese that participants brought to the lab. For participants in Experiment 1, Chinese was as novel as the artificial language was to participants in A&R. Participants in Experiment 2, on the other hand, had had 5-7 weeks of classroom instruction in Chinese immediately prior to the experiment. We hypothesized that their knowledge of the carrier phrase and the general concept of classifiers would allow them to home in on the noun-object associations more quickly, thus making their processing of the input in the sentence block more similar to that in the noun block. As a consequence, we expected that the difference between the two learning conditions may be diminished in Experiment 2. In line with this prediction, results from Experiment 2 showed no effect of learning condition, contrary to what we observed in Experiment 1, and contrary to A&R's original findings. It is important to bear in mind, however, that the critical result from Experiment 2 consists of a null effect, which should be treated with appropriate caution. We must point out that when we analyzed the results from the two experiments in a single model, Experiment emerged as a significant predictor (reflecting overall better performance by participants in Experiment 2), but the interaction between experiment and learning condition did not explain a significant proportion of additional variance.

In the absence of a significant interaction term, we cannot confidently conclude that prior experience with Chinese *diminished* the original order-of-learning effect. Nevertheless, we observe that while we were able to replicate A&R's effect in Experiment 1, we were not able to do so in Experiment 2. This at least suggests that once prior language learning experience outside a laboratory context plays into the picture, that picture becomes a considerably noisier one.

The absence of an order-of-learning effect in our Experiment 2 also aligns with the absence of such an effect in Siegelman and Arnon's (2015) second experiment, where the article was semantically informative. Taken together, the findings from these studies indicate that while blocking as a result of order of learning is a real and replicable effect in both artificial and natural language learning, this effect is exquisitely sensitive to small changes in the nature of the learning task (Siegelman & Arnon, 2015) and the nature of learners' prior knowledge (our Experiment 2). This is important to bear in mind when considering the potential implications of effects of selective attention and blocking, such as those observed by A&R, for language instruction and curriculum design. Siegelman and Arnon (2015) discuss their findings in the context of ultimate attainment in L2 acquisition, and propose that attention to different unit sizes may offer an explanation for some of the differences that have been observed in the ultimate attainment of first versus second language learners. The authors focus primarily on the theoretical implications of their proposal, but also write that "[f]rom an applied perspective, our proposal has the potential to significantly alter how we teach adults a second language" (p. 73). While we are intrigued and excited about the theoretical contribution of their proposal, we believe considerable caution is needed with regard to the applied implications that can be drawn from these findings at this point. On the one hand, exposing L2 learners to larger units and less segmented input early on may mimic the learning environment of young children more closely,

and may, in principle, be beneficial in terms of forcing learners to direct their attention to associative relations and co-occurrence patterns in the input, which in turn should strengthen associative relations in their linguistic representations. On the other hand, the neutralization of this effect when only minor factors in the learning task or prior experience were changed raises questions about the feasibility of constructing learning environments for adult L2 learners that would meet all the conditions for learning benefits to emerge from initial exposure to larger units. More specifically, the findings from Experiment 2 suggest that even if the teacher of the Chinese class attended by participants in Experiment 2 had decided to expose students to classifiers in week 7 through unsegmented input, these learners would likely not have done any better at learning classifier-noun associations than if they had learned nouns and classifiers in a more typical segmented, word-by-word fashion. What we cannot know based on these findings, however, is what would happen if they were consistently exposed to such input from week 1. Future research involving classroom intervention studies, more extensive input manipulations, and assessment of long-term learning outcomes will be needed to more fully assess the implications of laboratory-based training studies on selective attention and blocking for L2 instruction and curriculum design.

In the meantime, we believe it is important to understand, on the one hand, the benefits of carefully controlled laboratory studies, including studies employing artificial language paradigms, for providing proof of concept of learning mechanisms that are likely to underlie language learning in a much wider set of circumstances. On the other hand, it is equally important to understand the limitations of the implications that can be drawn from such studies for applied concerns such as language teaching and curriculum design. The results from Experiment 2 suggest that it would be premature to conclude from the order-of-learning effects

observed by A&R and in Experiment 1 that Chinese language textbooks should be rewritten to introduce all nouns together with their classifiers, or more generally that the teaching of second languages to adults should be significantly altered. At the same time, the fact that early exposure to larger units can, under certain circumstances, produce better learning outcomes should be taken seriously by applied linguists, especially as it goes against widespread practices of teaching words before sentences, and perhaps more generally against the intuition that starting with small units and progressing to bigger ones is the natural and best course of learning. Much remains to be done to understand the specific circumstances under which early exposure to larger units is beneficial for L2 learning, and in particular how such circumstances can be created in real-world language learning and instruction contexts.

Endnotes

[1] A&R collected reaction times (RTs) on participants' responses in the forced choice task. Due to technical limitations, it was not possible to collect RTs in our study.

[2] Preliminary data inspection during the course of testing had raised the suspicion that participants with knowledge of a gender-marking language may generally perform better in this experiment. For example, one (particularly savvy) participant observed that the task reminded him of learning articles in Italian. For this reason, care was taken to distribute participants with knowledge of a gender-marking language equally over the two learning conditions. Otherwise group assignment was random.

[3] Overall accuracy among the 8 participants with prior exposure to Chinese (classifier trials: M = 57%, noun trials: 96%) was very similar to that observed among the remaining 40 participants

(classifier trials: M = 61%, noun trials: 94%), indicating that their inclusion in the final sample did not distort the results.

[4] When the 8 participants with prior exposure to Chinese were excluded, the same pattern of results emerged, except that the significance of the interaction term became marginal (p = .058).

[5] When the participants with prior exposure to Chinese were excluded, the comparison for classifier trials in the sentence-first group became marginal, $t(20) = 1.86$, p = .08.

[6] When the 8 participants with prior exposure to Chinese were excluded, the same pattern of results emerged, except that the significance of the interaction term became marginal (p = .06). The significant interaction between experiment and trial type in the full dataset suggests that the difference in performance by Exp1 vs. Exp2 participants was modulated by trial type. Inspection of mean accuracies across experiments and trial types did not reveal the nature of this modulation: Collapsing across learning condition, Exp1 participants showed mean accuracies of 53% (*SD*=16) and 88% (14) on classifier and noun trials, respectively; Exp2 participants showed mean accuracies of 60% (19) and 95% (9) on classifier and noun trials, respectively. In both experiments, the difference between classifier and noun trials in terms of percent accuracy was 35. We therefore hesitate to draw conclusions from the significance of this interaction term in the mixed-effects model.

References

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition, 122*, 292-305. doi:10.1016/j.cognition.2011.10.009

Bates, D., Maechler, M, & Bolker, B. (2011). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-39.

Barr, D. J. (2013, March 27). Coding categorical variables when analyzing factorial experiments with regression. Retrieved from http://talklab.psy.gla.ac.uk/tvw/catpred/

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. doi:10.1016/j.jml.2012.11.001

Boersma, P., & Weenink, D. (2014). Praat: doing phonetics by computer (Version 5.3.52) [Computer program]. Retrieved from http://www.praat.org/

DeKeyser, R. (2013). Age effects in second language learning: stepping stones toward better understanding. *Language Learning, 63*, 52-67. doi:10.1111/j.1467-9922.2012.00737.x

Ellis, N. C. (2006a). Language acquisition as rational contingency learning. *Applied Linguistics, 27*, 1-24. doi:10.1093/applin/ami038

Ellis, N. C. (2006b). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics, 27*, 164-194. doi:10.1093/applin/aml015

Ellis, N. C., & Sagarra, N. (2010). The bounds of adult language acquisition: Blocking and learned attention. *Studies in Second Language Acquisition, 32*, 553-580. http://dx.doi.org/10.1017/S0272263110000264

Ellis, N. C., & Sagarra, N. (2011). Learned attention in adult language acquisition: A replication and generalization study and meta-analysis. *Studies in Second Language Acquisition, 33*, 589-624. http://dx.doi.org/10.1017/S0272263111000325

Ellis, N. C., Hapeez, K., Martin, K. I., Chen, L., Boland, J., & Sagarra, N. (2014). An eye-

tracking study of learned attention in second language acquisition. *Applied Psycholinguistics, 35*, 547-579. doi:10.1017/S0142716412000501

Grüter, T., Lew-Williams, & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research, 28*, 191-215. doi:10.1177/0267658312437990

Guillelmon, D., & Grosjean, F. (2001). The gender marking effect in spoken word recognition: The case of bilinguals. *Memory and Cognition, 29*, 503-511. doi:10.3758/BF03196401

Haselgrove, M., Esber G. R., Pearce, J. M., & Jones, P. M. (2010). Two kinds of attention in Pavlovian conditioning: Evidence for a hybrid model of learning. *Journal of Experimental Psychology: Animal Behavior Processes*, 36, 456-470. http://dx.doi.org/10.1037/a0018528

Jia, G., & Aaronson, D. (2003). A longitudinal study of Chinese children and adolescents learning English in the United States. *Applied Psycholinguistics, 24*, 131–161. http://dx.doi.org/10.1017/S0142716403000079

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell, & R. M. Church (Eds.), *Punishment and aversive behaviour* (pp. 276-296). New York: Appleton-Century-Crofts.

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review, 7*, 636-645. doi:10.3758/BF03213001

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

Le Pelley, M. E., Haselgrove, M., & Esber, G. R. (2012). Modeling attention in associative learning: two processes or one? *Learning & Behavior, 40*, 292-304. doi:10.3758/s13420-012-0084-4

Li, C. N., & Thompson, S. A. (1981). *Mandarin Chinese: a functional reference grammar*. Berkeley: University of California Press.

Liang, S.-Y. (2009). *The acquisition of Chinese nominal classifiers by L2 adult learners* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3369365)

Liu, Y., Yao, T., Bi, N., Ge, L., & Shi, Y. (2009) *Integrated Chinese*. Boston, MA: Cheng & Tsui.

Llanes, À., & Muñoz, C. (2013). Age effects in a study abroad context: children and adults studying abroad and at home. *Language Learning,* 63, 63-90. doi:10.1111/j.1467-9922.2012.00731.x

Long, M. (2013). Maturational constraints on child and adult SLA. In G. Granena, & M. Long (Eds.), *Sensitive periods, language aptitude, and ultimate attainment* (pp. 3-41). Amsterdam: John Benjamins.

Luk, Y. P., & Shirai, Y. (2009). Is the acquisition order of grammatical morphemes impervious to L1 knowledge? Evidence from the acquisition of plural *–s,* articles and possessive '*s*. *Language Learning, 59*, 721-754. doi: 10.1111/j.1467-9922.2009.00524.x

MacWhinney, B. (1987). The competition model. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition* (pp. 249-308). Hillsdale, NJ: Lawrence Erlbaum.

MacWhinney, B. (2001). The Competition Model: The input, the context, and the brain. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 69-90). New York: Cambridge University Press.

MacWhinney, B. (2008). A Unified Model. In P. Robinson, & N. Ellis (Eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition* (pp. 341-371). New York, NJ: Routledge.

MacWhinney, B., Bates, E., & Kliegl, R. (1984). Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning & Verbal Behavior, 23*, 127–150. doi:10.1016/S0022-5371(84)90093-8

Penfield, W., & Roberts, L. (1959). *Speech and brain mechanisms*. Princeton, NJ: Princeton University Press.

Ramscar, M., Dye, M., Popick, H. M., & O'Donnell-McCarthy, F. (2011). The Enigma of number: Why children find the meanings of even small number words hard to learn and how we can help them do better. *PLoS ONE, 6*, e22501. doi: 10.1371/journal.pone.0022501

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology, 66*, 1-5. http://dx.doi.org/10.1037/h0025984

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black, & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Shanks, D. R. (1995). *The psychology of associative learning.* New York: Cambridge University Press.

Siegelman, N., & Arnon, I. (2015). The advantage of starting big: Learning from unsegmented input facilitates mastery of grammatical gender in an artificial language. *Journal of*

Memory and Language, 85, 60-75. http://dx.doi.org/10.1016/j.jml.2015.07.003

Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.

Table 1

Summary of procedure in both learning conditions

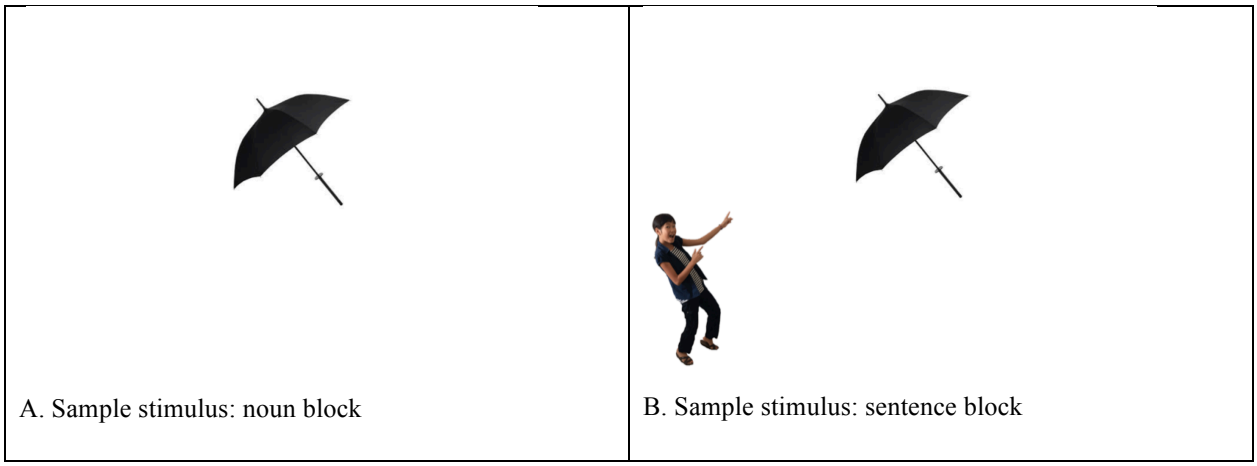
Condition	Word-first condition	Sentence-first condition
Learning phase	1. Block 1: noun trials	1. Block 1: sentence trials
	2. Block 2: sentence trials	2. Block 2: noun trials
	3. Distractor block	3. Distractor block
Test phase	4. Forced-choice	4. Forced-choice
	5. Production	5. Production

FIGURE CAPTIONS

Figure 1. Example of visual stimuli in the learning phase.

Figure 2. Forced Choice task: Mean accuracy by trial type and learning condition in Experiments 1 and 2. Error bars indicate 95% confidence intervals of the means adjusted for repeated measures.

Figure 3. Production task: Mean accuracy for classifier-noun pairs by learning condition in Experiments 1 and 2. Error bars indicate 95% confidence intervals.



A. Sample stimulus: noun block

B. Sample stimulus: sentence block

Figure 1. Example of visual stimuli in the learning phase.

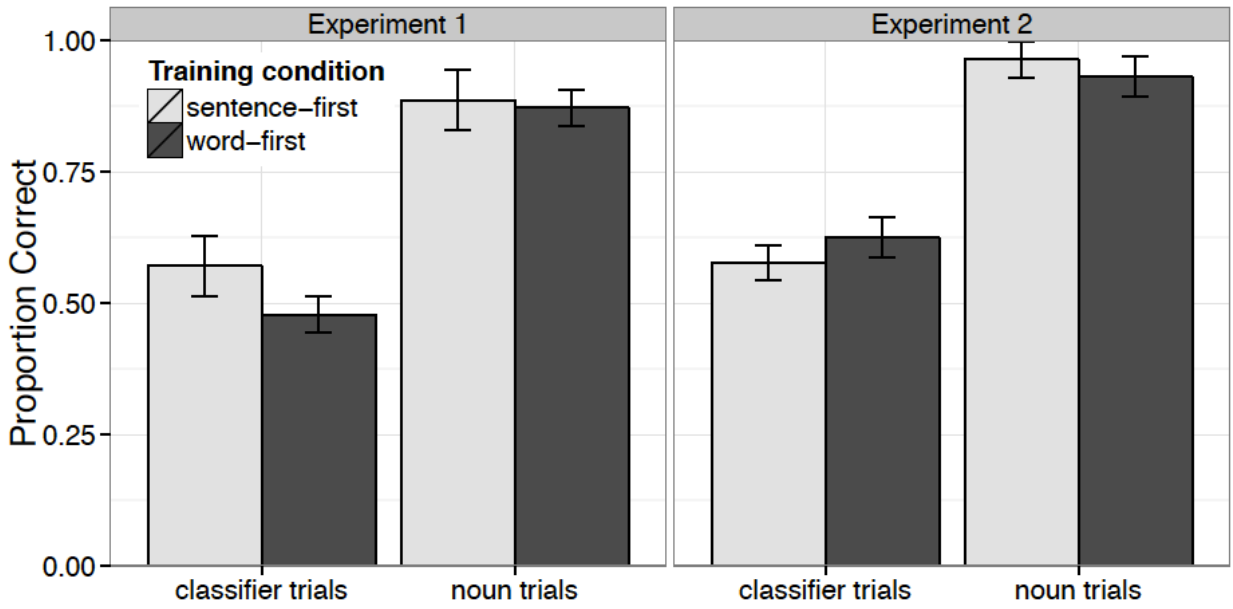


Figure 2. Forced Choice task: Mean accuracy by trial type and learning condition in Experiments 1 and 2. Error bars indicate 95% confidence intervals of the means adjusted for repeated measures.

Figure 3. Production task: Mean accuracy for classifier-noun pairs by learning condition in Experiments 1 and 2. Error bars indicate 95% confidence intervals.