LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

REGISTERED REPORT

Does Early Unit Size Impact the Formation of Linguistic Predictions? Grammatical Gender as a Case Study

Rana Abu-Zhaya[a] and Inbal Arnon[b]

[a]University College London [b]The Hebrew University of Jerusalem

Abstract: Making adults learn from larger linguistic units can facilitate learning article–noun agreement. Here we ask whether initial exposure to larger units improves learning by increasing the predictive associations between the article and noun. Using an artificial language learning paradigm, we taught 106 Hebrew-speaking participants novel article–noun associations with either segmented input first or unsegmented input first, and tested their learning of the article–noun association and their ability to use articles to predict nouns. Our results showed that participants exposed to unsegmented input first were more likely to treat the article–noun unit as one word and were more accurate at learning the correct article–noun associations. However, participants in the unsegmented-first condition did not show increased gaze to the target compared to those in the segmented-first condition. We discuss how these findings inform our understanding of the challenges that adults face when learning a second language.

Keywords linguistic predictions; grammatical gender; order of exposure; eye tracking

A one-page Accessible Summary of this article in nontechnical language is freely available in the Supporting Information online and at https://oasis-database.org

Correspondence concerning this article should be addressed to Rana Abu-Zhaya, Division of Psychology and Language Sciences, University College London, Chandler House, 2 Wakefield Street, London WC1N 1PF, United Kingdom. Email: ranaabuzhaya@gmail.com

The handling editor for this manuscript was Naoko Taguchi.

DOI: 10.1111/lang.12638

Introduction

Language learning and processing are impacted by predictions based on prior linguistic experience (Pickering & Garrod, 2013): Speakers keep track of how predictable linguistic elements are and use this information to facilitate further processing (Huettig, 2015). Linguistic predictions are based on the accumulation of distributional information at multiple levels of representation (Kuperberg & Jaeger, 2016). The effect of such predictions on language processing is often tested by exposing speakers to more and less predictable elements, and showing that distributional information (accumulated via exposure to natural language or to prior experimental trials) impacts the strength of predictions, and that differences in predictability impact learning and processing patterns (e.g., Kamide et al., 2003). For example, reading times for infrequent syntactic structures become faster when these structures are seen more often in the experimental setting (Fine et al., 2013).

Less work, however, explores the impact of the order in which distributional information is presented on the formation of linguistic predictions and their facilitatory effect on processing. Theoretically, several learning models predict that *order of exposure* to distributional cues will impact the predictions that learners form. Some go further in suggesting that the order in which cues are experienced is critical (the highlighting effect; Kruschke, 2005), and that, by shifting learners' attention, cues that are experienced and learned earlier have precedence over those learned later (see Yoshida & Burling, 2012, for the relevance of the highlighting effect for word learning). Discriminative learning models explicitly articulate such effects and propose that order of exposure to predictive relations between elements strongly impacts learning, such that earlier learned cues can block the learning of later associations (Kamin, 1969; Ramscar et al., 2010; Rescorla & Wagner, 1972). The attentional blocking of later cues by earlier learned ones has been proposed as an explanation for adults' difficulty in learning a second language (L2): What they have learned in their first language (L1) blocks their attention to new cues relevant for learning the L2 (Ellis, 2007; see Ellis & Sagarra, 2010, 2011, for more recent experimental evidence).

The current study extends this line of work and examines whether manipulating order of exposure to segmented versus unsegmented input impacts the linguistic associations created between words[1] by changing learners' reliance on words versus *multiword units* (MWUs; we use the term MWU to refer to a sequence larger than one lexical word (Arnon & Christiansen, 2017)).

Background Literature

The Starting Big Hypothesis: The Impact of Unit Size and Order of Exposure on Learning

The idea that order of exposure to larger versus smaller units (via unsegmented vs segmented speech, respectively) will impact the predictive associations formed between words has been advocated in the *starting big hypothesis* (Arnon, 2010; Arnon & Christiansen, 2017; Arnon & Ramscar, 2012). This hypothesis proposes that (some of) the difference in language learning between children and adults is related to the linguistic units that they attend to and rely on. The claim is that both words and MWUs serve as building blocks for language (Arnon, 2010; Arnon & Christiansen, 2017; Arnon & Cohen Priva, 2013; Arnon & Snider, 2010). However, whereas children draw on MWUs in the process of learning their L1 (Arnon & Clark, 2011; Bannard & Matthews, 2008), adults learning a L2 do so less often (Arnon & Ramscar, 2012), possibly due to their existing knowledge of words (as a concept) and of the specific words in their L1 (Arnon & Christiansen, 2017). Importantly, learning from MWUs is predicted to facilitate mastery of certain grammatical relations between words by increasing the predictive relations between them (Arnon & Ramscar, 2012; Siegelman & Arnon, 2015). This approach emphasizes the importance of early building blocks: Starting with MWUs and then segmenting them into separate words should lead to better learning compared to starting with individual words and learning to combine them. The prediction is that exposing learners to less segmented input will increase their reliance on MWUs (compared to words), which will consequently influence learning outcomes: Learners will show better mastery of certain relations between words when they are initially learned as part of a larger MWU. An example of such a relation, which serves as the case study we use herein, is article–noun agreement in languages with grammatical gender. This is discussed in the next section.

Grammatical Gender as a Case Study

Grammatical gender is a system that assigns nouns to classes that condition agreement with other elements in the sentence. Such assignment to classes involves considerable arbitrariness, and nouns designating the same object can be assigned different genders across different languages. We focus here on the agreement patterns between articles and nouns (as in the Spanish contrast between feminine *la pelota* "the ball" and masculine *el zapato* "the shoe"). Whereas native speakers master grammatical gender with ease (Grüter et al., 2012; Hopp, 2012, 2016; Hopp & Lemmerth, 2016), adult L2 learners struggle in learning such systems. Their ability to master grammatical gender is more variable (Hopp, 2012, 2016), and is impacted by the presence or absence of gender marking in their L1 (Bordag & Pechmann, 2007). This L1 impact seems to be task-dependent: Whereas highly proficient learners of a L2, irrespective of their L1, correctly assign gender to nouns in a judgment task, their ability to use grammatical gender agreement (or lack thereof) in a sentence to judge its grammaticality is more dependent on the presence of grammatical gender in their L1 (Sabourin et al., 2006).

Further, unlike native speakers who use articles to efficiently process the upcoming noun (Brouwer et al., 2017; Dussias et al., 2013; Grüter et al., 2012; Hopp, 2012, 2016; Hopp & Lemmerth, 2016; Lew-Williams & Fernald, 2007, 2010), adult L2 learners' ability to do so is inconsistent and dependent on a range of variables. Earlier studies indicated that adult learners do not use articles to facilitate processing of upcoming nouns, even when they are proficient in the L2 (Lew-Williams & Fernald, 2010); and that when such ability emerges, it is still not as efficient as that of native speakers (Grüter et al., 2012). However, later studies examining adult learners from a broad proficiency spectrum showed that use of gender marking to facilitate further processing is dependent on L2 proficiency: Highly proficient and early L2 learners use articles predictively, whereas low-proficiency late learners do not show this ability (Dussias et al., 2013). Use of grammatical gender to facilitate processing is also related to the accuracy of assigning grammatical gender, to the consistency and timing of producing correct gender inflections (Hopp, 2012, 2016), and to gender marking in the L1 and whether a noun has the same or different grammatical gender across languages (Hopp & Lemmerth, 2016). In sum, these findings demonstrate that adults show difficulty in learning gender systems and that their ability to use gender-marked articles to facilitate processing is variable and dependent on a wide range of influences.

The Impact of Initial Unit Size on Learning Grammatical Gender

The starting big hypothesis points to an additional variable that could impact L2 learners' general difficulty in learning grammatical gender and their less persistent and more variable use of gender marking in processing: the size of early linguistic units and how they influence learning the associations between words. Under this account, native speakers' greater ability to use the article to facilitate their processing of the upcoming noun is related to their greater reliance on article–noun MWUs during learning. L2 learners, in contrast, tend to learn from individual words, leading to a weaker association between the article and the noun, and hence less consistent and accurate predictions.

Evidence for this proposal comes from artificial language learning studies, where researchers manipulated the order of exposure to segmented versus unsegmented input as a way of manipulating learners' reliance on MWUs. In the first study by Arnon and Ramscar (2012), learners were exposed to an artificial grammatical gender system in two exposure conditions. Participants learned 14 novel labels for concrete objects; the nouns were divided into two "grammatical" classes such that each noun followed one of two articles (*bol*, *sem*). Sentences in this language followed the order carrier phrase–article–noun (e.g., *os-ferpel-en bol viltord*). Participants were either exposed to full sentences first and then to noun labels on their own, or the other way around. The prediction was that exposure to full sentences first would increase treatment of the article–noun sequence as one unit and enhance learning the relation between them. Importantly, in both conditions, learners were speakers of the same language (namely, English) and were exposed to the same input, with only the order of input differing. During testing, participants completed a forced-choice task followed by a production task. In the forced-choice trials, participants saw a picture, heard two sentences, and were asked to choose which sentence best matched the picture. In half of these trials, the incorrect sentence had the right noun but the wrong article; in the other half, the incorrect sentence had the right article but the wrong noun. In the production task, participants saw a picture and had to produce a sentence to describe it based on the language they had learned. If learning was solely based on distributional information, article–noun pairings should be learned equally well in both conditions given that participants eventually received the same amount of input. However, results showed that participants who were first exposed to full sentences and then heard noun labels were significantly better at learning the article–noun pairing than participants who were first exposed to noun labels. Specifically, participants who first heard full sentences were more accurate at choosing the correct article and correct noun, and were more likely to produce the correct article–noun sequences.

In a follow-up study, using a design better suited to manipulating and evaluating the linguistic units that learners extract, Siegelman and Arnon (2015) provided more direct evidence that exposure to unsegmented input leads to an increased reliance on MWUs. A similar artificial language was used. Participants were exposed to either segmented sentences first (with 250-ms pauses between words), which should increase reliance on wordlike units, or unsegmented sentences (without pauses) first, which should increase reliance on MWUs. Participants heard 12 novel labels for concrete objects that consistently followed either one of two articles (*fo*, *se*); sentences in this language also followed the order carrier phrase–article–noun (e.g., *os-ferpel-ti fo etkot*). Here too, participants in both conditions heard exactly the same input, but in flipped order across the two conditions. During the exposure phase, participants were asked to type some of the sentences they heard, allowing researchers to directly test whether article–noun sequences were treated as one word or two separate words. The prediction was that participants who heard unsegmented sentences first would be more likely to treat the article–noun sequence as one word and would show better learning of the relation between article and noun. Participants were then tested in a forced-choice task and a production task that were identical to those in the previous study. Here too, if learning was based solely on distributional information, then participants in both conditions should learn the article–noun pairs equally well. However, this study showed that exposure to unsegmented input first, as opposed to segmented input, shifted adults' attention to larger input units, causing them to treat article–noun sequences as single units and leading to better learning of the article–noun association.

The proposed mechanism for why MWUs facilitate learning is that treating the article–noun sequence initially as one unit will increase the predictive relations between the article and the noun, and enable learners to use the article to facilitate the processing of the upcoming noun. This prediction was tested in a study by Shantz (2018), in which English and German speakers were taught an artificial language modeled on that used by Arnon and Ramscar (2012). Using a visual world eye-tracking paradigm, Shantz tested whether the learning context and the order of exposure to individual words versus sentences influenced adults' ability to use grammatical gender predictively. Contrary to what one might predict, the results did not show that exposure to sentences first facilitated earlier gaze toward the noun upon hearing the article. However, it is difficult to evaluate and interpret this lack of effect for several reasons. First and most importantly, participants struggled to learn the language and did not show the expected offline effect: Participants were not better at learning the article–noun pairings in the sentence-first condition and showed very low accuracy overall (36% correct article–noun matching compared to 60% in the study by Arnon & Ramscar, 2012). This was probably caused by the greater complexity of the artificial language: Participants were taught 24 nouns compared to 12 in the study by Siegelman and Arnon (2015) and 14 in that by Arnon and Ramscar. The lack of an effect of order of exposure on gaze patterns could reflect the overall difficulty in learning. In addition, Shantz manipulated early units by comparing learning between words-first and sentences-first conditions (as in the study by Arnon & Ramscar, 2012). Shantz did not implement the improved manipulation used by Siegelman and Arnon where early unit size is manipulated by changing the order of exposure to segmented versus unsegmented input. Hence, Shantz's study does not provide an adequate test of the impact of early unit size on the predictive relations between articles and nouns.

In sum, the evidence we have to date does not tell us whether order of exposure to unsegmented versus segmented input indeed changes the strength of association between the article and the noun. In the current study, we tested this hypothesis. We used the same artificial language learning task as in the study by Siegelman and Arnon (2015) with a few minor modifications to the task and the addition of an eye-tracking component. These changes are detailed in the Method section; given that they are introduced intentionally and based on the recommendations detailed by Marsden et al. (2018), this study is a conceptual replication of the studies reviewed above (Arnon & Ramscar, 2012; Shantz, 2018; Siegelman & Arnon, 2015), as well as a significant extension. The replication refined the theoretical constructs upon which these studies were built, and the extension allowed us to examine whether using a stronger (and better) manipulation of early unit size leads to an increased association between the article and noun and facilitates processing.

The Present Study

In our study, Hebrew-speaking participants were exposed to an artificial language in two exposure conditions: segmented-first and unsegmented-first. As in the study by Siegelman and Arnon (2015), they were asked to type some of the sentences they heard during exposure so that we could assess the units they extracted. The facilitatory relations between the article and the noun were assessed after the learning phase by examining participants' gaze patterns to two objects while hearing a sentence in the artificial language. Specifically, we compared gaze patterns between same-gender trials, where the two nouns have the same grammatical gender in the artificial language, and different-gender trials, where the two nouns differ in grammatical gender (this design is based on the same-/different-gender trials that are a common experimental test of predictive processing in studies using real language, e.g., Lew-Williams & Fernald, 2010). The article can facilitate noun processing only in different-gender trials and not in same-gender trials. In previous studies utilizing this paradigm, native speakers have shown higher proportions of looks and faster reactions to the target noun in different-gender trials compared to same-gender trials, suggesting that the article facilitates the detection of the correct noun. We used the same paradigm to examine whether learners rely on the article to facilitate processing of the upcoming noun and whether their ability to do so depends on the order of exposure to unsegmented versus segmented input. These trials were followed by forced-choice trials in which the incorrect sentence included the right noun but the wrong article, mimicking the "article trials" used by Siegelman and Arnon (2015).

As mentioned above, our participants were all native speakers of Hebrew, a language that inflects nouns and adjectives for gender but has no article–noun gender agreement marking (Deutsch & Bentin, 2001), making this particular aspect of the artificial language unfamiliar to all participants. Further, based on the results obtained from participants in the study by Siegelman and Arnon (2015), who were all native speakers of Hebrew, we expected that our participants would learn the associations between the articles and nouns. Although it is possible that the lack of article–noun gender agreement marking in our participants' L1 would make the task harder for them (compared to participants whose L1 has such agreement), we did not expect this to interact with input condition. Any difficulty (or facilitation) caused by the L1 should be similar in the two learning conditions. Importantly, we ensured that the gender of nouns in Hebrew could not provide a cue to the gender class in the artificial language (half of the nouns in each artificial gender class correspond to masculine nouns in Hebrew and half to feminine ones). We return to the issue of L1 influence on gender learning and predictive gaze in the Discussion section.

Research Questions

In the experiment reported herein, we aimed to extend and replicate the study by Siegelman and Arnon (2015). Specifically, we asked the following questions:

1. To what extent does varying the order of exposure to segmented and unsegmented sentences (as a way to manipulate early unit size) impact learning of the association between the article and the noun?
   (a) Are participants more likely to treat the article–noun sequence as one word in the unsegmented-first condition compared to the segmented-first condition?
   (b) Are participants more accurate in learning the association between the article and noun in the unsegmented-first condition compared to the segmented-first condition?
   (c) Does the unsegmented-first condition lead to increased facilitation in processing the noun compared to the segmented-first condition?

These questions translate into the following predictions. First, we predicted that participants exposed first to unsegmented sentences would be more likely to type out the article and noun as one word than participants in the segmented-first condition, suggesting that the former treated these as a single unit. Second, we predicted that exposure to unsegmented sentences first would facilitate a stronger association between the article and the noun, such that participants would be more accurate in matching the correct article to the noun. Finally, we predicted that the ability to use the article to facilitate noun processing would be impacted by order of exposure to segmented and unsegmented input, such that when learners started out from unsegmented input, they would be better able to use the article to facilitate the processing of the upcoming noun. Finding that early exposure to larger units, via unsegmented speech, impacts the strength of the associations formed between words would support the idea that MWUs facilitate learning by increasing the association between the words that comprise them (in this instance, the article and the noun), and would point to early unit size as an additional variable impacting L2 learners' ability to form associative relations in a L2. More broadly, such a finding would highlight how the order in which information is learned impacts the formation of linguistic associations. In contrast, finding that the associative relations between the article and noun are similar regardless of whether learners are exposed to unsegmented or segmented input first would indicate that the better learning from MWUs found in previous studies is not mediated by the formation of stronger associations between the elements of the MWU.

Method

The experiment described below is an extension and conceptual replication of Experiment 1 in the study by Siegelman and Arnon (2015) and was closely modeled on it, with several important modifications. We first detail the aspects that are identical to Siegelman and Arnon's study and then the modifications. All study materials—consent form and questionnaire (in Hebrew and English), images and sound files for all the phases of the experiment (training, distractor block, and test), and the exact instruction pages for participants (in Hebrew and English)—are publicly available via the Open Science Framework (OSF; https://osf.io/98ak3/).

We used the exact same artificial language as used by Siegelman and Arnon (2015), with 12 novel labels for 12 objects presented with the same carrier phrase in the same exposure conditions: segmented-first and unsegmented-first. We also used the same sound files as Siegelman and Arnon for all sentences throughout the experiment. We used typing trials to assess the size of linguistic units that learners form, as in the original study.

We made several modifications to the study. The first, and most important, is the addition of an eye-tracking component after the learning phase to test whether the article facilitates the processing of the upcoming noun. This required the creation of a novel testing phase where participants heard a sentence and saw two objects while we tracked their gaze. The comparison of interest was between same-gender trials (where the nouns for both objects were from the same gender class) and different-gender trials (where the nouns for the two objects were from different gender classes). In addition, we made some minor modifications that were not expected to change the results. First, we used new images for each of the objects to ensure that they were of equal size and quality (which is important for the eye-tracking component). Second, we replaced the article *se* with *si* to control for similarity to the Hebrew demonstrative *ze* (as was done in Experiment 2 in the study by Siegelman & Arnon, 2015). Third, we created two lists with different article–noun pairings in each condition to ensure that the effects were not specific to particular article–noun pairings. Fourth, because we added the eye-gaze testing phase, we shortened the distractor block to include only four novel nouns instead of six. We did not think this last change would impact our results given that the distractor block only serves as a buffer between the exposure and testing phases (to ensure both groups hear the same stimuli just before testing). Given the addition of the eye-tracking test and the time needed to calibrate and validate eye-gaze data, shortening the distractor block ensured that the experiment was not too long. Fifth, we removed the production task that followed the forced-choice test trials in the original experiments, as it would not add new data pertinent to our research questions. And finally, in the forced-choice task, we tested only participants' knowledge of the article–noun pairings (because noun knowledge was tested during the eye-tracking phase when participants were asked to select the noun described by the sentence).

Participants

Participant recruitment was conducted through an online platform for experiments offered to undergraduate students at the Hebrew University of Jerusalem. One hundred and six undergraduate students participated in the study in exchange for course credit or payment. They were all native speakers of Hebrew (exposed to Hebrew from birth), with no history of speech, reading, hearing, or learning disability and no history of attention deficit disorder or attention deficit hyperactivity disorder. Students who were nonnative speakers of Hebrew or who reported a history of any of the mentioned disorders through the online subject recruitment system were not able to participate in the study. Participants with glasses or contact lenses were able to participate in the experiment if they passed the calibration and validation phases.

Participants were randomly assigned to one of two experimental conditions and were randomly tested on one of two lists within each condition. They were briefly informed of the goals of the study and signed a consent form prior to participation. In the consent form and participant questionnaire, participants were asked to list their native languages and other languages that they had learned and to confirm that they had not been diagnosed with any of the disorders mentioned above. Participants who wore glasses were asked to clean their glasses well before the experiment. Out of the 106 participants we tested, we excluded five: three due to experimental errors and two for failure to pass the calibration and validation in the eye tracking phase. Out of the 101 participants who completed the study, one participant's behavioral data file was corrupted and could not be retrieved. Hence, that participant was excluded from all analyses, yielding a sample of 100 participants (72 female; 28 male) with a mean age of 23.69 years (range: 19–30 years; $SD = 2.04$; 95% CI [23.28, 24.09]). We tested 50 participants in each condition: 25 in each of the lists within each condition.

Power Analyses

Given that no previous study had used the specific design used here (a combination of artificial language learning and eye tracking) with the same dependent measure (proportion of fixations to target) and statistical model (linear mixed-effects model with two fixed effects), we based our power analyses on the behavioral data from the study by Siegelman and Arnon (2015). We conducted all analyses with an alpha of .05 and power of .80 using the power analysis calculator available through the software G*Power (Faul et al., 2009). We based our sample choice on the largest of these estimations. To address Research Question 1a (Are participants more likely to treat the article–noun sequence as one word in the unsegmented-first condition compared to the segmented-first condition?), we used the typing trial results from Siegelman and Arnon's study. Although they did not use a mixed-effects model to analyze these data as we did, these results are the closest we have to our current design. The difference between the two experimental conditions (unsegmented-first vs. segmented-first) in the proportion of one-word responses, as reported using a *t* test, yielded a large effect size, $d = 1.742$. Obtaining such an effect for a simple *t* test examining the difference between the two conditions would require a total sample of 14 participants (seven per condition).

To address Research Question 1b (Are participants more accurate in learning the association between the article and noun in the unsegmented-first condition compared to the segmented-first condition?), we used the “article trials” from Experiment 1 in the study by Siegelman and Arnon (2015) and the reported *t* test to inform our sample size estimation. The difference in accuracy score on the article trials between the two experimental conditions yielded an effect size of $d = 0.56$. Obtaining such an effect for a similar *t* test examining the difference in accuracy for forced-choice trials between the two conditions would require a sample of 41 participants. Unfortunately, we do not have any previous data allowing us to estimate the sample for Research Question 1c (Does the unsegmented-first condition lead to increased facilitation in processing the noun compared to the segmented-first condition?); hence, our sample size estimation for this question was informed by the sample size needed for the previous two research questions. Thus, we aimed to recruit a sample of 45 participants per condition, giving a total of 90 participants; this should be adequate for obtaining the desired effect size while allowing for expected attrition due to data loss.

Materials

We used a slightly modified version of the artificial language used in Experiment 1 in the study by Siegelman and Arnon (2015). The language has one carrier phrase (*os-ferpel-ti*), two articles (*fo*, *si*[2]), and 12 novel nouns that label concrete inanimate objects. The language has a fixed word order of carrier phrase–article–noun (e.g., *os-ferpel-ti fo gorok*). Nouns are divided into two classes, and each noun appears with only one article. The only cue to class membership is distributional; there are no phonological or semantic cues for which noun is paired with which article. Importantly, half of the labels in each artificial class correspond to feminine nouns in Hebrew and half to masculine ones (see Table 1); hence, Hebrew gender cannot be used as a cue to class membership. All novel labels are bisyllabic, and the objects they name are designated by high-frequency, early-acquired Hebrew nouns.

All elements (nouns, articles, carrier phrase) were recorded for Siegelman and Arnon’s (2015) experiments by a female Hebrew speaker, and then synthesized to a frequency of 170 Hz to remove any cues to word boundaries. Throughout our experiment, we used the same recorded token of each noun, article, and carrier phrase as used by Siegelman and Arnon. The carrier phrase is 1,200 ms long, articles are 281 ms long (as in previous studies of article–noun agreement; e.g., Lew-Williams & Fernald, 2010), and nouns are between 650 ms and 1,000 ms long ($M = 785$ ms). Sentences were concatenated using

Table 1 The 12 novel labels and articles used in the experiment

Article in List 1	Article in List 2	Novel label	English noun	Gender of Hebrew noun
fo	si	gorok	pan	F
		panjol	television	F
		toonbot	clock	M
		fertsot	bed	F
		perdip	table	M
		etkot	key	M
si	fo	hekloo	bath	F
		hertin	iron	M
		geesoo	hat	M
		slindot	piano	M
		jatree	cup	F
		sodap	spoon	F

Note. In each list, each set of six nouns was matched with a different article. F = feminine; M = masculine.

Praat (Boersma & Weenink, 2005), such that segmented sentences included a 250-ms pause between their elements, but unsegmented sentences did not. Images were chosen from the web to fit a 400-×-400-pixel square, as commonly used in previous studies (Borovsky et al., 2016; Ellis et al., 2015). For each of the 12 objects we chose six images in which the objects appeared on a white background. During each presentation of the objects throughout the experiment, one of the six images was selected.

In addition to the experimental items described above, a second set of sentences served as a distractor block between the training and testing phases in both conditions. The distractor block comprises stimuli from an artificial language whose articles, labels, and objects differ from those of the test language. This is the same language used by Siegelman and Arnon (2015), with sentences composed of a carrier phrase (*os-ferpel-en*) followed by one of the articles *ped* or *gab* and one of four novel labels (see Appendix S1 in the Supporting Information online).[3] The stimuli for this set were recorded for Siegelman and Arnon's experiment by a male speaker, and we use these same recordings in this project; the choice of a male was intended to maximize the difference from the test language. Unlike sentences in the test language, the distractor-block sentences have no consistent mapping between articles and nouns: Each of the four nouns occurs with each of the two articles (i.e., sometimes with one article and sometimes with the other).

Design

To address the research questions in this study, we adopted a true experimental design. The experiment had two exposure conditions: segmented-first and unsegmented-first. In the unsegmented-first condition, participants first heard unsegmented sentences (with no pauses between elements) and then segmented sentences (with pauses), whereas in the segmented-first condition, the order of exposure was flipped. The conditions presented learners with the same input but in a different order. Two experimental lists were created to counterbalance the matching between nouns and articles across conditions (e.g., in List 1 *gorok* is matched with *fo*, whereas in List 2 it is matched with *si*; see Table 1).[4]

Procedure

The study took place in a quiet room at a research lab at the Hebrew University of Jerusalem. The experiment was composed of two phases: training and testing. Participants were briefed before the start of the experiment that they would be participating in a project studying the mechanisms underlying language learning. At the beginning of the training phase, participants were told they would see objects and hear sentences describing them in a novel language, and that they should try to remember those as well as possible because they would be occasionally required to type the last sentence they heard. Prior to the testing phase, they were instructed to call the experimenter, who initiated the eye-tracking system and performed the calibration and validation procedures as detailed below. Participants were told that they had to look at the screen during the entire testing phase as we would be tracking their eye gaze. At the beginning of the testing phase, they were told that they would see two objects and hear one sentence that described only one of them, and that at the end of each trial, and only once the sentence was finished, they would have to choose the object that they thought best matched the sentence they had heard. In the second part of the testing phase, participants were told that they would see one object and hear two sentences. They were told that only one of the sentences would be correct and that they would have to choose which of them best described the picture.

Training Phase

Participants saw pictures of objects and heard sentences describing them in the novel language. Each participant heard 120 sentences: 60 segmented and 60 unsegmented in two separate blocks, whose order depended on condition. Each of the 12 objects was presented five times in unsegmented sentences and five times in segmented sentences. The order of presentation of objects and their corresponding sentences was randomized within each of the five repetitions in each of the two blocks. Trials were presented consecutively without any lag in between. The length of each trial was set at 3,000 ms to ensure all sentences were played in their entirety.

The typing trials proceeded as follows. Every few sentences, participants were asked to type the last sentence they heard. There were 15 trials for each sentence type (a total of 30 typing trials across the segmented and unsegmented sentences). As in the study by Siegelman and Arnon (2015), the position of the typing trials in each block was the same for all participants and was based on 15 pregenerated positions that disallowed any sequence of more than two typing trials (4, 9, 10, 13, 14, 21, 23, 24, 27, 30, 36, 44, 45, 47, 58). Importantly, given that the order of sentences was randomized within each block for each participant, the specific sentences to be typed in were different for each participant. Only after participants had typed their responses did the experiment proceed to the presentation of the next sentence in the training phase (there was no time limit on typing a response). These typing responses were recorded and analyzed for the purpose of examining how learners segment the novel speech stream and whether they treat the article–noun sequences as one word or two.

Following the two learning blocks, participants heard 24 sentences from the distractor language. These sentences were unsegmented, and, as mentioned above, they were composed of four different novel labels that were interchangeably paired with either one of two articles, *ped* and *gab*. Each of the eight sentences created from such pairings was repeated three times. As in the test language trials, trials in the distractor block lasted 3,000 ms each and were presented consecutively. As in the original experiment carried out by Siegelman and Arnon (2015), the distractor block allowed us to control for any recency effects and ensured that all participants heard the same items before the testing phase regardless of the condition they were tested on.

Testing Phase

Learning was assessed identically in both conditions; we used eye tracking to assess which condition facilitates a stronger article–noun association and forced-choice trials to evaluate learning of the article–noun pairings.

Eye-gaze trials. The testing phase started with 24 same-/different-gender trials, during which we collected eye-gaze data. On each of these trials, participants saw two objects and heard a sentence that described one of them. On same-gender trials, both objects' labels were from the same noun class (e.g., in List 1, *gorok* and *panjol* are both paired with *fo*), and hence it was not possible to use the article to facilitate processing of the upcoming noun. On different-gender trials, on the other hand, objects' labels were from two different classes (e.g., in List 1, *toonbot* is paired with *fo*, whereas *geesoo* is paired with *si*), and it was possible to use the article to facilitate processing of the upcoming noun. Importantly, in all trials, regardless of their type, we matched objects on the gender of the nouns that designate them in Hebrew to ensure that any differences in fixation patterns were not based on gender knowledge from the L1. For example, in one trial, participants saw a pan and a TV, both of which are designated by feminine nouns in Hebrew. In another trial, participants saw a clock and a hat, which are both designated by masculine nouns in Hebrew. Each object appeared four times: twice as a target (once in a same-gender trial and once in a different-gender trial) and twice as a distractor object. Each object appeared as a target once on the left side and once on the right side. Similarly, each object appeared as a distractor once on the left side and once on the right side.

Importantly, even though all participants were exposed to both segmented and unsegmented sentences during learning, all sentences in the same-/different-gender trials were segmented (i.e., they included a 250-ms gap between the article and the noun). This was done to allow for a more accurate assessment of predictive gaze (in unsegmented sentences the noun follows the article immediately, which is not enough time for predictive gaze to be formed). At the end of each trial, participants were asked to choose the object described in the sentence using the keyboard ("1" for left, "2" for right). This allowed us to examine whether participants learned the labels for the objects. As in previous studies (e.g., Brouwer et al., 2017; Grüter et al., 2012; Lew-Williams & Fernald, 2010), the two images were presented side by side for 2 s prior to the speech signal, and stayed on screen for the duration of the speech signal and until the participants chose the object that they thought matched the sentence (there was no time limit for making such a choice). Trials were separated by a fixation screen, as detailed below in the section on eye-tracking methods.

Forced-choice trials. These trials tested participants on their offline mastery of the article–noun pairings and were an exact replication of the "article trials" used by Siegelman and Arnon (2015). In these 12 trials, participants saw an object and heard two sentences separated by 1,000 ms of silence. One of these sentences included the correct article, and the other one included the incorrect article. As in Siegelman and Arnon's study, half of the trials included segmented sentences and the other half included unsegmented sentences; this ensured that both types of sentences from the exposure phase were represented at testing. Participants were tested on each object only once. The order of trials was randomized for each participant, and the order of presentation of the two sentences within each trial was counterbalanced to ensure that participants heard the correct sentence first in half of the trials, and second in the other half. After hearing the two sentences, participants were presented with an instruction screen prompting them to choose the sentence that best described the object by using the keyboard ("1" or "2"; there was no time limit for making such a choice). The next trial was presented 250 ms after a choice had been made.

Eye-Tracking Methods

The eye-tracking component of the study was initiated following the distractor block. We recorded eye movements monocularly from the dominant eye at a rate of 1,000 Hz using the Desktop Mount (without head support) of the EyeLink 1000 Plus Eye Tracker (SR Research, Mississauga, Ontario, Canada) fitted with a 25-mm camera lens. Eye-tracking data were acquired and stored on a Dell Inspiron laptop that performed all online detection and automatic classification of fixations, saccades, and blinks. The experiment was presented on a Dell Desktop via E-Prime. The two computers were connected via an Ethernet cable, allowing data to be shared.

A small target sticker with a high-contrast pattern was placed on participants' foreheads to aid the eye tracker in identifying the location of the eyes and measuring eye movements. Before the initiation of the testing phase, we performed a camera setup to define the thresholds for the pupil size and corneal reflection, followed by a manual 13-point calibration and validation routine. Prior to the initiation of each same-/different-gender trial, participants were prompted to fixate toward a central stimulus in the shape of a cross, and once they did, the stimulus disappeared, and the trial began. This fixation part of the trial also served as a drift checking and correction that aimed to evaluate whether the model of the eyes that was created during the calibration and validation phases had become invalidated. Of our sample of 100 participants, 56 had a dominant right eye, and the remaining 44 had a dominant left eye.

Predictions, Coding and Scoring

Typing Trials

Based on prior work, we expected participants in the unsegmented-first condition to (a) be more likely than those in the segmented-first condition to treat the article–noun sequence as one word and (b) show a positive correlation between the proportion of article–noun chunks (out of the total typing trials) and the accuracy of learning the article–noun pairing (as measured in the forced-choice trials). To test these predictions, we coded each typing trial for whether the article–noun sequence was typed as one word without spaces (= 1), or as two words (= 0). We then calculated a chunk ratio for each participant as the number of one-word responses out of the total number of trials.

Forced-Choice Trials

The 12 forced-choice trials assessed the offline mastery of the article–noun pairings. We predicted that participants in the unsegmented-first condition would show better learning of the article–noun associations than those in the segmented-first condition (as in the study by Siegelman & Arnon, 2015). For each trial, we obtained a binary response, correct (= 1) or incorrect (= 0), and calculated an accuracy score for each participant as the number of correct responses out of the total number of trials.

Same-/Different-Gender Trials: Accuracy

At the end of each of the 24 same-/different-gender trials, participants were asked to choose the object that was best described by the sentence they heard. This allowed us to examine how well they learned the novel labels. We predicted that participants in the two conditions would learn the noun labels with similar accuracy (this was around 90% in the study by Siegelman & Arnon, 2015), because order of exposure to segmented and unsegmented sentences should not influence the learning of the object–label mapping. Here too, for each trial we obtained a binary response, correct (= 1) or incorrect (= 0), and calculated an accuracy score for each participant as the number of correct responses out of the total number of trials.

Same-/Different-Gender Trials: Eye Gaze

We analyzed the eye-tracking data to see if the proportion of fixations to the target varied as a function of trial type (same-gender, different-gender) and condition (segmented-first, unsegmented-first). We predicted (a) an effect of trial type, such that participants would show a higher proportion of fixations to the target in different-gender trials compared to same-gender trials, and (b) a condition by trial-type interaction, such that the difference in proportion of fixations to the target between the two trial types would be smaller, or even not found, in the segmented-first condition compared to the unsegmented-first one. This would suggest that participants in the unsegmented-first condition develop stronger associations between the article and the noun that lead to this facilitation effect.

Results

Behavioral data (typing responses during typing trials, and accuracy during forced-choice trials and same-/different-gender trials, i.e., responses coded 0 or 1) were extracted from E-prime and filtered to include only the needed measures. All eye-gaze data were processed using the batch processing option in EyeLink Data Viewer (SR Research, 2018). Our interest areas were predefined as the location of the target and distractor images, and we defined our interest period as detailed below. We then used the Analysis menu in Data Viewer to generate the following output reports: trial, interest area, and sample. The trial report provides a summary of participants' gaze in each individual trial, and the interest area report provides information about each of the interest areas (target and distractor images) in each trial separately. The sample report offers the most fine-grained level of detail about the data, such that each row represents a separate sample of looking data; each sample is time-stamped in milliseconds and contains information about the position of the tracked eye. Because we used the EyeLink 1000 Plus Eye Tracker, our sample report included 1,000 samples per second. Our output reports included fixations that occurred during our window of analysis; fixations that occurred outside this window of interest were not extracted or analyzed. All files were then converted into Excel comma-separated files to facilitate data analysis. All data sheets are available in the Results folder on the project OSF page (https://osf.io/98ak3/), along with a README file that includes notes on each file and variable names.

All analyses were conducted in R (R Core Team, 2018) using the lme4 (Bates et al., 2015) and eyetrackingR (Dink & Ferguson, 2015) packages. Our analysis code is available on the project OSF page. In all analyses, we fitted the maximal linear mixed-effects models that are possible and justified given our design and hypotheses (as recommended by Barr et al., 2013). For all statistical tests reported below, we interpreted the findings as statistically significant if $p < .05$, unless stated otherwise (i.e., in the case of multiple comparisons).

In line with Siegelman and Arnon (2015), we planned to exclude participants whose accuracy score was less than two standard deviations from the mean accuracy on the forced-choice trials, because such low accuracy is an indication of difficulty in learning the artificial language. On average, most participants ($n = 79$) achieved accuracy scores higher than 0.5 in the forced-choice trials (range: 0.25–1.00; $M = 0.67$; $SD = 0.16$; 95% CI [0.64, 0.70]), and there was only one participant whose score was two standard deviations below the mean (0.25). This participant was excluded from *all* subsequent analyses and was not replaced with a new one; this brought our sample down to

99 participants. Further exclusionary criteria were registered and implemented separately for the analyses relating to each research question, as detailed below.

Typing Trials

As registered, to test our predictions, we coded each typing trial for whether the article–noun sequence was typed as one word without spaces (= 1), or as two words (= 0). Coding was completed by an undergraduate research assistant in the lab and by the first author (there was 100% agreement between the two coders). Cases in which the article and noun were not clearly identified in the typed response were coded as NA. We then calculated a one-word ratio for each participant (one-word typing trials/total typing trials). This ratio was calculated out of the total number of typing trials (not excluding responses that were coded as NA). There were 30 typing trials in the entire experiment, half in the first exposure block and half in the second exposure block. For participants in the unsegmented-first condition, the first block consisted of unsegmented sentences and the second block consisted of segmented sentences. For participants in the segmented-first condition, the order was reversed. If a participant typed responses for all 30 trials, but some of those were coded as NA, their one-word ratio was still calculated out of 30 trials. Following Siegelman and Arnon (2015), we excluded participants who *always* typed article–noun units as one word, because this would indicate that they did not accurately segment the sentences by the end of the study. These participants were excluded *only* from the analyses of the typing trials. There were two such participants (leaving us with 97 in the sample). One other participant typed in only English words for all sentences (e.g., “I know what time it is”) and was also excluded from the analyses. Thus, our final sample for the typing trial analyses was 96 participants.

We had a total of 2,880 typing trials (96 participants × 30 trials); 49 of these (∼2%) were coded as NA because they only included the carrier phrase or an indecipherable version of the article–noun unit. We were able to code participants’ responses to the article–noun unit as either one or two words in 2,831 trials (∼98%). Participants typed the article–noun sequence as one word in 26.21% of the trials. As predicted, participants in the unsegmented-first condition were more likely to treat the sequence as one word: In the segmented-first condition, only 8.8% of article–noun combinations were typed as one word ($M = 0.08$ $SD = 0.17$; 95% CI [0.04, 0.13]), whereas in the unsegmented-first condition, 44% of article–noun combinations were typed as one word ($M = 0.43$; $SD = 0.35$; 95% CI [0.33, 0.54]).

In order to test whether exposure condition had an effect on participants’ performance in the typing trials, we ran a mixed-effects logistic regression model to predict whether the article–noun sequence was typed as one word or two words, with condition (segmented-first, unsegmented-first) as a fixed effect, and by-subject and by-item random intercepts (registered model syntax: one-word-ratio $\sim$ condition + (1|subject) + (1|item)). Item corresponds to the different objects that appeared in each typing trial (i.e., a total of 12 different items). Participants in the unsegmented-first condition were more likely to treat the article–noun sequence as one word compared to those in the segmented-first condition (b = 3.68; SE = 0.63; 95% CI [2.49, 5.03]; z = 5.82; $p < .001$). These analyses replicate the previous findings reported by Siegelman and Arnon (2015): Participants who were exposed first to unsegmented sentences were more likely to treat the article–noun sequence as one chunk.

To unpack this effect further, in additional unregistered analyses, we examined the differences between the conditions in the first and second blocks of the exposure. In the segmented-first condition, the first 15 typing responses were in response to segmented sentences, whereas in the unsegmented-first condition, the first 15 responses were in response to unsegmented sentences. We fitted a mixed-effects logistic regression model with condition (segmented-first, unsegmented-first) and block (1, 2) as fixed effects, along with the interaction term of condition and block, and by-subject and by-item random intercepts, as well as by-subject random slopes for block (model syntax: one-word-ratio $\sim$ condition * block + (1+block|subject) + (1|item)). We found a significant effect of condition (b = 4.82; SE = 0.9; 95% CI [3.13, 6.72]; z = 5.35; $p < .001$), block (b = −2.81; SE = 0.71; 95% CI [−4.42, −1.52]; z = −3.94; $p < .001$), and their interaction (b = −2.88; SE = 1.13; 95% CI [−5.16, −0.56]; z = −2.55; p = .01). When we unpacked these effects further, we found that the one-word ratio was low in the segmented-first condition, and did not change between blocks, as would be expected if participants were treating the article and noun as two words from the start (first block: M = 0.07; SD = 0.17; 95% CI [0.02, 0.12]; second block: M = 0.1; SD = 0.2; 95% CI [0.04, – 0.15]; t(48) = 0.82; 95% CI [−0.03, 0.07]; p = .41; Cohen's d = 0.12; 95% CI [−0.16, 0.4]). In contrast, in the unsegmented-first condition, the one-word ratio was higher in the first block (M = 0.58; SD = 0.41; 95% CI [0.46, 0.70]) compared with the second block (M = 0.28; SD = 0.40; 95% CI [0.16, 0.40]; t(46) = 5.00; 95% CI [0.18, 0.42]; $p < .001$; with a medium effect size: Cohen's d = 0.73; 95% CI [0.4, 1.05]). That is, participants in this condition started out typing the article–noun sequence as one word, and after exposure to segmented sentences started to segment the sequence into two words. We examine the relation between chunking and better learning of the article–noun mapping in the next section, after we analyze accuracy on the forced-choice trials.

Forced-Choice Trials

Participants were above chance in choosing the correct article on the forced-choice trials: $t(98) = 11.12$; 95% CI [0.64, 0.71]; $p < .001$ (see Appendix S2 in the Supporting Information online for registered analyses that examined the effect of list and article type on performance). To test the effect of condition, we ran a mixed-effects logistic regression model with condition (segmented-first, unsegmented-first) as a fixed effect, and random by-subject and by-item intercepts (registered model syntax: accuracy $\sim$ condition + (1|subject) + (1|item)). Item corresponds to the target object that appeared in each trial (i.e., 12 different items). As predicted, accuracy was higher in the unsegmented-first condition ($M = 0.71$; $SD = 0.16$; 95% CI [0.66, 0.76]) than in the segmented-first condition ($M = 0.65$; $SD = 0.15$; 95% CI [0.60, 0.69]; see Figure 1; $b = 0.32$; $SE = 0.15$; 95% CI [0.02, 0.64]; $z = 2.11$; $p = .03$). These results replicate the effect reported by Siegelman and Arnon (2015) showing that the learning condition significantly impacts participants' accuracy in choosing the correct article, with participants who are exposed to unsegmented input first performing better than those exposed to segmented input first.

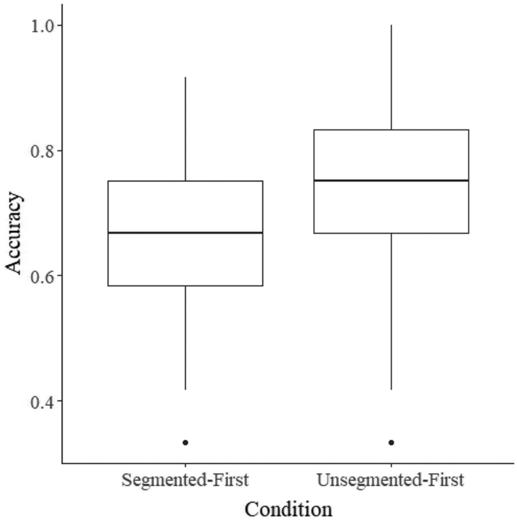

Figure 1 Accuracy in the forced-choice trials by condition.

Finally, we return to the prediction that more chunking would lead to better learning of the article–noun mapping. To examine it, we looked at the correlation between the one-word ratio in the typing trials and accuracy scores on the forced-choice trials. When looking across all participants, the correlation (Spearman's rho) was positive, but not significant ($\rho = .12$; $p = .12$). In contrast with the results reported by Siegelman and Arnon (2015) and our predictions, more chunking (i.e., treating the article–noun as one unit) was *not* correlated with better accuracy on the forced-choice trials. However, further inspection of our typing-trial data revealed a high number of zero chunkers (i.e., participants who *never* treated the article–noun sequence as one word) in the segmented-first condition ($n = 25$). This led us to make a more refined prediction that was not tested by Siegelman and Arnon and that we did not register: If early chunking leads to better accuracy, then the correlation between accuracy in learning the article–noun association and chunking the article–noun as one unit should be positive *only* in the unsegmented-first condition (because participants in the segmented-first condition are segmenting from the start). These unregistered analyses were not statistically significant but revealed a small positive correlation between chunking and accuracy in the unsegmented-first condition ($n = 47$; $\rho = .22$; $p = .06$), and a small negative correlation in the segmented-first condition ($n = 49$; $\rho = -.25$; $p = .07$. These findings demonstrate that the relationship between the early units of learning and learning outcomes varies by condition.

Same-/Different-Gender Trials: Accuracy

At the end of each of the 24 eye-tracking trials, participants were asked to select the correct object of the two that appeared on the screen, allowing us to test how well they learned the noun labels. On same-gender trials, the labels for both objects were from the same noun class, whereas on different-gender trials, the labels for objects were from two different classes. Participants were above chance in choosing the correct object: $t(98) = 26.57$; 95% CI [0.82, 0.88]; $p < .001$ (see Appendix S3 in the Supporting Information online for registered analyses that examined the effect of list, article type, and trial type on performance). To examine whether accuracy differed by condition, we fitted a mixed-effects logistic regression model with condition (segmented-first, unsegmented-first) as a fixed effect, and by-subject and by-item random intercepts (model syntax: accuracy $\sim$ condition + (1|subject) + (1|item)). Item corresponds to the target object that appeared in each trial (i.e., a total of 12 different items). As predicted, accuracy did not differ by condition ($b = -0.01$; $SE = 0.25$; 95% CI [-0.52, 0.49]; $z = -0.05$; $p = .96$): Learning the noun

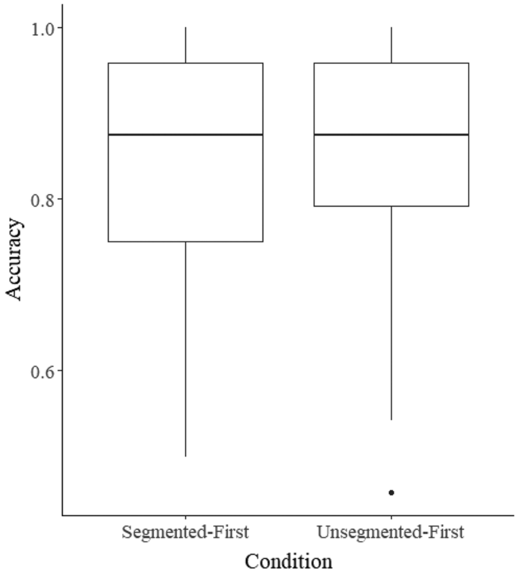

Figure 2 Accuracy in the same-/different-gender trials by condition.

labels was similar in the two conditions (segmented-first: $M = 0.85$; $SD = 0.13$; 95% CI [0.81, 0.89]; unsegmented-first: $M = 0.85$; $SD = 0.13$; 95% CI [0.81, 0.89]; see Figure 2).

Same-/Different-Gender Trials: Eye Gaze

Our dependent variable was the proportion of fixations to the target out of the total fixations. We analyzed participants' eye gaze by using (a) a broad time window starting 200 ms from article onset to noun offset and (b) a fine-grained cluster analysis. These analyses provide a comprehensive investigation of the impact of order of exposure on processing patterns. Although we believed the broader time window would show evidence of differences in facilitation by trial type and condition, it was possible that this time window would not be sensitive enough (given the short exposure to a completely novel artificial language) to capture differences. Conducting the fine-grained analyses allowed us to explore in detail the impact of order of exposure on the article–noun association and the exact timing in which the proportion of fixations to the target exceeds those to the distractor between trial types and conditions.

Prior to performing any analyses of eye-tracking data, we excluded trials in which participants selected the wrong noun. As mentioned earlier, participants were mostly accurate in selecting the right noun at the end of each of the same-/different-gender trials; however, out of a total of 2,376 trials (24 trials × 99 participants), there were 350 trials in which participants selected the wrong noun. These trials came from 79 participants, who, on average, selected the wrong noun in 4.43 trials (*SD* = 2.93; range: 1–13). These 350 trials were removed from all subsequent analyses, leaving us with 2,026 trials; this proportion of removed trials (350/2,376 = 15%) was higher than that reported in Siegelman and Arnon's (2015) study, wherein accuracy was at 90%, suggesting that only 10% of trials would be removed for wrong noun selection. That is, participants in the present study were slightly less accurate in learning the noun labels than in prior work.

We defined our spatial regions of interest as the two 400-×-400-pixel squares in which the target and distractor objects appeared.

Broad Time Window

We predicted an effect of trial type, such that participants would show a higher proportion of fixations to the target in different-gender trials compared to same-gender trials, and a condition by trial-type interaction, such that the difference in proportion of fixations to the target between the two trial types would be smaller, or even not found, in the segmented-first condition compared to the unsegmented-first one.

Facilitation was measured as the proportion of fixations to the target out of the total fixations to the target and distractor during the window of analysis (i.e., interest period). Given that it takes about 200 ms to plan and initiate a saccadic eye movement (Altmann, 2011), we started measuring eye movements 200 ms from the onset of the article; the window of analysis included the *remaining* length of the article (about 80 ms), the 250 ms of pause that follow it (in segmented sentences), and the duration of the noun (as was done in previous comparable work; Grüter et al., 2012; Lew-Williams & Fernald, 2007, 2010); hence, our broad window of analysis started at 200 ms from the beginning of the article and lasted until the end of the noun. Only fixations that occurred during this time window were included in the analyses. Variations in the duration of the noun led to small differences in the duration of the interest period (*M* = 1,183.34 ms; *SD* = 96.14; range: 1,033–1,418). These differences should not impact our dependent variable as it is calculated as the proportion of fixations to the target out of the total fixations within the interest period for each trial.

When inspecting the data, we found that there were two trials for which the summation of all fixation durations exceeded the duration of the interest period, and six trials in which the total amount of fixations (to the screen) during the window of analysis was zero; we removed these eight trials from all subsequent analyses. Further, there were 66 trials in which there were no fixations to the target and distractor during the window of analysis (note that during these trials, the eye tracker detected fixations, but none were judged to be located on the areas of interest we defined, i.e., the target and distractor). These trials were also removed from all subsequent analyses. Trials in which participants' total fixation time—time of fixation on the target *and* distractor—was less than 20% of the time window specified above (i.e., the interest period) were also excluded from subsequent analyses (Borovsky et al., 2016; Borovsky & Peters, 2019); there were 40 trials that fulfilled this criterion, leaving us with a total of 1,912 trials (80% of the data) from 99 participants (number of valid trials per participant: $M = 19.31$; $SD = 3.64$; 95% CI [18.58, 20.03]).

Participants' proportion of fixations to the target averaged 0.58 ($SD = 0.10$; 95% CI [0.57, 0.6]) and was significantly higher than chance: $t(197) = 11.03$; $p < .001$ (see Table 2 for data by condition and trial type; see Appendix S4 in the Supporting Information online for registered analyses that examined the effect of list and article type on performance).

Table 2 Proportion of fixations to the target by condition and trial type

Condition	Trial type	Mean (*SD*)	95% CI	*t* test against chance
Segmented-first	Different-gender	0.588 (0.109)	[0.55, 0.62]	$t(48) = 5.63$; $p < .001$
	Same-gender	0.581 (0.103)	[0.55, 0.61]	$t(48) = 5.46$; $p < .001$
Unsegmented-first	Different-gender	0.587 (0.101)	[0.55, 0.61]	$t(49) = 6.12$; $p < .001$
	Same-gender	0.572 (0.107)	[0.54, 0.60]	$t(49) = 4.75$; $p < .001$

Prior to fitting our mixed-effects model, we inspected the data again, but this time without averaging the proportion of fixations per subject by trial type, as this is not necessary for a mixed-effects model. Here, we treated the 1,912 trials as individual data points. The data in the four different condition-by-trial-type combinations were not normally distributed (Shapiro Wilk's test: $p < .001$), but the variances within each of the conditions (by trial type) were homogeneous (Levene's test: $p > .68$). When we inspected the histogram for each condition, we saw that the data were slightly left-skewed; however, since the skewness values are between −0.5 and 0.5 (segmented-first: −0.22;

unsegmented-first: −0.24), the data in each condition can be seen as approximately symmetric and do not require transformation.

To test our prediction that the proportion of gaze would be impacted by exposure condition and trial type, we registered a mixed-effects linear regression model to predict participants' proportion of fixations to the target. The model included condition (segmented-first, unsegmented-first) and trial type (same-gender, different-gender) as fixed effects, along with the interaction term of trial type and condition, and by-subject and by-item random intercepts, and by-subject and by-item random slopes for trial type (model syntax: proportion to target ∼ condition * trial-type + (1+trial-type|subject) + (1+trial-type|item)). As in previous models, item corresponds to the target object that appeared in each trial (i.e., a total of 12 different items). However, upon further reflection and inspection of our data, we realized that we should not have included a by-item random slope for trial type in our model, given that each of the 12 target objects appeared only once as a target in each of the trial types. Thus, our revised model includes condition (segmented-first, unsegmented-first) and trial type (same-gender, different-gender) as fixed effects, along with the interaction term of trial type and condition, and by-subject and by-item random intercepts, and by-subject random slopes for trial type (model syntax: proportion to target ∼ condition * trial-type + (1+trial-type|subject) + (1|item)). In contrast with our predictions, we found no effect of condition ($b = -0.0083$; $SE = 0.015$; 95% CI [-0.03, 0.02]; $t = -0.55$; $p = .58$), trial type ($b = 0.013$; $SE = 0.015$; 95% CI [-0.01, 0.04]; $t = 0.904$; $p = .36$), and no interaction between condition and trial type ($b = -0.0021$; $SE = 0.03$; 95% CI [−0.06, 0.05]; t = −0.068; $p = .94$). That is, we do not see facilitation (higher proportion of gaze to the target) in the unsegmented-first condition compared to the segmented-first one, or in the different-gender trials compared to the same-gender trials (see Figure 3).

Fine-Grained Cluster Analysis

In order to inspect participants' looking patterns during the experiment in more detail and explore whether there were periods within our window of analysis in which fixations to the target differed by trial type and condition, we conducted fine-grained analyses using the sample report. These analyses follow the guidelines laid out for the eyetrackingR package (Dink & Ferguson, 2015; see http://www.eyetracking-r.com/).

Our initial sample data set included sample data only from the 1,912 trials that we reported on in the previous section. Given the fine-grained nature of the data we are using here, we examined the amount of data loss that happened because the participants were blinking, or the eye tracker lost track of

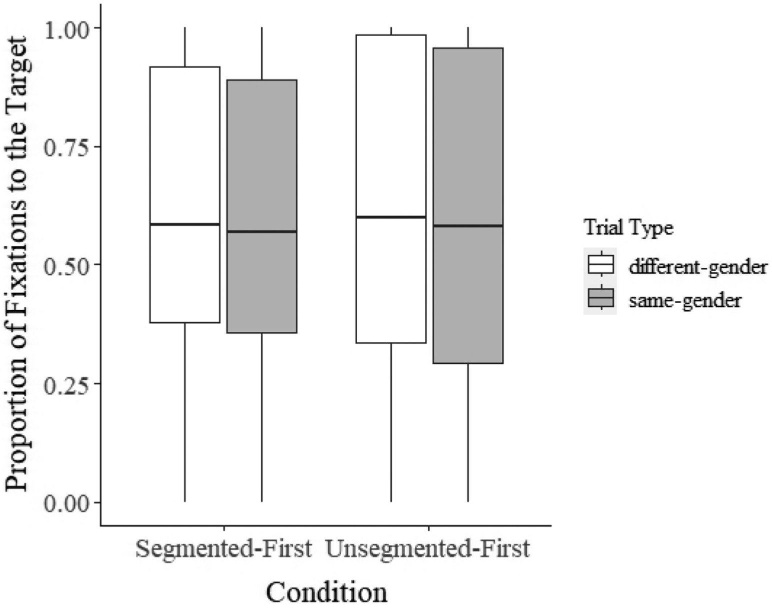

Figure 3 Proportion of fixations to the target by condition and trial type (within each condition, data from different-gender trials are on the left side and data from the same-gender trials are on the right side).

their eyes (i.e., track loss). On average, participants contributed 1,123 samples (range: 258–1,418). However, track loss (proportion of samples in which the eye tracker lost track of the participant's eye) ranged from 0% to 70%. In cleaning our data further, we removed all trials in which the track loss was higher than 20%. There were 64 such trials, leaving us with a total of 1,848 trials from 99 subjects with a track-loss proportion that ranged from 0% to 10%. The mean proportion of samples contributed per trial across our participants was 98% (*SD* = 2.1%).

We calculated the proportion of time spent fixating on the target and distractor images in each 50-ms bin within our window of analysis (Borovsky et al., 2016; Borovsky & Peters, 2019), which started 200 ms after the onset of the article (1,650 ms from the onset of the sentence). We then performed fine-grained analyses of rapid fixations by using a nonparametric cluster test procedure to characterize the exact timing of participants' preference to look at the target. Specifically, we used this approach to identify the exact time window(s) across our broad window of analysis during which the proportion of

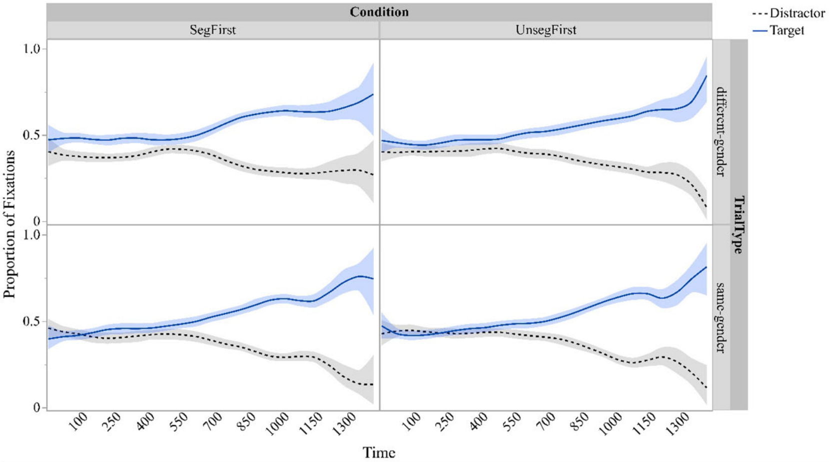

Figure 4 Proportion of fixations to the target (straight line) and distractor (dashed line) by condition and trial type. Time is shown in milliseconds. The time window plotted in this figure reflects our window of interest and *not* the duration of the entire trial; thus, the zero point is the beginning of the window of analysis, which started 200 ms from the onset of the article (i.e., 1,650 ms into the sentence). The shaded areas around the lines represent the bootstrap confidence region of the fit for 95% of the smoother fits. SegFirst = segmented-first; UnsegFirst = unsegmented-first.

fixations to the target differed by condition and trial type. This cluster-based permutation approach (see Groppe et al., 2011; Maris & Oostenveld, 2007, for detailed tutorials on this method) was aimed at identifying clusters (i.e., temporally adjacent time points) in which there is a statistically significant difference between the comparison conditions. No cluster-based differences emerged when we compared participants' proportion of fixations to the target by trial type in each of the conditions separately, and when we compared the proportion of fixations by condition to each of the trial types separately. That is, although participants looked longer at the target compared to the distractor as the sentence unfolded, this pattern of gaze was similar in the two exposure conditions and for both different-gender and same-gender trials (see Figure 4). This result supports our prior analyses that did not show any effect of condition, trial type, or the interaction between them on participants' fixations to the target.

Discussion

We set out to investigate the impact of linguistic unit size on learning and to ask if the order in which distributional information is presented influences the formation of linguistic predictions and their facilitatory effect on processing. In particular, we wanted to replicate and extend previous findings on the facilitative effect of larger units on learning grammatical gender agreement, and to ask if the facilitation is driven by an increased association between the article and noun that would be reflected in eye-gaze patterns. In order to test this, we added an eye-tracking component to an existing artificial language learning paradigm (Siegelman & Arnon, 2015) to see (a) whether the learning of article–noun agreement would be improved when participants were exposed first to unsegmented input (leading to the extraction of MWUs) and (b) whether this improvement would lead to increased predictive relations between the article and the noun that would be reflected in more facilitative gaze. Our behavioral results replicated the facilitative effect of larger units on learning: Participants in the unsegmented-first condition showed better learning of the article–noun mapping and were more likely to treat the article–noun sequence as one unit than those in the segmented-first condition. These findings provide additional support for the facilitative effect of multiword building blocks on learning grammatical relations (Arnon, 2010, 2021; Arnon & Christiansen, 2017).

However, in contrast with our predictions, our eye-tracking results did not indicate increased predictive relations between the article and the noun in the unsegmented-first condition: Participants' proportion of fixations to the target was not higher in the unsegmented-first condition compared to the segmented-first condition. That is, although accuracy in forming the article–noun association was improved in the unsegmented-first condition, this was not reflected in gaze patterns. Interestingly, and contrary to our predictions as well, there was also no difference in gaze patterns between the different-gender and same-gender trials: Even though the different-gender trials were more informative (because only one of the presented objects could appear with the heard article), the proportion of gaze to the correct target was not higher compared to same-gender trials (where both objects could appear with the article). This result differs from what is found in the processing of grammatical gender for native speakers (children and adults, e.g., Brouwer et al., 2017; Dussias et al., 2013; Grüter et al. 2012; Hopp & Lemmerth, 2016; Lew-Williams & Fernald, 2007), but is aligned with the findings of previous studies testing adult learners in their L2. For instance, Lew-Williams and Fernald (2010) also reported that even though the adults were familiar with the nouns they were tested on (which in our design is exemplified by the high accuracy of learning the object–noun mapping regardless of condition and trial type), they did not show different looking patterns in the same- and different-gender trials.[5] Importantly, we do see evidence of predictive gaze in this paradigm: Participants started looking at the target more than the distractor before the end of the noun was heard (see Figure 4; around 500–550 ms from the start of our window of analysis; this time point in the window of analysis is about 200 ms from the beginning of the noun), indicating the formation of linguistic predictions in the artificial language. However, as evident in our analyses, these gaze patterns did not differ by condition, trial type, or the interaction between them.

There are several ways to interpret the lack of effect of exposure condition on gaze patterns. One possibility is that the mechanism by which larger units facilitate learning does not involve increased predictive relations. This conclusion, however, would seem premature given other considerations. In particular, it is possible that the language was not learned well enough, or for long enough, to support the development of knowledge-sensitive gaze patterns. That is, participants may have not learned the nouns and the article–noun mappings well enough to generate online predictions. This possibility is supported by the lack of difference in gaze patterns between the different-gender and same-gender trials. Even though the different-gender trials should have been more informative, participants did not look more (or earlier) at the target in these trials (regardless of exposure condition) and were not more accurate on these trials. Participants were somewhat less accurate in selecting the correct noun in this study (around 80% accuracy) compared to the study by Siegelman and Arnon (2015; around 90% accuracy), which may have also impacted their ability to use the article information in real time. More importantly, participants were not at ceiling in learning the article–noun mappings: Accuracy on the forced-choice trials was 70% in the unsegmented-first condition and 65% in the segmented-first condition (these accuracy rates are similar to those found in prior work, and not surprising given that the language was taught for a mere 20 min). If participants did not learn the article–noun mapping well enough, we cannot expect this knowledge to facilitate the processing of the upcoming noun in real-time processing. The idea that gaze patterns are impacted by the entrenchment and stability of linguistic knowledge is supported by findings showing more nativelike predictive gaze in more proficient L2 speakers (e.g., Hopp & Lemmerth, 2018; though there is a debate on the presence of predictive gaze in L2 processing in general: Kaan, 2014), more predictive gaze in adults compared to children (Aumeistere et al., 2022), and more in children with larger productive vocabularies (Mani & Huettig, 2012).

To further explore this, we looked at gaze patterns only for trials where participants had selected the correct article–noun mapping in the forced-choice trials. These analyses yielded the exact same patterns we reported on earlier (see the R code for these unregistered analyses on the OSF page at https://osf.io/98ak3/), suggesting that the lack of effect we observed cannot be explained by the accuracy of learning the article–noun association. However, since we only had one observation for each article–noun mapping (i.e., there were only 12 forced-choice trials), our ability to reliably assess each individual's knowledge of each article–noun mapping is inherently limited. An additional property of our design that could have limited the detection of prediction-sensitive gaze is that the length of the window between the article (the element that could be used to form predictions) and the noun (the predicted element) was relatively short (250 ms of silence between the article and the noun). In studies of natural language, researchers often add intervening linguistic elements in order to lengthen the window where predictions can be formed: For example, in a paper looking at processing of grammatical gender in Dutch, an adjective was added between the article and the noun to enable more accurate detection of predictive gaze (Loerts et al., 2013).

Our behavioral results support the accumulating evidence that early unit size impacts learning, that learning from MWUs can improve learning, and that manipulating adults' input can increase their reliance on MWUs, leading to better outcomes (Arnon, 2010; Arnon & Ramscar, 2012; Havron et al., 2018; Siegelman & Arnon, 2015). These findings are part of a bigger picture in which MWUs are important building blocks in language learning and use (the starting big approach; Arnon, 2021). Under this approach, infants use both words and MWUs as early building blocks, whereas adults—because of their existing knowledge of words—rely more on individual words, which hinders the learning of certain grammatical relations. The reliance on MWUs in L1 acquisition is supported by findings showing that preverbal 11-month-olds are sensitive to multiword frequency and can distinguish between high- and low-frequency trigrams (e.g., *clap your hands* vs. *take your hands*; Skarabela et al., 2021), and that adults are faster to process MWUs that they learned early in life (multiword age-of-acquisition effect; Arnon et al., 2017). Adults' lesser reliance on MWUs in learning a L2 is backed up by findings that show how literacy enhances the use of words as units of processing and influences segmentation strategies (Havron & Arnon, 2017a, 2017b).

The combined findings have implications for how we teach L2s: They suggest that encouraging adults to extract and use MWUs could enhance learning. One way of doing this is by exposing them to unsegmented auditory input, as we did in this study. Another way is to modify their written input to reduce the differentiation of words. In a recent study, we taught adult Hebrew speakers article–noun agreement in a miniature language based on real Greek. The language consisted of 12 nouns, half neuter and half feminine in gender. Participants were exposed to written sentences in the language, and we manipulated whether the article–noun sequences were written as two words (as in real Greek) or as one word. Participants showed better learning of the mapping when the sequences were written as one word, suggesting that manipulating orthography could lead to increased chunking (Kimchi, et al., under review).

Limitations and Future Directions

Although we found better learning of article–noun associations when participants were exposed to unsegmented input first, we did not find evidence for increased predictive gaze in this condition. One limitation of the current study is that participants were not faster to orient to the correct noun in different-gender trials compared to the less informative same-gender trials. This suggests that they had not learned the language well enough to use the article information predictively. To test whether an effect of condition will emerge if there is better learning of the language, and if the window for forming predictions is longer, future work should (a) increase exposure time, (b) increase the window between the article and the noun, (c) add additional testing of the article–noun mapping so that the knowledge of each is reliably assessed, and (d) look at eye-gaze data only for trials where the article–noun mapping has been learned correctly. An additional avenue for future work is to use natural languages instead of artificial ones, and to see whether a stronger association between the article and noun is formed when the auditory stimulus is accompanied by an unsegmented orthographic representation.

Conclusions

In sum, we have presented results from an artificial language learning study, including an eye-tracking component, in which we investigated the effect of early unit size on learning. This paper supports the accumulating evidence that reliance on larger units early on can improve learning outcomes; however, we did not find evidence that this is driven by an increase in the predictive relations between the words.

Final revised version accepted 28 November 2023

OPEN RESEARCH BADGES

This article has earned Open Data and Open Materials badges for making publicly available the digitally-shareable data and the components of the research methods needed to reproduce the reported procedure and results. The article has also earned a Preregistered Research Designs badge for having a preregistered research design. Data and materials design are available at https://osf.io/98ak3/ and preregistered is available at https://osf.io/9javf. All proprietary materials have been precisely identified in the manuscript.

Notes

[1] Within the predictive processing literature, there is ongoing debate about whether the processing effects seen in existing studies are driven by active prediction (e.g., Kuperberg & Jaeger, 2016) or by facilitation (e.g., Huettig, 2015). Here, we do not address this debate but are interested only in asking whether exposure to unsegmented input first creates stronger associations, and hence increased facilitation for processing the upcoming word.

[2] The article *si* replaced *se* from Experiment 1 in the study by Siegelman and Arnon (2015). That is due to findings by Siegelman and Arnon showing an effect of article type on participants' performance; such effect may have stemmed from the resemblance to the Hebrew demonstrative *ze*. The article *si* was used in Experiment 2 in Siegelman and Arnon's study, and we used the recorded sentences from that study.

[3] The distractor block used by Siegelman and Arnon (2015) included six objects and their novel labels. Our change from six to four was necessary to maintain a reasonable length for the experiment without drastically changing its design.

[4] These two lists were not used in the original experiment carried out by Siegelman and Arnon (2015), but we believe they are necessary to ensure a fully balanced design.

[5] We reviewed this literature extensively in our introduction, and the emerging picture is that adults' ability to use gender-marked articles to predict the upcoming noun in a L2 is highly variable and is impacted by proficiency level, the specific L1 and L2 involved, and various other influences.

References

Altmann, G. T. M. (2011). Language can mediate eye movement control within 100 milliseconds, regardless of whether there is anything to move the eyes to. *Acta Psychologica*, *137*(2), 190–200. https://doi.org/10.1016/j.actpsy.2010.09.009

Arnon, I. (2010). *Starting big: The role of multiword phrases in language learning and use*. Stanford University.

Arnon, I. (2021). The Starting Big approach to language learning. *Journal of Child Language*, *48*(5), 937–958. https://doi.org/10.1017/S0305000921000386

Arnon, I., & Christiansen, M. H. (2017). The role of multiword building blocks in explaining L1–L2 differences. *Topics in Cognitive Science*, *9*(3), 621–636. https://doi.org/10.1111/tops.12271

Arnon, I., & Clark, E. V. (2011). Why *brush your teeth* is better than *teeth*: Children's word production is facilitated in familiar sentence-frames. *Language Learning and Development*, *7*(2), 107–129. https://doi.org/10.1080/15475441.2010.505489

Arnon, I., & Cohen Priva, U. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech*, *56*(3), 349–371. https://doi.org/10.1177/0023830913484891

Arnon, I., McCauley, S. M., & Christiansen, M. H. (2017). Digging up the building blocks of language: Age-of-acquisition effects for multiword phrases. *Journal of Memory and Language*, *92*, 265–280. https://doi.org/10.1016/j.jml.2016.07.004

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, *122*(3), 292–305. https://doi.org/10.1016/j.cognition.2011.10.009

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82. https://doi.org/10.1016/j.jml.2009.09.005

Aumeistere, A., Bultena, S., & Brouwer, S. (2022). Wisdom comes with age? The role of grammatical gender in predictive processing in Russian children and adults. *Applied Psycholinguistics*, *43*(4), 867–887. https://doi.org/10.1017/S0142716422000170

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*(3), 241–248. https://doi.org/10.1111/j.1467-9280.2008.02075.x

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*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Package lme4. *Journal of Statistical Software*, *67*(1), 1–91.

Boersma, P., & Weenink, D. (2005). *Praat: Doing phonetics by computer*. https://www.fon.hum.uva.nl/praat/

Bordag, D., & Pechmann, T. (2007). Factors influencing L2 gender processing. *Bilingualism: Language and Cognition*, *10*(3), 299–314. https://doi.org/10.1017/S1366728907003082

Borovsky, A., Ellis, E. M., Evans, J. L., & Elman, J. L. (2016). Semantic structure in vocabulary knowledge interacts with lexical and sentence processing in infancy. *Child Development*, *87*(6), 1893–1908. https://doi.org/10.1111/cdev.12554

Borovsky, A., & Peters, R. E. (2019). Vocabulary size and structure affects real-time lexical recognition in 18-month-olds. *PloS One*, 1–21. https://doi.org/10.1371/journal.pone.0219290

Brouwer, S., Sprenger, S., & Unsworth, S. (2017). Processing grammatical gender in Dutch: Evidence from eye movements. *Journal of Experimental Child Psychology*, *159*, 50–65. https://doi.org/10.1016/j.jecp.2017.01.007

Deutsch, A., & Bentin, S. (2001). Syntactic and semantic factors in processing gender agreement in Hebrew: Evidence from ERPs and eye movements. *Journal of Memory and Language*, *45*(2), 200–224. https://doi.org/10.1006/jmla.2000.2768

Dink, J. W., & Ferguson, B. (2015). *eyetrackingR: An R library for eye-tracking data analysis*. Retrieved from http://www.eyetrackingr.com

Dussias, P. E., Valdés Kroff, J. R., Guzzardo Tamargo, R. E., & Gerfen, C. (2013). When gender and looking go hand in hand: Grammatical gender processing in L2 Spanish. *Studies in Second Language Acquisition*, *35*(2), 353–387. https://doi.org/10.1017/S0272263112000915

Ellis, E. M., Borovsky, A., Elman, J. L., & Evans, J. L. (2015). Novel word learning: An eye-tracking study. Are 18-month-old late talkers really different from their typical peers? *Journal of Communication Disorders*, *58*, 43–157. https://doi.org/10.1016/j.jcomdis.2015.06.011

Ellis, N. C. (2007). Blocking and learned attention in language acquisition. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *29*(29), 965–970.

Ellis, N. C., & Sagarra, N. (2010). The bounds of adult language acquisition: Blocking and learned attention. *Studies in Second Language Acquisition*, *32*(4), 553–580. https://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*(4), 589–624. https://doi.org/10.1017/S0272263111000325

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE*, *8*(10), e77661. https://doi.org/10.1371/journal.pone.0077661

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology*, *48*(12), 1711–1725. https://doi.org/10.1111/j.1469-8986.2011.01273.x

Grüter, T., Lew-Williams, C., & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research*, *28*(2), 191–215. https://doi.org/10.1177/0267658312437990

Havron, N., & Arnon, I. (2017a). Minding the gaps: Literacy enhances lexical segmentation in children learning to read. *Journal of Child Language*, *44*(6), 1516–1538. https://doi.org/10.1017/S0305000916000623

Havron, N., & Arnon, I. (2017b). Reading between the words: The effect of literacy on second language lexical segmentation. *Applied Psycholinguistics*, *38*(1), 127–153. https://doi.org/10.1017/S0142716416000138

Havron, N., Raviv, L., & Arnon, I. (2018). Literate and preliterate children show different learning patterns in an artificial language learning task. *Journal of Cultural Cognitive Science*, *2*(1), 21–33. https://doi.org/10.1007/s41809-018-0015-9

Hopp, H. (2012). Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research*, *29*(1), 33–56. https://doi.org/10.1177/0267658312461803

Hopp, H. (2016). Learning (not) to predict: Grammatical gender processing in second language acquisition. *Second Language Research*, *32*(2), 277–307. https://doi.org/10.1177/0267658315624960

Hopp, H., & Lemmerth, N. (2016). Lexical and syntactic congruency in L2 predictive gender processing. *Studies in Second Language Acquisition*, *40*(1), 1–29. https://doi.org/10.1017/S0272263116000437

Hopp, H., & Lemmerth, N. (2018). Lexical and syntactic congruency in L2 predictive gender processing. *Studies in Second Language Acquisition*, *40*(1), 171–199. https://doi.org/10.1017/S0272263116000437

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118–135. https://doi.org/10.1016/j.brainres.2015.02.014

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different? *Linguistic Approaches to Bilingualism*, *4*(2), 257–282. https://doi.org/10.1075/lab.4.2.05kaa

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133–156. https://doi.org/10.1016/S0749-596X(03)00023-8

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. Campbell & R. Church (Eds.), *Punishment and aversive behaviour* (pp. 279–296). Appleton-Century-Crofts.

Kimchi, I., Reshef, N., Vasileva, A., & Arnon, I. (under review). Enhancing second language learning through orthographical changes. *Language Learning*.

Kruschke, J. K. (2005). Learning involves attention. In G. Houghton (Ed.), *Connectionist models in cognitive psychology* (pp. 113–140). Psychology Press.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science*, *18*(3), 193–198. https://doi.org/10.1111/j.1467-9280.2007.01871.x

Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language*, *63*(4), 447–464. https://doi.org/10.1016/j.jml.2010.07.003

Loerts, H., Wieling, M., & Schmid, M. S. (2013). Neuter is not common in Dutch: Eye movements reveal asymmetrical gender processing. *Journal of Psycholinguistic Research*, *42*(6), 551–570. https://doi.org/10.1007/s10936-012-9234-2

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(4), 843–847. https://doi.org/10.1037/a0029284

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning*, *68*(2), 321–391. https://doi.org/10.1111/lang.12286

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347. https://doi.org/10.1017/S0140525x12001495

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*(6), 909–957. https://doi.org/10.1111/j.1551-6709.2009.01092.x

R Core Team. (2018). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing.

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). Appleton-Century-Crofts.

Sabourin, L., Stowe, L. A., & de Haan, G. J. (2006). Transfer effects in learning a second language grammatical gender system. *Second Language Research*, *22*(1), 1–29. https://doi.org/10.1191/0267658306sr259oa

Shantz, K. (2018). *Why wait? Psycholinguistic investigations of the roles of learning condition and gender stability in L2 gender-based anticipation* [Unpublished doctoral dissertation]. University of Illinois at Urbana-Champaign.

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. https://doi.org/10.1016/j.jml.2015.07.003

Skarabela, B., Ota, M., O'Connor, R., & Arnon, I. (2021). 'Clap your hands' or 'take your hands'? One-year-olds distinguish between frequent and infrequent multiword phrases. *Cognition*, *211*, 104612. https://doi.org/10.1016/j.cognition.2021.104612

SR Research. (2018). *EyeLink data viewer 3.2.1* [Computer software].

Yoshida, H., & Burling, J. M. (2012). Highlighting: A mechanism relevant for word learning. *Frontiers in Psychology*, *3*, 1–12. https://doi.org/10.3389/fpsyg.2012.00262

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Accessible Summary

Appendix S1. The Article–Noun Combinations and the Objects Corresponding to Each Noun in the Distractor Block.

Appendix S2. Additional Analyses for the Forced-Choice Trials.

Appendix S3. Additional Analyses for the Same-/Different-Gender Trials: Accuracy.

Appendix S4. Additional Analyses for the Same-/Different-Gender Trials: Eye Gaze.