Exploring variation between artificial grammar learning experiments: Outlining a meta-analysis approach

Antony S. Trotter

University College London

Padraic Monaghan

Lancaster University & University of Amsterdam

Gabriël Beckers

Utrecht University

Morten H. Christiansen

Cornell University, Aarhus University, & Haskins Labs

Word count: 5706

Address for correspondence:

A. S. Trotter

Division of Psychology, Speech, Hearing, and Phonetic Sciences

University College London

2 Wakefield Street, WC1N 1PF

UK

Email: t.trotter@ucl.ac.uk

Abstract

Artificial grammar learning (AGL) has become an important tool used to understand aspects of human language learning and whether the abilities underlying learning may be unique to humans or found in other species. Successful learning is typically assumed when human or animal participants are able to distinguish stimuli generated by the grammar from those that are not at a level better than chance. However, the question remains as to what subjects actually learn in these experiments. Previous studies of AGL have frequently introduced multiple potential contributors to performance in the training and testing stimuli, but meta-analysis techniques now enable us to consider these multiple information sources for their contribution to learning – enabling intended and unintended structures to be assessed simultaneously. We present a blueprint for meta-analysis approaches to appraise the effect of learning in human and other animal studies for a series of artificial grammar learning experiments, focusing on studies that examine auditory and visual modalities. We identify a series of variables that differ across these studies, focusing on both structural and surface properties of the grammar, and characteristics of training and test regimes, and provide a first step in assessing the relative contribution of these design features of artificial grammars as well as species specific effects for learning.

Introduction

Artificial grammar learning (AGL) studies present learners with sequences of stimuli that inhere particular structural properties (Miller, 1958) of differing complexity (e.g., Reber, 1967), and then test learners on their ability to respond to sequences that incorporate aspects of this structure. Such an approach has been a very powerful method enabling investigations within a species into the possibilities and constraints on structural learning, such as distinctions between phrase-structure grammars or finite state grammars (e.g., Bahlmann, Schubotz, & Friederici, 2008), or the extent to which adjacent or non-adjacent dependencies in sequences are available to the learner (e.g., Conway et al., 2010; Gomez & Gerken, 1999; Jamieson & Mewhort, 2005; Lai & Poletiek, 2011; Vuong, Meier & Christiansen, 2016). The paradigm is also of great potential use across species, and has been extensively used to address questions about what structures are learnable by which species, and under what conditions (e.g., Abe & Watanabe, 2011; Chen et al., 2015; Fitch & Hauser, 2004; Saffran et al., 2008).

There has already been substantial progress made in addressing these questions, resulting in an intensive array of studies of learning in birds (e.g., Abe & Watanabe, 2011; Chen & ten Cate, 2015; Gentner et al., 2006; Spierings et al., 2015, 2017), non-human primates (e.g., Endress et al., 2010; Heimbauer et al., 2018; Wilson, Smith, & Petkov, 2015), as well as human children and adults (e.g., Frost & Monaghan, 2017; Gomez & Gerken, 1999; Saffran et al., 2008), addressing acquisition of multiple grammatical structures across these species. The other papers in this special issue provides a host of further examples of the paradigm in use.

However, testing different structures and different species raises substantial methodological problems when it comes to direct comparisons between grammars and between species. Potential confounds both within and across studies have caused substantial concern in the past in terms of the validity of conclusions being drawn from studies (e.g., Beckers et al., 2012, 2017; de Vries et al., 2008; Perruchet & Pacteau, 1990; Perruchet et al., 2004), such as determining exactly what aspect of the structure is being responded to – whether that be the actual structures themselves, or some other feature of the stimuli (see, e.g., Knowlton & Squires, 1996). However, by using current meta-analysis techniques, the presence of these potential confounds can actually provide valuable opportunities for teasing apart some of the multiple factors that may contribute to learning. Thus, the pattern of such confounds across studies provides a backdrop against which the contribution of specific experimental design decisions can be assessed in terms of their effect on participant learning. Critically, meta-analysis permits researchers to quantify the effects of different kinds of stimuli within a species, but also differences across species in how they may respond to different grammatical structures. In the present study, we present an analysis of a subset of AGL studies, providing a framework that more comprehensive analyses can follow.

In cross-species comparisons, a key topic of interest is to determine which grammatical structures are potentially learnable by distinct species (Fitch & Friederici, 2018; Ghirlanda et al., 2017). The prospect of such discoveries has broad repercussions for the evolution of communicative systems, and the human specificity of language structure. The stakes are thus high. As one influential example, Fitch and Hauser (2004) conducted a study that required human adults and cotton-top tamarins to distinguish between strings generated by a phrase-structure and a finite-state grammar. Only the humans were able to make this distinction when trained on strings from the phrase-structure grammar. Subsequent research, however, has revealed several confounds in this study, suggesting that the humans may have relied on other sources of information to make their responses instead of the intended structural information (e.g. de Vries et al., 2008; Perrruchet & Rey, 2005).

An ideal, perfectly-controlled methodological study would isolate a particular grammatical structure and test learning of that particular structure without influence from other properties of the stimulus. However, the complexity of language structure and the practical challenges of training and testing different species on language-like structures introduces variation into the actual tasks being conducted. Ensuring that only one particular aspect of language structure is tested, and tested in the same way across studies involving different species, remains a substantial, potentially insoluble, challenge.

In a recent small-scale review of cross-species studies of artificial grammar learning, Beckers et al. (2017) identified several characteristics that could have biased learning toward accepting the grammatical structure being tested without necessarily indicating learning of the structure. These included the extent to which the test sequence had previously occurred in the same form during exposure to the training sequences (either wholly or in part), whether the test sequence shared the same onset as the training sequences, and whether the test and training sequences were cross-correlated even if they did not contain exactly the same sequences or subsequences. Thus, in a study containing one or more of these specific properties, it would be impossible to conclusively demonstrate that the grammatical rule was acquired by the learner. Such questions have been raised for almost as long as artificial grammar learning studies have been conducted – the extent to which learning is of particular grammatical structures or instead responding to lower-level fragments in the sequences (cf. Knowlton & Squire, 1996; Perruchet & Pacteau, 1990—see Frost, Armstrong, Siegelman & Christiansen, 2015, for a review).

Artificial grammars also differ on fundamental structural properties. Some AGL studies contain dependencies between adjacent stimuli, whereas others contain dependencies between non-adjacent elements in the stimuli. Furthermore, artificial grammars may differ in terms of the number of distinct stimulus elements that sequences contain, and the number of different categories to which these stimulus elements belong. An artificial grammar with a larger versus a smaller vocabulary, or a larger versus smaller set of grammatical categories, may affect learning distinctly. Learning studies can also vary in terms of the modality of the stimuli – whether they are auditory or visual (Heimbauer et al., 2018). For example, whilst cotton-top tamarins are often trained on auditory (e.g. human non-words, monkey calls; Neiworth et al., 2017) and visual materials (e.g. structured visuospatial sequences; Locurto, Fox, & Mazzella, 2015), zebra finches only receive auditory materials consisting of manipulations of species-specific birdsong (e.g. Chen and ten Cate, 2015; van Heijningen et al., 2009). Modality is known to have distinctive effects on learning sequence structure (for reviews, see Frost et al., 2015; Milne, Wilson & Christiansen, 2018), and for these reasons modality is taken as a focus of the literature that we will analyse.

Artificial grammar learning studies also differ in terms of how training and testing is conducted. Studies of complex sequences with non-human primates and birds may require substantial training time – several thousand trials over several weeks – whereas studies with human adults are typically constrained to short training sessions with a constrained set of training trials. Testing also varies in terms of how the effects of learning are measured. For instance, in testing human adults and children there is frequently a distinction between explicit, reflection-based tasks for adult responses, such as alternative forced choice, or go/no-go responses, and implicit, processing-based tasks such as head-turn preferences or looking times. These tasks may tap into different mechanisms, with processing-based tasks more effective for assessing processing-based learning, such as acquisition of grammatical structures (Christiansen, in press; Frizelle, O'Neill, & Bishop, 2017; Isbilen et al., 2018).

As we have summarised, studies of artificial grammar learning may vary along several of these dimensions simultaneously. In this paper, we present a blueprint for how a meta-analysis approach could proceed to quantify how various design features of AGL studies might influence performance. We analyse a subset of AGL studies that have focused on presenting stimuli in either auditory or visual modalities, as reflected in the key words used within these articles. As we focus only on a subset of AGL studies, the conclusions drawn within the analysis may not generalise to the wider literature. The primary aim of our study is thus to provide a meta-analytic framework that a more comprehensive study may adopt. We show how meta-analytical methods enable us to measure the relative contributions of multiple potential confounds – reconsidered here as moderators – in influencing the size of the observed effects. This means that what was once considered a confound can actually be reinterpreted as providing a valuable and interesting source of data towards determining the limits and constraints on learning within and across species.

Method

Literature Search

We conducted the literature search and meta-analysis in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher, Liberati, Tetzlaff, & Altman, 2009), pre-registering the encoding and analysis to be conducted (https://aspredicted.org/wf2uk.pdf). The literature search was conducted on the SCOPUS database (Scopus, 2019) on articles published up to March 2019. In order to focus our literature review, we searched for studies that considered explicitly the modality of presentation in artificial grammar learning. We therefore conducted two searches of keywords appearing in titles, keywords, and abstracts of articles. In the first, we searched the keywords "artificial grammar learning" and "vision" OR "visual". In the second, we used the keywords "artificial grammar learning" and "auditory" or "audio" or "audiovisual". The results were then merged into a master list, and submitted to study selection criteria.

The search we performed avoided bias in selecting publications for analysis, in accordance with PRISMA guidelines, but it is important to note that the results of the search were not comprehensive in including all papers that conducted AGL studies with auditory or visual stimuli. The literature search for instance failed to include several influential artificial grammar learning studies (e.g., Gentner et al., 2006; Hauser & Fitch, 2004; Reber, 1967; Saffran et al., 2001, 2008). Our approach therefore outlines a blueprint for conducting meta-analyses of potential design differences in AGL research, rather than to provide a final, comprehensive answer as to the size of effects of learning in AGL studies.

Study selection

The literature search resulted in 91 records. Of these, 11 were duplicates. Of the 80 articles remaining, 8 were review articles, 3 presented computational modelling and no behavioural data, 1 study reported neuroimaging data of primates with no behavioural data, and 2 reported a case study on an aphasic population with no control group. These articles were removed, and the remaining 66 articles contained 78 studies involving 3559 subjects (this includes subjects tested more than once in the same article

– see Results section for how the analysis took into account multiple studies within articles). Figure 1 shows the PRISMA literature search flowchart. The list of studies included are reported in the Supplementary Materials.

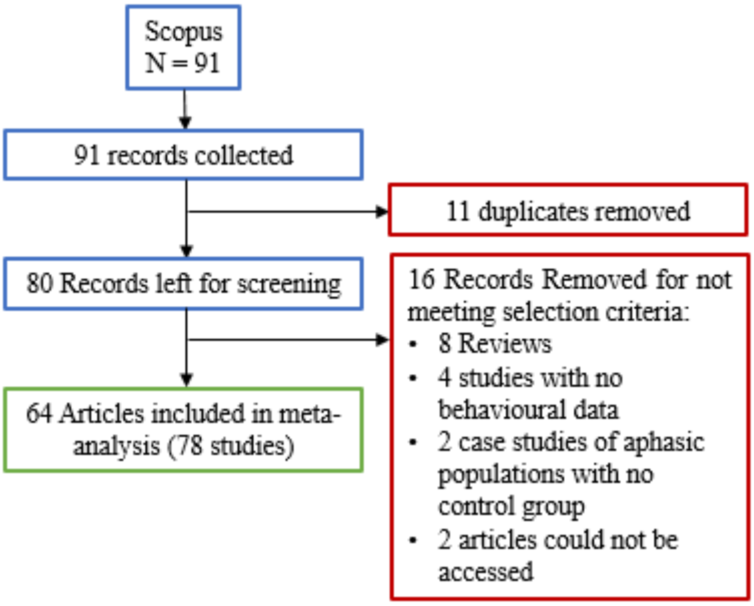
Scopus
N = 91
91 records collected
11 duplicates removed
80 Records left for screening
16 Records Removed for not meeting selection criteria:
8 Reviews
4 studies with no behavioural data
2 case studies of aphasic populations with no control group
2 articles could not be accessed
64 Articles included in meta-analysis (78 studies)

Figure 1. Flowchart of the PRISMA literature search criteria used in the current meta-analysis.

Data extraction and effect size calculation

The effect size for each study was initially computed as Cohen's d, and subsequently corrected to Hedge's g, with the variance of g computed in accordance with Borenstein et al. (2009). Formula (1) provides correction factor *J*, which is multiplied with Cohen's d to provide Hedge's g (2). The variance of Hedge's g, V_g, was provided by (3), where the variance of Cohen's *d* is computed, and corrected by *J*.

$$(1)\ J = \left(1 - \frac{3}{4df - 1}\right)$$

$$(2)\ g = J \times d$$

$$(3)\ V_g = \left(\frac{1}{n} + \frac{d^2}{2 \times n}\right) \times J^2$$

Cohen's d was derived for each type of dependent variable, the dependent variable for each study is shown in the Supplementary Materials. For studies reporting the number correct, numbers endorsed or responded to, or go/no-go responses as dependent variable, the effect size was computed from the difference to chance responding in a one sample test (see Equation 4):

$$(4)\ d = \frac{Mean - Chance}{SD_{Within}}$$

In cases where tests and language structures were similar over different test sessions or conditions (e.g. Cope et al., 2017; Goranskaya et al., 2016; Mueller et al., 2010), we combined the means and SDs from each of the multiple test sessions, and computed the one sample difference from chance. The pooled mean was simply computed as the arithmetic mean across the sessions, weighted by number of participants in the session. For pooled SD, we took the average SD using equation (5), where n_1 is the number of items in test session 1, n_2 is the number of items in test session 2, etc., and SD_1 is the observed standard deviation of the test session 1 response accuracy, etc. (see van Witteloostuijn, Boersma, Wijnen, & Rispens, 2017):

$$(5)\ SD_{Average} = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2 + (n_3 - 1)SD_3^2 + (n_4 - 1)SD_4^2}{n_1 + n_2 + n_3 + n_4 - 4}}$$

Subsequently, we computed *d* using equation (4), with the pooled mean, 50% as chance, divided by the SD $_{\text{Average}}$. In serial reaction time studies, the effect was measured as the standardised mean difference in RT between presentations of a trained vs. an untrained structure, with SD$_{\text{Average}}$ computed as in (5), which assumes conservatively that there is a correlation of 1 between the trained and untrained structure responses across participants (a lower correlation would result in a lower SD, so this formula provides a conservative upper limit for the effect size). For instance, for Kemeny and Nemeth's (2017) data represented in Figure 3, presenting the mean response time (RT) and SEM per testing block. In this case, we pooled the mean RT for the grammatical blocks 4 and 6 weighted by the number of participants in the session, and computed *d* as the difference to the mean RT for the ungrammatical block 5, with SD computed as the SD $_{\text{Average}}$ across blocks 4, 5, and 6, using (5).

For sequence reproduction tasks, the effect size was computed as difference in mean accuracy for grammatical sequences and ungrammatical sequences, with *SD* as the SD $_{\text{Average}}$ computed using (5).

In head-turn preference paradigms (e.g. Gomez & Gerken, 1999), effect size was the proportion of trials where the participant turned towards the grammatical violation sequences over the grammatical sequences, indicating observation of the violation. These values were compared to chance and *d* computed in the same way as for response accuracy measures.

For looking time paradigms (e.g. Milne et al., 2018), the effect size was computed as the difference in fixation duration between grammatical and ungrammatical sequences, computed using the same approach as that for sequence reproduction paradigms. Positive effects were generally computed as longer looking to ungrammatical than grammatical sequences (a novelty effect). However, in cases where the interpretation of the authors suggested that longer looking times to grammatical stimuli (or preferences in head-turn to grammatical sequences) reflected greater learning (i.e., a familiarity effect), we re-signed these effects.

In studies where means and variance were reported only in figures, we contacted authors for data, and utilized the Digitizeit digitizer software (available from: http://www.digitizeit.de/) when such data was not available, to extract the means and SDs. In cases where graphs displayed the mean and 95% confidence intervals (Hall et al., 2018), confidence intervals were converted into SDs according to (6), which assumes that the authors had computed the confidence intervals using the t-distribution (which is more conservative than assuming confidence intervals based on the Z-distribution), where tcrit is the critical value of the *t*-distribution for n-1 degrees of freedom at p = .05:

$$(6)\ SD = \sqrt{n} \times \frac{upperlimit - lowerlimit}{2 \times tcrit[n-1]}$$

Each study was encoded for several features in order to test their influence on learning performance. We encoded the animal class and species that was tested, and in the case of human studies, distinguished whether the study was on children (<18 years) or adults.

For properties of the AGL structure, we encoded whether the study contained at least some repetitions of the stimuli experienced during training in the testing, whether the artificial grammar contained adjacent dependencies or did not contain adjacent dependencies, and whether the artificial grammar contained non-adjacent dependencies or did not contain non-adjacent dependencies.

For characteristics of training and testing, we encoded the type of test response that was being collected – whether this was a Yes versus No judgment, a go or no-go task, a scale judgment, a forced choice test between two or more alternatives, serial reaction time, head-turn preference, looking time, sequence production, or frequency estimation task. We subsequently grouped these variables into whether they required reflection on the grammatical structure (reflection-based; forced choice tests, yes versus no judgement, go/no-go, scale judgement), or more directly tapped into the underlying processing of the grammatical structure (processing-based; looking time, head-turn preference, serial reaction time, sequence production) (Christiansen, in press). We encoded the amount of exposure to the artificial grammar that participants experienced in terms of the total number of stimulus tokens from the grammar during exposure (training length).

Importantly, we also encoded a number of surface features of the AGL, including whether the stimuli were visual, auditory, or a combination of both visual and auditory, in order to determine whether learning varied according to the modality of the task. Further, we also encoded the size of the artificial grammar in terms of the size of the vocabulary in the grammar (or the number of distinct items), as well as the number of different categories in the grammar (e.g., for a phrase-structure grammar with four nouns, two verbs, two adjectives, and two determiners, the number of categories is 4 (noun/verb/adjective/determiner) and the size of the vocabulary is 14.

Results

Evidence of acquisition of structure from AGL studies

The overall effect size across the studies, and the extent to which each of the encoded study variables predicted differences in effect sizes across the studies, was determined by conducting a random effects meta-analysis of effect sizes, using the R package metafor (Viechtbauer, 2010). This approach takes into account inconsistencies between the studies analysed, provides an estimate of sampling error, and also permits a measurement of the effects of each of the variables in moderating the size of the overall behavioural effect (Borenstein, Hedges, Higgins, & Rothstein, 2010; Borenstein, Higgins, & Rothstein, 2009). We encoded each experiment in an article and each test in an experiment as a separate study, and as these cannot be assumed to result in effect sizes independent from one another, we encoded article as a nested multilevel variable in the analysis (Konstantopoulos, 2011).

The model was run using the rma.mv function with the restricted maximum likelihood (REML) method. We utilised the *t* method to generate test statistics and confidence intervals. The model was run using the rma.mv function with restricted likelihood (REML) method, and the t-adjustment to calculate the model estimates of standard errors, p values and confidence intervals. Effect sizes for individual studies and the overall average weighted effect sizes are presented in Figure 2. A positive effect size indicates greater preference for stimuli conforming to the AGL structure, while a negative effect size indicates preference for non-conforming stimuli (except in the case of the looking studies, where a positive effect indicates longer looking to violating stimuli – as this was the predicted effect of such studies in reflecting AGL acquisition, e.g., Gomez & Gerken, 1999).

The meta-analysis resulted in the average weighted effect size = 1.069, SE = .130, 95% CI [.813, 1.326], p < .0001, indicating that overall there was strong evidence of learning in AGL studies.

3.2 Publication bias

To determine whether there was publication bias in the sample, we conducted a Peters’ test (Peters et al., 2006) on the random multilevel meta-regression model. The Peters’ test revealed a significant asymmetrical distribution, $t(154) = -2.290$, p = .023, indicating the presence of publication bias in our sample. The funnel plot (Figure 2) displays the standard error (a measure of study precision) against the effect sizes of the individual studies. In the absence of publication bias, studies should be symmetrically distributed around the average weighted effect size in a funnel shape, with high precision studies being closer to the average weighted effect size, and lower precision studies symmetrically distributed around the average weighted effect size. The distribution indicates that there are more large positive effect sizes for smaller sample sizes than would be expected from a standard distribution of studies, suggesting a potential publication bias. The size of the effect of AGL acquisition, and the sources of heterogeneity of the effects, should thus be considered in light of possible bias in the studies published.

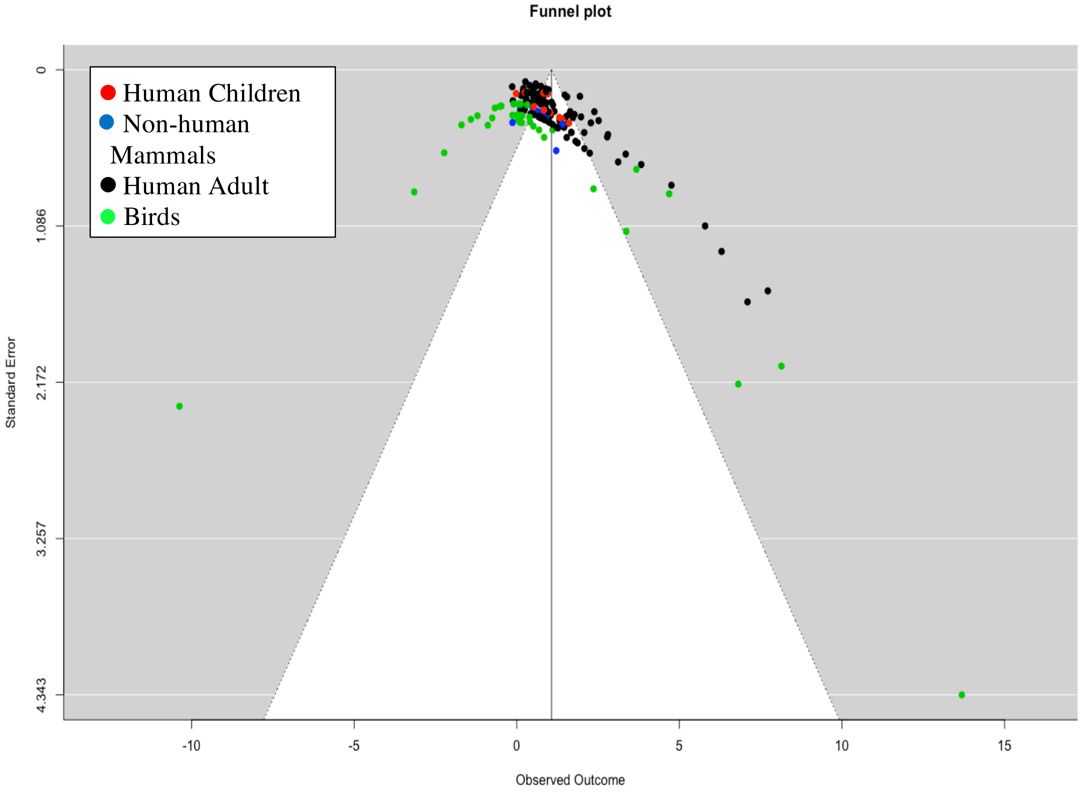

Figure 2. Funnel plot showing the relationship between the standard error and the effect size of the individual studies. Points are colour-coded according to animal class. Black points illustrate Human Adult Studies, blue illustrate Non-human mammals studies, red are Human Child studies, and green are Bird studies.

3.3 Heterogeneity in effect size variance associated with study variables

Cohran's Q-test for heterogeneity was significant ($Q(155) = 1185.657$, $p < .0001$), indicating that variance in the data cannot be explained by random measurement error, but that different aspects of studies are contributing to the effect size. We thus analysed the effects of each of the set of variables we encoded from each of the studies as moderators, shown in Table 1.

For the effect of animal class (but also distinguishing human adults and human children from non-human mammals), there were significant differences on the size of effect of learning between different species. For human adults, the overall effect size was 1.252 (SE = .148, 95% CI [0.958, 1.545], *p* < .0001). For human children, the overall effect size was 0.615 (SE = .231, 95% CI [.101, 1.129], *p* = .0237). For non-human mammals, the overall effect size was 0.626 (SE = .172, 95% CI [.221, .1.032], *p* = .008). For birds, the overall effect size was 0.428 (SE = 0.533) (95% CI [-0.653, 1.509], *p* = .427).

Properties of training and testing of AGL studies were found to produce significant differences in effect sizes. Log-transformed number of training trials related negatively to effect size, -0.188 (0.054) (95% CI [-0.295, -0.0815], *p* = .0006). Further, repetition of trained items at test resulted in larger effects 1.051 (SE = 0.279, 95% CI [0.499, 1.602], *p* = .0002).

Surface level features of the language did not significantly moderate the variance of effect sizes (see Table 1), and this included also the modality of stimulus delivery. The number of categories, the vocabulary size, and critically, whether the stimuli were visual or auditory were not found to affect the overall effect size.

For the structural properties of the language, there were moderating effects. The presence of repetition of items from training to test positively influenced effect sizes, with an overall effect of 1.051 (SE = 0.279) (95% CI [0.499, 1.602], *p* = .0002).

As there were different sized effects of learning for each animal class, and possible confounds between study design characteristics and animal class tested, we conducted further analyses of moderator variables for human adult, human child, birds, and non-human mammals separately.

Table 1. Contributions of each moderating variable to account for variance in effect sizes across studies.

Moderator		*F*	*Df1, Df2*	*p*
Population				
	Animal Species	2.613	(10, 145)	< .0001***
	Animal Class	5.811	(3, 152)	.0009***
	Human vs. Non-human	7.555	(2, 153)	.0007***
Training and testing				
	Log Training Length	12.149	(1, 154)	< .0001***
	Stimulus Modality	0.095	(2, 153)	.909
	Test Response	1.624	(10, 145)	.105
	Test Type	3.698	(1, 154)	.056
Surface level properties				
	Categories in Language	0.0001	(1, 154)	.992
	Number of unique vocabulary items	3.021	(1, 154)	.084
Structural Properties				
	Repetition of items	14.162	(1, 154)	.0002**
	Adjacent dependencies	0.238	(1, 154)	.627
	Non-adjacent dependencies	0.118	(1, 154)	.608

Note. *F* is the statistic for testing whether the moderator accounts for some heterogeneity between studies; p is the significance for the *F*-test *** p < .001, ** p < .01, **p* < .05. Note that Animal Class distinguishes birds, non-human mammals, human adult, and human child. Animal species also distinguishes human adult and human child.

3.4 Moderator Analysis of Human Adults

There was significant heterogeneity of variance in the effect size in studies testing human adults (*Q*(99) = 707.273, *p* < .001), so we analysed the effect of each moderator (see Table 2 for the significance of each moderator). There was a significant effect of the presence of non-adjacent dependencies (effect = 0.582, SE = 0.259, 95%

CI [0.068, 1.096], *p* = .027), suggesting that adult human participants are overall successful in learning non-adjacencies in artificial grammars.

Table 2. Contributions of each moderating variable to account for variance in effect sizes in Human Adult studies.

Moderator		*F*	*Df1, Df2*	*p*
Training and testing				
	Log Training Length	0.415	(1, 98)	.521
	Stimulus Modality	0.306	(2, 97)	.737
	Test Response	0.671	(8, 91)	.716
	Test Type	1.884	(1, 98)	.173
Surface level properties				
	Categories in Language	0.319	(1, 98)	.574
	Number of unique vocabulary items	1.023	(1, 98)	.305
Structural properties				
	Repetition of items	0.036	(1, 98)	.851
	Adjacent dependencies	1.745	(1, 98)	.190
	Non-adjacent dependencies	5.050	(1, 98)	.027*

3.5 Moderator Analysis of Human Children

There was significant heterogeneity (*Q*(10) = 49.953, *p* < .0001), so we further analysed the effect of each moderator (see Table 3). In this analysis, the only significant moderator was the test response participants made. This analysis indicated that head-turn preference paradigms produced an overall effect of 1.301 (SE = 0.1663, 95% CI [0.772, 1.831], *p* = .004). Sequence production paradigms, by comparison, produced an effect that failed to statistically differ from 0 (effect size = 0.150, SE = 0.144, 95% CI

[-0.433, 0.721], *p* = .395). Finally, binary yes-no judgement tasks produced an overall effect of 0.822 (SE = 0.099. 95% CI [0.506, 1.137], p = .004).

Table 3. Contributions of each moderating variable to account for variance in effect sizes in human child studies.

Moderator		*F*	*Df1, Df2*	*p*
Training and Testing				
	Log Training Length	0.214	(1, 9)	.654
	Stimulus Modality	3.427	(1, 9)	.097
	Test Response	15.978	(2, 8)	.002*
	Test Type	0.271	(1, 9)	.615
Surface level properties				
	Categories in Language	0.059	(1, 9)	.813
	Number of unique vocabulary items	0.862	(1, 9)	.377
Structural properties				
	Repetition of items	2.503	(1, 9)	.148
	Adjacent dependencies	0.023	(1, 9)	.884
	Non-adjacent dependencies	0.012	(1, 9)	.917

3.6 Moderator Analysis of Non-human Mammals

There was significant heterogeneity (*Q*(7) = 15.928, *p* < .026), therefore we analysed the effect of each moderator (see Table 4). Non-human mammals only took part in studies delivered in the auditory modality, and all of which were processing based, included adjacent dependencies, and did not include repetitions at test, and hence we did not include a moderator analysis of testing modality, repetition of items, adjacency, and testing type. No moderator accounted for a significant proportion of variance in this dataset.

Table 4. Contributions of each moderating variable to account for variance in effect sizes in non-human mammals studies.

Moderator		*F*	*Df1, Df2*	*p*
Training and testing				
	Log Training Length	1.121	(1, 6)	.331
	Test Response	1.262	(1, 6)	.304
Surface level properties				
	Categories in Language	0.760	(1, 6)	.418
	Number of unique vocabulary items	0.365	(1, 6)	.567
Structural properties				
	Non-adjacent dependencies	0.111	(1, 6)	.750

3.7 Moderator Analysis of Birds Studies

There was again significant heterogeneity ($Q(36) = 259.498$, $p < .0001$), therefore we analysed the effect of each moderator (see Table 5). Birds, however only took part in classification-based tasks, and thus, we did not analyse the effect of test type. Log training length accounted for a significant portion of the variance, increased training resulted in a lower effect size -0.739 (SE = .268, 95% CI [-1.283, -0.195], p = .009). Increased vocabulary sizes tended to increase effect sizes (effect size = 0.099, SE = 0.038, 95% CI [0.022, 0.177], p = .014). Stimulus modality explained a significant portion of variance, with visual stimuli producing larger effects (effect size = 1.993, SE = 0.788, 95% CI [0.395, 3.592], p = .016) than auditory stimuli. The response task used also accounted for a significant portion of variance of effect sizes, however, the meta-analytic estimate for both 2AFC tasks (effect size = 2.288, SE = .135, 95% CI [-0.488, 5.065], p = .090) and go/no-go tasks (effect size = -0.042, SE = 0.294, 95% CI [-0.642, 0.559], p = .889) failed to significantly differ from 0. This reflects the fact that variance of effect sizes in birds was large; to properly account for the moderating effect of task type on the variance in effect size for bird studies, a larger set of studies for inclusion would be helpful. Finally, the repetition of items accounted for a significant portion of the variance of effect sizes, whereby repeating items at test resulting in an effect size of 5.013 (SE = 0.740, 95% CI [3.511, 6.515], *p* < .0001). This effect is explained by the only study including repetitions of whole strings at test (Spierings & ten Cate, 2016) produced large effect sizes.

Table 5. Contributions of each moderating variable to account for variance in effect sizes in birds studies.

Moderator		*F*	*Df1, Df2*	*p*
Training and testing				
	Log Training Length	7.609	(1, 35)	.009**
	Stimulus Modality	6.407	(1, 35)	.016*
	Test Response	6.407	(1, 35)	.016*
Surface level properties				
	Categories in Language	0.053	(1, 35)	.819
	Number of unique vocabulary items	6.712	(1, 35)	.014*
Structural properties				
	Repetition of items	45.926	(1, 35)	< .0001***
	Adjacent dependencies	2.462	(1, 35)	.126
	Non-adjacent dependencies	1.661	(1, 35)	.206

Discussion

We presented a focused literature search analysing AGL studies that address the modality of stimulus presentation, taking into account the varieties of designs, as well as species, that are tested across these studies. This approach provides a blueprint for how meta-analysis in AGL studies can assess the influence of multiple moderators on learning, providing insight into the conditions under which learning of regularities in artificial grammars can be observed. Confounds and differences between studies – both intended and unintended (and previously viewed as adding opacity to the field of research) – can be considered sources of information for disentangling multiple contributors to learning of artificial grammar stimuli, rather than serve only as an impediment to comparison between studies. Heterogeneity of design can actually be analysed through an estimate of heterogeneity of variance which can then be associated with the presence or absence of differences across studies.

The current analysis was conducted to provide a framework for how future, more comprehensive meta-analyses might robustly identify patterns in the artificial grammar learning literature. However, our literature search was constrained by a restricted set of keywords that selected only papers where AGL and modality of presentation were explicitly tagged as features of the study. We know that influential studies in the literature were omitted by our approach. Whereas our focus here was to avoid bias in selecting the papers for inclusion in our analysis by conducting an objective keyword search, this absence of key studies highlights that there are relevant papers that are not included in the current analysis, and so the comprehensiveness of our search cannot be assumed. Consequently, the precise results of the meta-analysis and the moderator analysis should not be taken as the final word on this topic. Instead, we have shown how a future analysis, on an even more comprehensive set of studies, may help move the field forward. Such a study will be a considerable undertaking; a Scopus search with the keywords “artificial grammar learning” or “statistical learning”, for instance, resulted in 6,511 records and still failed to include the landmark studies by Fitch and Hauser (2004), Gentner et al. (2006), and Reber (1967), mentioned in the Introduction, though the search did succeed in including the key studies by Saffran (2001) and Saffran et al. (2008). Finding principled ways to limit the literature search, without omitting key articles, presents an additional interesting challenge in this field of research.

This shortcoming raises concerns about terminological specificity in the field of artificial grammar learning. If we take Fitch and Hauser’s (2004) study, this paper explicitly implements an AGL method, however, it instead describes it as a “familiarization/discrimination paradigm” in its abstract. Gentner and colleagues (2006) do not describe their method in the abstract, and in text specify it as a go/no-go operant conditioning procedure of AB^n and A^nB^n grammars. Similarly, Saffran’s (2001) and Saffran et al.’s (2008) methods are variously described as statistical learning, grammatical pattern learning, or familiarization-discrimination.

Cumming (2014) provided a compelling argument for favouring magnitude estimation over null hypothesis significance testing in assessing experimental effects. A tenet of this approach is to employ meta-analytic thinking throughout the research process, including writing, reporting, and publication. The diversity of terms utilised to describe related methods makes it difficult to devise a singular, constrained set of search terms that would gather them together in a given search. Moving forward, we would suggest that using informative, umbrella keywords will ameliorate this issue, facilitating meta-analyses, and in Cumming’s (2014) view, support research integrity.

In terms of the results of our focused meta-analysis in terms of what can be learned across animal classes, the analyses showed that the size of learning effects varies according to the species tested, though the evidence of publication bias and the potential lack of comprehensiveness in the search mean that interpretations based on size of effects must be treated with caution. The overall largest effect was observed for studies involving adult humans, but there were also overall significant effects of learning associated with child humans, non-human mammals, though not for birds. However, there are many differences between studies designed to appraise learning in different species, and heterogeneity of the variance within studies addressing each species points to ways in which these design differences may have profound effects on learning. The analyses of moderator effects within each animal class demonstrated that multiple variables were affecting learning, highlighting potential distinctions across species.

The size of the observed effects for human children was affected by the test response required, with similar effect sizes for head-turn preference and Yes/No judgement tasks. Whilst sequence production tasks did not significantly differ from 0, this likely reflects the small number of child studies included in the present analysis. For birds, the presence of training items at test produced large effects, perhaps unsurprising given the large amount of training they receive. Intriguingly, a greater number of training trials related negatively to effect size. This is likely correlated with the specific species of bird tested, and thus represents an important variable to focus on in a comprehensive meta-analysis. For adult humans, larger effects were produced by grammars containing non-adjacent dependencies than sequences without those dependencies, which have traditionally been difficult to observe in individual studies (e.g., Frost & Monaghan, 2016; Lai & Poletiek, 2011; Perruchet et al., 2004), see

Wilson et al. (in press) in this issue for further discussion. The absence of a significant effect of adjacent dependencies was unexpected, but highlights the variation that can occur in the effect sizes across studies testing these structures.

Further meta-analytical techniques can help determine the additional sources of information that might support such learning, such as use of reflection- versus processing-based test measures (Vuong et al., 2016). In order to measure the effect of learning on processing, rather than explicit decision-making based on the structures experienced by the learner, a task that probes processing is proposed to be more effective (Christiansen, in press; Frizelle et al., 2017; Isbilen et al., 2018), however, in the present analysis there was no statistically reliable difference between the two. This may be a consequence of the comparatively large number of reflection-based effects (135) relative to processing-based effects (21) included in this analysis, or of the range of grammars that tend to be tested in AGL studies, a large number of studies use Reber-style (1967) grammars, where explicit testing may produce a similar magnitude of effects. Moreover, the effect of reflection-based measures may also have been inflated by including the non-human animal data as they are unlikely to engage in the kind of conscious reflections often observed in human studies. Finally, the presence of a potential publication bias combined with the much longer use of reflection-based assessments in AGL studies going more than half a century may further explain this pattern.

A key issue that emerged during our analysis was that individual stimuli within a test may contain alternative structures or vary in the presence of surface features. The analyses in this paper report effect sizes and features of the stimuli across sets of stimuli, which can obscure the individual influence of these features. Making raw data sets publicly available would enable this by-items analysis to reveal the precise contribution of multiple variables to learning behaviour (e.g., Beckers et al., 2017).

The studies included here were selected from an objective literature search on SCOPUS, intending to avoid bias in our selection of tests, focusing on studies of AGL that describe the modality of the stimuli. Interestingly, except in the case of birds, modality was not found to affect the results, but this may also have been affected by observed publication bias. Expanding further to a literature search of an even broader literature would help to determine more clearly which moderators are affecting performance, and which are orthogonal to artificial grammatical learning. There are, for instance, other structures that are of key interest to both language acquisition research, and cross-species investigations of the limits of grammar learning – such as distinctions between phrase structure and finite-state grammars (Fitch & Friederici, 2012; Fitch & Hauser, 2004), or focused on hierarchical centre-embedded structures (Lai & Poletiek, 2011). Debates on the learnability of these structures (e.g., de Vries et al., 2008) will be facilitated by a wider survey of the published literature. In our blueprint for a meta-analysis approach in this field, we have made an illustrative first step toward providing a perspective on what is learned and what is learnable within and across species.

Acknowledgements

AT, PM, and MHC were supported by the International Centre for Language and Communicative Development (LuCiD) at Lancaster University, funded by the Economic and Social Research Council (UK) [ES/L008955/1]. GJLB is part of the Consortium on Individual Development (CID), which is funded through the Gravitation Program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO; grant number 024.001.003).

References

Abe, K., & Watanabe, D. (2011). Songbirds possess the spontaneous ability to discriminate syntactic rules. *Nature Neuroscience, 14(8),* 1067-1076.

Bahlmann, J., Schubotz, R. I., & Friederici, A. D. (2008). Hierarchical artificial grammar processing engages Broca's area. *Neuroimage*, *42*(2), 525-534.

Beckers, J. L., Berwick, B. C., Okanoya, K., & Bolhuis, J. J. (2012). Birdsong neurolinguistics: Songbird context-free grammar claim is premature. *Neuroreport, 23(3),* 139-145.

Beckers, J. L., Berwick, R. C., Okanoya, K., & Bolhuis, J. J. (2017). What do animals learn in artificial grammar studies? *Neuroscience and Biobehavioral Reviews*, *81(Part B)*, 238-246.

Bormann, I. (2012). Digitizeit [computer software]. Retrieved from https://www.digitizeit.de/

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). Effect sizes based on means. In M. Borenstein, L. V. Hedges, J. P. Higgins, and H. R. Rothstein (Eds.) *Introduction to Meta-Analysis* (21 – 32). doi: 10.1002/9780470743386.ch4.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*(2), 97-111.

Chen, J., & ten Cate, C. (2015). Zebra finches can use positional and transitional cues to distinguish vocal element strings. *Behavioural Processes*, *117*, 29 – 34.

Chen, J., van Rossum, D., & ten Cate, C. (2015). Artificial grammar learning in zebra finches and human adults: XYX versus XXY. *Animal Cognition, 18(1),* 151-164.

Christiansen, M.H. (in press). Implicit-statistical learning: A tale of two literatures. *Topics in Cognitive Science.* https://doi.org/10.1111/tops.12332

Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition, 114(3),* 356-371.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1). 7 – 29.

de Vries, M. H., Monaghan, P., Knecht, S. & Zwitserlood, P. (2008). Syntactic structure and artificial grammar learning: the learnability of embedded hierarchical structures. *Cognition, 107*, 763-774.

Endress, A. D., & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition, 105*(2), 247–299.

Endress, A. D., Carden, S., Versace, E. & Hauser, M. D. (2010). The apes' edge: positional learning in chimpanzees and humans. *Animal Cognition*, *13(3)*, 483-495.

Fitch, W. T., & Friederici, A. D. (2012). Artificial grammar learning meets formal language theory: an overview. *Philosophical Transactions of the Royal Society B*, *367*(1598), 1933-1955.

Fitch, W. T., & Hauser, M. D. (2004). Computational constraints on syntactic processing in a nonhuman primate. *Science*, *303*, 377-380.

Frizelle, P., O'Neill, C., & Bishop, D. V. (2017). Assessing understanding of relative clauses: A comparison of multiple-choice comprehension versus sentence repetition. *Journal of Child Language*, *44*(6), 1435-1457.

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: the paradox of statistical learning. *Trends in Cognitive Sciences, 19*(3), 117-125.

Gentner T. Q., Fenn K. M., Margoliash D., & Nusbaum, H. C. (2006). Recursive syntactic pattern learning by songbirds. *Nature,* 440: 1204–1207.

Ghirlanda, S., Lind, J., Enquist, M. (2017) Memory for stimulus sequences: a divide between humans and other animals? *Royal Society Open Science, 4:161011*, http://doi.org/10.1098/rsos.161011

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70(2)*, 109-135.

Heimbauer, L. A., Conway, C. M., Christiansen, M. H., Beran, M. J., Owren, M. J. (2018). Visual artificial grammar learning by rhesus macaques (Macaca mulatta): exploring the role of grammar complexity and sequence length. *Animal Cognition, 21(2),* 267-284.

Isbilen, E S., Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (2018), Bridging artificial and natural language learning: Comparing processing- and reflection-based measures of learning. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1856-1861). Austin, TX: Cognitive Science Society.

Jamieson, R. K., & Mewhort, D. J. K. (2005). The influence of grammatical, local, and organizational redundancy on implicit learning: An analysis using information theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31(1)*, 9-23.

Knapp, G. & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine, 22(17)*, 2693-2710.

Knowlton, B. J., & Squire, L. R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22(1)*, 169-181.

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods, 2(1),* 61–76.

Lai, J., & Poletiek, F. H. (2011). The impact of adjacent-dependencies and staged-input on the learnability of center-embedded hierarchical structures. *Cognition*, *118*(2), 265-273.

Locurto, C., Fox, M., & Mazzella, A. (2015). Implicit learning in cotton-top tamarins (*Saguinus Oedipus*) and pigeons (*Columba livia*). *Learning & Behavior, 43(2),* 129-142.

Miller, G.A. (1958). Free recall of redundant strings of letters. *Journal of Experimental Psychology. 56,* 485–491.

Milne, A. E., Wilson, B. & Christiansen, M. H. (2018). Structured sequence learning across sensory modalities in humans and nonhuman primates. *Current Opinion in Behavioural Sciences, 21,* 39-48.

Moher, D., Liberati, A., Tetzlaff, & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine, 151*, 264–269.

Neiworth, J. J., London, J. M., Flynn, M. J., Rupert, D. D., Alldritt, O., & Hyde, C. (2017). Artificial grammar learning in Tamarins (*Saguinus Oedipus*) in varying stimulus contexts. *Journal of Comparative Psychology, 131(2),* 128-138.

Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of experimental psychology: General*, *119*(3), 264-275.

Perruchet, P. & Rey, A. (2005) Does the mastery of center-embedded linguistic structures distinguish humans from nonhuman primates? *Psychonomic Bulletin and Review, 12*, 307–313.

Perruchet, P., Tyler, M. D., Galland, N., & Peereman, R. (2004). Learning nonadjacent dependencies: no need for algebraic-like computations. *Journal of Experimental Psychology: General*, *133*(4), 573-583.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. *JAMA*, *295*(6), 676-680.

Reber, A.S. (1967). Implicit learning of artificial grammars. *Verbal Learning and Verbal Behavior, 5*, 855–863.

Saffran, J. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language, 44*(4), 493–515.

Saffran, J., Hauser, M., Seibel, R., Kapfhamer, J., Tsao, F., & Cushman, F. (2008). Grammatical pattern learning by human infants and cotton-top tamarind monkeys. *Cognition, 107(2),* 479-500.

Schwarzer, G. (2012). *Package 'meta' for R software: General package for meta-analysis.* https://cran.r-project.org/web/packages/meta/meta.pdf.

Scopus (2019). https://www.scopus.com Accessed 18 March 2019.

Spierings, M., de Weger, A., & ten Cate, C. (2015). Pauses enhance chunk recognition in song element strings by zebra finches. *Animal Cognition, 18(4),* 867-874.

Spierings, M. J., Hubert, J., & ten Cate, C. (2017). Selective auditory grouping by zebra finches: testing the iambic-trochaic law. *Animal Cognition, 20(4),* 665-675.

van Heijningen, C. A. A., Chen, J., van Laatum, I., van der Hulst, B. (2013). Rule learning by zebra finches in an artificial grammar learning task: which rule? *Animal Cognition*, *16(2)*, 165-175.

van Heijningen, C. A. A., de Visser, J., Zuidema, W., & ten Cate, C. (2009). Simple rules can explain discrimination of putative recursive syntactic structures by a songbird species. *Proceedings of the National Academy of Sciences of the Unites States of America*, *106(48)*, 20538-20543.

van Witteloostuijn, M., Boersma, P., Wijnen, F., & Rispens, J. (2017). Visual artificial grammar learning in dyslexia: A meta-analysis. *Research in Developmental Disabilities, 70,* 126-137.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36(3),* 1-48.

Vuong, L.C., Meyer, A.S. & Christiansen, M.H. (2016). Concurrent learning of adjacent and nonadjacent dependencies. *Language Learning, 66,* 8-30.

Wilson, B., Smith, K., & Petkov, C. I. (2015). Mixed-complexity artificial grammar learning in humans and macque monkeys: evaluating learning strategies. *European Journal of Neuroscience, 41(5),* 568-578.

Wilson, B., Spierings, M, Ravignan, A., Mueller, J.L., Mintz, T.H., Wijnen, F., van der Kant, A.., Smith, K., & Rey, A. (in press). Non-adjacent dependency learning in humans and other animals. *Topics in Cognitive Science, in press*.

Zaccarella, E., & Friederici, A. D. (2017). The neurobiological nature of syntactic hierarchies. *Neuroscience and Behavioral Reviews, 81,* 205-212.