Neuroscience and Biobehavioral Reviews 81 (2017) 238–246

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

Review article

What do animals learn in artificial grammar studies?

Gabriël J.L. Beckers[a,*], Robert C. Berwick[b,c], Kazuo Okanoya[d], Johan J. Bolhuis[e,f]

[a] Cognitive Neurobiology and Helmholtz Institute, Department of Psychology, Utrecht University, Utrecht, The Netherlands
[b] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA
[c] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA
[d] Department of Cognitive and Behavioral Sciences, The University of Tokyo, Tokyo, Japan
[e] Cognitive Neurobiology and Helmholtz Institute, Departments of Psychology and Biology, Utrecht University, Utrecht, The Netherlands
[f] Department of Zoology and St. Catharine's College, University of Cambridge, Cambridge, UK

ARTICLE INFO

Article history:
Received 1 August 2016
Received in revised form
29 November 2016
Accepted 16 December 2016
Available online 22 December 2016

Keywords:
Animal cognition
Artificial grammar learning
Auditory memory
Biolinguistics
Bird
Rule learning
Primate
Syntax

ABSTRACT

Artificial grammar learning is a popular paradigm to study syntactic ability in nonhuman animals. Subjects are first trained to recognize strings of tokens that are sequenced according to grammatical rules. Next, to test if recognition depends on grammaticality, subjects are presented with grammar-consistent and grammar-violating test strings, which they should discriminate between. However, simpler cues may underlie discrimination if they are available. Here, we review stimulus design in a sample of studies that use particular sounds as tokens, and that claim or suggest their results demonstrate a form of sequence rule learning. To assess the extent of acoustic similarity between training and test strings, we use four simple measures corresponding to cues that are likely salient. All stimulus sets contain biases in similarity measures such that grammatical test stimuli resemble training stimuli acoustically more than do non-grammatical test stimuli. These biases may contribute to response behaviour, reducing the strength of grammatical explanations. We conclude that acoustic confounds are a blind spot in artificial grammar learning studies in nonhuman animals.

Contents

* Corresponding author at: Cognitive Neurobiology and Helmholtz Institute, Department of Psychology, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands.
E-mail address: g.j.l.beckers@uu.nl (G.J.L. Beckers).

http://dx.doi.org/10.1016/j.neubiorev.2016.12.021

1. Introduction

A key goal of the cognitive (neuro)sciences is to develop an account of the human language capacity, presumed to be an internal computational system. Since the latter part of the 20th century, a traditional approach to this problem, following Chomsky (1975) and much other work, is to characterize this capacity *via* generative grammars. Since grammars are part of neural computational systems, their activity is typically not directly observable, for example, in the sentences or language that comprise external behavior. The investigation of human, or "natural" generative grammars has thus proceeded by drawing on many kinds of experimental methods and data to indirectly infer the properties of human grammars–linguistic examples, sentence processing, language acquisition, brain imaging, and the like. An additional barrier is the lack of non-human model organisms, which impedes comparative work, since so far as it is known only humans possess full-fledged generative grammars (Berwick et al., 2013).

Artificial grammar learning (AGL) is one methodology that has been advanced in an attempt to overcome hurdles like these. Roughly, the idea is that one can construct deliberately simplified, hence artificial grammars (AGs) that focus on just a few syntactic properties, and then calculate what these simplified systems might yield in the way of observable external forms, what are sometimes called the grammar's language, the set of strings defined ('generated') by that grammar, as described in the following section. Note that for an artificial generative grammar, the grammar itself is internal to the computational system, while some of the representations the grammar generates are "externalized," like the sequence of sounds in speech. Since by design the experimenter knows both the internal form (the AG) and the observable, external forms the AG yields, the AG can be used in experimental paradigms where either humans or other animals can be tested to see whether the particular properties highlighted by the AG can be acquired, and so represented and used. This remains one of the few direct ways to ascertain whether nonhuman animals possess grammatical abilities, and if so, which level of complexity they can master.

2. Artificial grammar learning

2.1. What is an artificial grammar?

An artificial grammar is a particular subset of the full class of generative grammars. For our analysis in the remainder of this article, it is useful to define such grammars more carefully. In general, a generative grammar consists of a finite set of rules along with some computational (recursive) procedure to generate or *derive* possible sentences. Here for illustration we focus on one narrow type of generative grammar used most often in AGL studies, so-called *context-free* grammars. These consist of *production* or *rewrite* rules built out of a finite set of *nonterminal* and *terminal* tokens or symbols. Terminal tokens are analogous to externally observable words or sounds in a human language, like *the* or *apple*; nonterminals correspond to phrases like *Noun Phrase* or *Prepositional Phrase.* Each rewrite rule consists of two related parts, a lefthand side and a righthand side, where a specified symbol(s) on the lefthand side is to be replaced with the symbols on the righthand side. Additionally, given a set of rules, there is a set procedure to generate or *derive* possible sequences of terminal symbols, beginning with a designated nonterminal starting symbol, and successively replacing left-hand side symbols in rules with their right-hand sides until no more rules can apply. To illustrate, consider the following simplified grammar designed to reflect certain aspects of the syntactic structure of English Noun Phrases, which consists of five rewrite rules, each with a left- and right-hand sides separated by an arrow→, four nonterminal symbols, which are capitalized, and three terminals, which are in lower case (the rules are numbered for convenience):

(1) Start → NounPhrase
(2) NounPhrase → Determiner Noun
(3) Noun → apple
(4) Noun → bird
(5) Determiner → the

Beginning with the nonterminal *Start* symbol, this miniature grammar can generate, for example, the sequence *the apple via* the successive replacement of symbols according to the grammar rules as follows: *Start*; *NounPhrase* (via Rule 1); *Determiner Noun* (via Rule 2); *the Noun* (via Rule 5); *the apple* (via Rule 3). (At this point no more rules apply since the last string has no symbols that appear on the left-hand side of any rule, so the generation halts, have produced the sequence noted.) This grammar will generate exactly one other form consisting of just terminal symbols, *the bird.*

Importantly, AGs are "artificial" and crucially different from natural grammars in at least two ways: (1) AG rule tokens are unlike those found in natural grammars, *e.g.*, they use abstract symbols such as "B" or "Z" rather than, say, part of speech or phrase tokens like *Noun Phrase* or *Noun*; and (2) AG rules themselves are typically deliberately schematized and simplified as compared to those of natural human generative grammars. Note that despite the general claim under the so-called contemporary "Minimalist Program" that there is a single operation, Merge, that builds syntactic structure, Merge interacts with the features of lexical items and other linguistic principles to yield what amounts to a very large, complex set of equivalent context-free rules in any particular human language (Barton et al., 1987). However, AGs are often deliberately designed to reflect only certain key abstract properties of natural grammars as part of their methodological role in experimental manipulations. Even in this sense, however, the miniature example grammar just presented is "artificial" in that it does not fully reflect the properties of English Noun Phrases, for example, the fact that Noun Phrases may themselves be modified by sentence-like phrases, as in *the apple the bird ate*. One tenet of some current accounts of human generative grammar holds that the chief property of human language syntax is its arbitrarily deep hierarchical structure, rather than sequential left-to-right order (Everaert et al., 2015) but a glance at the miniature grammar above and most AGs shows that they typically do not partition syntactic information in this way, and embed sequential order directly into their rule systems along with hierarchical structure, which is nearly unavoidable (*the apple* but not *apple the*), while finding it challenging to focus on just hierarchical structure.

2.2. Learning tokens and sequence rules

An important consideration in the interpretation of AGL studies is that the learning of rules is implicit. That is, the only information available to the subjects is the auditory input itself, from which regularities should be spontaneously recognized and memorized. Subjects are not explicitly told what the grammar is, nor are they explicitly trained in a way that should promote the learning of grammatical rules. Indeed, an important motivation behind many studies is to test whether or not grammars can be acquired implicitly from exemplars of strings that are produced by them. Since non-human animals cannot explain *a posteriori* what strategy they apply to differentially respond to stimuli, and humans appear often not to be aware of their strategies (*e.g.* Knowlton and Squire, 1996), it is up to the researchers to provide a convincing case when they claim that subjects learned and apply grammatical rules.

What information do perceptual systems require to learn artificial grammars? A grammar not only consists of sequence rules but

also of a lexicon of tokens, which are particular sounds, or classes of sounds, from which the rules generate sequences called 'strings'. Non-human subjects in particular cannot be expected to know *a priori* that particular segments of auditory input streams should be considered as tokens. As is the case for sequence rules, this is information that perceptual systems should infer implicitly from regularities in the auditory input itself.

To see how this can lead to ambiguities in the interpretation of what is learned, consider a thought experiment where an aardvark is habituated to the string *abcd* by repeated passive playback, where the tokens *a*, *b*, *c*, and *d* stand for the human speech sounds 'klor', 'biff', 'cav', 'dupp'. How does it habituate to this stimulus? The aardvark may first learn to recognize the tokens {*a*, b, *c, d*} as these are often-recurring sounds in its environment, and then learn the sequence rules that describe their temporal order. In this case, the rules are simple because forward transition probabilities between terminal tokens are always 1. Such an account may seem intuitive to the human reader, especially when string stimuli are visually presented as a sequence of letter symbols. We use such representations all the time. However, the aardvark can only learn from auditory sensory input as such, without any prior knowledge of tokens or sequences. And, given that it is unknown to the aardvark what the tokens are, one can question whether or not in this case it should be expected to learn them. The stimulus *abcd* may have been generated according to grammatical rules, but are these same rules needed to habituate to it?

How does the aardvark's brain recognize that the *abcd* sound has been experienced before and requires no attention? Although the underlying mechanisms are not known, and we do not propose a specific mechanistic model for this phenomenon here, it may be useful to consider the type of information and the kind of integration that would be required. At a minimum, current acoustic input, integrated over some time window, should sufficiently match (parts of) a memory trace of the temporal stream of sensory features that the *abcd* sound corresponds to. As long as the input matches, no attention is required. For how long should the match last? This would depend on what other input the aardvark normally experiences, but in general, integration over longer durations enables the recognition of more specific temporo-spectral patterns that enable source identification with more confidence. For example, in humans, a 1-s window may be sufficient to accurately recognize the sound of a piano, but a 5-s window may be required to recognize that the piano plays the *Goldberg Variations* by Bach. A match over longer durations thus provides more evidence that a current source is similar to a specific type of source that has been experienced before and to which an appropriate response, or the inhibition of a response, has been learned.

We argue that in order to recognize the occurrence of *abcd* in the current auditory scene, it would be sufficient to compare salient acoustic features of the input stream to those of a memory representation of *abcd* based on these same acoustic features, and integrate the results of such comparisons over sufficiently long time windows. It would not be required to identify tokens and verify whether or not their sequence structure conforms to grammatical or other string-based rules, even simple ones. Moreover, note that even when one assumes tokenization and analysis of string structure, this cannot be seen as an *alternative* mechanism. The sound of token *a* ('klor') is shorter than that of the *abcd* stimulus, but it is still hundreds of milliseconds long. Cochlear neuron firing patterns track acoustic changes on a millisecond scale and recognition of this token therefore already requires neural circuits to perform extensive temporal integration based on feature comparisons. Thus, a grammatical explanation would depend on the same mechanisms for temporo-spectral integration, memorization, and matching, and would additionally require mechanisms for lexical segmentation and for rule detection.

Are auditory systems capable of recognizing longer specific acoustic patterns without tokenization? We think that this is very likely the case. Indeed their performance in this regard is impressive as nicely shown by Agus et al. (2010), who created a large set of unique white noise fragments (*i.e.* random waveform) and then asked human subjects to discriminate between a normal 1-s noise and a 1-s noise that consisted of a repetition of a half-second noise fragment played twice contiguously. The subjects were indeed able to do this. Importantly, performance was specific to the particular white noise sounds to which they were exposed and not transferred to novel white noise sounds. That is, discrimination appeared not to depend on an abstract rule based on repetition, but on memory representations of the particular white noise sounds as such. These memories could last for several weeks. It should therefore be uncontroversial that auditory systems, in humans at least, are very good at detecting short-term acoustic regularities in extremely complex sounds, forming long-term memories of such sounds even after limited exposure and without reinforcement, and using them for discrimination.

The aardvark thought experiment above is based on only one string stimulus, and is therefore simpler than real artificial grammar learning studies. However, it should illustrate that there can be at least two very different though not mutually exclusive views of how string sounds like *abcd* are learned and processed. One is based on memory of tokens and the sequence rules that describe their order. The other one is based on memory of an integrated stream of sensory features that is closely related to the acoustic signal as such. The former appears often tacitly assumed in auditory artificial grammar learning experiments, but we will argue below that the latter can provide alternative explanations that may in part, or fully, explain response behavior. This may especially be the case when the tokens used to generate experimental strings are particular sounds rather than sound classes.

3. Syntax or surface?

In a previous paper (Beckers et al., 2012), we have provided a critical analysis of the string sets that were used in an artificial grammar learning study that claims to demonstrate context-free grammar learning in Bengalese finches, *Lonchura striata* var. *domestica* (Abe and Watanabe, 2011). Our conclusion was that there were biases in acoustic similarity between familiarization and test strings that could alternatively explain the experimental results without the use of grammatical rules. That is, large fragments of the grammatical test and familiarization strings were acoustically identical, while this was to a lesser extent the case for ungrammatical test strings. For example, their grammar- conforming test string stimulus $A_1A_2C_1F_2F_1$ was acoustically similar to three training stimuli, $A_1A_2C_2F_2F_1$, $A_1A_2C_3F_2F_1$, $A_1A_2C_4F_2F_1$ (letter-subscript symbols stand for particular Bengalese finch sounds), having only a mismatch in one sound element. All their grammar-conforming test stimuli had this level of similarity with training stimuli, whereas this was never the case for a grammar-violating stimulus, which showed such a match (four elements shared at the same positions) at best with only one training stimulus but more often with none.

Since then, we noticed that also in other auditory artificial grammar studies in non-human animals, including more recent ones, acoustic biases may be present and may have played a role in learning. These studies do not claim grammatical competence at a context-free level, and some (*e.g.* van Heijningen et al., 2012) report only very modest levels of abstract rule learning such as the recognition of repeated syllables, in an attempt to critically evaluate earlier farther reaching claims. Nevertheless, both from evolutionary and mechanistic points of view we think it is impor-

tant to distinguish between the learning of grammatical rules, even the simplest ones, and the recognition of familiar sounds on the basis of the extent of similarity in acoustic features.

We therefore set out to systematically analyze acoustic similarity within string sets in a sample of 9 studies, and to compare these to the behavioral results obtained. Studies that claimed or suggested the use of grammatical rules, potentially even very simple ones, by a nonhuman animal were selected on an *ad hoc* basis, provided they were based on published string sets that we could analyze. It was decided *a priori* to include in this review every study that we analyzed, and hence not to exclude studies *a posteriori* if they did not contain any biases in acoustic similarity.

We emphasize that we do not question the value of the experimental studies that we discuss. It requires much effort to collect this type of data, and even if the interpretation of the data is in part challenged, the results stand and constitute a valuable contribution in an understanding of how perception of complex vocalizations works. We hope that discussing acoustic similarity-based explanations will lead to future studies ruling them out by design, or considering them explicitly as a strategy for neural systems in their recognition of string stimuli.

It is important to note that all alternative interpretations that we present in this paper are based on strings that are composed of particular sounds, not sound classes. This is the case in many published artificial grammar learning studies, but certainly not all (*e.g.* Spierings and Cate, 2016). Further, although we argue the concept of 'token' is not required in our alternative interpretation, and hence neither is the concept 'string', we will still use the token/string notation for brevity and clarity in discussion. Thus, when we refer to, *e.g.*, the *abc* string matching the *abcd* string at the trigram level, we do *not* suggest that the subjects necessarily parse the strings into *a*, *b*, *c*, and *d* tokens and then analyze for trigram sequences. Instead, this is to be read that the sound that the symbols *abc* together refer to is completely contained within the sound that the symbols *abcd* refer to, and is thus acoustically identical to a large extent.

In the following section, we discuss four types of confound that we have identified in the extent of acoustic similarity between training and test strings. A schematic overview of these confounds with examples is shown in Fig. 1.

4. Types of acoustic similarity biases in artificial grammar learning studies

4.1. Test string is fully contained within familiarization string

We hypothesize that when a subject is habituated by passive playback to the familiarization string *abcde*, it may also show a diminished response to the test string *bcd* because it is acoustically fully present in the familiarization string. The reason for this is that there are no acoustically novel parts in the test string that may trigger a dishabituation response. We argue that this type of similarity may have played a role in an experiment with cotton-top tamarin monkeys (*Saguinus oedipus*) by Saffran et al. (2008).

In this study, tamarin monkeys were shown to more easily learn strings from a Predictive (P)-language, in which predictive dependencies mark phrase units, as compared to strings from a Non-predictive (NP)-language, which lacks predictive dependencies (their Experiment 3). The tokens consisted of particular speech syllable sounds ("biff", "cav", "dupp, 'klor', and "jux"). After familiarization with strings from either the P-language or the NP-language, the monkeys were exposed to grammatical and ungrammatical strings of the corresponding language, while orienting responses to the test stimuli presented from a concealed loudspeaker were measured. The subjects responded differently to grammatical and ungrammatical test strings when they had been familiarized with the P-language, but not when they had been familiarized with the NP-language (see their Fig. 3). The authors conclude that this suggests "that tamarins are able to detect regularities in a simple grammar, written over individual word tokens, when predictive dependencies are present." The authors have ruled out that familiarization corpora differed in n-gram properties up to trigrams relative to the test items.

However, there is a bias in how often test strings are fully contained within familiarization strings: ungrammatical test strings never fully occur in any familiarization string, but the grammatical test strings do, and not in a balanced way (see our Supplementary Analysis 1; Table S1). For example, their grammatical test string *bcd* occurs only once fully in a grammatical test string of the NP-language (*bcd*, the same string), but thrice in different grammatical test strings of the P-language (*bcd*, *bcdc*, and *bcdcj*). The same goes for a second grammatical test string (*bkcd*), while the two other grammatical test strings (*bkcjd* and *bcjd*) occur equally often in each language. Thus, assuming that different familiarization strings were all played equally often, the subjects had overall been familiarized more extensively with full test sounds of the P-language than with full test sounds of the NP-language. This could have resulted in corresponding differences in the strength of sensory feature-based memory representations. Such an alternative non-grammatical explanation, however, depends on the assumption that the tamarins do not contribute much weight to the long silence after the final syllable of the test strings, as the corresponding syllables in the familiarization strings were followed by additional syllables.

A similar bias can be found in Wilson et al., 2015a (see our Supplementary Analysis 1, Table S2), a study that we will discuss below in the context of a different type of similarity.

4.2. Test string shares long substrings with familiarization string

If a subject is habituated to an *abcde* string and then tested with *ebcda* and *ebcad* strings, it may respond less to the *ebcda* test string than to the *ebcad* test string because the former shares a longer uninterrupted sound fragment that is identical to the familiarization string. The longer the match between current input and a memory representation of a familiarization sound, the more likely their source is the same, and requires no attention. For this to work in this case, temporal integration windows over the current input stream that are used for memory matching should be longer than the silent intervals (if any) between the lexical items but shorter than the duration of the sound of three consecutive items.

We found that such a bias in the duration of continuous sound fragments that are shared between familiarization or training strings is common in animal artificial grammar studies. For example, Wilson et al. (2013) habituated macaque (*Macaca mulatta*) and marmoset (*Callithrix jacchus*) monkeys to strings conforming to the grammatical rules of a Reber grammar, and then presented them with test strings that were either consistent with these rules or that contained illegal violations. The strings were in part similar to the ones used by Saffran et al. (2008), although the experimental design was different. Because looking behavior in macaques was more extensive during grammar violating strings than during grammar consistent strings, and occurred at multiple places in the strings, the authors conclude that the results "... provide evidence for a previously unknown level of AGL complexity in Old World monkeys ...".

However, test strings in this experiment are biased with respect to the maximum length of the substrings they share with the training sounds. We highlighted all matching substrings in our Supplementary Analysis 2, Table S3, and show the maximum length and duration of shared substrings between training and test strings in Fig. S1. Grammar consistent test strings share between

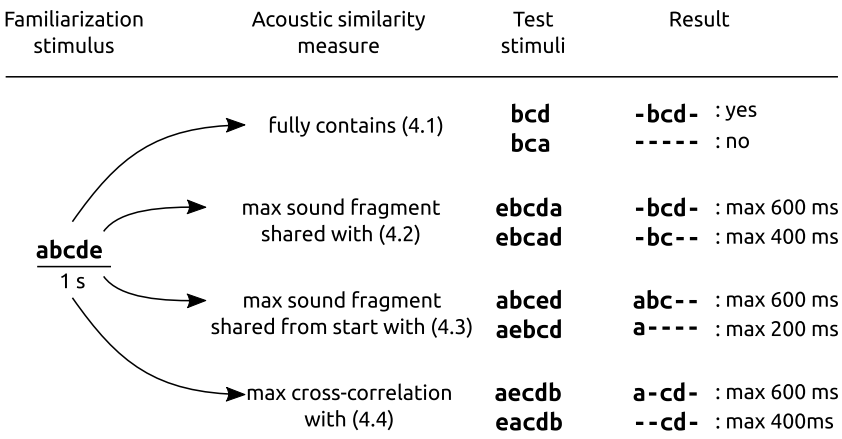

Fig. 1. Schematic overview of the four different measures that have been used in this review to assess forms of acoustic similarity between familiarization (training) and test stimuli. Numbers between brackets refer to the sections where specific measures are discussed. Although the example stimuli are represented as strings, letters stand for particular sounds and it is important to note that similarity is directly based on sound fragments in which pressure waveforms are identical. The extent of similarity in 4.2, 4.3, and 4.3 is thus considered a function of the duration in which pressure waveforms overlap, and its assessment does not necessarily require tokenization.

4 and 6 length-substrings with at least one of exposure strings, whereas grammar violating test strings do not share more than 1–2 length-substrings with any of the exposure strings. This bias in the corresponding duration of matching sound fragments provides a nongrammatical cue that could alternatively explain the differential looking behavior.

We have analyzed the string sets of two later studies from the same laboratory, based on similar but not identical experimental designs, the results of which led the authors to claim that rhesus macaques are sensitive to grammatical features of habituation strings. In Wilson et al. (2015b) it is concluded that the "... results suggest that humans and macaques are largely comparably sensitive to the adjacent AG relationships and their statistical properties", and in Wilson et al. (2015a) it is concluded that "... the form of statistical learning that the behavioral results in both species support is the implicit learning of the relationships between elements in a sequence as a function of the probabilities with which they occur in the exposure corpus". Here too, however, we find that the same alternative strategy, based on the maximum length of shared substrings between familiarization and test strings, can alternatively explain the behavioral results (Supplementary Analyses 3 and 4, respectively; Tables S4 and S5, and Figs. S2 and S3). Furthermore, all grammar conforming test strings in Wilson et al. (2015b) are also fully contained within a familiarization string (Table S2), whereas this is not the case for any of the grammar violating strings, a type a resemblance that has been discussed above, and that could also contribute to or explain the behavioral results in this study.

In addition to these habituation studies in rhesus monkeys, we found this type of bias in an operant Go-NoGo study, investigating whether zebra finches (*Taeniopygia guttatta*) use positional or transitional cues in recognizing strings of zebra finch natural vocal syllables (Chen and ten Cate, 2015). These authors carried out two experiments, the second of which consisted of three separate tests, and conclude on the basis of the results that zebra finches "... can attend to both transitional and positional cues and that their sequential coding strategies can be biased toward transitional cues depending on the learning context."

In their Experiment 1, birds are conditioned to respond differentially to the Go stimulus *abcdefabefcdab* and the NoGo stimulus *abdfcedfabceab*. The shared *ab* bigrams at the beginning and end of both strings should prevent the birds from using the start or final part of the strings as a basis of discrimination. On average, the birds correctly respond with a key peck in 87% of the cases to the Go stimulus, and only in 10% of the cases to the NoGo stimulus, showing a high level of discrimination. To one of the test strings, A1: *abefcdabcdefab*, they respond in 70% of the cases, while to a different test string, B2: *abcedfabdfceab*, they respond in only 49% of the cases. We notice that this difference in response behavior corresponds to a difference in shared substring length: A1 shares maximally 8 consecutive syllables (*e.g. abefcdab*) with the Go string and only maximally three (*e.g. fab*) with the NoGo string. The overlapping sound fragment between A1 and Go is thus 5 syllables longer than that between A1 and NoGo, which could explain the relatively high response percentage to A1 of 70%. B2, by contrast, shares a sound fragment of maximally three syllables with the Go string (*e.g. abc*), and sound fragments of maximally six syllables with the NoGo string (*e.g. abdfce*), corresponding to a lower response percentage of 49%. Our analyses of all of the strings in this study reveal similar biases in shared substring length (and corresponding shared sound fragment duration) in all of the tests, in a way that corresponds to the behavioral responses (SupplementaryAnalysis 5; Tables S6–S9; Figs. S5–S7).

4.3. Test string and familiarization string share the same beginning

Some parts of a memorized stimulus may be more important than others, and the beginning may receive particularly much weight in experimental paradigms. For example, we expect that in an habituation-dishabituation experiment where subjects are familiarized with a sound stimulus *abcde*, exposure to the novel stimulus *abcfg* will lead to fewer dishabituation responses than exposure to the novel stimulus *fgcde*, even though substrings shared with the familiarization stimulus are equally long. This is because from a perceptual point of view, evidence that *fgcde* is novel is available, and increasing, immediately after the start of the stimulus. A decision that the source of this stimulus is novel and requires an involuntary attention switch can thus be made right away. While it is true that the last part of the *fgcde* contains familiar sound and provides evidence that attention may not be required, an attentional switch may already have taken place or be well underway before this is taken into account. By contrast, in the case of *abcfg*, evidence has been accumulation during the first three token sounds that the source is known and does not require attention. Of course, here too the last part of that stimulus contains information that contradicts that evidence, but a decision to respond is now based all information available, which is mixed on whether or not there is an event that requires attention. This may overall lead to a reduced dishabituation response. For the same reason, we hypothesize that in a Go-NoGo experiment where the subjects are trained with *abcde*

as a Go stimulus and *fghij* as a NoGo stimulus, subjects may respond to the test string *abckl* more often than to a test string *klabc*, even though the shared substrings with the Go-stimulus are the same. The decision to respond or not may predominantly depend on differences between the start of the training strings because these are first available and sufficient for an appropriate response. This reasoning is consistent with findings of Chen et al. (2015a) who show that the zebra finches in a Go-NoGo experiment involving artificial sequences consisting of four syllables base their response behaviour predominantly on the first part.

More generally, time is an inherently crucial parameter in auditory-behavioural tests, as neural systems start to perform computations to reconstruct potentially important events in the environment before the input is 'finished'. For this reason, we think that a systematic bias in similarity between the beginning of training and test stimuli should be avoided.

One study where such a bias may have played a role, and we think provides the basis for a non-grammatical explanation of the results, is that of van Heijningen et al. (2012).

In their Experiment 1, these authors trained zebra finches to discriminate the Go-sounds *ABA* and *BAB* from the NoGo sounds *BBA*, *AAB*, *BAA*, and *ABB*. The idea behind this experiment was to test if the zebra finches spontaneously use a rule like "if the first and the last element is the same then respond, and when they are different, do not respond". When the birds achieved sufficiently high discrimination levels, test strings *ACA*, *CAC*, *ABBA*, and *BAAB* (test string set 1) were used to test this hypothesis. The hypothesis was rejected because response levels to all of the test strings were similar to that of the NoGo training stimuli instead of the Go training stimuli (their Fig. 2a).

The same birds were subsequently tested with four new test strings based on the same tokens: *ABAB*, *BABA*, *AABB*, and *BBAA* (test string set 2). The birds responded to the *ABAB* and *BABA* stimuli as Go-stimuli, and while responses to the *AABB* and *BBAA* stimuli were more similar to, and even lower than, the NoGo-stimuli (their Fig. 2b). This led the authors to conclude that the zebra finches attended to the presence or absence of adjacent *AA* or *BB* repeats. That is, they apply a repetition rule, which appears to be the basis for the authors' conclusion that the birds have a limited degree of abstract rule learning.

Is there an alternative explanation of these results that is consistent across test string sets and that does not require abstract rule learning? If we assume that the birds rote memorized (parts of) the training stimuli on the basis of their acoustic content alone, then they should at least wait for sound of the third element to occur before deciding on the appropriate response. The first two elements of the training stimuli are not sufficient to distinguish the Go and NoGo training stimuli: strings starting with *AB* and *BA* are present in both groups. Only when (part of) the third element has been perceived, sufficient information has accumulated to perform the correct response to every training string. Comparing the number of elements from the start of the string that are identical between training stimuli and test stimuli, we find that test strings *ABAB* and *BABA* share the first three elements with a Go training stimulus, and only the first two with a NoGo training stimulus. These test stimuli are exactly the ones that the birds responded to. By contrast, the other test string stimuli, *ABBA*, *BAAB*, *AABB*, and *BBAA*, the ones the birds did overall not respond to, all share the first three elements with a NoGo training stimulus, and only the first one or two elements with a Go training stimulus. Thus, the birds' response behavior can also be explained by the extent of acoustic similarity between the beginning of the strings: if the sound over the duration of the first three elements (or at least a sound fragment longer than the first two element), which is the fastest cue to distinguish all Go from all NoGo training stimuli, is the same as a Go stimulus, then go, and if it is the same as a NoGo stimulus, do not go.

van Heijningen et al. (2012) did consider a similar, though not identical, alternative explanation of their results that does not require abstract rule learning, as they pointed out the possibility that "... the birds [...] responded to the [test strings] according to whether these contained the exact three-element sequences from the training instead of using an abstract rule". However, they dismiss this alternative explanation because the response levels to test strings *AABB* and *BBAA* are even lower than the response levels to the training NoGo stimuli, whereas the response levels to the test strings ABBA and BAAB is similar to the training NoGo stimuli (their Fig. 2b). This is not predicted by their alternative nongrammatical explanation. The authors therefore prefer a rule learning explanation, although based on a limited level of abstraction, and explain the very low response levels to the *AABB* and *BBAA* test strings by the birds treating them as 'super negative' stimuli: they contain *two* AA or BB repeats instead of the single one that is present in the NoGo stimuli and that they interpret (in test string set 2) as the characteristic, abstract feature that the birds use for discrimination. However, the lower-than-NoGo response levels to test strings *AABB* and *BBAA* is also consistent with our non-gramatical explanation above, because both these test stringss share only the first element with a Go stimulus, whereas two of the four NoGo stimuli share the first two elements with a Go stimulus, and the other two only the first. Half of the NoGo stimuli are thus more similar to the Go stimuli than the *AABB* and *BBAA* test stringss are, explaining the overall differences between these groups.

Another argument that the authors put forward against rote memorization of the training stimuli is that in a second experiment (Experiment 2, not discussed here because the results were overall negative) the same birds responded initially less to the original *ABA* and *BAB* Go stimuli when additional strings were added to the training string sets. However, the strings that were added to the NoGo set were *ABAB* and *BABA*, which fully contain these sounds from the start, and a response to which was suddenly punished. A reduction in response levels to *ABA* and *BAB* is thus also expected on the basis of a acoustic memory-based explanation.

Taken together, we suggest our acoustic memory-based explanation is the more parsimonious one because i) it does not depend on abstraction, but only on matching sensory features with memory, and ii) it is consistent with previous operant work with similar stimuli in the same species that shows that the start of string sounds carry particular weight in the discrimination. Both in the authors' and our interpretations of what is learned, it remains unexplained why the overall response levels to the *CAC* and *ACA* test strings are similar to that of the NoGo stimuli (their Fig. 2a). It may be the new element C causes the birds to perceive this as a novel sound. They are forced to make a 'choice' and most birds may prefer to avoid punishment over the opportunity of immediate access to food most of the time.

Another study that we identified in which a bias in the duration of shared sound fragments from the beginning of the string may have played a role investigated positional learning in chimpanzees (Endress et al., 2010). The subjects were exposed to the habituation strings XABXXX, XAXBXX, XAXXBX and XXXABX. The lexicon consisted of three different particular chimpanzee vocalizations. All habituation strings start and end with an X, but differ in the positions of A, B, and the X items in between. The idea behind the design is that the apes can either use positional or transitional ("chaining") cues to memorize the sequence of the items. Test strings, which could violate positional or transitional regularities, or both, or none, were used to measure the level of dishabituation, which in turn was used to determine the cues that had been used. The test strings were AXXXXB (*p*), BXXXXA (*p,c*), XXABXX, XXBAXX (*c*), XXAXBX, XXBXAX (*c*), with letters between the parenthesis indicating when a test string violates a positional (*p*) or chaining (*c*) regularity. Overall, the subjects dishabituated more strongly to positional violating

strings than to chaining violation strings (their Fig. 2), and the authors suggest "that when given the opportunity to learn either a chaining regularity, positional regularity or both, chimpanzees initially and spontaneously extracted the positional regularity under the test conditions presented." They rule out other explanations, among which the idea that the chimpanzees may have restricted their attention to the first or last element of the strings.

However, notice that all strings that do not contain position violations (*i.e.* XXABXX, XXBAXX, XXAXBX, XXBXAX, which are the ones that they dishabituated less to) all share a two-item length beginning with one of the habituation strings (XXXABX), whereas the position-violation test strings are different from the start. Hence, even if one agrees with the authors' arguments why it is unlikely that the subject just considered the first or the last element of the sequence, there is still a bias in a two-element length time frame at the beginning of the test strings that may partly or completely explain the differential responses.

Further, the behavioral results of the studies of Wilson et al. (2015a,b, 2013), discussed above in relation to biases in shared substring durations, and Saffran et al. (2008), discussed above in relation to biases in test strings that are fully contained in training strings, can also be alternatively explained on the basis of a bias in the duration of shared substrings from the beginning of the stimuli (Supplementary Analysis 6, Figs. S8 and S9).

4.4. Test string has high cross-correlation with familiarization string

So far we discussed three ways in which we assessed the extent of acoustic overlap between strings, all of which are based on identifying a shared sound fragment that is continuous. *I.e.*, there is no discontinuity in similarity within the substring as the acoustic wave form is identical from the beginning to the end. Discontinuities in similarity, however, may not completely nullify the accumulation of evidence that current input is caused by a sound source that has been experienced earlier. Consider the string *abcdef*. A subject that has habituated to the corresponding sound may show less dishabituation to a test string *abcgef*, than to a test string *abcgfe*, even though the maximum shared substring length in both cases is three, and they both share an equally long sound fragment at the beginning of the stimulus. This could be caused by neurons that integrate over longer timescales and whose firing rates are not strongly affected by relatively brief mismatches. If so, *abchefg* is largely similar to *abcdefg*, but *abchgfe* less so.

Tolerance by perceptual systems to some degree of mismatch is likely because in natural auditory scenes repeated vocalizations are rarely identical and parts of them may also be intermittently masked by other sounds. Auditory systems should still recognize such sounds as originating from the same source.

One straightforward way to identify multiple longer substrings that are shared at matching positions between strings is to calculate their cross-correlation, which is a similarity measure of two signals as a function of the lag of one relative to the other. The input to cross-correlation algorithms could be sound waveforms, spectrograms, or cochleagrams. But since these are rarely available in published work, and since we focus on studies in which token symbols stand for particular sounds, we use as input the string symbols themselves, as they correspond to a particular sound waveform, and thus spectrogram and cochleagram. Fig. 1 illustrates the cross-correlation between two strings. For our purposes, the *maximum* cross-correlation is the most informative value, which essentially counts how many string items are identical at corresponding positions if one is allowed to freely shift one string over the other and choose the optimal lag. The fact that some corresponding items may be adjacent, and others not, has no effect on this measure.

We previously provided a critical analysis (Beckers et al., 2012) of the strings used in Abe and Watanabe (2011), who claimed that Bengalese finches (*Lonchura striata* var. *domestica*) spontaneously learn a context-free grammar from the string examples they were presented with. These strings consisted of natural Bengalese finch syllables. Our conclusion was that the behavioral results could be alternatively explained in terms of acoustic similarity matching, if the birds allowed for a short mismatch of one syllable. However, we did not provide cross-correlation analyses in that critique, which for completeness are provided here in Supplementary Analysis 7, Fig. S10, in the form of image plots, and which illustrate the biases in the string sets in an alternative way.

We find that a bias in this type of similarity can also provide an alternative interpretation of the results in Chen et al. (2015b). These authors asked if zebra finches and human adults can detect the difference between a XYX and a XXY structure, where X and Y denote arbitrary tokens. The tokens consisted of six zebra finch song elements: {*A,B,C,D,E,F*}. In a series of Go-NoGo tests (their Experiment 1a), the birds overall failed to learn the abstract rule, but the authors conclude from additional results (their Experiment 1b, shown in their Fig. 5), that some of the birds may have learned to recognize the abstract XYX structure when the test strings are generated with acoustically familiar elements rather than novel ones. *E.g.*, when trained to respond to Go training stimuli like *CDC*, birds respond similarly often to test stimuli like C*BC*, but they respond significantly less often to test strings like *CCB*, although not as rarely as to NoGo training stimuli, like *CCD*. On the basis of this finding, the authors suggest that the birds learned to attend to a 'repetition rule': "if it starts with *AA*, *BB*, *CC* or *DD*, treat the stimulus as a NoGo stimulus; if these specific bigrams are not present, treat stimuli as a Go stimulus." Overall they conclude on the basis of the response patterns that "zebra finches show evidence of simple rule abstraction related to positional learning, suggesting stimulus-bound generalization".

However, in this example the maximum cross-correlation between test string *CBC* and Go training string *CDC* is 2, whereas that between test string CCB and Go training string *CDC* is only 1. Cross-correlation analyses of all the strings in their Experiment 1b show that there is an overall bias in this type of resemblance between test and training strings, which provides an alternative, non-grammatical explanation for the difference in response behavior (Supplementary Analysis 8, Fig. S11). All XYX test strings show 2–3 times more often a high maximum cross-correlation with XYX Go training strings than with XXY NoGo training strings. Two of the four XXY test strings, in contrast, are balanced in this regard, while the other two show 2 times more often a high maximum cross-correlation with XXY NoGo training strings.

If the birds used this bias in level of acoustic similarity, then there is no need to suggest the use of XYX rule abstraction or that the birds learned a repetition rule. Such an acoustic similarity explanation has the added advantage that it is also consistent with the results in their Experiment 1a (Supplementary Analysis 8, Fig. S12). The authors suggest for these results that "The most likely explanation is that the birds learned to use the final elements of the XYX strings to discriminate 'good' and 'bad' sounds, demonstrating a stimulus-bound generalization". However this explanation is different from the one the authors suggest for Experiment 1b, even though these are the same type of stimuli.

A bias between test strings in the level of cross-correlation with training strings can further also provide an alternative explanation in some of the other studies that we already discussed above, namely Chen and ten Cate (2015), Experiments 2 (all tests) and 3, but not Experiment 1, and Wilson et al. (2015a,b, 2013). The corresponding quantifications are shown in Supplementary Analysis 9, Fig. S12.

Note that one caveat of calculating cross-correlation between strings is that the sounds that the tokens refer to may not be equally long in duration. Thus, the maximum cross-correlation between *abc* and *adc* is 2, which is considerable given that the highest possible value is 3. However, if *a* and *c* stand for very short sounds and *b* or *d* for a very long sound, then this measure does not reflect the acoustic similarity very well.

5. General discussion and conclusions

Our analyses revealed that in a number of published artificial grammar learning (AGL) studies in nonhuman animals there appear to be biases in acoustic similarity between training and test stimuli. Following on from a previous re-analysis (Beckers et al., 2012), in the present review we assessed the extent of such confounds, and considered a sample of 8 additional studies that based their experiments on strings of tokens that represent particular sounds. In these studies, we identified 11 experiments that led authors to suggest or claim the learning and use of grammatical or at least string-based rules, which ranged from simple repetition or positional rules to language-like recursion. After quantification, we found that none of these 11 experiments appears to be free from acoustic similarity confounds that we argue can contribute to or alternatively explain the observed behaviors in terms of acoustic recognition memory (Table 1).

We do not suggest that our acoustic similarity measures are necessarily closely analogous to the computations that the subjects' neural systems used for learning and recognition of the stimuli, or that subjects cannot have used other or additional cues, including grammatical ones. However, despite such uncertainties about how precisely neural systems assess surface feature-based similarity we believe that it is a significant finding that all the studies that we have analyzed contain biases in acoustic similarity in their design. This renders interpretations in terms of rule learning in all the studies we looked at as debatable, if one agrees that it may be easier for neural systems to assess similarity in acoustic features than to assess conformity to grammatical rules. We argue that this is the case, because the latter depends on the former, but not the other way around.

A number of the studies contained biases in multiple acoustic similarity measures. This is unsurprising because the measures we looked at are not independent. For example, if strings share long continuous substrings they are also highly cross-correlated. Nevertheless, there are differences between the measures that may correspond to the involvement of different neural mechanisms, and it would be interesting to test which measure explains the observed results best. We did not do so here because response data to individual strings are usually not published, and these are necessary for a meaningful comparison. Even better would be to perform direct experimental tests of the measures discussed here, and perhaps additional ones, in order to assess which ones predict response behaviour more strongly than other ones in different learning paradigms. We should emphasize that the role of acoustic similarity in perceived similarity and response behaviour remains untested in the context of animal AGL studies, and is only hypothesized by us, and at this point presented as an alternative explanation for grammatical rule-based accounts. However, from a rich history of psychoacoustic studies in animals, sometimes carried out in the same laboratory that now investigate sequence rule learning (*e.g.* Beckers and ten Cate, 2001; Beckers et al., 2003), it is clear that acoustic feature-based perception can be used in learning and discriminating between stimuli, and hence should be taken fully into account.

More generally, we think that it is difficult to avoid biases in acoustic similarity completely, at least when tokens are particular sounds, and it may be wise to consider explicitly multiple explanations about what subjects learn, including acoustic similarity-driven learning strategies, in addition to one or more grammatical hypotheses of interest (*e.g.* van Heijningen et al., 2009). By creating *a priori* predictions about the behavioral response to specific strings for each explanation, one can test which is the best one. Such an approach would combine well with recent developments in applied Baysian statistics and informative hypotheses testing, because a positive outcome of an experiment is not just the rejection of one null-hypothesis but rather a value for how much more likely one explanation is over another, given the data (Van De Schoot et al., 2011).

In addition, we believe that there are other considerations for future work that would improve the interpretability of AGL experiments. We noticed that in studies that are based on an operant conditioning paradigm, it often remains unclear what association(s) have been learned during the training phase. *E.g.*, is a test stimulus not responded to because it resembles the NoGo training stimulus or because it does not resemble the Go training stimulus, or both? Interpretation would be easier if it were known what the response level is to a stimulus that resembles neither the Go nor the NoGo stimulus (set). If the level is intermediate, then the subject must have learned specific aspects of both categories, but if the level is similar to the NoGo stimulus (set), then the subject may have learned predominantly or exclusively specific aspects of the Go stimulus (set). Including a neutral test stimulus like a white noise will likely help in understanding which training stimuli are specifically recognized, and which ones are simply seen as part of an excluded set.

Overall, we conclude that interpretation in the studies that we looked at is overwhelmingly 'string-focused' and often ignores strategies based on lower-level sensory features. We think that acoustic similarity matching per se can provide a parsimonious explanation for observed response behaviors that does not involve tokenization or sequence rules. Nevertheless, what if subjects do tokenize string stimuli and learn their sequence or positional rules? First, even though some neural circuits may process stimuli on the basis of tokens and sequence or positional rules, these may not be involved in the particular response behaviour that is measured in habituation or operant conditioning experiments, which may depend on sensory feature-based representations instead. Analogously, it is well-known, for example, that speech sound input is processed in parallel by different neural circuits that analyze prosodic and semantic content. Indeed, there may exist many parallel auditory processing streams, each of which may in principle determine or influence the outcome of a behavioral response. Second, even when response behavior is based on string representations, then in many cases the subjects may process training stimuli as 'chunks', *i.e.* particular sequences of tokens that are fixed, without using the grammatical rules that have produced them. If so, similarity judgments may be based on the length of overlap between token chunks (Perruchet and Pacteau, 1990), and our analyses show that there is often a bias in this respect.

What are the consequences of our findings for our understanding of artificial grammar learning in nonhuman animals in general? We cannot extend our interpretations to studies we have not investigated here, but we believe that if studies are based on tokens that represent particular sounds, rather than classes of sound, then these should be critically examined as to whether or not acoustic similarity explanations can be ruled out. However, not all studies on grammaticality in animals are based on this design. For example, in a recent study Spierings and ten Cate (2016) found that budgerigars (*Melopsittacus undulatus*) can generalize the discrimination between three-element sound stimuli with an XYX or an XXY structure to stimuli with the same structure but acousti-

Table 1
Biases in acoustic similarity between test and training strings in the reviewed studies that can alternatively explain results. The nature of these biases is discussed in different sections in the text. *Contains*: test string occurring fully in training strings (Section 4.1); *Shared start*: test and training strings sharing same start (Section 4.2); *Shared Substring*: test and training strings share substrings (Section 4.3); *Cross-correlation*: test and training strings share tokens at corresponding positions (Section 4.4).

Experiment	Contains	Shared Start	Shared Substring	Cross-correlation
Abe and Watanabe (2011) Fig. 3				•
Chen et al. (2015b) Exp. 1b				•
Chen and ten Cate (2015) Exp. 1			•	
Chen and ten Cate (2015) Exp. 2			•	•
Chen and ten Cate (2015) Exp. 3	•		•	•
Endress et al. (2010) Exp. 1		•		
van Heijningen et al. (2012) Exp. 1		•		
Saffran et al. (2008) Exp. 3	•	•		
Wilson et al. (2013)		•	•	•
Wilson et al. (2015a)		•	•	•
Wilson et al. (2015b)	•	•	•	•

cally novel elements. The present review here did not address such findings and has no direct consequences for their interpretation.

From our analyses of string design in AGL studies in nonhuman animals we conclude that acoustic confounds are a blind spot. Such confounds should be excluded or carefully addressed if the aim is to obtain meaningful results pertaining to possible grammatical abilities in non-human animals. It may well be that this problem is not specific to animal studies, and that similar confounds may occur in human studies. Indeed, non-grammatical explanations for response behaviour in human AGL experiments have been proposed (for a review see, *e.g.*, Pothos, 2007). The grammars used in animal experiments are modeled after, or sometimes identical to, the ones used in human studies, and many animal studies follow the general design of human studies quite closely for comparative reasons. We therefore suggest to critically evaluate string sets in human AGL experiments for the types of acoustic similarity we discussed here. To facilitate this, we make the software library we wrote for quantification and visualization in the supplementary information of this paper freely available (aglcheck, https://github.com/gjlbeckers-uu/aglcheck). This may help avoiding unintended cues in future studies, and a posteriori estimating their extent in existing studies so that their potential involvement in response behavior can be evaluated.

Acknowledgements

G.J.L.B and J.J.B. are 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). We thank Carel ten Cate and Christopher Petkov for fruitful discussions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.neubiorev.2016.12.021.

References

Abe, K., Watanabe, D., 2011. Songbirds possess the spontaneous ability to discriminate syntactic rules. Nat. Neurosci. 14, 1067–1074, http://dx.doi.org/10.1038/nn.2869.

Agus, T.R., Thorpe, S.J., Pressnitzer, D., 2010. Rapid formation of robust auditory memories: insights from noise. Neuron 66, 610–618, http://dx.doi.org/10.1016/j.neuron.2010.04.014.

Barton, E., Berwick, R.C., Ristad, E., 1987. Computational Complexity Theory and Natural Language. MIT Press, Cambridge, MA.

Beckers, G.J.L., ten Cate, C., 2001. Perceptual relevance of species-specific differences in acoustic signal structure in Streptopelia doves. Anim. Behav. 62, 511–518, http://dx.doi.org/10.1006/anbe.2001.1768.

Beckers, G.J.L., Goossens, B.M.A., ten Cate, C., 2003. Perceptual salience of acoustic differences between conspecific and allospecific vocalizations in African collared-doves. Anim. Behav. 65, 605–614, http://dx.doi.org/10.1006/anbe.2003.2080.

Beckers, G.J.L., Bolhuis, J.J., Okanoya, K., Berwick, R.C., 2012. Birdsong neurolinguistics: songbird context-free grammar claim is premature. Neuroreport 23, 139–145, http://dx.doi.org/10.1097/WNR.0b013e32834f1765.

Berwick, R.C., Friederici, A.D., Chomsky, N., Bolhuis, J.J., 2013. Evolution, brain, and the nature of language. Trends Cogn. Sci. 17, 89–98, http://dx.doi.org/10.1016/j.tics.2012.12.002.

Chen, J., ten Cate, C., 2015. Zebra finches can use positional and transitional cues to distinguish vocal element strings. Behav. Processes Cause Funct. Behav. Biol. A Tribute Jerry Hogan 117, 29–34, http://dx.doi.org/10.1016/j.beproc.2014.09.004.

Chen, J., Jansen, N., ten Cate, C., 2015a. Zebra finches are able to learn affixation-like patterns. Anim. Cogn. 19, 65–73, http://dx.doi.org/10.1007/s10071-015-0913-x.

Chen, J., van Rossum, D., ten Cate, C., 2015b. Artificial grammar learning in zebra finches and human adults: XYX versus XXY. Anim. Cogn. 18, 151–164, http://dx.doi.org/10.1007/s10071-014-0786-4.

Chomsky, N., 1975. The Logical Structure of Linguistic Theory. Springer, New York.

Endress, A.D., Carden, S., Versace, E., Hauser, M.D., 2010. The apes' edge: positional learning in chimpanzees and humans. Anim. Cogn. 13, 483–495, http://dx.doi.org/10.1007/s10071-009-0299-8.

Everaert, M.B.H., Huybregts, M.A.C., Chomsky, N., Berwick, R.C., Bolhuis, J.J., 2015. Structures, not strings: linguistics as part of the cognitive sciences. Trends Cogn. Sci. 19, 729–743, http://dx.doi.org/10.1016/j.tics.2015.09.008.

Knowlton, B.J., Squire, L.R., 1996. Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. J. Exp. Psychol. 22, 169–181.

Perruchet, P., Pacteau, C., 1990. Synthetic grammar learning: implicit rule abstraction or explicit fragmentary knowledge? J. Exp. Psychol. Gen. 119, 264–275, http://dx.doi.org/10.1037/0096-3445.119.3.264.

Pothos, E.M., 2007. Theories of artificial grammar learning. Psychol. Bull. 133, 227–244, http://dx.doi.org/10.1037/0033-2909.133.2.227.

Saffran, J., Hauser, M., Seibel, R., Kapfhamer, J., Tsao, F., Cushman, F., 2008. Grammatical pattern learning by human infants and cotton-top tamarin monkeys. Cognition 107, 479–500, http://dx.doi.org/10.1016/j.cognition.2007.10.010.

Spierings, M.J., ten Cate, C., 2016. Budgerigars and zebra finches differ in how they generalize in an artificial grammar learning experiment. Proc. Natl. Acad. Sci. U. S. A., 201600483, http://dx.doi.org/10.1073/pnas.1600483113.

Van De Schoot, R., Hoijtink, H., Romeijn, J.-W., 2011. Moving beyond traditional null hypothesis testing: evaluating expectations directly. Quant. Psychol. Meas. 2, 24, http://dx.doi.org/10.3389/fpsyg.2011.00024.

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. Proc. Natl. Acad. Sci. U. S. A. 106, 20538–20543, http://dx.doi.org/10.1073/pnas.0908113106.

van Heijningen, C.A.A., Chen, J., van Laatum, I., van der Hulst, B., ten Cate, C., 2012. Rule learning by zebra finches in an artificial grammar learning task: which rule? Anim. Cogn. 16, 165–175, http://dx.doi.org/10.1007/s10071-012-0559-x.

Wilson, B., Slater, H., Kikuchi, Y., Milne, A.E., Marslen-Wilson, W.D., Smith, K., Petkov, C.I., 2013. Auditory artificial grammar learning in macaque and marmoset monkeys. J. Neurosci. 33, 18825–18835, http://dx.doi.org/10.1523/JNEUROSCI.2414-13.2013.

Wilson, B., Kikuchi, Y., Sun, L., Hunter, D., Dick, F., Smith, K., Thiele, A., Griffiths, T.D., Marslen-Wilson, W.D., Petkov, C.I., 2015a. Auditory sequence processing reveals evolutionarily conserved regions of frontal cortex in macaques and humans. Nat. Commun. 6, 8901, http://dx.doi.org/10.1038/ncomms9901.

Wilson, B., Smith, K., Petkov, C.I., 2015b. Mixed-complexity artificial grammar learning in humans and macaque monkeys: evaluating learning strategies. Eur. J. Neurosci. 41, 568–578, http://dx.doi.org/10.1111/ejn.12834.